An introduction to CarTHaGène

An integrated genetic/radiated hybrid/comparative mapping tool

T. Schiex[1], M. Bouchez[1], P. Chabrier[1], T. Faraut[1], C. Gaspin[1], S. de Givry[1], D. Milan[1], J.C. Nelson[2], B. Servin[1]

[1] INRA (Toulouse), France
[2] Kansas State University, USA

March 30, 2009

Genetic and RH maps

Given a dataset (`dsload`) of genetic/RH data on a set of markers M, a genetic/RH map is defined by:

- a set of markers $N = \{m_1, \dots m_n\} \subseteq M$
- which is ordered (eg. $m_1 < \dots < m_n$)
- with a distance between each pair of adjacent markers $(d(m_i, m_{i+1}))$

The genetic/RH mapping problem: find a map (order+distances) that best explains the data set.

What is a good map ?

- Non parametric approach: minimizes the number of compulsory crossovers/breaks, maximizes the sum of 2-points LOD...
- Parametric approach: maximizes the probability of the data (likelihood) under a probabilistic model.

  Parameters: probability of recombination/breakage between 2 adj. markers θ_{ij} (probability of retention r)

CarTHaGène criteria: multipoint maximum likelihood. May use non parametric approaches to guide the search.

Probabilistic models in CARTHAGÈNE

1. Backcross: as in MapMaker. Dedicated EM.
2. RIL (sib/self): as in MapMaker. Dedicated EM.
3. F2 Intercross: as in MapMaker.
4. Phase known outbreds (1:1, 1:2:1, 1:1:1:1 seg. ratio)
5. Haploid RH: Dedicated EM.
6. Diploid RH.

"Dedicated EM": can run more than 2 orders of magnitude faster than existing EM implementation (MapMaker, RHMAP).

Working with multiple populations

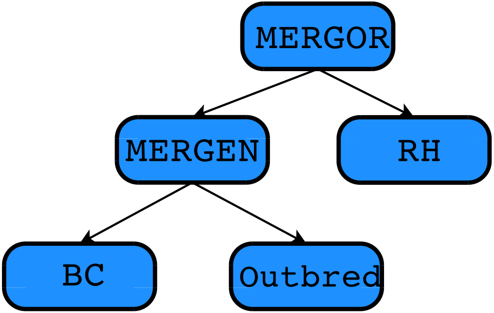
1. Consensus mapping (`dsmergen`): one map for all populations.

   All populations share the same marker ordering and distances. Can be used only for similar population types (eg. backcross with outbreds. RIL sib/self merge with RIL sib/self resp. only).

2. Simultaneous mapping (`dsmergor`): one order for all populations.

   No assumption that distances are the same. Can be used to merge eg. genetic and RH data, or RIL with BC.

Can be combined



Computing linkage groups

As in MapMaker: given a distance threshold θ_{max} and a LOD threshold ℓ_{min}, pool markers that have:

1. a pairwise distance below θ_{max}
2. a LOD above ℓ_{min}

Weakness: 2 unrelated markers can be pooled just if they are enough related to one marker (`group`, `groupget`).

2-points information is computed on loading (`dsload`, `mrklod2p`, `mrkfr2p`).

Ordering markers and the TSP

In a group of n markers, there are $\frac{n!}{2}$ differents orders.

For $n = 10$: 1.8 10^{6} orders, $n = 20$: 1.2 10^{18}, $n = 100$: 4.7 10^{157}!

Under strong hypothesis (BC, RIL, RH, no missing or untyped marker in a population), maximum likelihood ordering is equivalent to the

Symmetric Traveling Salesman Problem

Given n cities and inter-cities distances, find a path that passes once through each city and that minimizes the overall distance.

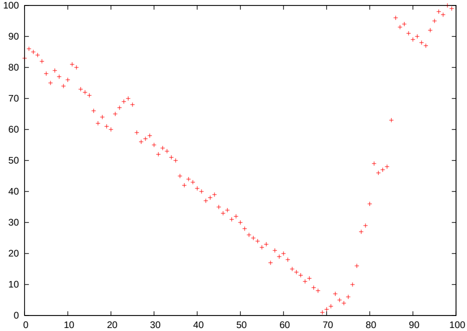
One of the most studied optimization problem in computer science. Know to be potentially very hard (NP-hard).

