Theor Appl Genet (2003) 107:1102–1112
DOI 10.1007/s00122-003-1305-1

D. I. Mester · Y. I. Ronin · Y. Hu · J. Peng · E. Nevo · A. B. Korol

Efficient multipoint mapping: making use of dominant repulsion-phase markers

Received: 8 August 2002 / Accepted: 27 January 2003 / Published online: 20 August 2003

Abstract The paper is devoted to the problem of multipoint gene ordering with a particular focus on "dominance" complication that acts differently in conditions of coupling-phase and repulsion-phase markers. To solve the problem we split the dataset into two complementary subsets each containing shared codominant markers and dominant markers in the coupling-phase only. Multilocus ordering in the proposed algorithm is based on pairwise recombination frequencies and using the well-known travelling salesman problem (TSP) formalization. To obtain accurate results, we developed a multiphase algorithm that includes synchronized-marker ordering of two subsets assisted by re-sampling-based map verification, combining the resulting maps into an integrated map followed by verification of the integrated map. A new synchronized Evolution-Strategy discrete optimization algorithm was developed here for the proposed multilocus ordering approach in which common codominant markers facilitate stabilization of the marker order of the two complementary maps. High performance of the employed algorithm allows systematic treatment for the problem of verification of the obtained multilocus orders, based on computing-intensive bootstrap and jackknife technologies for detection and removing unreliable marker scores. The efficiency of the proposed algorithm was demonstrated on simulated and real data.

Keywords Multilocus ordering · Synchronized optimization algorithm · Dominant marker · Repulsion phase · Bootstrap

Communicated by J.W. Snape

D. I. Mester · Y. I. Ronin · Y. Hu · J. Peng · E. Nevo ·
A. B. Korol (✉)
Institute of Evolution, University of Haifa, Mt. Carmel,
Haifa 31905, Israel
e-mail: korol@esti.haifa.ac.il
Tel.: +972-48240-449
Fax: +972-48240-449

Introduction

Mapping numerous markers has become central to genetic analysis in the molecular-genomic era. An important step in generating multilocus genetic maps based on results of linkage analysis is in determining the true order of the genetic loci mapped, e.g. Mendelian genes or DNA markers. One of the possibilities in addressing this problem is to recover the linear marker order from the experimentally derived pairwise marker distance matrix d_{ij}. A primary difficulty in ordering genetic loci using linkage analysis is the large number of possible orders: for n loci on a chromosome, $n!/2$ distinct orders should be compared. In real problems, n might vary from dozens to many hundreds of markers. Clearly, even for $n \sim 30$–50, it would not be feasible to evaluate all $n!/2$ possible orders using two-point linkage data. This is why multilocus ordering is considered as an NP-hard combinatorial problem (Wilson 1988; Olson and Boehnke 1990; Falc 1992; Ellis 1997). A solution to this problem based on the maximum-likelihood approach employing Mapmaker software takes an hour on a Pentium-IV (1,500 Mhz) computer even for a modest case of $n = 10$. A second group of complications in the marker-ordering problem derives from various genetic and experimental obstacles like dominant markers, marker misclassification, negative and positive interference, and missing data.

Several methods have been proposed for determination of marker order (Lathrop et al. 1985; Lander and Green 1987; Knapp et al. 1995; Newell et al. 1995; Liu 1998), and implemented in software packages like LINKAGE (Lathrop and Llouel 1984), MapMaker (Lander and Green 1987), FastMap (Curtis and Gurling 1993), JoinMap-3.0 (VanOoijen 2002) and OutMap (Whitaker and Williams 2001). Historically, the main approach to ordering markers within linkage groups was based on multipoint maximum-likelihood analysis. Various optimization techniques for such analysis, including the branch and bound method (Lathrop et al. 1985), simulated annealing (Thompson 1984; Weeks and Lange 1987; Stam 1993;

Jansen et al. 2001) and seriation (Buetow and Chakravarti 1987) were employed.

Olson and Boehnke (1990) compared several methods for marker ordering, including multilocus likelihood and more simple criteria based on two-point linkage data (by minimizing the sum of adjacent recombination rates or adjacent genetic distances). The simple criteria are based on the biologically reasonable assumption that the true order of a set of linked loci will be the one that minimizes the total map length of the chromosome segment.

The accuracy of any ordering method may depend on the distribution of recombination frequencies (presence of large gaps), the percentage of missing data, noise caused by marker misclassification and genetic interference. That is why there is a tendency to test the obtained solution by some quality control analysis, allowing for verification of the derived multilocus order (Liu 1998). The most popular approach to do that is based on re-sampling procedures like bootstrap and/or jackknife analysis. Such a methodology increases the reliability of the results, but due to its computing-intensive nature it can be implemented only for cost-efficient ordering algorithms.

Recently, Mester et al. (2003 revised) developed a new, very fast and highly reliable algorithm for multilocus ordering based on two-locus linkage data that employs the Evolutionary Optimization Strategy (ES). The present manuscript continues and adapts the proposed approach for multilocus ordering with an excess of dominant markers complicated by the presence of repulsion-phase configurations. In linkage analysis, situations, when each of the parents provides to the F1 progeny one (and only one) dominant allele at two linked loci (A/a and B/b) (i.e. the resulting hybrid is F_1 = Ab/aB), are referred to as the “repulsion-phase”, in contrast to the “coupling phase” when F_1 = AB/ab (Bailey 1961). From the viewpoint of information content, dominance brings a loss of mapping information that is manifested in an increase of variance of the estimates of recombination rate, $V(r)$ (Sall and Nilsson 1994). For example, the ratio of variances $V(r)$ for a pair of dominant coupling-phase markers to that for co-dominant markers will be approximately 2.0 for $r = 0.05$, whereas the same ratio for repulsion-phase markers increases manifold. Even more important is another negative consequence of the repulsion phase: a downward bias of the estimates of recombination rate increasing with reduction of sample sizes. This bias derives from the fact that the expected number of recombinant phenotypes ab/ab in the F_2 progeny of the repulsion heterozygote F_1 = Ab/aB is $Nr^2/4$, where N is the sample size. If $Nr^2/4 < 1$, it may happen that no recombinants will appear in some samples resulting in estimates of $r = 0$, whereas in other samples, with recombinants, $\bar{r} \neq 0$ will be obtained, causing high between-sample variation. For example, for $N = 100$ and $r = 0.05$, the estimated value will be 0.0132 ± 0.0503 (average of 10,000 Monte-Carlo runs), instead of the expected 0.05.

