BIOINFORMATICS APPLICATIONS NOTE Vol. 30 no. 11 2014, pages 1623–1624
doi:10.1093/bioinformatics/btu091

Genome analysis

Advance Access publication February 14, 2014

LPmerge: an R package for merging genetic maps by linear programming

Jeffrey B. Endelman[1,*] and Christophe Plomion[2,3]

[1]Department of Horticulture, University of Wisconsin, Madison, WI 53706, USA, [2]INRA, UMR1202, BIOGECO, F-33610 Cestas and [3]Université de Bordeaux, UMR1202 BIOGECO, F-33170 Talence, France

Associate Editor: John Hancock

ABSTRACT

Summary: Consensus genetic maps constructed from multiple populations are an important resource for both basic and applied research, including genome-wide association analysis, genome sequence assembly and studies of evolution. The LPmerge software uses linear programming to efficiently minimize the mean absolute error between the consensus map and the linkage maps from each population. This minimization is performed subject to linear inequality constraints that ensure the ordering of the markers in the linkage maps is preserved. When marker order is inconsistent between linkage maps, a minimum set of ordinal constraints is deleted to resolve the conflicts.

Availability and implementation: LPmerge is on CRAN at http://cran.r-project.org/web/packages/LPmerge.

Contact: endelman@wisc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on August 5, 2013; revised on January 5, 2014; accepted on February 8, 2014

1 INTRODUCTION

Broadly speaking, two types of strategies have been used to construct genetic maps across multiple populations. One is to minimize an objective function based on the observed recombination frequencies between markers, analogous to the strategy used for linkage mapping in a single population. Examples of this approach include the software packages JoinMap (Van Ooijen, 2006) and MultiPoint (Ronin *et al.*, 2012). The second strategy is to work directly with the component linkage maps instead of the underlying recombination frequencies, which can lead to significant gains in computational efficiency without compromising map accuracy (Wenzl *et al.*, 2006). This second strategy is used by the software package MergeMap (Wu *et al.*, 2011), which has been used in several different species (Gautami *et al.*, 2012; Khan *et al.*, 2012; Muñoz-Amatriaín *et al.*, 2011).

Endelman (2011) identified a weakness in MergeMap and proposed an alternative algorithm for merging linkage maps based on linear programming (LP), which was incorporated into an R package called DAGGER. The accuracy of the LP algorithm was validated on simulated data, and in comparison with MergeMap using real barley data was found to have lower mean-squared error on a genome-wide basis (Endelman, 2011). A significant limitation of DAGGER was that it required the component linkage maps to have a consistent marker order. In practice, the linkage maps from different populations often have ordering conflicts, even when there is a single physical order, because of genotyping errors and statistical errors arising from the use of small populations. Another weakness of DAGGER was that the length of the consensus genetic map was shrunken compared with the original linkage maps.

Both of these limitations have been overcome to create a new R package called LPmerge, which is available on CRAN. The LP algorithm has been reformulated without graph theory to eliminate shrinkage of the consensus map, and a new algorithm for resolving ordinal conflicts between linkage maps has been added. The objectives of this Applications Note are to describe these modifications and to illustrate the performance of the conflict resolution algorithm.

2 ALGORITHM

2.1 Consensus map error

In the first step of LPmerge, markers are assigned to bins. If in every map where two markers were jointly mapped they co-segregated (had the same map position), they are placed in the same bin. Because of some markers co-segregating in one map but not in another, a single linkage map bin may be represented by multiple consensus map bins. Map positions in the ith linkage map are denoted by y_i, and consensus map positions are denoted by x. Within linkage map i, the markers are ordered from $j=1$ to M_i, and the map distance between the jth and $(j+q)$th markers is $y_i(j+q)–y_i(j)$. Letting $u(j;i)$ denote the consensus map bin containing marker j from map i, the corresponding distance in the consensus map for these two markers is $x(u(j+q;i))–x(u(j;i))$. The absolute error between the consensus map and the ith linkage map for this interval is thus

$$E_{i,j,q} = \left|[x(u(j+q;\,i)) - x(u(j;\,i))] - [y_i(j+q) - y_i(j)]\right| \quad (1)$$

The total error between the consensus map and the linkage maps is a sum over maps (i), markers (j) and interval sizes (q):

$$E = \sum_{i=1}^{T} W_i N_i^{-1} \sum_{q=1}^{K} \sum_{j=1}^{M_i} E_{i,j,q} \quad (2)$$

where T is the number of linkage maps and K is the maximum interval size. At the end of a linkage map, when the sum $j+q$ exceeds M_i, this expression is evaluated as if the linkage maps were circular rather than linear. For example, when $j=M_i-1$

*To whom correspondence should be addressed.

and $q = 3$, the expression $j+q$ evaluates to 2. These 'wrap-around' error terms keep the total consensus map length commensurate with the average linkage map length. The normalization factor

$$N_i = K\sum_{j=1}^{M_i} 1$$

is the number of error terms for map i, and W_i is a set of possible weights for the average ($W_i = 1$ for unweighted). The maximum interval size K can be varied to produce different consensus maps, and additional criteria can be used to select one. A tutorial illustrating this process is available at http://potatobreeding.cals.wisc.edu/software.

