CARTHAGENE: Constructing and Joining Maximum Likelihood Genetic Maps*

Thomas Schiex & Christine Gaspin
Institut National de la Recherche Agronomique
Chemin de Borde Rouge BP 27
Castanet-Tolosan 31326 Cédex, France
{tschiex,gaspin}@toulouse.inra.fr

Abstract

Genetic mapping is an important step in the study of any organism. An accurate genetic map is extremely valuable for locating genes or more generally either qualitative or quantitative trait loci (QTL). This paper presents a new approach to two important problems in genetic mapping: automatically ordering markers to obtain a *multipoint maximum likelihood* map and building a multipoint maximum likelihood map using pooled data from several crosses.

The approach is embodied in an hybrid algorithm that mixes the statistical optimization algorithm EM with local search techniques which have been developed in the artificial intelligence and operations research communities. An efficient implementation of the EM algorithm provides maximum likelihood recombination fractions, while the local search techniques look for orders that maximize this maximum likelihood. The specificity of the approach lies in the neighborhood structure used in the local search algorithms which has been inspired by an analogy between the marker ordering problem and the famous traveling salesman problem.

The approach has been used to build joined maps for the wasp *Trichogramma brassicae* and on random pooled data sets. In both cases, it compares quite favorably with existing softwares as far as maximum likelihood is considered as a significant criteria.

Introduction

The aim of genetic mapping is to locate genetic markers at *loci* on the chromosomes. The specific content of a locus in a given chromosome is termed the *allele*. Given a locus on a chromosome, individuals of diploid species have two alleles, one on each of the corresponding chromosomes contributed by the parents. An individual is said to be *homozygous* at a locus if the two alleles are identical at this locus, else the individual is said to be *heterozygous*.

Each chromosome contributed by each parent is built, during the meiosis, by using sections of either member of each pair of chromosomes of the parent, changes in the chromosome used being called crossovers. At a given locus, there is a 50% chance of having either one of the parental allele. However, the "closer" two loci are on the chromosome, the higher the probability that both alleles on this chromosome will appear together on the contributed chromosome. Two loci (or genetic markers) are thus said to be *linked* if the parental allele combinations are preserved more often than would be expected by random choice. When a parental allele combination between two loci is *not* preserved, a recombination event is said to have occurred (this corresponds to an odd number of crossovers between the two loci).

The degree of linkage, or genetic distance, between two loci is a function of the frequency of recombinations. The measurement of genetic distance is expressed in Morgan (or more usually cM for centiMorgan) and is defined as the expected number of crossovers between the two loci on the chromosome.

In the simplest case, the process of building a genetic map goes as follows: starting from usually incomplete observations on the alleles of the loci of interest on several *related members* of families, one first builds *linkage groups*, composed of loci which are significantly linked together. For a given linkage group, a map is defined by locating each loci on a linear chromosome *i.e.*, by a linear order of the loci and a genetic distance between each adjacent pair of loci.

Given N genetic markers in a linkage group, the marker ordering problem is therefore to find an order of the markers that best respects the available evidence. The task is difficult in two aspects: it is not always obvious to say when an order best respects the available evidence and the number of possible orders becomes rapidly tremendous since $\frac{N!}{2}$ different orders exist.

Existing approaches to the markers ordering problem varies along two aspects: the criteria used to qualify what the best map is and the algorithmic machinery used to actually find the order that maximizes the criteria. Several packages, such as GMendel (Echt, Knapp, & Liu 1992) or Rapid chain delineation (Doerge 1996), Seriation (Buetow & Chakravarti 1987), *etc.* exploit two point measures

(measures taking only into account two markers simultaneously). The criteria defined can be computed very efficiently under a given order which enables the use of sophisticated local search techniques (such as simulated annealing in GMendel). But these criteria may be meaningless for data sets which contain a large part of missing (some two-point estimations may be indefinite). Other packages consider that multipoint maximum likelihood (that exploits the data on all markers simultaneously) is the criteria that defines the best order. Examples of such packages are MAPMAKER (Lander *et al.* 1987), LINKAGE (Lathrop *et al.* 1985)... The criteria is theoretically firmly grounded, exploits all the available evidence and can therefore perform better than the previous criteria when several missing exist. However, its computation may be much more expensive and order optimization techniques have been limited to the use of simple heuristics or enumerative search.

CARTHAGENE combines the multipoint maximum likelihood criteria with local search techniques. The choice of maximum likelihood as the criteria limits the approach to reasonably simple pedigree (only backcross data and recombinant inbred lines are allowed actually, an extension to intercross is currently being worked out) but makes it possible to tackle data sets with several missing in the best possible way. The use of adequate local search techniques makes it possible to enlarge the maximum number of markers that could be previously handled using exhaustive search while still offering most of the usual services, for example a collection of the maps whose likelihood lies in the vicinity of the best map's likelihood.

The paper goes as follows: the first section introduces the notion of maximum likelihood orders. This section does not contain anything essentially original ; its aim is simply to bring to light the strong theoretical connection that exists between the traveling salesman problem and the *multipoint maximum likelihood* marker ordering problem in the simple case of backcross with no missing data. Then, we rapidly introduce the local search techniques and the structure of the neighborhood used in CARTHAGENE, which was inspired by the previous connection and show how these techniques can be extended to tackle data sets with missing data and to build joined maps. We finally present some results that confirm pragmatically the interest of the approach.