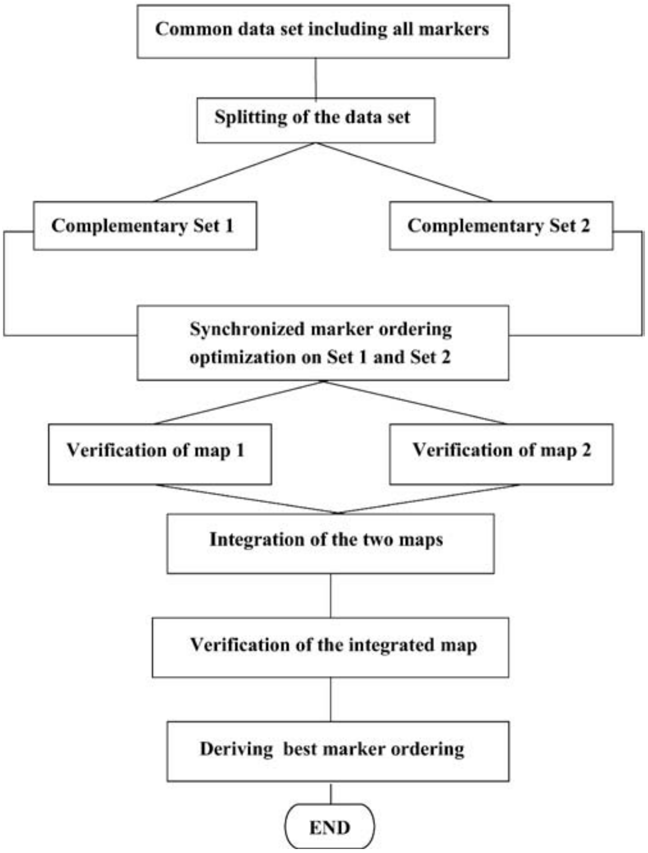

Common data set including all markers
Splitting of the data set
Complementary Set 1
Complementary Set 2
Synchronized marker ordering optimization on Set 1 and Set 2
Verification of map 1
Verification of map 2
Integration of the two maps
Verification of the integrated map
Deriving best marker ordering
END

Fig. 1 Algorithm of the problem decision

Therefore, the dominance complication acts differently in conditions of coupling-phase and repulsion-phase. As was shown elsewhere (Mester et al. 2003, revised), when all dominant markers were in the coupling phase, the proportion of dominant and co-dominant markers had no effect on the quality of marker ordering. A dramatically different result will be obtained with dominant markers in the repulsion phase (see below). It appeared that the higher the proportion of repulsion-phase markers, the lower the quality of multilocus ordering. High precision of ordering in the coupling-phase data and low precision in the repulsion-phase data justify splitting the data into two subsets, each containing all co-dominant markers and coupling-phase dominant markers only, and generating two complementary multipoint maps for each linkage group (Knapp 1995; Peng et al. 2000). Despite the general tendency of using co-dominant markers for genetic mapping (like RFLP, microsatellites and, recently, SNP), for many organisms dominant markers (mainly, PCR-based AFLP) remain a major tool, and RFLP or microsatellites, if available, are used as anchor markers (Peng et al. 2000; Menz et al. 2002; Parsons and Shaw 2002; Pérez-Enciso and Roussot 2002). Moreover, AFLP markers have been employed as a cost-efficient tool for constructing integrated physical and genetic maps (Klein et al. 2000). In many organisms, the resource populations should be provided with high-density maps including many dozens and even hundreds of markers (Harushima et al. 1998; Hall et al. 2001; Cone et al. 2002; Menz et al. 2002).

Once the strategy of generating two complementary multipoint maps is adopted for mapping dominant markers, the next step should be integration of the two maps. This may encounter difficulties caused by local and global map disturbances affecting the ordering of co-dominant markers common for both maps, if the density of such co-dominant markers is relatively low (e.g. in cases when co-dominant markers serve as anchors). In fact, the availability of shared co-dominant markers enables mutual control during multilocus ordering (for maintaining the same order of the shared markers) which, together with computing-intensive jackknife and bootstrap techniques (Efron 1979), may significantly improve the quality of the resulting map. To implement the idea of parallel ordering of two subsets of markers with shared co-dominant markers, we developed a new "synchronized ES algorithm" which optimizes both complementary maps simultaneously with an additional restriction of shared order of co-dominant markers in both maps. The algorithm is represented in Fig. 1.

Materials and methods

Synchronized Evolution Strategy algorithm

Evolutionary Strategy foundations

Usually the optimization process of an objective function $f(x)$ with n real-value variables $x = (x_1, x_2,\ldots,x_n)$ can be represented as an evolution of the solution vector $\boldsymbol{x}\in R^n$. Evolution Strategy is a heuristic algorithm mimicking natural population processes. The numerical procedures in such optimization are based on simulation of mutation and reproduction, followed by selection of the fittest "genotypes", employing the obtained values of the optimization criterion.

Together with the Genetic Algorithm (Holland 1975) and Evolutionary Programming (Fogel 1992), Evolution Strategy forms the class of Evolutionary Algorithms (Nissen 1994). The evolutionary strategies were proposed in the 1970s (Rechenberg 1973; Schwefel 1977, 1987) to solve optimization problems with real-value variables (for a recent survey of search strategies for combinatorial problems see Muhlenbein et al. 1998). Evolution Strategies define the size of a population and rules for the selection process. Clearly, the multilocus ordering problem cannot be directly represented in terms of ES with real-value formulation. Combinatorial versions of ES differ from the real-value formulation by specific representation of the solution vector x and mutation mechanisms (Homberger and Gehring 1999; Mester 1999; Mester et al. 2003 revised). An ES algorithm employs the following steps:

(1) Create λ individuals ($\boldsymbol{x}^k$) of initial population P^0.
(2) Compute the fitness $f(\boldsymbol{x}^k)$, $k = 1,\ldots, \lambda$.
(3) If the optimization process is terminated, then stop.
(4) Select the $\mu \leq \lambda$ best individuals (selection phase).
(5) Create λ/μ offspring x^{k+1} of each of μ individuals by small variation (mutation phase).
(6) Return to Step 2.

On each iteration, the mutation operator (referred to hereafter as the *mutator*) changes the vector $\boldsymbol{x}^k$ thereby producing a new solution vector $\boldsymbol{x}^{k+1}$. Our version of the combinatorial ES algorithm includes several mutators that mutate the solution vector via removing and inserting β coordinates of $\boldsymbol{x}^k$ (Mester 1999, 2003 submitted). In the *mutation stage*, the chosen mutator $\mathbf{M}(\boldsymbol{x}^k)$ produces an offspring from the parent. If the first offspring appeared to surpass the parent, the same mutator is again applied to the new parent, and so on. If the offspring does not surpass the parent, then to generate the new offspring, the algorithm uses the next mutator. After mutation, the vector $\boldsymbol{x}^{k+1}$ is "improved" by standard combinatorial procedures of order $O(n^2)$:

(1) 2-Opt (Lin and Kernighan 1973),
(2) Or-forward and Or-backward (Or 1976),
(3) 1-interchange (Osman 1993).

A more detailed description of the ES algorithm for multipoint marker-ordering as a TSP problem is presented by Mester and co-authors (2003, revised).

Synchronized multipoint marker-ordering as a double TSP problem

For the non-synchronized version of the ordering problem, we consider n markers enumerated arbitrarily by n coordinates $x_i\in\boldsymbol{x}$ and for each n-1 marker pairs (x_i, x_j) a "distance" d_{ij}. As d_{ij}, either pair-wise recombination fractions r_{ij} or map distances c_{ij} (e.g. in Haldane or Kosambi metrics) will be employed. In combinatorial formulation, the solution (individual) can be represented as a vector $x = (x_1, x_2, \ldots, x_n)$ that consists of n ranked discrete coordinates (chromosomes) or as a directed graph $G(A, B)$ with a set of nodes $A = \{a_1, a_2, \ldots, a_n\}$ and a set of arcs $B = A \times A$, where node a_j, $j > 0$ represents the chromosome. The fitness function assigns to each of the $n(n - 1)/2$ arcs (a_i, a_j) [or pair of coordinates (x_i, x_j)] a non-negative d_{ij} cost of moving from element i to element j. The problem is symmetric if and only if, $d_{ij} = d_{ji}$ for all arcs. For optimization of a combinatorial problem, one needs to define such an order of the vector coordinates (or nodes) that will provide minimum total cost. Different criteria can be used to discriminate between competitive orders; for example, total distance measured as a sum of distances between consecutive adjacent markers. These criteria are based on a biologically reasonable assumption that the true order of a set of linked loci will be the one that minimizes the total length of the chromosomal map (see also Kirkpatrick et al. 1983; Press et al. 1986; Week and Lange 1987; Falk 1992; Schiex and Gaspin 1997).

