Bioinformatics Advance Access published December 14, 2004

CarTHaGene: multipopulation integrated genetic and radiated hybrid mapping

Simon de Givry, Martin Bouchez, Patrick Chabrier, Denis Milan, Thomas Schiex

INRA, Biométrie et Intelligence Artificielle/Génétique Cellulaire, BP 27, 31326 Castanet-Tolosan Cedex, FRANCE

ABSTRACT

Summary: CarTHaGene is an integrated genetic and radiated hybrid mapping tool which can deal with multiple populations, including mixtures of genetic and RH data. CarTHaGene performs multipoint maximum likelihood estimations with accelerated EM algorithms for some pedigrees and has sophisticated algorithms for marker ordering. Dedicated heuristics for framework mapping are also included.

CarTHaGene can be used as a C++ library, through a shell command and through a graphical interface. XML output for companion tools is integrated.

Availability: The program is available free of charge on www.inra.fr/bia/T/CarthaGene for Linux, Windows and Solaris machines (with Open Source).

Contact: carthagene @nospam@ ossau.toulouse.inra.fr

INTRODUCTION

Genetic mapping aims at locating polymorphic markers on chromosomes by exploiting a probabilistic model of crossing-over. Many genetic mapping tools are available to analyze data in experimental crosses. Most of them are designed to analyze line crosses, one family at a time; few can integrate data from several crosses to build consensus maps (see linkage.rockfeller.edu/soft).

Radiation hybrid (RH) mapping is a somatic cell technique with the same aim. It nicely complements the genetic mapping technique, allowing for finer resolution. Most existing RH mapping packages are listed at compgen.rutgers.edu/rhmap.

Although capable of handling line crosses, CarTHaGene has been designed to create consensus maps from multiple populations. Instead of directly integrating existing maps, CarTHaGene computes maximum multipoint likelihood maps, taking into account all the available information, offering additional reliability. It can integrate RH and genetic data together.

METHODS

Parametric probabilistic HMM based models for crossover during meiosis (Lander *et al.* 1987) and for chromosome breakage and retention during RH panel construction (Lange *et al.* 1995) involve parameters such as recombination and breakage probabilities between adjacent markers as well as retention probability. Given experimental data, and assuming some marker ordering, the values of these parameters can be estimated by maximizing the probability that the data observed has been generated by the model. This likelihood allows therefore to simultaneously evaluate the assumed marker ordering and to estimate the corresponding parameters (distances). This is done by a so-called Expectation-Maximization (EM) algorithm. Models for backcross, f2 intercross, recombinant inbred lines (self and sibs) and phase-known outbreds are available. For RH estimation, the equal retention model both in its haploid and diploid forms is used. The EM forward-backward algorithm used in CarTHaGene has been accelerated by taking into account specific properties of backcross and haploid RH data (see (Schiex *et al.* 2001)). Compared to usual EM implementations, the accelerated algorithm can run one or two orders of magnitude faster with no loss of precision.

CarTHaGene provides two ways to merge data files.

(1) If one assumes that the data merged represents a single map (same order and distances, either genetic or RH), a so-called *genetic merging* is done. Untyped markers in a population/panel are considered as missing data and one consensus map is produced. RH and genetic data cannot be merged under this model because they use different distances.

(2) Otherwise, it is assumed that the data files are representative of maps with a common marker ordering but specific distances per data-set. Here, a single consensus ordering is produced but a specific set of distances is estimated for each model merged. Any type of data, genetic or RH, can be merged here. This is called *order merging*. These two methods can be combined freely.

For genetic merging, the EM implementation deals with combined data sets by performing E computation on each data set and then using a M step that takes into account the merging performed. For data-sets merged by order, independent log-likelihoods are obtained per data-set and summed up.