Maximum Likelihood Orders

Given a probabilistic model of recombination for a given family structure (or pedigree), a genetic map of a linkage group and the set of available observations on the alleles of the loci of the linkage group, one can define the probability that the observations may have occurred given the map. This is termed the likelihood of the map. The likelihood in itself is not interesting, it is only meaningful when compared to the likelihood of other maps. In the sequel, we consider, as usual, that the interesting maps are the maps with a maximum likelihood *i.e.*, which best explain the observations. To later justify our approach for the combinatorial aspect of the genetic mapping problem, we first introduce some basic definitions.

The simplest pedigree, which is routinely used for plants and animals is the *backcross* structure defined by the crossing of one homozygous parent with an heterozygous one. Since one parent is homozygous, the alleles of one member of each chromosome pair of any descendant is known. The alleles of the other member of each pair will be used to track down recombination events in the heterozygous parent.

In this section, we assume that so-called phases are known (*i.e.*, it is known which allele appears on which member of the chromosome pairs) and that no data is missing. The alleles carried on one member of the pairs of the heterozygous individual are denoted 0, the others are denoted 1. In this case, the observation on a descendant are simply defined by the alleles carried on the member of each chromosome pair which is contributed by the heterozygous parent.

Assume we have N loci and K descendants. For a given loci ℓ and a given descendant i, the allele contributed by the heterozygous parent, which can be either 0 or 1 will be denoted X^i_ℓ. This defines the observations.

A genetic map is defined by a linear order on the loci *i.e.*, a one-to-one mapping ϕ from the set $\{1,\ldots,N\}$ to $\{1,\ldots,N\}$ and a probability of recombination between two adjacent loci $\phi(\ell)$ and $\phi(\ell+1)$, denoted $\hat\theta_{\ell,\ell+1}$. A recombination (resp. non-recombination) event occurs when the allele contributed by the heterozygous parent on a given individual changes for two adjacent markers *i.e.*, when $|X^i_{\phi(\ell)} - X^i_{\phi(\ell+1)}|$ is equal to 1 (resp. 0). If we assume that there is no interference *i.e.*, that recombination events occur independently in each interval, the probability of the observations given the map is simply obtained by multiplying the probabilities of all the events observed which yields the following formula:

$$\prod_{\ell=1}^{\ell=N-1}\prod_{i=1}^{i=K}\Big[(1-\hat\theta_{\ell,\ell+1}).(1-|X^i_{\phi(\ell)}-X^i_{\phi(\ell+1)}|)+ \hat\theta_{\ell,\ell+1}.|X^i_{\phi(\ell)}-X^i_{\phi(\ell+1)}|\Big]$$

Obviously, a maximum likelihood map is also a maximum log-likelihood map and the previous formula can be simply transformed by taking the logarithm.

$$\sum_{\ell=1}^{\ell=N-1}\sum_{i=1}^{i=K}\log\Big[(1-\hat\theta_{\ell,\ell+1}).(1-|X^i_{\phi(\ell)}-X^i_{\phi(\ell+1)}|)+ \hat\theta_{\ell,\ell+1}.|X^i_{\phi(\ell)}-X^i_{\phi(\ell+1)}|\Big]$$

We are looking for maximum log-likelihood maps. For a given order of the loci, we can easily compute the probabilities θ^* that maximize the log-likelihood. Looking at first and second-order derivatives of the previous formula, the $\hat{\theta}$ that maximize the log-likelihood can be easily obtained[1]:

$$\hat{\theta}^*_{\ell,\ell+1} = \frac{\sum_{i=1}^{i=K} |X^i_{\phi(\ell)} - X^i_{\phi(\ell+1)}|}{K}$$

So, in this case, when all alleles are known, and as far as recombination fractions are considered, multipoint likelihood computation comes down to simple two-points likelihood computation. The log-likelihood for optimal recombination fractions can therefore be rewritten:

$$\sum_{\ell=1}^{\ell=N-1} \underbrace{K.\left[\hat{\theta}^*_{\ell,\ell+1}\log(\hat{\theta}^*_{\ell,\ell+1}) + (1-\hat{\theta}^*_{\ell,\ell+1})\log(1-\hat{\theta}^*_{\ell,\ell+1})\right]}_{\text{elementary contribution to log-likelihood}}$$

The maximum log-likelihood of an order is equal to a sum of elementary contributions which depend only on two loci. One can therefore precompute all these numbers for all pairs of loci and the problem of finding an order that maximizes this multipoint maximum likelihood is in essence identical to problems defined using criteria such as the sum of two-points estimations (eg. SAR for sum of adjacent recombinations or SAL for sum of adjacent LODs). All these problems are actually obvious instances of the *symmetric wandering salesman problem* (a variant of the famous *symmetric traveling salesman problem*): given n cities and the distances between each pair of cities, find a path that goes once through each city and that minimizes the overall distance. The choice of the first and last cities in the path is free. One can simply associate one imaginary city to each marker, and define as the distance between two cities the opposite of the elementary contribution to the log-likelihood defined by the corresponding pair of markers. We would like to acknowledge the fact that the similarity between marker ordering and the TSP for two-points approaches is mentioned in (Liu 1995).