Building heuristics: Nearest Neighbor selection (2-pt LOD)

Simulated **backcross** data:

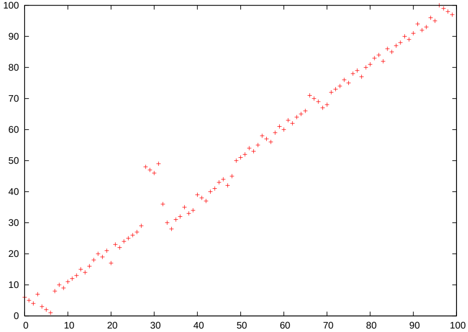
100 markers positioned at random on a 3-Morgan chromosome

200 individuals, 15% missing data, 10% genotyping errors



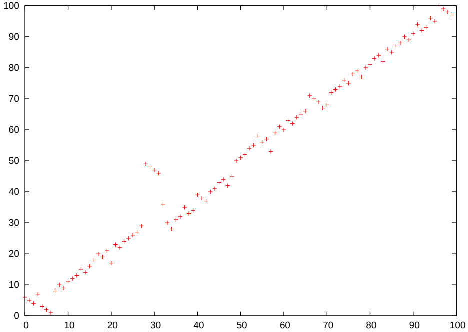
Log10-likelihood: -3749.44

Improving heuristics: Submap Reversals (2-opt)



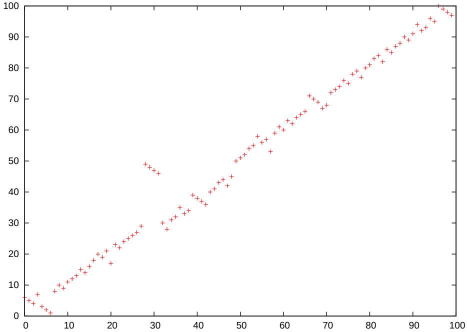
Log10-likelihood: −3668.36 in 23 seconds (PC 2.8 GHz)

Imp. heuristics: Exhaustive Search on Small(5) Submaps



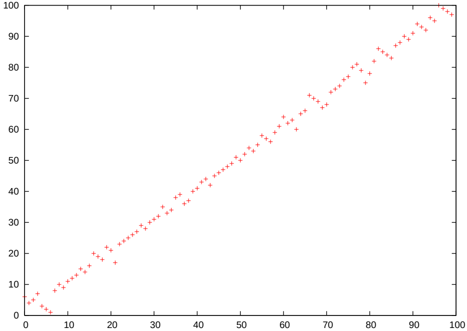
Log10-likelihood: −3667.87

Improving heuristics: Marker Reinsertion



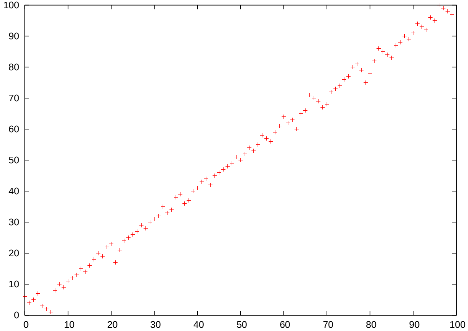
Log10-likelihood: -3667.37

Improving heuristics: Submap Swaps (3-opt) and more



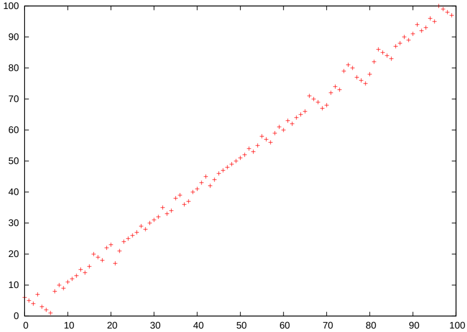
Log10-likelihood: -3658.21 in 2.3 seconds (LKH)