2.2 Resolving marker order conflicts

When minimizing the error [Equation (2)], it is desirable that the consensus map be as consistent as possible with the linkage maps in terms of marker order. This is achieved through the use of linear inequalities. For a pair of adjacent linkage map bins (v,w), each with a single consensus map bin, the corresponding constraint is $x(w) - x(v) \geq 0$. When the linkage map bins contain multiple consensus map bins, constraints are added for every combination of the consensus map bins. The total set of ordinal constraints can be written in matrix notation as $\mathbf{Ax} \geq \mathbf{0}$.

If there are conflicts in marker order between the maps being merged, the linear system $\mathbf{Ax} \geq \mathbf{1}$ will be infeasible. Finding the minimum number of constraints to remove to achieve feasibility is NP-hard (Amaldi and Kann, 1995). LPmerge uses a polynomial-time approximation from Chinneck (2001) (Algorithm 1), which is based on the idea of elasticizing constraints. The elastic LP corresponding to $\mathbf{Ax} \geq \mathbf{1}$ is

$$\begin{aligned} \min_{\mathbf{x},\mathbf{b}} \sum_i b_i \\ \mathbf{Ax} + \mathbf{b} \geq \mathbf{1} \\ \mathbf{b} \geq \mathbf{0} \end{aligned} \tag{3}$$

where $\mathbf{b}$ represents the elastic variables, and the objective is to minimize their sum. The algorithm proceeds by finding which constraint (row of $\mathbf{A}$) leads to the lowest elastic sum when removed. This constraint is then removed, and the procedure is repeated. When the algorithm eliminates a constraint and finds the elastic sum is zero, then that subsystem is feasible and the algorithm stops.

2.3 LP problem

After removing the ordinal conflicts in $\mathbf{A}$, linear programming is used to find a consensus map with minimum error (see Supplementary Material online).

3 RESULTS

Figure 1 illustrates the performance of the conflict resolution algorithm on a toy problem from the LPmerge reference manual. Four linkage maps have been merged (I–IV), each with seven markers (A–G). Map I represents the true marker order, and in the other maps the order of two adjacent markers has been inverted (highlighted in red) to represent mapping errors from the single population analyses. LPmerge correctly identifies the outliers from each map and eliminates them, as documented in the session log output:

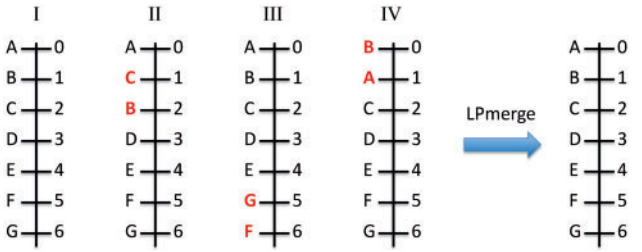

Fig. 1. Toy problem illustrating the ability of LPmerge to resolve marker order conflicts (shown in red). Map I represents the correct order, which is recovered by LPmerge

```
Eliminated following constraints
Map II: C < B
Map III: G < F
Map IV: B < A
```

Unlike other map-merging algorithms, LPmerge removes inequality constraints rather than markers to resolve conflicts. The consensus map returned by LPmerge for this toy problem has the correct marker order and distances (Fig. 1).

Funding: J.B.E. acknowledges support from the Bill and Melinda Gates Foundation. C.P. acknowledges support from EU Project No. 289841 (PROCOGEN).

Conflict of Interest: none declared.

REFERENCES

Amaldi,E. and Kann,V. (1995) The complexity and approximability of finding maximum feasible subsystems of linear relations. *Theor. Comput. Sci.*, **147**, 181–210.

Chinneck,J.W. (2001) Fast heuristics for the maximum feasible subsystem problem. *INFORMS J. Comput.*, **13**, 210–223.

Endelman,J.B. (2011) New algorithm improves fine structure of the barley consensus SNP map. *BMC Genomics*, **12**, 407.

Gautami,B. *et al.* (2012) An international reference consensus genetic map with 897 marker loci based on 11 mapping populations for tetraploid groundnut (*Arachis hypogaea* L.). *PLoS One*, **7**, e41213.

Khan,M.A. *et al.* (2012) A multi-population consensus genetic map reveals inconsistent marker order among maps likely attributed to structural variations in the apple genome. *PLoS One*, **7**, e47864.

Muñoz-Amatriaín,M. *et al.* (2011) An improved consensus linkage map of barley based on flow-sorted chromosomes and single nucleotide polymorphism markers. *Plant Genome*, **4**, 238–249.

Ronin,Y. *et al.* (2012) Two-phase analysis in consensus genetic mapping. *G3 (Bethesda)*, **2**, 537–549.

Van Ooijen,J.W. (2006) *JoinMap 4: software for the calculation of genetic linkage maps in experimental populations*. Kyazma B.V., Wageningen, The Netherlands.

Wenzl,P. *et al.* (2006) A high-density consensus map of barley linking DArT markers to SSR, RFLP and STS loci and agricultural traits. *BMC Genomics*, **7**, 206.

Wu,Y. *et al.* (2011) Accurate construction of consensus genetic maps via integer linear programming. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **8**, 381–394.