This connection is interesting in several aspects: the WSP is known to be an NP-hard problem and this shows that the marker ordering problem may be difficult in some cases (algorithm theory tells us that the tremendous number of existing orders is not sufficient to conclude this). More interestingly, all the techniques which have been developed for the TSP, and which can easily be adapted to the WSP, can also be applied here.

[1] This is obtained by solving the simple equations stating that first-order derivatives are equal to 0. At this point, one can further notice that the matrix of second order derivatives is diagonal negative on the domain of optimization, which shows that this point is a maximum.

Tackling the ordering problem

Now that the maximum likelihood ordering problem is known to be equivalent to the WSP, the considerable experience that has been accumulated on this problem can be exploited to provide markers ordering efficient algorithms.

Branch and Cut is probably the best known optimal procedure for the TSP. It has been used to solve instances with several thousand cities to optimality (Applegate *et al.* 1995). But Branch and Cut finely exploits the mathematical properties of the salesman problem and could probably not be extended to the more complex case of maximum likelihood with missing data. However, it could be used to tackle the problems defined by two-points measures such as SAR and SAL.

When Branch and Cut (or Branch and Bound) cannot be used, an alternative can be found in heuristics. By heuristics, we mean procedures that experimentally work in acceptable time and usually find good quality solutions. In no case heuristics allow to compute solutions with a guarantee of optimality. A number of heuristics algorithms have been experimented on the TSP and could be directly used for the markers ordering problem when the criteria reduces to a sum of elementary contributions.

Among TSP dedicated heuristics, some are known to work rapidly. For instance, the nearest neighbor algorithm begins with a partial tour consisting of a single city, adds the other cities one after the other by selecting the nearest next city not already in the tour. The algorithm terminates when all cities are in the tour. Several existing programs associate the nearest neighbor algorithm with reshuffling procedures to tackle the markers ordering problem (Stam 1993; Doerge 1996). The double minimum spanning tree, Greedy, Christofides *etc.* are other examples of such heuristics which could be used as well (Johnson 1990).

Local search is a more general heuristic which is largely used in difficult optimization problems. Given one initial feasible solution, the algorithm generates a sequence of solutions by repeatedly searching the so-called *neighborhood* of the current solution for a configuration that will become the new current solution. The configuration chosen is usually the configuration which maximizes the criteria in the neighborhood. A local optimum is a point which neighborhood contains only configurations which are worse than the current solution. Most local search procedures include a special mechanism that enables them to break out of local optima.

For markers ordering, a solution is an order of the markers, and the criteria is its maximum likelihood. CARTHAGENE strongly relies on the power of the *2-change* neighborhood, a successful well-known neighborhood structure introduced in (Lin & Kernighan 1973) to tackle the TSP. Adapted to the WSP, the 2-change neighborhood of a map is the set of all maps obtained by an in-

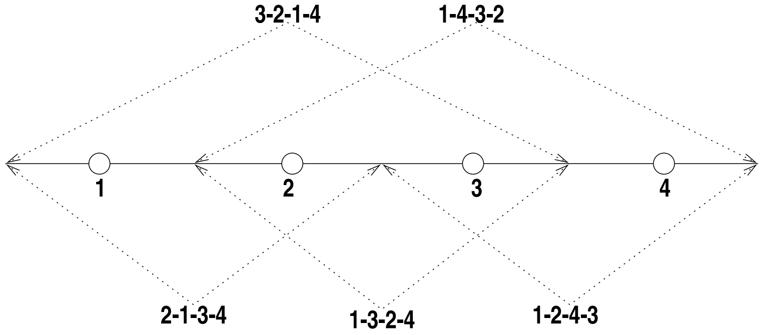

Figure 1: The five maps in the neighborhood of an initial map with four markers labelled 1, 2, 3 and 4. Each map is obtained by flipping a subsection of the initial map.

version of a subsection of the map. Thus, for N markers, the neighborhood has a size of $\frac{N.(N-1)}{2} - 1$. This is illustrated in figure 1 in the case of 4 markers.

In CARTHAGENE, this neighborhood is exploited by two different local search procedures:

- Tabu search (TS (Glover 1989; 1990)) repeatedly scans the current neighborhood, selecting the best neighbor to be the new solution. To avoid being stucked in local optima, the content of the neighborhood of the current solution is influenced by a memory mechanism which may forbid some moves (which are said to be tabu) in the neighborhood. In CARTHAGENE, the tabu moves are the recent moves (*i.e.*, subsections of the map which have been recently inverted). The precise definition of "being recent or not" varies stochastically during search as advocated by (Taillard 1991). A tabu move may eventually be chosen if it leads to a map which improves the best likelihood known (this is called "aspiration" in tabu terminology).