In the problem with two subsets of dominant markers (with coupling-phase linkage within each of the sets and repulsion-phase between the sets) one needs to optimize two separate, albeit related, TSPs simultaneously under the condition of identical ordering of shared co-dominant markers. Therefore, in this case we optimize two complementary sets (vectors) of markers $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$, respectively, with the above restriction. Then, using the resulting two maps $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$, we restore the true marker ordering on the full map. In our synchronized model, the minimum of the sum of distances between adjacent markers on both vectors $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$ was applied as the optimization criterion (OC):

$$OC = \sum_{ij\in x1} d_{ij}\delta_{ij} + \sum_{ij\in x2} d_{ij}\delta_{ij} \rightarrow \min, \qquad (1)$$

where $\delta_{ij} = 0$ or $\delta_{ij} = 1$ represent (in the criterion) only $u \leq n - 1$ distances out of all $n(n - 1)/2$ pairwise distances on each set of markers respectively; $d_{ij}\ \delta_{ij} > 0$. Figure 2 illustrates the structure of the synchronized parallel Evolution Strategy algorithm.

The program for simulations was written in Visual Basic 6.0. Monte-Carlo testing experiments were conducted on a double-processor Pentium-III (2 × 800 Mhz). In order to compare different situations, the following coefficient of restoration quality [proximity between the "true" (simulated) and estimated orders] was employed:

$$K_r = (n-1)/\sum_{i=1}^{n-1} |x_i - x_{i+1}|, \qquad (2)$$

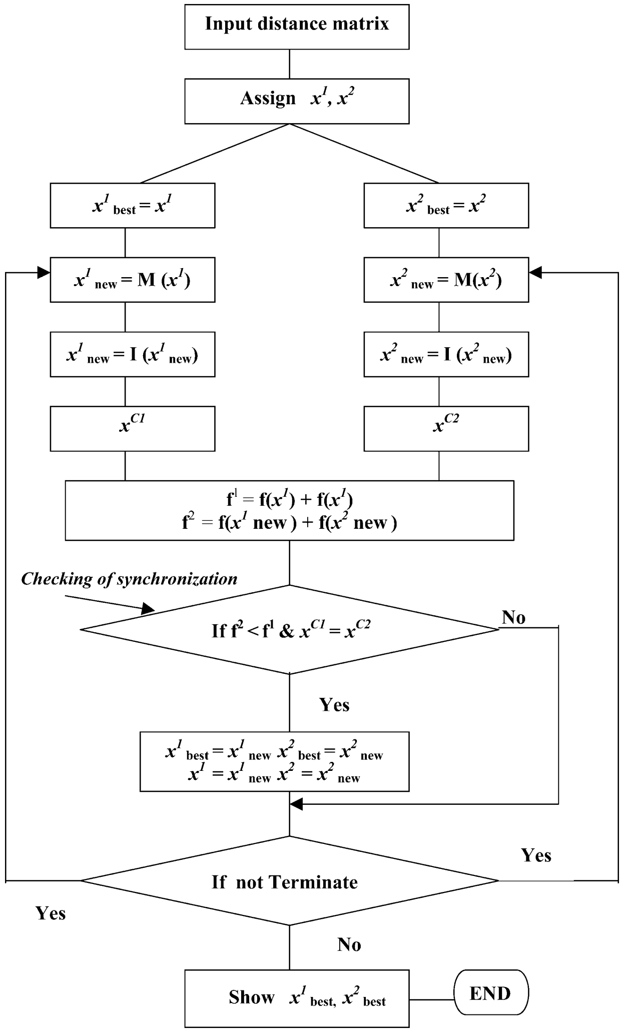

Input distance matrix
Assign x^1, x^2
$x^1_{best} = x^1$
$x^2_{best} = x^2$
$x^1_{new} = M(x^1)$
$x^2_{new} = M(x^2)$
$x^1_{new} = I(x^1_{new})$
$x^2_{new} = I(x^2_{new})$
x^{C1}
x^{C2}
$f^1 = f(x^1) + f(x^1)$
$f^2 = f(x^1 new) + f(x^2 new)$
Checking of synchronization
If $f^2 < f^1$ & $x^{C1} = x^{C2}$
No
Yes
$x^1_{best} = x^1_{new}$ $x^2_{best} = x^2_{new}$
$x^1 = x^1_{new}$ $x^2 = x^2_{new}$
If not Terminate
Yes
Yes
No
Show x^1_{best}, x^2_{best}
END

Fig. 2 Synchronized parallel Evolution Strategies algorithm with common restriction. In this scheme the following designations were accepted: $\boldsymbol{x}^1$, $\boldsymbol{x}^2$ are current solution vectors of each TSP respectively; $\boldsymbol{x}^1_{\text{new}}$, $\boldsymbol{x}^2_{\text{new}}$ are solution vectors after mutation M($\boldsymbol{x}$) and improving procedures I($\boldsymbol{x}$); $\boldsymbol{x}^1_{\text{best}}$, $\boldsymbol{x}^2_{\text{best}}$ are best solution vectors after last iteration; $\boldsymbol{x}^{C1}$, $\boldsymbol{x}^{C2}$ are common parts of solution vectors for last iteration; f^1, f^2 are fitness from current and new solutions respectively

where x_i is the digit code of the i-th marker in the currently ordered marker sequence.

Simulated data sets

The simulation algorithm repeatedly generated a single-chromosome mapping population F_2 for a chosen number of markers. The following are the numerical values (ranges) of the main parameters in the majority of experiments.

(1) The number of markers per chromosome: m = 50 or 100.
(2) Probability distributions for distances between adjacent markers: P [$3 < d$ (cM) < 5] = 0.8; P [$5 < d$ (cM) < 15] = 0.2, with even distribution within each of the two ranges, for 50 markers, and P [$3 < d$ (cM) < 5] = 0.8; P [$5 < d$ (cM) < 15] = 0.15, P [$15 < d$ (cM) < 35] = 0.05 for 100 markers.
(3) Proportions of co-dominant and the two types of dominant markers were 0.1, 0.35 and 0.55, respectively, for 50 markers (0.05, 0.475 and 0.475, respectively, for 100 markers).
(4) In case of arbitrary interference, the distribution of coincidence coefficient values was: P ($0 < c < 1$, positive interference) = 0.5, P ($1 < c < 2$, slight-to-moderate negative interference) = 0.25, and P ($2 < c < 20$, moderate-to-strong negative interference) = 0.25.
(5) Population size n = 200.
(6) The proportion of individuals employed in jackknife runs, 90%.

