Constructing and Joining Maximum Likelihood Genetic Maps

T. Schiex, C. Gaspin, V. Laurent & B. Mangin
Institut National de la Recherche Agronomique
Chemin de Borde Rouge BP 27, Castanet-Tolosan 31326 Cédex, France

Abstract

We present CAR$^{\text{T}}_{\text{H}}$AGENE a new piece of software for genetic mapping. Compared to existing genetic mapping softwares, CAR$^{\text{T}}_{\text{H}}$AGENE has both strengths and weaknesses. The most important weakness is certainly its actual limitation to backcross-like pedigree.

However, when CAR$^{\text{T}}_{\text{H}}$AGENE applies, it is able to order markers and produce maximum multipoint likelihood marker maps with a large number of markers (eg. > 30) and a large percentage of missing data (eg. $> 50\%$). This is done using optimization techniques originating from the artificial intelligence and operations research communities, hybridized with the usual statistical optimization algorithm EM.

When a large number of markers are present, CAR$^{\text{T}}_{\text{H}}$AGENE offers a service comparable to the `compare` exhaustive search command of MAPMAKER at reasonable cost. More interestingly, when several data sets on the same organism are available, each with a different set of informative markers, CAR$^{\text{T}}_{\text{H}}$AGENE allows to automatically build maximum multipoint likelihood joined maps. On this type of problems, CAR$^{\text{T}}_{\text{H}}$AGENE compares favorably to JOINMAP, a much more general package dedicated to this type of problem.

CAR$^{\text{T}}_{\text{H}}$AGENE has been used to build the first joined maps for the wasp *Trichogramma brassicae* and has also been assessed on randomly generated problems.

Introduction

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.

Several packages 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. 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. This criteria 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 sophisticated local search techniques, inspired by a strong analogy between the genetic markers ordering problem and the famous travelling salesman problem.

1 Principles

The problem of finding a maximum likelihood genetic map can be described as a double optimization problem: for a given order, there is the problem of finding recombination probabilities that yields a maximum multipoint likelihood, then one must find an order that maximizes this maximum likelihood.

The statistical optimization algorithm EM [Dempster et al.77] solves the first problem efficiently for simple pedigree in the case of missing data. It is used in the software MAPMAKER [Lander et al.87a, Lander et al.87b]. However, the second problem is more difficult because the number of possible orders to consider for N markers is $\frac{N!}{2}$ and one cannot simply enumerate all possible orders as soon as $N > 8$. However, this type of combinatorial problem is not unusual in Artificial Intelligence or Operations Research and several techniques are suited to this task.

The choice of the precise type of technique used and how it is actually applied is usually very important. In the case of genetic marker ordering, there is a crucial property that can be observed for backcross with no missing data: in this specific unrealistic case, the problem of finding an order that maximizes the maximum multipoint likelihood is equivalent to the problem of the wandering salesman problem (WSP), a variant of the famous travelling salesman problem (TSP) where we are given a set of n cities and a matrix of distance between each pair of cities, and one should find a path that goes exactly once through each city with a minimum length[1].

This connection shows that all the techniques which have been developed for the TSP can also be applied directly to the case of backcross with no missing. However, this case is obviously unrealistic. 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 techniques dedicated to the TSP and that can be extended to the marker ordering problem with missing data will probably work fairly well in this case, even if the connection with TSP is theoretically lost.

Among the several techniques used to solve the TSP, the only ones that seem to cope easily with the existence of missing data belong to the class of local search approaches such as simulated annealing, genetic algorithm, stochastic gradient, tabu search... These algorithms generate a sequence of solutions (markers orders) 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 (here, the maximum multipoint likelihood, computed using EM). 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.

A crucial point in the use of such techniques is the choice of the *neighborhood* used. Here, we have chosen to exploit a successful well-known neighborhood structure introduced in [Lin et al.73] to tackle the TSP. The 2-change neighborhood of a map is the set of all maps obtained by an inversion of a subsection of the map. Thus, for N markers, the neighborhood has a size of $\frac{N.(N-1)}{2} - 1$.

In CAR$^{T}_{H}$AGENE, this neighborhood is exploited by two different local search procedures: Tabu search [Glover90]) and Genetic Algorithm [Goldberg89]). A more detailed description of these algorithms appears in [Schiex et al.97]. They both have asymptotic convergence proofs, but in practice, there is no guarantee that an optimal order will be found even if they usually give excellent results. CAR$^{T}_{H}$AGENE is currently restricted to backcross data, for which EM can be highly optimized. An extension to intercross will soon be built, in which the execution of EM will probably be much slower.

In its actual version, CAR$^{T}_{H}$AGENE can tackle data sets with a large percentage of missing

[1] A similarity between marker ordering and the TSP for two-points approaches is already mentioned in [Liu95].

and can therefore be used to tackle the problem of finding a maximum likelihood map when several data sets for the same organism are available but each with a different set of informative markers. All existing data sets are simply merged and a marker which was uninformative or inexistent in a data set will be marked as missing. This yields a data set with a lot of missing data in large blocks, where several two point measures are undefined.