- Genetic algorithms (GA, (Goldberg 1989)) are based on an analogy with the genetic structure and behavior of chromosomes within a population of individuals. Individuals represent potential solutions to a given problem. A fitness score is associated to each individual and represents its adaptation ability. The algorithm makes the population of individuals evolve maintaining both diversity and favoring the existence of best individuals. Starting from an initial population, a new generation is created by randomly applying mutation and by crossing pairs of individuals (favoring crosses of good individuals). The hope is that the population will evolve towards one which contains optimal individuals w.r.t. the fitness. In CARTHAGENE, each individual represents one genetic map (an ordering of markers). The crossover operator computes two offspring individuals, I_1 and I_2 from two parents P_1 and P_2. The parents are cut into three sections by selecting randomly two markers. The middle section of P_1 is copied into the corresponding position of I_1, the rest of I_1 being filled in with values taken in order from the third, first and second section of $P2$, skipping values that have already been copied from the first parent. $I_2$is computed in the same way by reversing the parents. The mutation operator selects two markers and exchanges them. The evaluation of the fitness score of an individual is more original and complex because it will usually modify the individual under evaluation. Indeed, the 2-change neighborhood of the individual is searched for the order that maximizes the maximum likelihood. If an order is found which is better than the preceeding one, search for a better order is carried on in the 2-change neighborhood of this order. This process is repeated until no better order is found in the current neiborhood: a local optimum has been reached. When the evaluation stops, the individual is replaced by this local optimum and its likelihood is used as the fitness. With this approach, one skips from a local optimum to another throughout the crossover and mutation operators.

Finally, CARTHAGENE maintains a set of fixed size containing the S best different maps encountered during the search. A hash table (Cormen, Leiserson, & Rivest 1990) is used to efficiently test if the map is already present in the set, a heap structure is used to efficiently manage insertions/deletions (Cormen, Leiserson, & Rivest 1990). At the end of the search, the user can browse this set and check for the existence of other maps whose likelihood is close to the best map's likelihood in order to get an idea of how strongly the best map is supported.

Missing Data and Joined Maps

We have seen that two-points and multipoint approaches coincide in the specific case of backcross with no missing data. The underlying assumption of our approach to the more realistic case of missing data is that the problem is still *closely related in structure to the WSP* and that the previous techniques will probably work fairly well in presence of missing data *i.e.*, when two-points and multipoint estimations may differ and the previous connection with WSP is theoretically lost.

When all the alleles X_ℓ^i are not known, there is no short analytical form that can be exploited to compute the $\hat{\theta}$ that will maximize the likelihood and one usually relies on numerical algorithms. The statistical iterative optimization algorithm EM (Expectation/Maximization (Dempster, Laird, & Rubin 1977)) is well adapted to simple pedigree. It is currently used in MAPMAKER (Lander & Green 1987; Lander *et al.* 1987). The input of EM is an initial map and the output will be a map using the same ordering of loci, but with maximum likelihood recombination fractions (the likelihood of the map is also produced as a side effect). The algorithm works by iteratively going through two steps until the log-likelihood does not increase more than a given tolerance.

1. an **expectation** step where the expected number of recombination events between each pair of adjacent markers is computed, using the current values of the $\hat{\theta}$. This is done using a dynamic programming algorithm that exploits the linear structure of the problem. This also provides the log-likelihood of the current map.

2. a **maximization** step simply updates the $\hat{\theta}$ in each interval by dividing the expected number of recombination events in this interval (computed in the previous step) by the total number of individuals.

When EM is converged, the loglikelihood obtained indicates the quality of the order used and the extension of the previous discrete optimization algorithms is straightforward. Instead of computing the likelihood of an order by adding elementary contributions, we directly use EM. For backcross data, each iteration of EM is simply in $O(N.K)$. Rather than using the existing EM implementation of MAPMAKER, we decided to use our own implementation because several calls to this routine will be needed. It appears that this implementation, which is finely tuned to handle backcross data, is much faster than MAPMAKER implementation (we have observed ratio of more than 100 in speed on several data sets for a same quality of convergence, using the same machine, language and compiler. This ratio increases as the size of the data set increases). This efficiency is probably one of the nice properties of CAR$^{\text{T}}_{\text{H}}$AGENE but it is not obvious that such improvements will still be possible on pedigree such as F2 intercross data.

The power of local search combined with the generality of multipoint likelihood makes it possible to tackle data sets with a large part of missing data and more specifically to build joined maps using data from several backcross populations. The problem of building joined maps is specifically addressed by JOINMAP (Stam 1993) and also, to some extent, by GMendel. CAR$^{\text{T}}_{\text{H}}$AGENE is more limited in its scope than JOINMAP which can work on several data sets of different nature, but it uses a more powerful algorithmic machinery to provide maximum likelihood joined maps.

Merging two data sets is simply done by building a new data set that involves all the markers and individuals that appeared in the two original data sets. When an allelic information for a given marker is not available in one of the original data set, it is simply considered as missing in the new one. This yields data sets where missing are not randomly distributed on all individuals and markers but where large blocks of missing are introduced. When all crosses do not involve exactly the same set of markers (which is the usual case), some two-point estimations are simply impossible and it becomes essential to rely on multipoint likelihoods.

Because of the large amount of missing in such joined data sets, simple heuristic procedures usually fail to find an optimal map on such data, and this, whatever criteria they rely on. The local search techniques used in CAR$^{\text{T}}_{\text{H}}$AGENE usually allow to find maximum likelihood maps even in these difficult conditions. Here, the ability to also produce a set of maps in the vicinity of the optimum is highly valuable since it allows one to qualitatively assess how strongly the best order is supported by the available data.

Experiments

CAR$^{\text{T}}_{\text{H}}$AGENE and JOINMAP (rel. 1.4) have been applied both on simulated and real backcross-like data involving multiple populations. CAR$^{\text{T}}_{\text{H}}$AGENE has been used to build individual (intra-population) and joined maps and is compared with JOINMAP on this last problem.