Results

Improved multilocus ordering of dominant markers by splitting into two independent maps

As indicated above, a high proportion of repulsion-phase dominant markers may become a serious obstacle in multilocus ordering; hence, the necessity of splitting the data into two sets, each containing all co-dominant markers and coupling-phase dominant markers only for each linkage group (Knapp 1995; Peng et al. 2000). To illustrate how dramatic this effect could be, we generated an F_2 data set with 50 co-dominant markers. Our multilocus ordering algorithm (Mester et al. 2003, revised) easily manages with such a situation, resulting in a high quality restoring of marker order. The results shown in Fig. 3a represent the neighbourhood matrix based on 100 jackknife runs (with re-sampling 90% of individuals at each run) employing the previously described jackknife procedure of correction of double recombinants (Mester et al. 2003, revised). Let us consider now the results of application of the same approach to data with a majority of dominant markers (using the same data set but artificially converting a part of the co-dominant markers into dominant markers). In accordance to the description in the previous section, let the two types of dominant markers comprise 55% and 35% of the total number of markers. The consequences proved to be dramatic: in contrast to nearly deterministic neighbouring in the first example, a rather large uncertainty of ordering is detected by the jackknifing in the second example (Fig. 3b).

Synchronized multilocus ordering of two maps with shared co-dominant markers

Why synchronizing?

The entire set of co-dominant markers and the two types of dominant markers can be subdivided into two sets, each containing all co-dominant markers and coupling-phase dominant markers. Then, the same ordering procedure applied independently to each of the two sets gives much more reliable results. However, it could not be guaranteed that the obtained maps would have identical orders for the shared co-dominant markers. In fact, it appeared that for the range of parameters employed in our

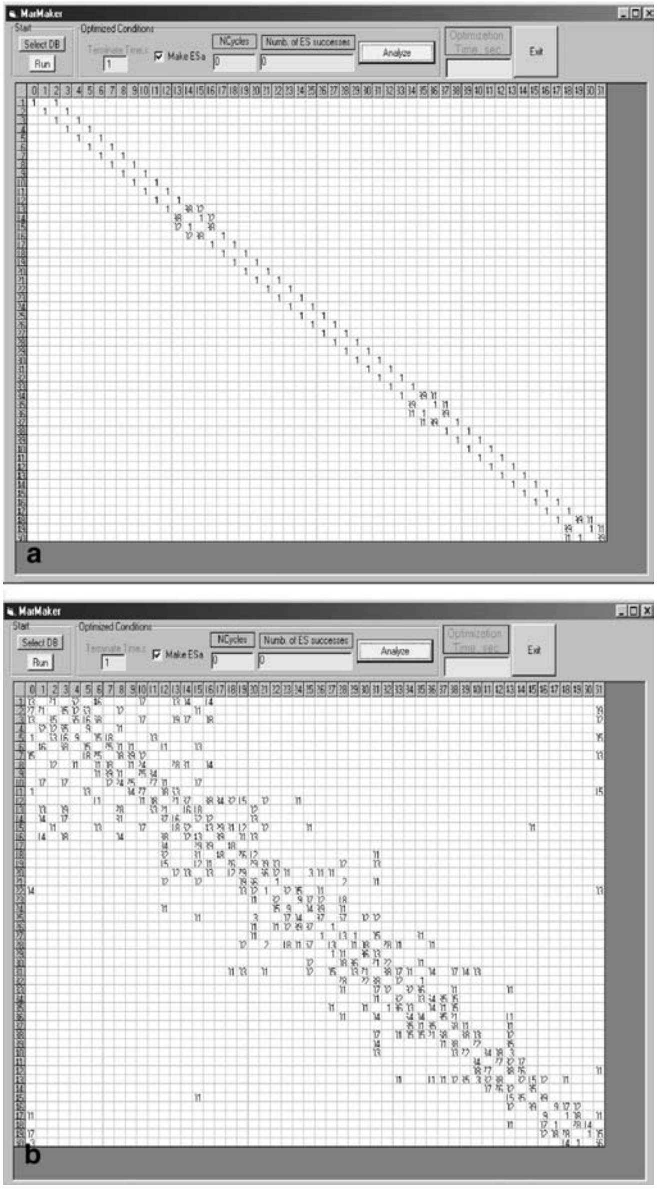


Fig. 3 Frequency neighborhood matrices for ordering of 50 markers. (**a**) co-dominant markers (**b**) co-dominant and dominant markers in repulsion phase

simulations (see previous section), such outcomes comprised about 10–25% (data not shown). This justifies the need for further sophistication of the TSP-based algorithms of multilocus ordering by including a new element, i.e. the *synchronized optimization* (see Fig. 2).

The proposed procedure of synchronized ordering of two parallel (complementary) maps aims to reduce the uncertainty caused by the presence of markers in the repulsion phase, as compared to the more traditional approaches of treating the entire linkage group as a whole. As was mentioned, in the majority (about 75–90%) of our Monte-Carlo simulated problems there was no need for synchronization at all, because the identical order of the shared co-dominant markers is maintained automatically. However, even for such cases, one cannot guarantee that this will always be the case when massive bootstrap or jackknife re-sampling iterations are applied for verification of the obtained ordering. Indeed, the matrices of pairwise recombination rates will change upon such re-sampling, resulting thereby in possible sporadic violation of the parallel ordering of the shared co-dominant markers. Morever, by splitting the map into two complementary maps we may, in fact, generate an additional problem: the appearance of gaps between nearby markers, if neighbouring dominant markers in linkage phase comprise a long chain. In such a case, synchronization becomes an especially helpful tool. Let us illustrate this aspect by two examples.

Example 1: the ordering of the markers of one of two complementary chains in one of our simulations was $DM_3CM_6DM_7DM_9\boldsymbol{DM_{14}CM_{19}}DM_{21}\ldots CM_{35}\ldots DM_{49}$, where DM_i, denotes the i^{th} (dominant) marker locus and CM_j denotes the j^{th} (co-dominant) marker locus. In our example, the rate of recombination R_{14-19} between the bold markers was 0.241 resulting in 89% of (correct) neighbourhoods between these loci when synchronized ordering was applied (using both complementary subsets of markers). The remaining 11% comprise all cases of excisions-transpositions with or without inversions (see below) that included markers from the interval 14–49. The importance of synchronization here can be shown by the results obtained after artificially converting CM_6 into DM_6 hence preventing its function as a shared (stabilizing) marker between the two subsets. In such a case, by applying the same algorithm, we obtained only 60% of (correct) neighbourhoods between markers 14 and 19, whereas miss-neighbourhoods with markers from the interval 14–49 are now three-fold (33%) compared to the previous case.

Example 2: in this example, one of two complementary subsets was $DM_2DM_4CM_5\ldots DM_{11}\ldots CM_{22}\ldots \boldsymbol{DM_{37}DM_{41}}CM_{42}DM_{44}DM_{46}CM_{47}DM_{49}DM_{50}$ with a gap between DM_{37} and DM_{41} (R_{37-41} was 0.295). Synchronized ordering resulted here in 65% of correct (or nearly correct) neighborhoods between DM_{37} and DM_{41} (or CM_{42}) loci due to the stabilizing effect of shared co-dominant markers (CM_{42} and CM_{47}). Indeed, by reducing the stabilization effect via converting $CM_{47} \rightarrow DM_{47}$ the proportion of correct neighbourhoods between DM_{37} and either DM_{41} or CM_{42} has reduced from 65% to 32%, correspondingly; this was accompanied by an increase of inversions involving DM_{37} and DM_{50} (or DM_{49}), from less than 10% to 38%.

