Theor Appl Genet (2000) 101:90–94

G. Charmet · P.F. Bert · F. Balfourier

A computerised algorithm for selecting a subset of multiplex molecular markers and optimising linkage map construction

Received: 17 March 1999 / Accepted: 23 August 1999

Abstract A computer algorithm is presented which allows selection of a subset of multiplex markers based on the minimisation of an optimality criterion for a genetic linkage map. It could be applied for choosing a subset of primers (e.g. RAPD, IMA or AFLP), each of which provides several unevenly spaced genetic markers. The goal is to achieve a saturated map of evenly spaced markers, using as few primers as possible to minimise cost and labour. Minimising the average map distance between markers is trivial, but simply leads to selection of those primers which provide the greatest number of markers. However, minimising the standard deviation of interval length ensures that weight is given both to the number of markers and to the evenness of their distribution on the linkage map. This criterion was found empirically to give a result fairly close to the optimum. A stepwise-like selection procedure is therefore implemented, which stops when the optimality criterion does not decrease any more. An example is given of a molecular map of perennial ryegrass with 463 markers obtained from 17 AFLP primers. It is demonstrated that this can be safely reduced to a 175 marker map with only 6 primers. Genetic diversity studies may also benefit from using such a subset of less-redundant markers in genetic distance estimation.

Key words Molecular map · AFLP · RAPD · Optimisation algorithm

Communicated by P. Langridge

G. Charmet (✉) · P.F. Bert · F. Balfourier
INRA station d'amélioration des plantes, 234 avenue du Brézet,
F-63039 Clermont-Ferrand, France
e-mail: charmet@valmont.clermont.inra.fr

Introduction

The advent of molecular markers such restriction fragment length polymorphism (RFLP) or random amplified polymorphic DNA (RAPD) has enabled the construction of saturated linkage maps in many animal and plant species. The main utilisation of such linkage maps is for locating genes involved in the determination of qualitative ("Mendelian") or quantitative traits. In the case of quantitative trait locus (QTL) location, is has been demonstrated that the number of recombinant lines used is more critical than the number of markers in achieving high detection power and narrow confidence intervals for QTL location and effect (Visscher et al. 1996). An optimum use of resources, i.e. labour and cost, therefore requires that the investigator should first select optimally located markers from a highly saturated reference map (e.g. evenly spaced every 5 cM) and then genotype several hundred segregating units for these markers. This is easy to achieve with, for example, microsatellites or RFLPs, as most primer pairs or probes, e.g. cDNA, will provide only one marker locus, at least in a diploid species. Other molecular technologies, such as RAPD or IMA (inter-microsatellite amplification), usually give several markers each. This "multiplex" ratio is even greater with the recent amplified fragment length polymorphism (AFLP) technology (Vos et al. 1995), which can provide several dozen polymorphic markers with a single selective primer pair and a single gel run. Moreover, AFLP markers are often located in "clusters" and thus show a high degree of redundancy.

This papers presents a computerised algorithm to select a subset from a set of multiplex markers, i.e. a set of genetic markers obtained together in a single experiment, with the aim of optimizing the linkage map obtained at a given cost. An illustration is given on perennial ryegrass. The previously published reference map comprised 463 AFLP markers obtained from 17 primer parts (Bert et al. 1999). Use of the proposed algorithm allowed us to identify a subset of 6 primer pairs, yielding 175 more evenly located markers, which can be used for mapping QTL in breeding populations.

Materials and methods

Optimisation algorithm

A computer algorithm has been devised to select a subsample of AFLP primers in order to achieve a reasonable map coverage at reduced cost. The procedure can be applied to any multiplex marker technology. Several optimization criteria have been tested: average interval length, standard deviation of interval length and proportions of interval length below or above a specified threshold. After preliminary tests, the standard deviation of interval length, which takes into account both the number of markers and the evenness of their distribution, was used as the optimisation criterion to be minimised.