Simulated data description

Several individual data sets to be joined must be generated for one common map. In our experiments, we considered the problem of joining two data sets. Let N^i the number of markers in each individual (also called intra-population) data set. Let N^c the number of markers common to the two individual data sets. For given values of N^i and N^c, we first build a map with $N = 2 \times N^i - N^c$ markers. The map is build by evenly distributing the N markers on a 200 cM chromosome. This map is called the "original map" and the order of the N markers in this map is called the "original order". For each pair of adjacent markers in this original map, the recombination probability between the two markers is computed using the inverse of Haldane's mapping function (Ott 1991).

The markers that are informative in each individual map are chosen as follows: first a set of N^c markers is selected randomly from the N markers. The set of $N - N^c$ remaining markers is randomly split in two sets with the same cardinality. These two sets, each merged with the set of the N^c common markers define the set of the markers which are informative in each individual data set.

Using the map built, a random individual data set for K individuals and for the subset of the N^i informative markers in the data set can be generated as follows:

1. for each individual, we first randomly choose the allele for the first marker in the set $\{0, 1\}$ with a uniform probability 0.5. Then for each successive marker, the allele is flipped with a probability equal to the recombination probability. This process is repeated K times to obtain a data set that represents allelic information for K individuals;

2. to incorporate "missing data", each of the allele generated in this way is, with probability p, deleted and replaced by a missing measure;

3. then, all the information on allele of markers which are not in the subset of informative markers for the individual map is deleted.

4. Finally, the order of the markers in each of these data sets is scrambled using a random permutation to avoid any bias.

Each individual data set is built using this process. The number of markers in each individual map was either 10 or 15. The number of common markers N^c was either 1, 2, 5 or 10. Globally, the global number of markers N in the initial original map varies from 10 ($N^i = 10, N^c = 10$) to 29 ($N^i = 15, N^c = 1$). The number of individuals in each individual data set is either 25, 50, 100 or 250. The probability p for an allele to be missing varies from 0 to 20% by 5% step. One hundred joining problems are solved for each combination of these parameters. Remember that these data sets represent an ideal case with no error.

TS runs with the following stopping criteria: it will stop when the number of iterations which have not improved the likelihood of the best solution is equal to twice the number of markers in the data set. This number of iterations corresponds to a choice of efficiency, needed because of the number of tests. On a Pentium Pro 200, the TS based algorithm run in less than $0.08''$ for intra-population data sets with 10 markers and 25 individuals, to roughly $3'$ for joined data sets with 29 markers and 250 individuals.

GA runs with the following stopping criteria : it will stop as soon as two generations of individuals produce the same best individual (the same order) or the maximum likelihood has not been improved (two different individuals may have the same maximum likelihood). At each generation, the number of individuals is $\frac{N}{2}$. On the same machine, the AG based algorithm run in less than $0.02''$ for intra-population data sets with 10 markers and 25 individuals, to roughly $15'$ for joined data sets with 29 markers and 250 individuals. Both approaches gave results of similar quality and the measures are given only in the case of TS.

Results on intra-population data sets

Figure 2 is dedicated to intra-populations data-sets with 10 markers. The curves give the percentage of original orders found by CARTHAGENE (by original orders we mean same order as the order used to generate the data-sets). Four number of individuals are reported: 25, 50, 100 and 250. It appears that 250 or even 100 individuals seems to be largely sufficient to find the good order with a reasonnable probability, even when missing data is present.

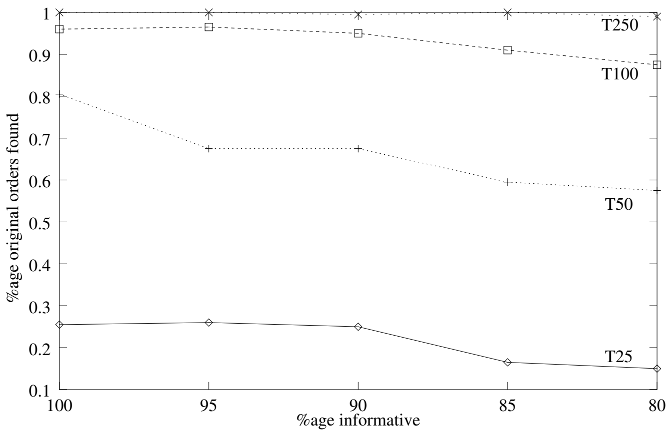

Figure 2: Percentage of original orders found (intra-pop).

Since CARTHAGENE maintains a set of the $S = 31$ best maps found, we measured the number of maps found in this set such that their loglikelihood lies within -3.0 of the likelihood of the best map found (Figure 3). The results show that a high percentage of success is "correlated" with a small number of maps found within -3.0 and therefore the number of maps found in the vicinity of the optimum gives an idea of how strongly the order found is supported by the data.