The main problem in genetic or RH mapping comes from the number of possible marker orders. For n markers, there exists $\frac{n!}{2}$ different possible orders. Since (Liu 1995; Schiex and Gaspin 1997) for genetic mapping and (Ben-Dor *et al.* 2000) for RH mapping, the connection between mapping and the *traveling salesman problem* (TSP) is well known. CarTHaGene relies on this connection and provides extensions of TSP solving algorithms:

- fast heuristics (eg. *nearest neighbor*) using 2-point information for guidance and multipoint estimations for final evaluation;
- more powerful meta-heuristics (eg. simulated annealing and taboo search) directly using multipoint maximum likelihood as the optimization criterion.
- even more efficient pure TSP heuristics such as the Lin-Kernighan heuristic (Lin and Kernighan 1973), using the LKH implementation (Helsgaun 2000). These algorithms can use either multiple 2-point maximum likelihood or *obligate chromosome breaks* (Ben-Dor *et al.* 2000; Agarwala *et al.* 2000) to define optimal maps. All the maps identified in this way are then evaluated using a multi-point EM based estimation.
- finally, dedicated heuristics for framework mapping, duplicated marker detection and map validation.

A unique feature of CarTHaGene is that, instead of producing a single supposedly optimal map, it produces an ordered set of alternative maps which allows to estimate the reliability of ordering of each marker. The set of all these maps can be explored manually and compared graphically (with a Postscript output). Dedicated automatic tools facilitate the identification of unreliable markers for further analysis.

Final maps can be produced under MapMaker (Lander *et al.* 1987) and XML formats for data exchange with MCQTL (Jourjon *et al.* 2004) (a multiple population QTL mapping software) and BioMercator (Arcade *et al.* 2004) (a software for integrating genetic maps and QTL detected in independent experiments).

IMPLEMENTATION CarTHaGene is implemented as a C++ library. A Tcl programmable shell command for automated mapping is available and a graphical Tcl/Tk interface for interactive mapping. Binaries for Windows, Solaris and Linux are provided with an open source distribution (using the G-Forge site mulcyber.toulouse.inra.fr).

ACKNOWLEDGMENTS Supported by GENOPLANTE project “Integrative Tools for Genetic Mapping”.

REFERENCES

Agarwala R., Applegate D.L., Maglott D., Schuler G.D., and Schaffer A.A. 2000. A fast and scalable radiation hybrid map construction and integration strategy. *Genome Res.*, 10:350–364.

Arcade A., Labourdette A., Falque M., Mangin B., Chardon F., Charcosset A., and Joets J. 2004. Biomercator: integrating genetic maps and qtl towards discovery of candidate genes. *Bioinformatics*. Epub ahead of print.

Ben-Dor Amir, Chor Benny, and Pelleg Dan. 2000. RHO – radiation hybrid ordering. *Genome Research*, 10:365–378.

Helsgaun K. 2000. An effective implementation of the lin-kernighan traveling salesman heuristic. *European Journal of Operational Research*, 126(1):106–130. http://www.dat.ruc.dk/˜keld/research/LKH.

Jourjon Marie-Françoise, Jasson Sylvain, Marcel Jacques, Ngom Baba, and Mangin Brigitte. 2004. Mcqtl: Multi-allelic qtl mapping in multi-cross design. *Bioinformatics*. To appear.

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

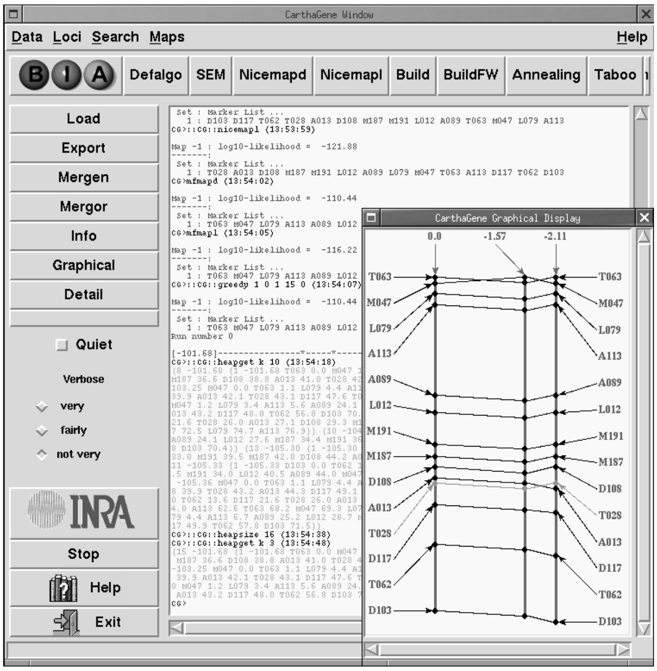
Lange Kenneth, Boehnke Michael, Cox David R., and Lunetta Kathryn L. 1995. Statistical methods for polyploid radiation hybrid mapping. *Genome Research*, 5(2):136–150.