Verification and correction of the complementary maps

In dealing with real data, one needs some tools to validate the obtained multilocus order, and it is hard to choose the solution from several (sometimes dozens) candidate solutions (like those provided by Mapmaker). To cope with this problem, some authors proposed computing-intensive procedures based on various combinations of

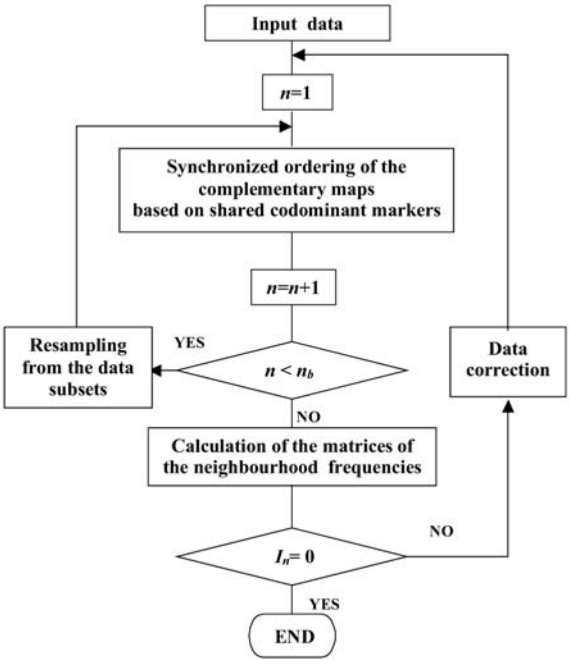

Input data
n=1
Synchronized ordering of the complementary maps based on shared codominant markers
n=n+1
Resampling from the data subsets
YES
$n < n_b$
NO
Data correction
Calculation of the matrices of the neighbourhood frequencies
NO
I_n= 0
YES
END

Fig. 4 Verification and correction of the complementary maps by jackknife analysis. n is a current jackknife number of n_j produced jackknives, I_n is termination criterion (1 – needs correction, 0 – does not need correction)

jackknifing and bootstrapping (Efron 1979; Wang et al. 1994; Liu 1998). To validate and correct the obtained pair of maps, the following analysis is conducted based on a series of jackknife runs (Fig. 4). In each run based on a sub-sample of individuals (e.g. 90%), we first order the markers in the two maps synchronically (as explained above in Fig. 2) and for each marker determine its two (left and right) neighbours. Then, for each marker, the frequency distribution of its closest left and right neighbours is calculated and the unstable neighbourhoods are detected using the entire set of generated jackknife runs. Note, that in dealing with real data, one may prefer to use the bootstrap technique instead of jackknifing, e.g. when the sample size is very limited.

Several factors can generate uncertain neighbourhoods among markers: (1) a small distance between markers, e.g. 1–3 cM for a sample size of about 200 or less, that in the presence of dominant markers on some bootstrap or jackknife iterations will give zero recombination rates; (2) a strong negative interference (Peng et al. 2000; Boyko et al. 2002; Esch and Weber 2002), which may violate the principle “the whole is larger than its parts” and result in local inversion of marker order; and (3) the presence of a big interval between nearby markers, derived from the division of the initial map into two complementary maps, if one-phase dominant markers comprised a contiguous chain. In the last case, one may observe map distortions of inversion-like type or excision-transposition with or without inversion.

Integrating the complementary maps

We consider here some simple ways of combining the verified complementary maps into an integrated map with subsequent testing of the reliability of the resulting map. This process should employ the marker ordering established on the previous steps, so that the integration step will not change the relative ordering of markers within each of the two subsets. However, the results of the integration step also need verification, in order to evaluate the reliability of the relative positioning of markers of the two complementary subsets in the final map.

Joining the two ordered marker subsets

In accordance with the algorithm of synchronized ordering, the complementary maps have identical orders of the co-dominant markers, although the summed-up lengths of intervals flanked by identical pairs of co-dominant markers may be different in the two complementary maps (because of different interior dominant markers). The combination of the dominant markers from the complementary maps is conducted for consequent pairs of intervals defined by neighbouring co-dominant markers. First, the recombination rates are transformed into genetic distances using some mapping function (e.g. Kosambi or Haldane). Consider an interior interval CM_{i}–CM_{i+1}. One of the complementary maps that displayed a shorter summed length of CM_i–$CM_{i\,+1}$ is normalized to have the same summed length of its subintervals for CM_i–CM_{i+1} as the first one. Then, we can combine the dominant markers of the two maps within the co-dominant markers flanking the considered interval. For the tail intervals we do not conduct scale transformations.

Verification, correction and testing the efficiency of the proposed procedure

Clearly, the reliability of the obtained joined map is the most important point of the entire algorithm, hence the need of tools for its verification (Fig. 5). Two major possibilities could be proposed to test the integrated map. First, using the established orders of the markers on the two complementary maps, one can employ bootstrap or jackknife re-sampling to evaluate the neighbourhood probabilities of the markers in the joined map. In addition, the decision can be based on the stability of the averaged ranks of the markers in an integrated map (if the stability is measured as an inverse of the variance of the ranks across the bootstrap runs). Markers that badly fit these two criteria should be removed from the map.

To illustrate the efficiency of the proposed algorithm of **s**ynchronized **o**rdering of split data sets with subsequent **i**ntegration and **v**erification of each step (the algorithm will be referred to as SOI-v) we generated two series of mapping examples: ten data sets with 50 markers and ten with 100 markers, with parameters

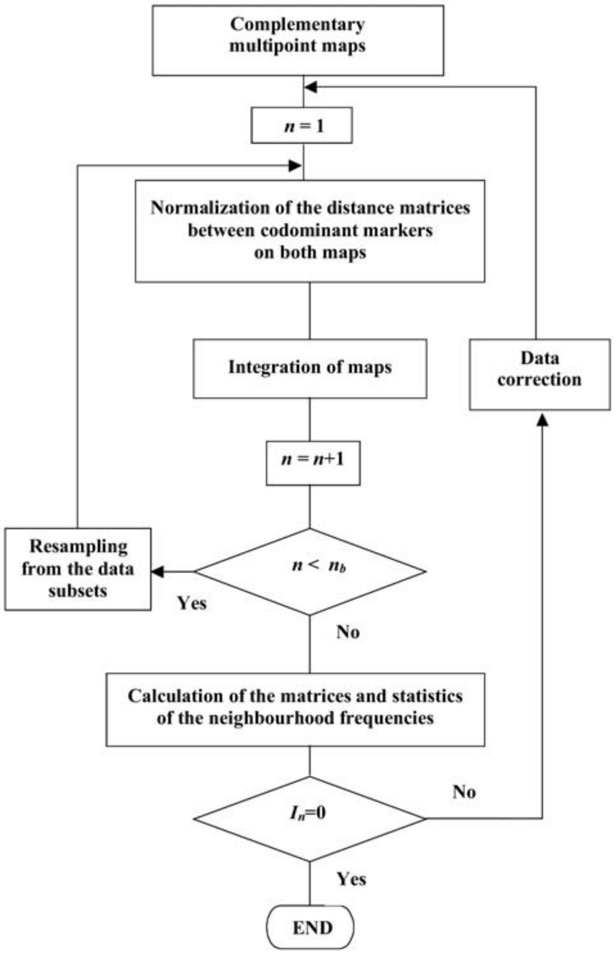

Complementary multipoint maps
n = 1
Normalization of the distance matrices between codominant markers on both maps
Integration of maps
n = n+1
Resampling from the data subsets
n < n_b
Yes
No
Data correction
Calculation of the matrices and statistics of the neighbourhood frequencies
I_n=0
No
Yes
END