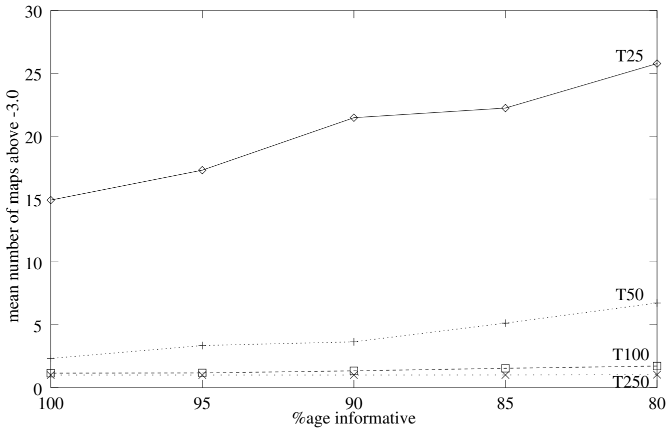

Figure 3: Mean cardinality of the set of maps with a loglike ratio above -3.0 w.r.t. to best map (intra-pop).

We do not report a last set of curves: the percentage of maps found with a loglikelihood equal to or better than the loglikelihood of the original map. This gives an idea of the effectiveness of the optimization algorithm. It was always equal to 100%.

Results on joined data sets

Figures 4 and 5 are dedicated to joined data-sets using two sets of 10 markers with 5 markers in common. The curves have the same meaning as in the previous case. These curves show that, even in the relatively nice case of 5 markers in common, building a merged map is more difficult. This is quite natural given the large number of missing in merged data sets and the increased number of markers.

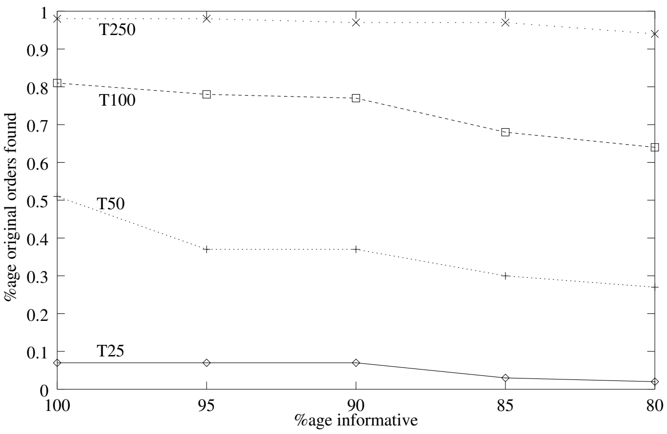

Figure 4: Percentage of original orders found (joined maps).

This is again reflected in the mean number of maps within -3.0 which appears pragmatically as an excellent indicator of how strongly the map found is supported by the data.

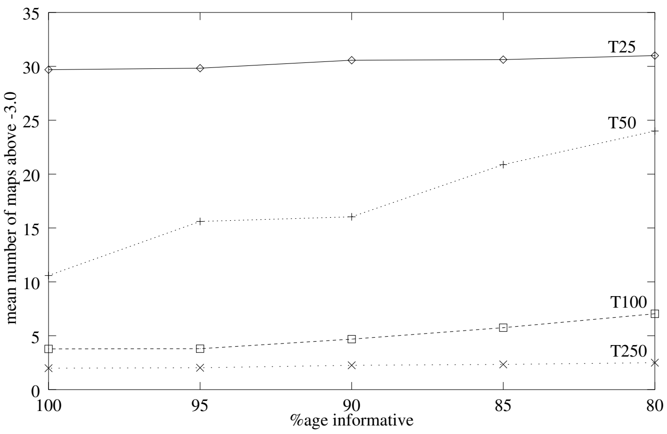

Figure 5: Mean cardinality of the set of maps with a loglike ratio above -3.0 w.r.t. to best map (joined maps).

We do not report the percentage of map found with a loglikelihood better than the loglikelihood of the original map which was again systematically equal to 100%.

Comparison with JoinMap on joined data sets

We now compare the results obtained with JoinMap (dotted curves) with the results obtained with CarTHaGene (filled curves). The joined data sets were obtained using 2 sets of 10 markers and a number of individuals of 250. The number of common markers considered are 1, 2, 5 and 10. The figure 6 represents the percentage of original orders found. CarTHaGene gives better results than JoinMap, which could be explained both by the criteria used and by the power of the local search algorithm of CarTHaGene. In the case of 10 markers in common, the joined data set has also 10 markers and no extra missing introduced because of the joining process: both systems work perfectly in this case (we have 500 individuals and only 10 markers).

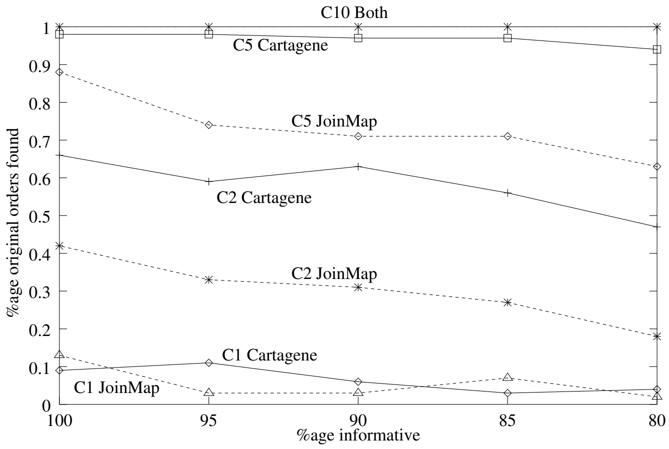

Figure 6: Percentage of original orders found.