Exhaustive Search in 2-pt approximation



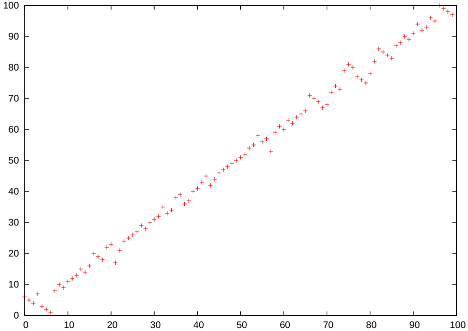
Log10-likelihood: −3658.83 in 0.15 seconds (Concorde)

Submap Reversals (2-opt) again



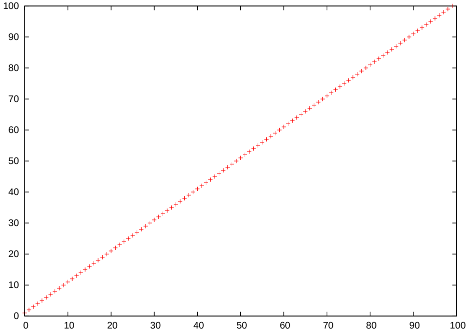
Log10-likelihood: -3650.36 in 14 seconds

Exhaustive Search on Small(5) Submaps again



Log10-likelihood: -3650.07

True map



Log10-likelihood: -3725.34

Good maps & the Heap

Good maps

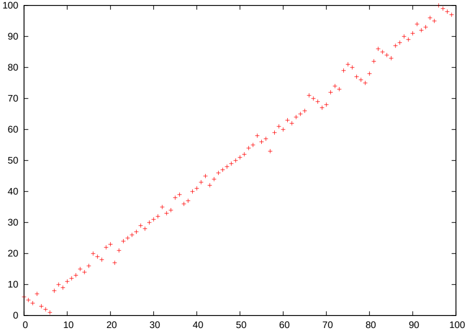
Not only max. likelihood maps. May be also reliable map (no alternative order has comparable likelihood).

For the same set of “active markers” (`mrkselset`, `mrkselget`), CARTHAGÈNE remembers the *k* best maps encountered by the ordering procedures.

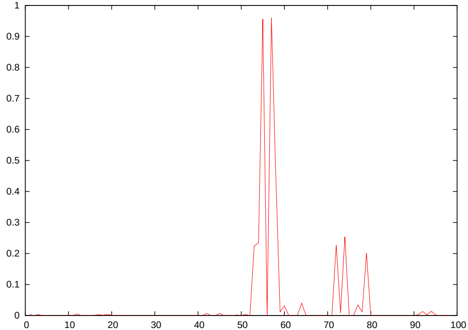
The heap

The set of the *k* best maps found. Gives a representation of the neighborhood of the current optimum map (`heaprints`, `heaprintds`, `heapget`. . .).

Check map reliability: MCMC algorithm



Log10-likelihood: -3650.07



$1-$ posterior probability of adjacent markers

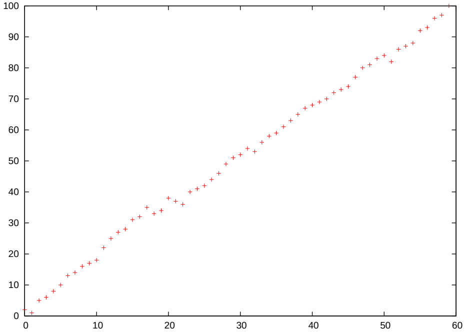
Best map posterior probability: 0.9781 (best map as reference order, $\lambda = 49$ breakpoints)

