Global optimization in QTL analysis

Kajsa Ljungberg[1], **Sverker Holmgren**[1], **Örjan Carlborg** [2]

Keywords: genetic mapping, quantitative traits, QTL, epistasis, global optimization

1 Introduction and problem formulation.

Many phenotypes of medical, economical and general scientific importance can be measured quantitatively. Quantitative traits are often affected by the joint effect of multiple genes and the environment, and one way to dissect the underlying genetic architecture is to identify quantitative trait loci, QTL, in the genome [2]. A QTL is a chromosomal region, locus, harboring one or several genes that affect the trait under study. Ten or more QTL can influence a single trait. Real examples show that the QTL can interact in nonlinear ways, and therefore it is desirable to simultaneously model their effects. A simultaneous search for n QTL can be regarded as a global optimization problem in n dimensions.

We consider QTL mapping in experimental populations. The exact form of the optimization problem depends on the choice of mapping method. With the widely used linear regression parametric method, searching for n QTL is equivalent to finding the n-element vector $\bar{x}$, representing the combination of n QTL positions, that minimizes

$$f(\bar{x}) = \min_b \|A(\bar{x})b - y\|_2^2,$$

where A is a matrix of indicator variables depending nonlinearly on $\bar{x}$ and of fixed effects coefficients used to remove the influence of non-QTL factors. The number of rows in A equals the number of individuals in the population. The vector y contains the phenotype values, and b is a vector of regression variables. Other common parametric methods result in a generalized least squares problem for every $\bar{x}$ [4].

2 Methods.

There are two parts to the computational problem. The first is the kernel problem, i.e. to solve the (generalized) least squares problem of the objective function for a given $\bar{x}$. In [4] we show how the special structure of the least squares problem can be efficiently exploited in an updating algorithm. We compare our kernel algorithm with the library least squares solver routines used in standard software. The relative gain in arithmetic operations, compared to the most suitable of the library routines, is roughly proportional to $(k_{fix}/k)^2$, the square of the ratio of the number of fixed columns in A to the total number of columns.

The second part of the computational problem is the global problem, i.e. finding the $\bar{x}$ that minimizes $f(\bar{x})$. The standard way of finding the global optimum of the objective function, i.e. the most likely positions of the QTL given a mapping method and model, is to perform an exhaustive search in a dense grid covering the search space. This ensures that the global optimum is found, but the method is computationally slow, and in practice infeasible for analyses in more than two dimensions. Already in two dimensions it is very

[1]Dept. of Information Technology, Uppsala University, Box 337, SE-751 05 Uppsala, Sweden. E-mail: kajsa.ljungberg@it.uu.se, sverker.holmgren@it.uu.se

[2]Roslin Institute (Edinburgh), Roslin, Midlothian EH25 9PS, United Kingdom. E-mail: contact@orjancarlborg.com

time-consuming to perform the thousands of searches required to obtain robust empirical significance thresholds for statistical evaluation of the results. In higher dimensions a more efficient global optimization algorithm is essential. In [5] we adapt an optimization algorithm called DIRECT [3] to the QTL search problem. We compare DIRECT with exhaustive search and with a genetic optimization algorithm previously used for QTL search [1].

3 Results.

We test DIRECT in 2-6 dimensional searches. To verify that the global optimum is found when testing on real data sets, an exhaustive search is performed in two and three dimensions. We also analyze simulated data with known QTL locations in up to six dimensions.

Global search method	Kernel algorithm	2D search, 191 pigs $(k_{fix}/k)^2 \approx 0.78$ [seconds]	3D search, 850 chickens $(k_{fix}/k)^2 \approx 0.03$ [seconds]
Exhaustive search	Library routine G02DAF	150,000	1,140,000,000
Exhaustive search	Library routine SQRDC	4,800	14,800,000
Exhaustive search	New updating algorithm	1,500	11,400,000
Genetic algorithm	New updating algorithm	40	8600
DIRECT	New updating algorithm	4	360

Table 1: Approximate CPU time in seconds required for one search to find the global optimum.

In Table 1 we show examples of results for searches in two and three dimensions using two farm animal real data sets. The reported CPU times are approximate, and the times for exhaustive search using G02DAF in 2 and 3 dimensions and SQRDC in 3 dimensions have been extrapolated from shorter runs. The kernel updating algorithm gives a substantial gain when the number of fixed columns in A is large compared to the total number of columns. The library routine G02DAF, used in standard software, is very slow due to several extra computations not needed for QTL mapping purposes.

Using DIRECT for the global optimization results in a one order of magnitude speed-up compared to the genetic optimization algorithm using a carefully tuned parameterization. DIRECT is 2-3 orders of magnitude faster than exhaustive search in two dimensions, and 4-5 orders of magnitude faster in three dimensions. This enables routine searches in at least three dimensions, including derivation of empirical significance thresholds.

References

[1] Carlborg, Ö., Andersson, L. and Kinghorn, B. 2000. The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics* 155:2003–2010.

[2] Doerge, R. 2002. Mapping and analysis of quantitative trait loci in experimental populations. *Nature Reviews Genetics* 3:43–52.

[3] Jones, D., Perttunen, C. and Stuckman, B. 1993. Lipschitzian optimization without the Lipschitz constant. *Journal of Optimization Theory and Applications* 79:157–181.

[4] Ljungberg, K., Holmgren, S. and Carlborg, Ö. 2002. Efficient algorithms for quantitative trait loci mapping problems. *Journal of Computational Biology* 9:793–804.

[5] Ljungberg, K., Holmgren, S. and Carlborg, Ö. 2003. Simultaneous search for multiple QTL using the global optimization algorithm DIRECT. Technical Report 2003-043, Department of Information Technology, Uppsala University, Sweden. Submitted to *Bioinformatics*.