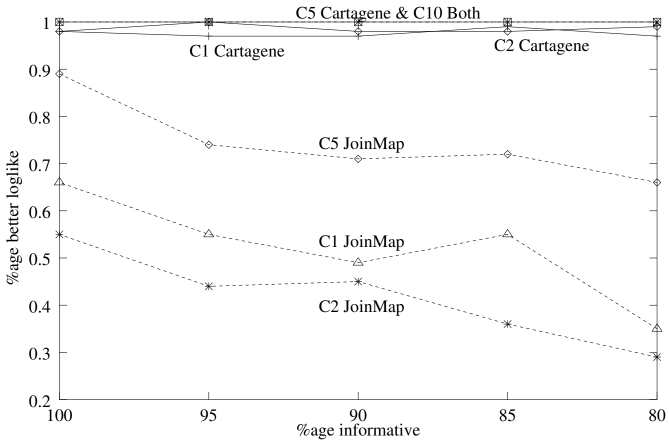

Figure 7: Percentage of map with a loglike better than the original map loglike.

In figure 7 the curves report the percentage of orders found whose loglikelihood was larger or equal to the loglikelihood of the original order used for generating the data (the maps produced by JOINMAP were optimized using EM for the comparison). The comparison is naturally advantageous for CARTHAGENE which precisely optimizes the loglikelihood. Note that here, a map with a likelihood better than the original map is not systematically found when few markers are common markers (down to 97% in the worst case).

Real data

The real data consists of backcross-like data obtained from segregation of RAPD markers in three F2 populations of haploid male progenies of several F1 female of the *Trichogramma brassicae* genome. Full results are described in (Laurent *et al.* 1997). Each intra-population involved around 90 individuals. Five groups were defined. Joined maps obtained for the group II using CARTHAGENE and JOINMAP are given at the end of the paper in Figure 8. One should note that the distances obtained using JOINMAP in (1) are not maximum likelihood distances and were therefore optimized using EM (2). Finally, the order obtained using JOINMAP is more than 14 time less likely than the map obtained using CARTHAGENE. For other groups, the difference in term of loglikelihood between orders obtained using CARTHAGENE and JOINMAP varied from 0 to 16, always in favor of CARTHAGENE. GA and TS gave the same map.

Discussion

CARTHAGENE efficiently builds multipoint maximum likelihood maps in the simple case of backcross data (this essentially dedicates CARTHAGENE to plant and animal studies). Its ability to tackle data sets with a large number of missing allows the construction of multipoint maximum likelihood joined maps with an apparently better probability of success than the package JOINMAP (which can tackle more complex pedigree).

The efficiency of local search also offers services which were previously offered by the exhaustive search `compare` command of MAPMAKER. However, these features can be used for a number of markers which currently exceeds the capacity of `compare`. Again, one should remember that MAPMAKER can handle CEPH and intercross pedigree which is yet impossible in CARTHAGENE.

An interesting feature of CARTHAGENE lies in its ability to produce a set of maps in the vicinity of the best map found. Our experiments show that the number of maps found with a loglikelihood within a constant of the best map likelihood is an indicator of how strongly the order is supported by the data. This is an essential feature for building joined maps.

CARTHAGENE shows the interest of exploiting the connection between the salesman problem and the markers ordering problem. There is still a large number of results on the salesman problem which could be exploited to further enhance the efficiency of CARTHAGENE and we intend to continue our work in this direction. CARTHAGENE also highlights the interest of using a theoretically grounded criteria such as multipoint maximum likelihood.

Another direction that we shall consider is the extension of CARTHAGENE to more complex pedigree. This is already well advanced for intercross data. We intend to make the program available to the public as soon as this is finished. People who have backcross-like data and who need CARTHAGENE rapidly can contact the authors by e-mail.

References

Applegate, D.; Bixby, R.; Chvátal, V.; and Cook, W. 1995. Finding cuts in the TSP (a preliminary report). Technical Report 95-05, DIMACS.

Buetow, K., and Chakravarti, A. 1987. Multipoint gene mapping using seriation. *Am. J. Hum. Genet.* 41:180–201.

Cormen, T. H.; Leiserson, C. E.; and Rivest, R. L. 1990. *Introduction to algorithms*. MIT Press. ISBN : 0-262-03141-8.

Dempster, A.; Laird, N.; and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. Ser.* 39:1–38.

Doerge, R. 1996. Constructing genetic maps by rapid chain delineation. *Journal of Quantitative Trait Loci* 2.

Echt, C.; Knapp, S.; and Liu, B.-H. 1992. Genome mapping with non-inbred crosses using GMendel 2.0. *Maize Genet. Coop. Newslett.* 66:27–29.

Glover, F. 1989. Tabu search – part I. *ORSA Journal on Computing* 1(3):190–206.

Glover, F. 1990. Tabu search – part II. *ORSA Journal on Computing* 2(1):4–31.

Goldberg, D. E. 1989. *Genetic algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Pubishing Company.

Johnson, D. 1990. Local optimization and the traveling salesman problem. In *Proc. 17th Int. Colloquium on Automata, Languages and Programming*, volume 443, 446–461. Springer-Verlag Lectures Notes in Computer Science.

Lander, E. S., and Green, P. 1987. Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA* 84:2363–2367.

Lander, E.; Green, P.; Abrahamson, J.; Barlow, A.; Daly, M. J.; Lincoln, S. E.; and Newburg, L. 1987. MAP-

MAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1:174–181.