Fig. 5 Verification and correction of the integrated map by jackknife analysis. n is a current jackknife number of n_j produced jackknives, I_n is the terminate parameter (1 – needs correction, 0 – does not need correction)

Table 1 Comparing the quality of multilocus mapping using the standard JoinMap algorithm and synchronized ordering with subsequent integration and verification algorithm (**SOI-v**). K_r is the coefficient of restoring marker ordering; the number of removed (unreliable) markers is shown in the brackets. Comparing the two ordering algorithms using ANOVA indicates highly significant superiority of SOI-v ($F_{1,36} = 49.33$, $p < 10^{-6}$) and a negative effect of marker density on the quality of ordering ($F_{1,18} = 7.06$, $p = 0.016$, for SOI-v, and $F_{1,18} = 18.59$, $p = 0.00042$, for JoinMap)

Set	Number of shared co-dominant markers	K_r using JoinMap	K_r using SOI-v
50 markers			
S1	7	0.816	0.960
S2	4	0.765	0.859
S3	3	0.830	0.830
S4	7	0.636	0.875
S5	5	0.662	0.859
S6	4	0.777	0.816
S7	8	0.765	0.890
S8	7	0.653	0.845
S9	5	0.753	0.859
S10	4	0.628	0.875
100 markers			
S1	4	0.303	0.660
S2	11	0.259	0.900
S3	8	0.414	0.779
S4	6	0.603	0.825 (4)
S5	3	0.607	0.825 (7)
S6	8	0.225	0.853 (3)
S7	3	0.520	0.839 (4)
S8	4	0.673	0.825 (3)
S9	4	0.700 (5)	0.687 (6)
S10	5	0.325 (5)	0.767 (18)

indicated above in the "Simulated data sets" section. For comparison, the same data sets have also been analyzed using the standard package JoinMap-3.0 (VanOoijen 2002). The quality of marker ordering in the resulting maps (Table 1) was defined by the coefficient of restoration ($0 \leq K_r \leq 1.0$), and proved to be very high for our algorithm. Nevertheless, the increase of marker density tended to reduce slightly the quality of ordering: $K_r = 0.887 \pm 0.013$ for maps with 50 markers and $K_r = 0.796 \pm 0.024$ for maps with 100 markers; $F_{1,18} = 7.06$, $p = 0.016$. For JoinMap, the effect of marker density was higher: $K_r = 0.729 \pm 0.024$ for maps with 50 markers and $K_r = 0.463 \pm 0.057$ for maps with 100 markers; $F_{1,18} = 18.59$, $p = 0.00042$.

An example of application on real data: mapping wheat chromosome 1B

We illustrate now the efficiency of the proposed approach using real data on a wheat mapping population (Peng et al. 2000). The experiment was performed on an F_2 progeny of a cross between wild emmer wheat *Triticum dicoccoides* (from the Mt. Hermon, Israel) and a *Triticum durum* cultivar, Langdon. The tetraploid *T. dicoccoides* is the progenitor of cultivated wheat; hence, the genetic dissection of quantitative trait differences between the wild species and the cultivated crop is of great interest from the viewpoint of domestication evolution (Peng et al. 2003). It is also important for the ever-increasing utilization of *T. dicoccoides* as a rich genetic resource for wheat improvement. The molecular markers (microsatellites and AFLPs) were scored on 150 F_2 individuals resulting in two versions of genetic maps each built on coupling phase dominant markers. Here we employ this data to illustrate the algorithms proposed in this article.

Let us start our example from the two maps build using Mapmaker (Peng et al. 2000) (in the list the co-dominant markers are denoted by c):

No.	*Chromosome* 1*BH*		*Chromosome* 1*BL*	
1.	Ws	(c)	Ws	(c)
2.	P56M50k		P55M56a	
3.	Xgwm550a	(c)	Xgwm550a	(c)
4.	Xgwm911	(c)	Xgwm911	(c)
5.	P55M53k		Xgwm264c	
6.	Xgwm273a	(c)	P57M51j	
7.	YrH52	(c)	P55M60v	
8.	Xgwm413	(c)	P55M60a	

Table (continued)

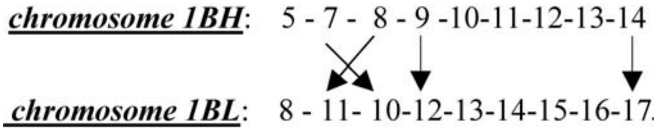
No.	*Chromosome* 1*BH*		*Chromosome* 1*BL*	
9.	Xgwm264a	(c)	Xgwm273a	(c)
10.	Xgwm11	(c)	YrH52	(c)
11.	Xgwm18	(c)	Xgwm413	(c)
12.	P56M50np	(c)	Xgwm264a	(c)
13.	Xgwm498a	(c)	Xgwm11	(c)
14.	P55M53b		Xgwm18	(c)
15.	P56M60ac		P56M50np	(c)
16.	P56M53m		Xgwm498a	(c)
17.	Xgwm403a		Xgwm131a	
18.	UBC199c		Xgwm153	
19.	P56M60k		Xgwm268	
20.	Xgwm124	(c)	Xgwm124	(c)
21.	UBC277a		UBC399	
22.	Xgwm131b		Xgwm259a	
23.	P57M52u		P56M52a	
24.	Xgwm140			

It is worth mentioning that Mapmaker suggests a few best solutions that differ in the likelihood level, and it is the user who should make the decision. The decision is "easy" when the best alternative is by orders of magnitude better than its competitors. But if the differences are not so high, it is difficult to make a justified choice without additional tools. Clearly, jackknife or bootstrap analysis would be useful but it is unpractical for Mapmaker (due to time limits) when the number of markers is higher than 12–15. To test the reliability of the obtained ordering, we employed to each of these maps the jackknifing procedure described in Mester et al. (2003, revised), with 90% of randomly chosen genotypes at each run. The obtained estimates of neighbourhood probabilities point to an indefinite region (marked by bold) in each of the two maps produced by Mapmaker. These regions include eight co-dominant markers (from Xgwm273a to Xgwm498a):

Chromosome 1*BH*												
No.	4	5	6	7	8	9	10	11	12	13	14	15
5	1			**1**								
6				0.17		**0.84**	0.92				0.07	
7		1	0.17		**0.83**							
8				**0.83**		1	0.14	0.03				
9			**0.84**		1		0.03	0.11	0.02			
10			0.92		0.14	0.03		**0.86**	0.03	0.02		
11					0.03	0.11	**0.86**		0.95	0.05		
12						0.02	0.03	0.95		1		
13							0.02	0.05	1		0.93	
14			0.07							0.93		1

Chromosome 1*BL*												
No.	7	8	9	10	11	12	13	14	15	16	17	18
8	1				0.99	0.01						
9				0.99		**0.6**	**0.4**	0.01				
10			0.99		1	0.1						
11		0.99		1		0.1						
12		0.01	**0.6**	0.01	0.01		0.98	**0.38**	0.01			
13			0.4			0.96		**0.62**				
14			0.01			**0.38**	**0.62**		0.9	0.09		
15						0.01		0.9		1	0.09	
16								0.09	1		0.91	
17									0.09	0.91		1

One may assume that the neighbourhood instability (that can be seen from the above matrices), derives from some non-concordance of markers-observed segregation (e.g. due to reading errors or negative interference, Mester et al. 2003, revised). This assumption is confirmed experimentally: by taking out marker Xgwm273a we obtain the orders with much higher neighbourhood stability (p = 1 for the H-version of the map and $p > 0.9$ for the L-version). It can be seen that independent analysis of the two groups resulted in non-identical relative ordering of the shared markers. This is exactly the reason why synchronous analysis keeping the same order of the shared markers is necessary.


chromosome 1BH: 5 - 7 - 8 - 9 -10-11-12-13-14
chromosome 1BL: 8 - 11- 10-12-13-14-15-16-17.