Framework building method

buildfw Δ_{min} Δ_{keep} S c $(S = \{\}, c = 0)$

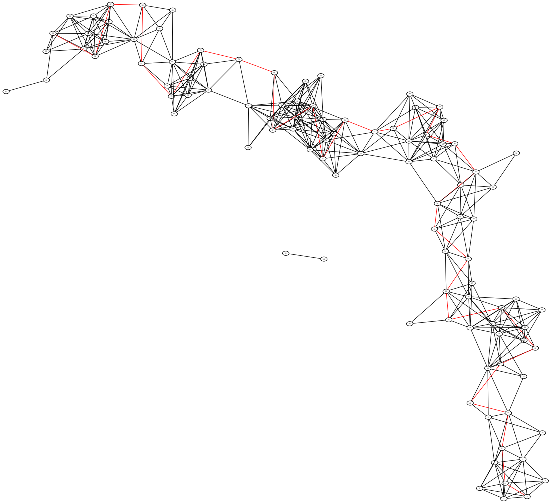
1. Start from all possible pairs of markers.
2. For all available maps, a new marker is inserted in all possible positions. The marker "reliability" is defined as the difference δ in loglike between the best and the second best insertion position. A marker can be inserted only if this difference is larger than Δ_{min}.
3. From all these new maps, keep only those such that $\delta \geq \Delta_{keep}$.
4. repeat to 2.

Framework building: `buildfw` $\Delta_{\min} = 3$ $\Delta_{keep} = 3$



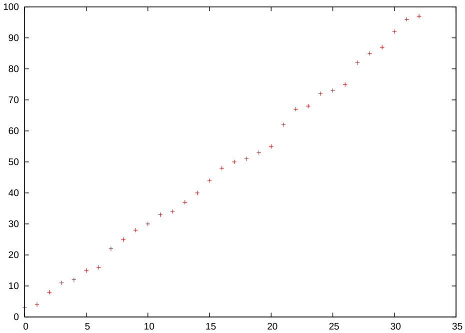
60 markers included. Log10-likelihood: -2257.13
Second best map (2opt+flip+reinsert) log10-likelihood: -2259.51

Framework building with δ guarantee



normalized 2-pt log10-likelihood contribution edge threshold at 49

Framework building with δ guarantee



33 markers included ($\delta = 2$). Log10-likelihood: -1321.35
Second best map (2opt+flip+reinsert) log10-likelihood: -1324.07

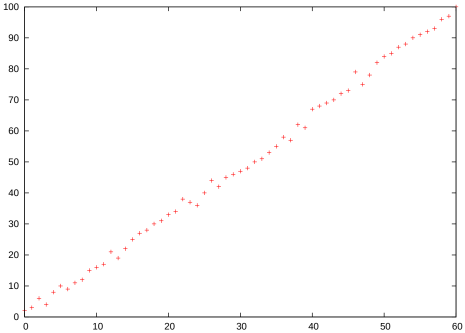
From framework to comprehensive

buildfw Δ_{min} Δ_{keep} S c

- S: a marker ordering to start from (rather than all pairs). Used to extend an existing "reliable" map.
- $c = 1$: when no marker with sufficient quality exists, tries to independently insert all remaining markers in all possible intervals.

  For each such marker: reports the best insertion position (+) and how far in loglike all other positions are (support for the best position).

Framework building: `frameworkn` 30 60 2;`buildfw` 3 3 *S*



61 markers included. Log10-likelihood: −2282.93

Second best map (2opt+flip+reinsert) log10-likelihood: −2284.22

Final points

Additional facilities

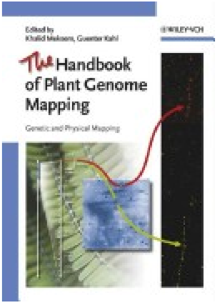
- Able to tackle very large data sets (thousand of markers)
- Graphical interface with map display/print (all commands accessible through the shell interface in the GUI).
- Full user documentation, Open source code with optional LKH code.
- Includes a complete interpreted programming langage (for developping mapping strategies and reusing them).
- Available under Linux, Solaris and Windows.

The software web site:
http://www.inra.fr/mia/T/Carthagene

References

S. de Givry, M. Bouchez, P. Chabrier, D. Milan, and T. Schiex. *CARTHAGENE: multipopulation integrated genetic and radiated hybrid mapping.* Bioinformatics, 21(8):1703-1704, 2005.

Khalid Meksem and Günter Kahl. *The Handbook of Plant Genome Mapping: Genetic and Physical Mapping.* Wiley-VCH, 2005.



Edited by
Khalid Meksem, Guenter Kahl

WILEY-VCH

The Handbook
of Plant Genome
Mapping

Genetic and Physical Mapping