Lathrop, G.; Lalouel, J.; Julier, C.; and Ott, J. 1985. Multilocus linkage analysis in humans: detection of linkage and estimation of recombination. *Am. J. Hum. Genet.* 37:482–488.

Laurent, V.; Vanlerberghe-Masutti, F.; Wajnberg, E.; Mangin, B.; Schiex, T.; and Gaspin, C. 1997. Construction of a composite genetic map of the parasitoid wasp *Trichogramma brassicae* using RAPD informations from three populations. *Submitted*.

Lin, S., and Kernighan, B. W. 1973. An effective heuristic algorithm for the traveling salesman problem. *Operation Research* 21:498–516.

Liu, B. H. 1995. The gene ordering problem, an analog of the traveling salesman problem. In *Plant Genome 95*.

Ott, J. 1991. *Analysis of human genetic linkage*. Baltimore, Maryland: John Hopkins University Press, 2nd edition.

Stam, P. 1993. Constructing of integrated genetic linkage maps by means of a new computer package:JOINMAP. *The Plant Journal* 3(5):739–744.

Taillard, E. 1991. Robust taboo search for the quadratic assignment problem. *Parallel computing* 17:443–455.

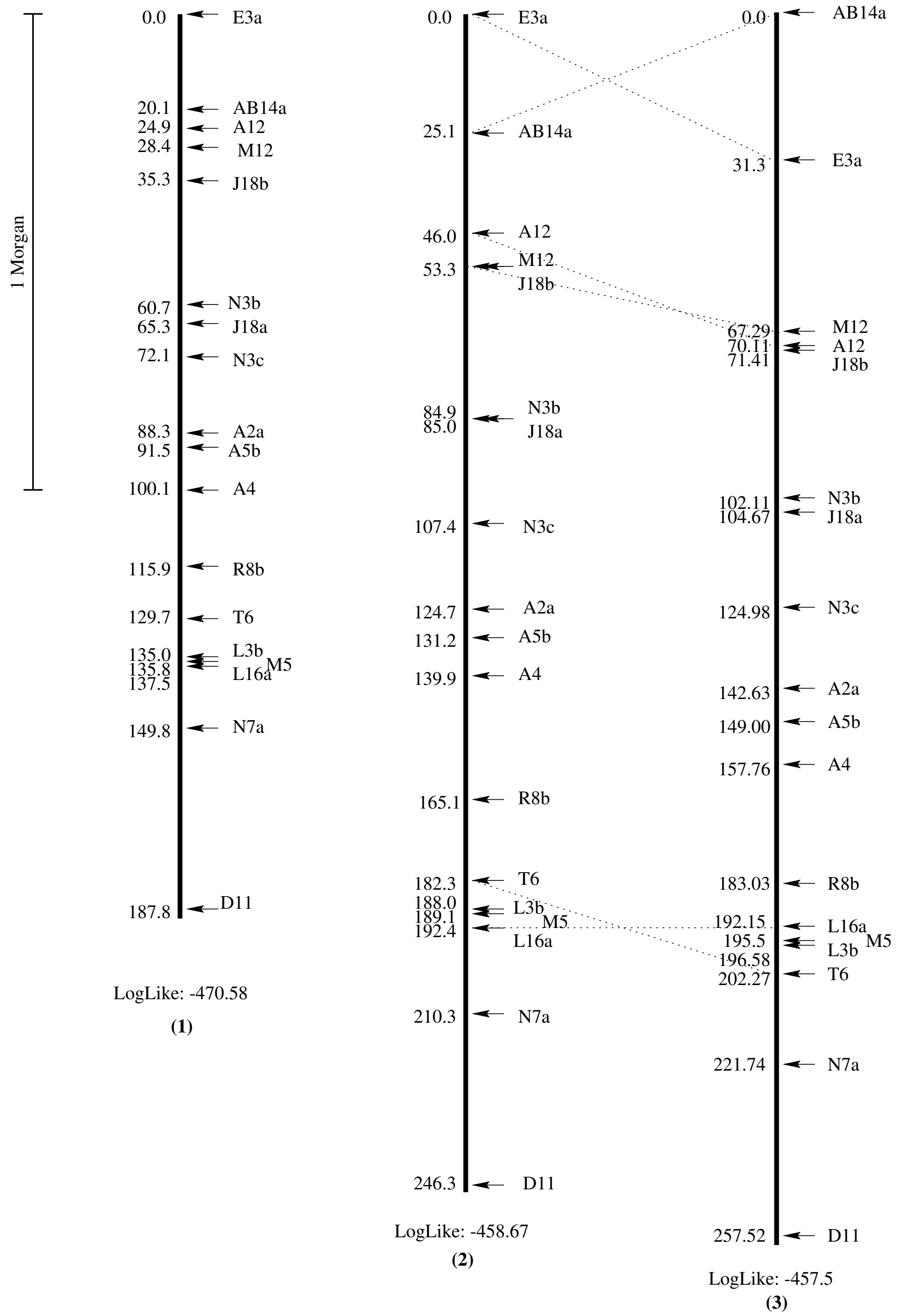

Figure 8: Maps for group II using (1) JoinMap, (2) JoinMap, distances optimized with EM, (3) CARTHAGENE.