The synchronous ordering of the initial map versions gives the orders with a problematic region identical to that revealed in the foregoing analysis:

(**Table** see page 1110)

After exclusion marker Xgwm273a, the following orders were obtained (each with $p > 0.95$):

chromosome 1BH: 5 - 7 - 8 - 9 -10 -11 -12-13 -14
chromosome 1BL : 8- 10 -11-12 -13 -14 -15-16 -17.

Chromosome 1*BH*

No.	4	5	6	7	8	9	10	11	12	13	14
5	0.98		0.05	**0.74**	0.19						
6		0.05		**0.45**	0.05	**0.5**	**0.67**	**0.23**	0.02		
7		**0.74**	**0.45**		**0.81**						
8		0.19	0.05	**0.81**		**0.8**	0.12	0.03			
9	0.01		**0.50**		**0.80**		**0.53**	0.14	0.02		
10			**0.67**		0.12	**0.53**		**0.66**	0.02		
11			**0.23**		0.03	0.14	**0.66**		**0.89**	0.05	
12			0.02		0.02	0.02	**0.89**		1	0.05	
13							0.05	1		0.95	
14								0.05	0.95		1

Chromosome 1*BL*

No.	7	8	9	10	11	12	13	14	15	16	17	18
8	1		0.06	**0.73**	0.19	0.01						
9		0.06		**0.45**	0.05	**0.5**	**0.67**	**0.23**	0.02			
10		**0.73**	**0.45**		**0.81**							
11		0.19	0.05	**0.81**		**0.8**	0.12	0.03				
12		0.01	**0.5**		**0.8**		**0.53**	**0.14**	0.02			
13			0.67		0.12	**0.53**		**0.66**	0.02			
14			**0.23**		0.03	**0.14**	**0.66**		**0.89**	0.05		
15			0.02			0.02	0.02	**0.89**		1	0.05	
16								0.05	1		0.95	
17									0.05	0.95		1

The final step is integration of the two maps. It results in the following order:

WS, P55M56a, P56M50k, Xgwm550a, Xgwm911, Xgwm264c, P57M51j, P55M60v, **P55M60a, P55M53k**, YrH52, Xgwm413, Xgwm264a, Xgwm11, Xgwm18, P56M50np, Xgwm498a, **P55M53b, Xgwm131a**, P55M60ac, P56M53m, Xgwm403a, UBC199c, Xgwm153, P56M60k, Xgwm268, Xgwm124, UBC277a, UBC399, Xgwm131b, Xgwm259a, P57M52u, P56M52a, Xgwm140.

Two (bold) marker pairs (**P55M60a-P55M53k** and **P55M53b-Xgwm131a**) displayed high uncertainty in the integrated maps and are shown in the most plausible order: in the first case, with p = 0.75 it is P55M60v-**P55M60a-P55M53k** (and P55M60v-**P55M53k-P55M60a,** with p = 0.25), and in the second case, with p = 0.53 it is Xgwm498a-**P55M53b-Xgwm131a** (and Xgwm498a-**Xgwm131a-P55M53b**, with p = 0.47). For the reminder markers nearly all neighborhood probabilities were p = 0.98–1.0. To resolve the uncertainty of the foregoing two islands (if repeated DNA analysis is impossible), one can exclude one of the problematic markers. In our example, after excluding **P55M53k** from the first pair and **Xgwm131a** from the second pair, we obtained the final order with neighbour probabilities $p \geq 0.99$.

Discussion

This paper continues and adapts the proposed approach for multilocus ordering with an excess of dominant markers complicated by the presence of repulsion-phase linkages. We considered situations complicated by a high proportion of dominant markers in the repulsion phase and a high negative interference. As was shown in our previous analysis (Mester et al. 2003, revised), when all dominant markers were in the coupling phase, the proportion of dominant and co-dominant markers had no serious effect on the quality of marker ordering. A dramatically different result was obtained with dominant markers in the repulsion phase. It appeared that the higher the proportion of repulsion-phase markers, the lower the quality of multilocus ordering. High precision of ordering in the coupling-phase data and low precision in the repulsion-phase data justify splitting the data into two sets, each containing all co-dominant markers and coupling-phase dominant markers only, and generating two complementary maps for each linkage group (Knapp 1995; Peng et al. 2000).

However, this approach may encounter difficulties caused by local and global map disturbances affecting the ordering of co-dominant markers common for both maps, if the density of shared co-dominant markers is relatively low (e.g. in cases when co-dominant markers serve as rare anchors). In fact, the availability of shared co-dominant markers enables mutual control during multilocus ordering, which together with computing-intensive jackknife and bootstrap techniques (Efron 1979) may significantly improve the quality of the resulting map. To implement this idea of parallel ordering of two subsets of markers with common co-dominant markers, we developed a new "synchronized ES algorithm" which optimizes both complementary subsets simultaneously with an additional restriction of the shared order of co-dominant markers in both maps. Clearly, after synchronized optimization two complementary maps are obtained, and one more major step is needed to obtain an integrated map.

In this paper we proposed a multi-phase algorithm (**SOI-v**) which includes splitting the data set on two complementary subsets (**S**), synchronized marker-ordering optimization (**O**) of the subsets, integration of the two maps into one (**I**) and verification of the integrated map (**v**). For synchronized optimization we successfully applied the Evolution Strategy algorithm that was recently adopted for multilocus ordering by Mester and co-authors (2003, revised). The efficiency of the SOI-v algorithm was checked on two generated series of mapping examples: with 50 and 100 markers (Table 1). The quality of marker ordering in the resulting maps was defined by the coefficient of restoration ($0 \leq K_r \leq 1.0$), and proved to be very high for the proposed algorithm, especially at higher marker density (Table 1).

Clearly, the K_r indicator is impossible to calculate for real data. However, using bootstrap analysis, one can easily control the reliability of the obtained multipoint ordering and detect markers and/or marker scores that are problematic, and should be either removed or repeatedly genotyped. Likewise, the proposed tools allow revealing chromosomal segments that require saturation by additional markers in order to achieve the desired reliability, or vise versa; the regions where the marker order cannot be recovered unequivocally due to high marker density and population size was not sufficient to resolve tight linkages. This formulation of jackknife- or bootstrap-based diagnosis of uncertain local orders may be especially useful when one becomes interested in targeting specific chromosomal regions, e.g. in map-based cloning. Beside the importance in joint mapping of dominant and co-dominant markers, the proposed approach of synchronized multipoint ordering may be useful in integrated genetic and physical mapping, phylogenetic comparisons of linkage groups, comparisons of conserved physical orders and other related subjects of genome analysis (Yang and Womack 1998; Klein et al. 2000; Hall et al. 2001; Bourque and Pevzner 2002).

Acknowledgements This study was supported by the Israeli Ministry of Absorption, the United States-Israel Binational Science Foundation (grant 4556), the German-Israeli Cooperation Project (DIP project funded by the BMBF and supported by the BMBFs International Bureau at the DLR), and the Ancell-Teicher Research Foundation for Molecular Genetics and Evolution. We thank the JoinMap software developers for the possibility to use the package for comparative testing with our new algorithm. The useful comments and suggestions of anonymous referees and J. Snape are acknowledged with thanks.