We first programmed a forward selection procedure: at each step, a new AFLP primer pair is entered, which minimises the standard deviation of interval length obtained from the primer subset (empty at the beginning). Note that map positions are taken from the reference map, and not recomputed at each step, which would take too long. The selection procedure can be stopped when

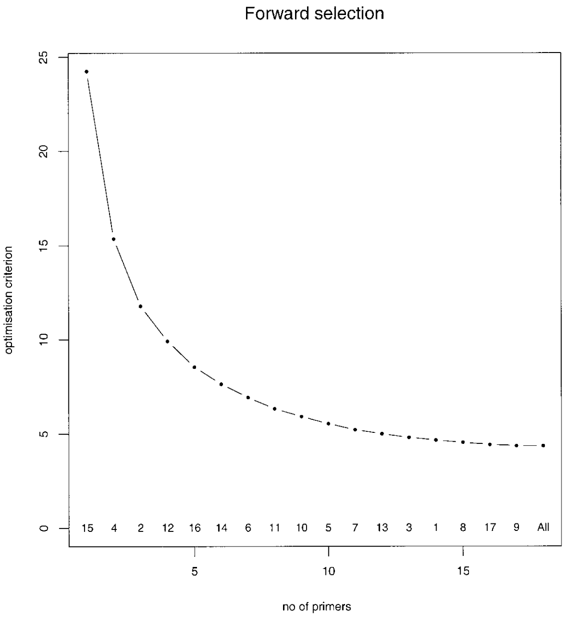

Fig. 1 Decrease in the optimisation criterion, i. e. the standard deviation of the interval lengths, against the introduction of primers in the forward procedure without the stopping rule