2 Functionalities

CARTHAGENE has been entirely written in C on a machine running Linux. It should be immediately portable to all Unix platforms and, with some more work, to all platforms with a C compiler. In its actual implementation, CARTHAGENE directly reads MAPMAKER raw files for backcross and recombinant inbred lines. It is entirely automated and will immediately start ordering all the markers found in the data file in order to build a maximum likelihood map. However, it has a nice additional feature that can be used to assess how strongly the best map found is supported by the evidence at hand. Using either local search algorithms, CARTHAGENE maintains a set of fixed size of all the best orders encountered during the search. At the end, not only the best map, but also all the other orders which were encountered and whose likelihood is not very different from the best likelihood is reported to the user. In practice, if several maps are found whose likelihood is close to the best likelihood then it is known that the best map is not strongly supported. Moreover, when one can observe which markers are stable or unstable in these maps, one can try to rerun CARTHAGENE without the unstable markers.

This feature is especially useful for the construction of joined maps where the reliability of the results, even under the assumption of perfect data, is absolutely unpredictable, depending on the size of each data sets and on the number of markers common to several data sets.

In the future, we expect to integrate linkage group computation, automatic phase detection when possible and a graphical interface (beyond additional improvement to the local search algorithms).

3 Application to real and simulated data

CARTHAGENE has been first used to build genetic maps for *Trichogramma brassicae*, a small parasitoid wasp. The data consist 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.97]. Each intra-population involved around 90 individuals. Globally, 84 markers were organized in five linkage groups and 2 pairs. The largest group contained 22 markers.

CARTHAGENE has been used both to build "intra-population" maps and joined maps. For intra-population maps, when the number of markers was small enough, the results have been systematically compared with MAPMAKER using the exhaustive search `compare` function and results have always been consistent.

For joined maps, CARTHAGENE has been compared to JOINMAP (rel. 1.4) [Stam93]. It appears that CARTHAGENE usually provides better results than JOINMAP in terms of likelihood. Actually, maps constructed with JOINMAP were found to be up to 10^{16} times less likely than the corresponding maps built by CARTHAGENE. Moreover, for all pooled maps, several maps were found with a likelihood close to the best likelihood. This tends to to show that these maps are not very reliable, an additional information which is essential. On a Pentium 133, the largest group was ordered in less then one minute using tabu search.

We performed further extensive tests on randomly generated data sets (using evenly distributed markers on a 2 Morgan long chromosome). Data sets with 10 and 15 markers were built, with a sample size varying from 25 to 250 individuals and a percentage of missing varying from 0 to 20%. On all these data sets, representing several thousands maps globally, CARTHAGENE always found a map whose likelihood was at least equal to the likelihood of the map used to generate the data sets. Running time was always lower than half a second.

Random joined data sets have also been built using pairs of data sets, varying the number of common markers in the two data sets from 1 to 10. The largest joined data sets (29 markers, 500 individuals, more than 50% missing) were ordered in few minutes. A systematic comparison with JOINMAP shows that CARTHAGENE probability to find the good order is significantly better than JOINMAP's one. However, we remind the reader that JOINMAP is much more general than CARTHAGENE, being able to handle other pedigree. Globally, the mean number of maps found within 3.0 units of theloglikelihood of the best map appears to be an excellent indicator of the probability of finding the good map. A more extensive analysis of these results can be found in [Schiex et al.97].

Conclusion

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 (see http://www-bia.inra.fr/T/schiex).

References

[Dempster et al.77] Dempster (A.P.), Laird (N.M.) et Rubin (D.B.). – Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. Ser.*, vol. 39, 1977, pp. 1–38.

[Glover90] Glover (F.). – Tabu search – part II. *ORSA Journal on Computing*, vol. 2 (1), Winter 1990, pp. 4–31.

[Goldberg89] Goldberg (D. E.). – *Genetic algorithms in Search, Optimization and Machine Learning*. – Addison-Wesley Pubishing Company, 1989.

[Lander et al.87a] Lander (E. S.) et Green (P.). – Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA*, vol. 84, April 1987, pp. 2363–2367.

[Lander et al.87b] Lander (E.S.), Green (P.), Abrahamson (J.), Barlow (A.), Daly (M. J.), Lincoln (S. E.) et Newburg (L.). – MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, vol. 1, 1987, pp. 174–181.

[Laurent et al.97] Laurent (V.), Vanlerberghe-Masutti (F.), Wajnberg (E.), Mangin (B.), Schiex (T.) et Gaspin (C.). – Construction of a composite genetic map of the parasitoid wasp *Trichogramma brassicae* using RAPD informations from three populations. *Submitted*, 1997.

[Lin et al.73] Lin (S.) et Kernighan (B. W.). – An effective heuristic algorithm for the traveling salesman problem. *Operation Research*, vol. 21, 1973, pp. 498–516.

[Liu95] Liu (B. H.). – The gene ordering problem, an analog of the traveling salesman problem. *In : Plant Genome 95*.

[Schiex et al.97] Schiex (T.) et Gaspin (C.). – Cartagene: Constructing and joining maximum likelihood genetic maps. *In : Proc. of ISMB'97*. – Available at http://www-bia.inra.fr/T/schiex.

[Stam93] Stam (P.). – Constructing of integrated genetic linkage maps by means of a new computer package:JOINMAP. *The Plant Journal*, vol. 3 (5), 1993, pp. 739–744.