References

Bailey NTJ (1961) Introduction to mathematical theory of genetic linkage. Calendon Press, Oxford

Bourque G, Pevzner PA (2002) Genome-scale evolution: reconstructing gene orders in the ancestral species. Genome Res 12:26–36

Boyko E, Kalendar R, Korzun V, Fellers J, Korol AB, Schulman AH, Gill BS (2002) A high-density cytogenetic map of the *Aegilops tauschii* genome incorporating retrotransposons and defense-related genes: insights into cereal chromosome structure and function. Plant Mol Biol 48:767–790

Buetow KN, Chakravarti A (1987) Multipoint gene mapping using seriation. Am J Hum Genet 41:189–201

Cone KC, McMullen MD, Bi IV, Davis GL, Yim YS, Gardiner JM, Polacco ML, Sanchez-Villeda H, Fang Z, Schroeder SG, Havermann SA, Bowers JE, Paterson AH, Soderlund CA, Engler FW, Wing RA, Coe EH Jr (2002) Genetic, physical, and informatics resources for maize: on the road to an integrated map. Plant Physiol 130:1598-1605

Curtis D, Gurling H (1993) A procedure for combining two-point lod scores into a summary multipoint map. Hum Hered 43:173–185

Efron B (1979) Bootstrap method: another look at the jackknife. Ann Stat 7:1–26

Ellis T (1997) Neighbour mapping as a method for ordering genetic markers. Genet Res 69:35–43

Esch E, Weber WE (2002) Investigation of crossover interference in barley (*Hordeum vulgare* L.) using the coefficient of coincidence. Theor Appl Genet 104:786–796

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

Fogel D (1992) Evolving artificial intelligence. PhD thesis, University of California, San-Diego

Hall D, Bhandarkar SH, Wang J (2001) ODS2: a multiplatform software application for creating integrated physical and genetic maps. Genetics 157:1045–1056

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 2,275 markers using a single F_2 population. Genetics 148:479–494

Holland J (1975) Adaptation in natural and artificial systems. MIT PRESS, Cambridge, USA

Homberger J, Gehring H (1999) Two evolutionary metaheuristics for a vehicle routing problem with time windows. INFOR 37:297–318

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

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

Klein PE, Klein RR, Cartinhour SW, Ulanch PE, Dong J, Obert JA, Morishige DT, Schlueter SD, Childs KL, Ale M, Mullet JE (2000) A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. Genome Res 10:789–807

Knapp SJ, Holloway JL, Bridges WC, Liu BH (1995) Mapping dominant markers using F_2 mating. Theor Appl Genet 91:74–81

Lathrop GM, Lloue1 JM (1984) Easy calculation of lod scores and genetic risks on small computers. Am J Hum Genet 36:469–465

Lathrop GM, Lloue1 JM, Julier C, Ott J (1985) Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. Am J Hum Genet 37:482–498

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

Lin S, Kernighan B (1973) An effective heuristic algorithm for the TSP. Operation Res 21:498–516

Liu BH (1998) Statistical genomics: linkage, mapping, and QTL analysis. CRC Press, New York

Menz MA, Klein RR, Mullet JE, Obert JA, Unruh NC, Klein PE (2002) A high-density genetic map of *Sorghum bicolor* (L.) Moench based on 2,926 AFLP, RFLP and SSR markers. Plant Mol Biol 48:483–499

Mester D (1999) The parallel algorithm for vehicle routing problem with time window restrictions. Scientific report, Minerva Optimization Center, Technion, Haifa, Israel

Mester D (2003) Evolutionary strategies algorithm for a large-scale vehicle routing problem with capacitate and time-windows restrictions. J Heuristics (revised)

Mester D, Ronin Y, Minkov D, Nevo E, Korol A (2003) Constructing large scale genetic maps using evolutionary strategy algorithm. Genetics (revised)

Muhlenbein H, Gorges-Scheuter MO, Kramer O (1998) Evolution algorithm in combinatorial optimization. Parallel Computing 7:65–85

Newell RW, Mott R, Beck S, Lehrach H (1995) Construction of genetic maps using distance geometry. Genomics 30:59–70

Nissen V (1994) Evolutionare algorithmen. Deutscher Universitats-Verlag, Wiesbaden

Olson JM, Boehnke M (1990) Monte Carlo comparison of preliminary methods of ordering multiple genetic loci. Am J Hum Genet 47:470–482

Or I (1976) Traveling salesman – type combinatorial problems and their relations to the logistics of regional blood banking. PhD thesis, Dept Industrial Engendering and Management Science, Northwestern University

Osman I (1993) Metastrategy simulated annealing and Tabu Search algorithm for VRP. Ann Operation Res 41:421–451

Parsons YM, Shaw KL (2002) Mapping unexplored genomes. A genetic linkage map of the Hawaiian cricket laupala. Genetics 162:1275–1282

Peng J, Korol A, Fahima T, Roder M, Ronin Y, Li Y, Nevo E (2000) Molecular genetic maps in wild emmer wheat, *Triticum dicoccoides*: genome-wide coverage, massive negative interference, and putative quasi-linkage. Genome Res 10:1509–1531

Peng J, Ronin Y, Fahima T, Röder MS, Li Y, Nevo E, Korol A (2003) Domestication QTLs in *Triticum dicoccoides*, the progenitor of wheat. Proc Natl Acad Sci USA 100:2489–2494

Pérez-Enciso M, Roussot O (2002) A method for computing identity by descent probabilities and quantitative trait loci mapping with dominant (AFLP) markers. Genetical Res 79:247–258

Press WT, Hannery BP, Teucolsky SA, Veterling WT (1986) Numerical recipes: the art of scientific computing. Cambridge University Press, London

Rechenberg I (1973) Evolutionstrategie. Fromman-Holzboog, Stutgart

Sall T, Nilsson NO (1994) The robustness of recombination frequency estimates in intercrosses with dominant markers. Genetics 137:589–596

Schwefel H-P (1977) Numeriche Optimierung von Computer-Modelen Mittels der Evolutions-strategie. Birkauser, Basel

Schwefel H-P (1987) Collective phenomena in evolutionary systems. Interne Berichte und Skripten, Fachbereich Informatic, University Dortmund

Schinex T, Gaspin C (1997) Carthagene: constructing and joining maximum-likelihood genetic maps. ISMB 5:258–267

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

Thompson EA (1984) Information gain in joint linkage analysis. IMA J Math Appl Med Biol 1:31–49

VanOoijen JW (2002) JoinMap 3.0: advanced computer software for the calculation of genetic linkage maps in experimental populations. j.w.vanooijen@plant.wag-ur.nl

Wang Y, Prade R, Griffith J, Timberlake W, Arnold J (1994) ODS_BOOTSTRAP: assessing the statistical reliability of physical maps by bootstrap re-sampling. CABIOS 10:625–634

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

Wilson S (1988) A major simplification in the preliminary ordering of linked loci. Genet Epidemiology 5:75–80

Whitaker D, Williams ER (2001) OutMap version 1.0. http://www.ffp.csiro.au/tigr/software/outmap/outmap.htm

Yang YP, Womack JE (1998) Parallel radiation hybrid mapping: a powerful tool for high-resolution genomic comparison. Genome Res 8:731–736