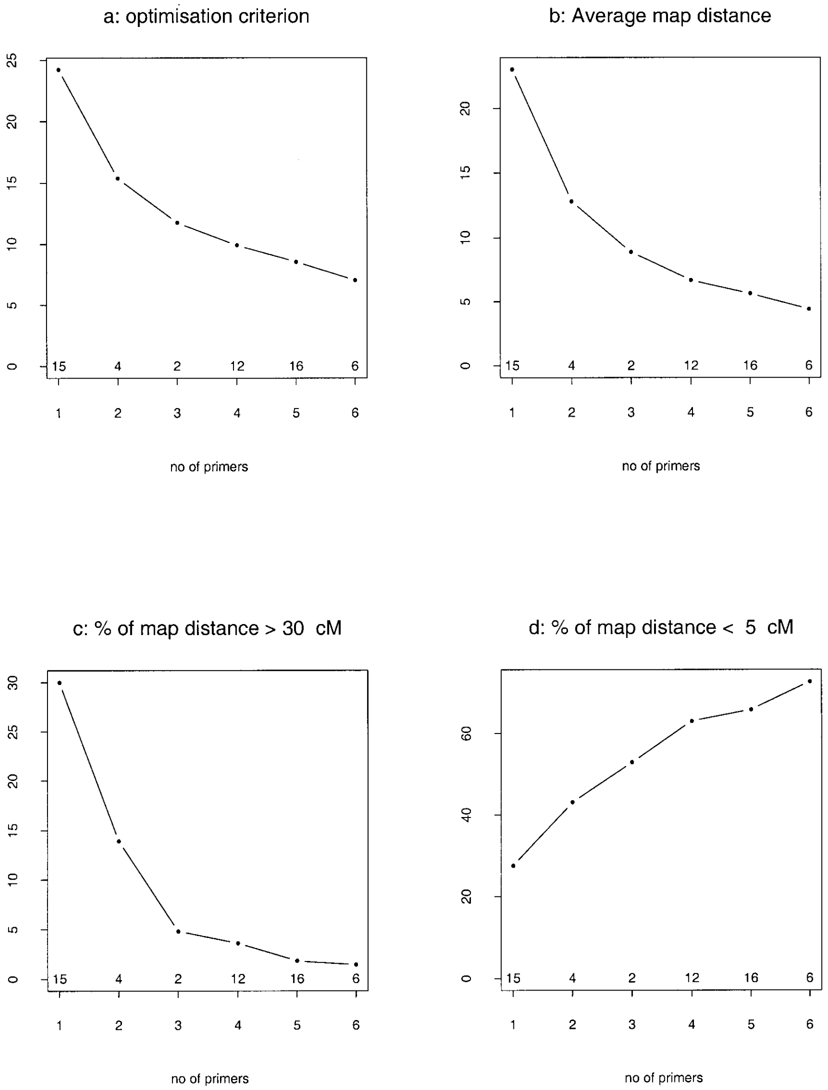

Fig. 2 Results of the stepwise-like selection procedure with a stopping threshold of 1 cM on the decrease in optimisation criterion. **a** optimisation criterion, i.e. the standard deviation of interval lengths, **b** average interval length, **c** percentage of map distances >30 cM, **d** percentage of map distances <5 cM