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

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

Schiex T. and Gaspin C. 1997. Cartagene: Constructing and joining maximum likelihood genetic maps. In *Proc. of ISMB’97*, Porto Carras, Halkidiki, Greece. Software available at www.inra.fr/bia/T/CartaGene.

Schiex T., Chabrier P., Bouchez M., and Milan D. 2001. Boosting EM for radiation hybrid and genetic mapping. In *Proc. of WABI’01*, volume 2149 of *LNCS*, pages 41–51.



```
 Set : Marker List ...
   1 : D103 D117 T062 T028 A013 D108 M187 M191 L012 A089 T063 M047 L079 A113
CG>::CG::nicemapl (13:53:59)

Map -1 : log10-likelihood =  -121.88
-------:
 Set : Marker List ...
   1 : T028 A013 D108 M187 M191 L012 A089 L079 M047 T063 A113 D117 T062 D103
CG>mfmapd (13:54:02)

Map -1 : log10-likelihood =  -110.44
-------:
 Set : Marker List ...
   1 : T063 M047 L079 A113 A089 L012
CG>mfmapl (13:54:05)

Map -1 : log10-likelihood =  -116.22
-------:
 Set : Marker List ...
   1 : T063 M047 L079 A113 A089 L012
CG>::CG::greedy 1 0 1 15 0 (13:54:07)

Map -1 : log10-likelihood =  -110.44
-------:
 Set : Marker List ...
   1 : T063 M047 L079 A113 A089 L012
Run number 0

[-101.68]---------------*-----*-----
CG>::CG::heapget k 10 (13:54:18)
{8 -101.68 {1 -101.68 T063 0.0 M047 1
M187 36.6 D108 38.8 A013 41.0 T028 42
103.25 M047 0.0 T063 1.1 L079 4.4 A11
39.9 A013 42.1 T028 43.1 D117 47.6 T0
M047 1.2 L079 3.4 A113 5.6 A089 24.1
013 43.2 D117 48.0 T062 56.8 D103 70.
21.6 T028 26.0 A013 27.1 D108 29.3 M1
7 72.5 L079 74.7 A113 76.9}} {10 -104
A089 24.1 L012 27.6 M187 34.4 M191 36
8 D103 70.4}} {13 -105.30 {1 -105.30
33.0 M191 39.5 M187 42.0 D108 44.2 A0
11 -105.33 {1 -105.33 D103 0.0 T062 1
.5 M191 34.0 L012 40.5 A089 44.0 M047
 -105.36 M047 0.0 T063 1.1 L079 4.4 A
8 39.9 T028 43.2 A013 44.3 D117 49.1
0 T062 13.6 D117 21.6 T028 26.0 A013
4.0 A113 62.6 T063 68.2 M047 69.3 L07
79 4.4 A113 6.7 A089 25.2 L012 28.7 M
17 49.9 T062 57.8 D103 71.5}}
CG>::CG::heapsize 16 (13:54:38)
CG>::CG::heapget k 3 (13:54:48)
{15 -101.68 {1 -101.68 T063 0.0 M047
 M187 36.6 D108 38.8 A013 41.0 T028 4
-103.25 M047 0.0 T063 1.1 L079 4.4 A1
 39.9 A013 42.1 T028 43.1 D117 47.6 T
0 M047 1.2 L079 3.4 A113 5.6 A089 24.
 A013 43.2 D117 48.0 T062 56.8 D103 7
CG>
```

Fig. 1. A typical graphical session of CarTHaGene, displaying possible maps with distance and loglikelihoods.