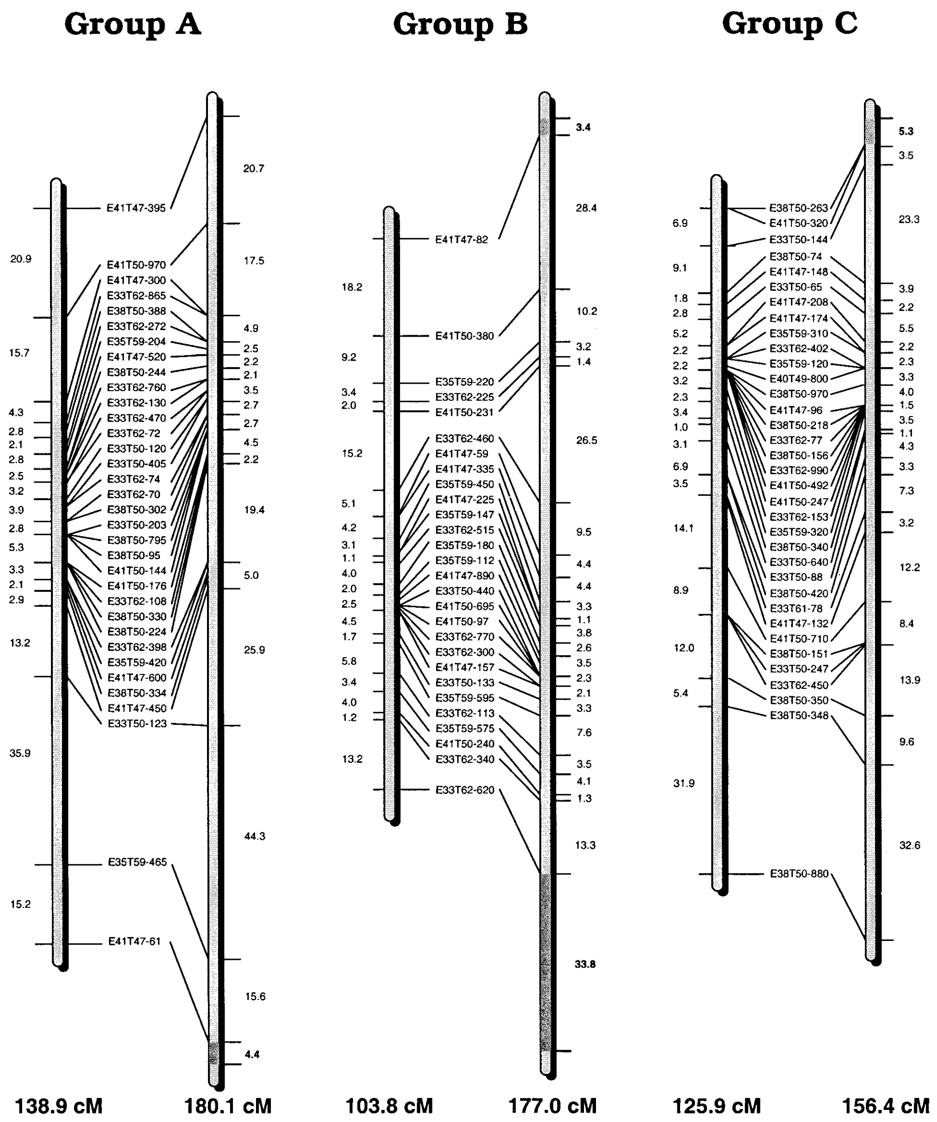

Fig. 3 Comparison of the new (optimised) map with the previously published one. Only the 175 markers from the 6 selected primers are plotted. *Right-hand side* Marker location on the reference map. Lost chromosome ends are indicated in dark grey. *Left-hand side* New map locations recalculated from data of the 175 selected markers only

a desired threshold for the decrease in the optimisation criterion is reached, or left running until all primers are included in the subset. The other map quality criteria can be plotted against the number of primers in the subset as well. We also tested a backward selection procedure, i.e. starting from the full set of primer sand eliminating, at each step, that primer the elimination of which gives the smallest increase in standard deviation of interval length. The results of both procedures were nearly identical. However, we cannot be certain that the best subset has been found.

Therefore, we implemented the following stepwise-like selection procedure, which is derived from that used in multiple regression (Draper and Smith 1981):

1) Initialise the unselected subset (Unsel) to the full set of primers and the selected subset (Sel) to empty.
2) Iteratively add each primer from Unsel to Sel and compute the optimisation criterion, i.e. the standard deviation of the interval length of the map using primers from Sel. If the reduction in the optimisation criterion exceeds a critical value, the primer is retained in Sel. If any terminal marker is missing in Sel, then an interval between the end of the chromosome and the first marker is considered. Similarly, if a chromosome does not have any marker, it is considered to be a single interval whose length equals that of the whole chromosome.
3) After each primer introduction, the number of primers now in Sel being defined as N, iteratively try to replace any primer in Sel by every primer from Unsel and make the replacement effective if the optimisation criterion is further reduced at constant N, else return to step 2. Step 3 is repeated until the decrease is below the desired threshold value.
4) Repeat step 2 until the decrease in optimisation criterion is below a desired threshold value for all primers remaining in Unsel.

Programming was carried out using the Splus language, and the source file is available upon request (charmet@clermont.inra.fr). Once the optimal subset of primers is obtained, various criteria of map quality can be plotted against the composition of Sel, either for the whole map or for each chromosome.

Example

This method has been applied to a perennial ryegrass map previously published (Bert et al. 1999), which comprised 463 markers obtained from 17 AFLP primer pairs. This map was developed on a subset of 95 plants from an intraspecific (*Lolium perenne* L) population constructed at the Institute of Grassland and Environmental Research, Aberystwyth, UK. This population derived from a cross between a di-haploid plant (DH290) and a Hybrid F1 [Romanian collection no. Ba 9982×(a plant from a poylcross with cv. Melle and North Italian collection)].

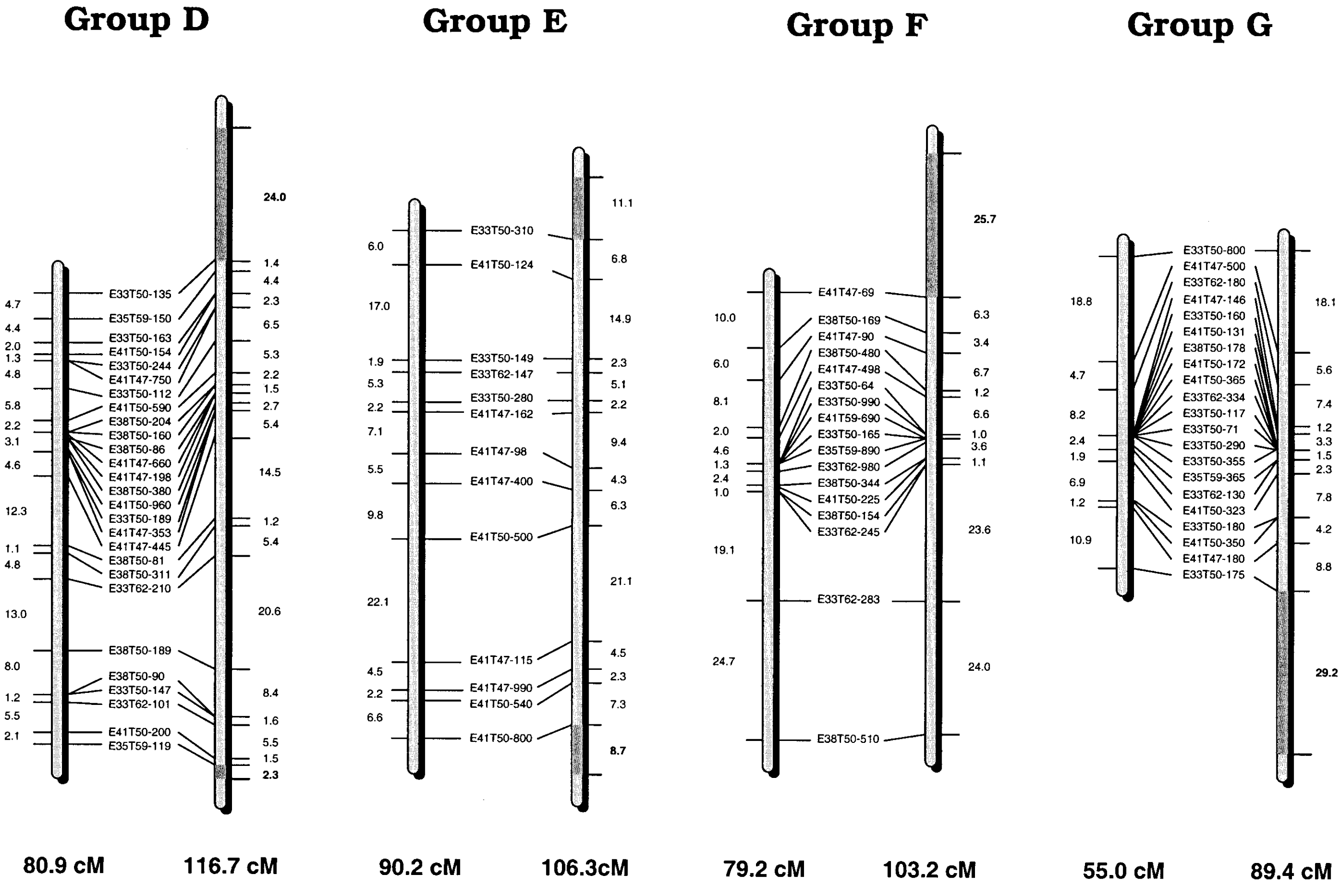

Fig. 3D–G

Results

The optimization criterion was plotted against selected primers in the forward method (Figure 1). This method was used without a termination threshold, and thus the lower asymptote represents the value of the optimisation criterion for the published map. From these asymptotic shapes, it seems sufficient to retain a subset of 5–8 primers, which leads to an average marker spacing of about 6 cM. Results from the backward selection procedure are nearly symmetrical and therefore not shown. A threshold of 1 cM was then assigned to the standard deviation of interval length, which was used as optimisation criterion. The ”stepwise” selection was identical to the forward procedure until step 5, then stopped after the 6th primer, with the only permutation being that of primer 14 by primer 6. This led to a slightly better map than the forward procedure alone. The 6 selected primers provide 175 markers.

Results of this stepwise procedure are presented in Fig. 2. Three other criteria of map quality are shown in addition to the optimisation criterion: average marker spacing, proportion of intervals with length less than 5 cM (indicating redundant markers for QTL scan) or greater than 30 cM (indicating gaps in the map). The resulting map is given in Fig. 3 and compared with the previously published map. Using the initial marker locations from the reference map, we find the total length of the new map to be 782.1 cM, i.e. 84% of the reference map (930 cM). It should be noticed that this fairly good coverage could have been obtained at a cheaper cost (theoretically only 35% of the initial cost). Moreover, marker spacing is more even, as part of the redundancy, particularly in the centromeric regions, has been eliminated. However the negative aspect is the loss of some chromosome ends, particularly of four segments with lengths between 20 and 33 cM on groups B, D, F and G. When map positions are recalculated using data of the 175 markers from the 6 selected primers, the map length is reduced again to 673.9 cM. This phenomenon of map reduction (the symmetric of map expansion when new markers are added) mostly affects the longer linkage groups.

Discussion and conclusion

The proposed algorithm succeeded in establishing a core subset of AFLP primers for mapping purposes. This procedure is not exactly what is usually termed ”stepwise”, as there is no possibility of simply removing a primer. It seems unlikely that this option should be used; obviously, it is not very useful to refine an existing reference map. However, marker selection is often desirable for mapping a new population, as for example in QTL analysis or marker-assisted selection. Our procedure may thus be useful for optimising this choice. Other applications can be found for the algorithm with only slight adaptations of parameter range. One may wish to select or core set of primers to evaluate genetic diversity in genetic resource collections. Thompson and Nelson (1998) recently described a procedure to achieve this goal. However, their method is based on selecting those primers that are significantly correlated to the first principal components of a dissimilarity matrix and then iteratively removing those primers that do not modify the clusters obtained from the new dissimilarity matrix. Although this method is likely to reduce marker redundancy, map information is not taken explicitly into account. Indeed, the use of linked markers will cause biases in estimates of genetic distance estimates. Although methods have been proposed to correct this bias (Dillman et al. 1997), it may appear simpler to select a subset of (nearly) unlinked markers. This can be achieved by using an appropriate optimisation criterion, which could be the proportion of interval lengths above a given threshold. Besides reducing bias in distance estimate, our method may allow the selection of a primer set which is small enough to allow screening of large collection, such as most genebanks.

Another possible application concerns fine mapping. AFLP technology is often used to saturate chromosome regions around QTLs or genes of interest to allow chromosome walking. Again an optimum choice of a set of primers can be achieved, for example by setting the proportion of interal lengths below a given threshold and then running the programme on a single linkage group.

Several optimisation criteria can be proposed, and three options are available for selecting primers, namely forward, backward and stepwise, as described above. The proposed algorithm can thus be used to deal with a range of applications, where selecting a subset of primers from a redundant set is needed to save labour and cost.

References

Bert PF, Charmet G, Sourdille P, Balfourier F (1999) High density molecular map for ryegrass (*Lolium perenne* L.) using AFLP markers. Theor Appl Genet 99:445–452

Dillman C, Bar-Hen A, Guérin D, Charcosset A, Murigneux A (1997) Comparison of RFLP and morphological distances between maize *Zea mays* L. inbred lines. Consequences for germplasm protection purposes. Theor Appl Genet 95:92–102

Draper NR, Smith H (1981) Applied regression analysis. 2nd edn. John Wiley and Sons, London

Thompson JA, Nelson RL (1998) Core set of primers to evaluate genetic diversity in soybean. Crop Sci 38:1356–1362

Visscher PM, Thompson R, Haley CS (1996) Confidence intervals in QTL mapping by bootstrapping. Genetics 143:1013–1020

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M, Frijters A, Pot J, Peleman J, Kuiper M, Zabeau M (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4444