



UPPSALA
UNIVERSITET

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Science and Technology 133*

Numerical Algorithms for Mapping of Multiple Quantitative Trait Loci in Experimental Populations

KAJSA LJUNGBERG



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2005

ISSN 1651-6214
ISBN 91-554-6427-0
urn:nbn:se:uu:diva-6248

Dissertation presented at Uppsala University to be publicly examined in 2446, MIC, Uppsala, Friday, January 13, 2006 at 10:15 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract

Ljungberg, K. 2005. Numerical Algorithms for Mapping of Multiple Quantitative Trait Loci in Experimental Populations. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 133. 61 pp. Uppsala. ISBN 91-554-6427-0.

Most traits of medical or economic importance are quantitative, i.e. they can be measured on a continuous scale. Strong biological evidence indicates that quantitative traits are governed by a complex interplay between the environment and multiple quantitative trait loci, QTL, in the genome. Nonlinear interactions make it necessary to search for several QTL simultaneously. This thesis concerns numerical methods for QTL search in experimental populations. The core computational problem of a statistical analysis of such a population is a multidimensional global optimization problem with many local optima. Simultaneous search for d QTL involves solving a d -dimensional problem, where each evaluation of the objective function involves solving one or several least squares problems with special structure. Using standard software, already a two-dimensional search is costly, and searches in higher dimensions are prohibitively slow.

Three efficient algorithms for evaluation of the most common forms of the objective function are presented. The computing time for the linear regression method is reduced by up to one order of magnitude for real data examples by using a new scheme based on updated QR factorizations. Secondly, the objective function for the interval mapping method is evaluated using an updating technique and an efficient iterative method, which results in a 50 percent reduction in computing time. Finally, a third algorithm, applicable to the imputation and weighted linear mixture model methods, is presented. It reduces the computing time by between one and two orders of magnitude.

The global search problem is also investigated. Standard software techniques for finding the global optimum of the objective function are compared with a new approach based on the DIRECT algorithm. The new method is more accurate than the previously fastest scheme and locates the optimum in 1-2 orders of magnitude less time. The method is further developed by coupling DIRECT to a local optimization algorithm for accelerated convergence, leading to additional time savings of up to eight times. A parallel grid computing implementation of exhaustive search is also presented, and is suitable e.g for verifying global optima when developing efficient optimization algorithms tailored for the QTL mapping problem.

Using the algorithms presented in this thesis, simultaneous search for at least six QTL can be performed routinely. The decrease in overall computing time is several orders of magnitude. The results imply that computations which were earlier considered impossible are no longer difficult, and that genetic researchers thus are free to focus on model selection and other central genetical issues.

Keywords: Scientific computing

Kajsa Ljungberg, Department of Information Technology, Box 337, Uppsala University, SE-75105 Uppsala, Sweden

© Kajsa Ljungberg 2005

ISSN 1651-6214

ISBN 91-554-6427-0

urn:nbn:se:uu:diva-6248 (<http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-6248>)

And there's so much work to do,
so many puzzles to ignore.

Och
det finns så mycket jobb att ta sig an,
så många gåtor att strunta i.

John Ashbery,
ur *Little Sick Poem*

(övers: Tommy Olofsson
& Vasilis Papageorgiou)

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals.

- I Ljungberg, K., Holmgren, S., Carlborg, Ö. (2002) Efficient algorithms for quantitative trait locus mapping. *Journal of Computational Biology*, 9:793-804⁽¹⁾
- II Ljungberg, K., Holmgren, S., Carlborg, Ö. (2004) Simultaneous search for multiple QTL using the global optimization algorithm DIRECT. *Bioinformatics*, 20:1887-1895⁽²⁾
- III Ljungberg, K., Mishchenko, K., Holmgren, S. (2005) Efficient algorithms for multi-dimensional global optimization in genetic mapping of complex traits. Tech. Rep. 2005-035, Division of Scientific Computing, Department of Information Technology, Uppsala University. Under review for *Optimization methods and software*.
- IV Ljungberg, K. (2005) Efficient evaluation of the residual sum of squares for quantitative trait locus models in the case of complete marker genotype information. Tech. Rep. 2005-033, Division of Scientific Computing, Department of Information Technology, Uppsala University. Under review for *Bioinformatics*.
- V Jayawardena, M., Ljungberg, K., Holmgren, S. (2005) Using parallel computing and grid systems for genetic mapping of multifactorial traits. Tech. Rep. 2005-036, Division of Scientific Computing, Department of Information Technology, Uppsala University.

⁽¹⁾ Reprinted with permission from Mary Ann Liebert, Inc. publishers.

⁽²⁾ Reprinted with permission from Oxford University Press.

Contents

1	Introduction	9
2	Basic genetics	11
2.1	Building blocks of hereditary information	11
2.2	Crossover	12
2.3	Genetic linkage and distance measures	13
2.4	Phenotypes and genotypes	14
2.5	Mendelian and quantitative traits	14
2.6	Genetic effects	15
2.7	Experimental populations	15
2.8	QTL mapping in inbred or outbred populations	17
2.9	Glossary	17
3	QTL models for experimental populations	19
3.1	Choosing a model and quantifying effects	19
3.2	QTL models for experimental populations	20
4	The computational problem	23
5	Search space	25
6	Summary of contributions	27
7	Evaluating the objective function	29
7.1	Missing genotype data	29
7.2	Numerical methods for linear regression	31
7.2.1	Continuous indicator variables	31
7.2.2	Polynomials for g^a and g^d	33
7.2.3	Complete genotype information	34
7.3	Numerical methods for interval mapping	35
8	Finding the global optimum	39
8.1	Global optimization algorithms	39
8.1.1	Stochastic algorithms	39
8.1.2	Deterministic methods	40
8.1.3	Hybrid methods	42
8.1.4	The DIRECT algorithm	42
8.2	DIRECT applied to QTL mapping	43
9	Conclusions and outlook	49
10	Acknowledgments	51
11	Summary in Swedish	53
	Bibliography	57

1. Introduction

Traits that vary continuously, e.g. the height of a tree, the bone density of a mouse and the blood pressure of a human, are called quantitative. They are in general affected by an interplay between multiple genetic factors and the environment. In contrast, qualitative traits fall into one of a small number of categories, for example ABO blood type, and normally depend on a single gene. Most animal and plant traits that are medically or economically important, such as cholesterol levels, body weight, susceptibility to infections, agricultural crop yield and milk production, are quantitative. Consequently, understanding the genetic factors behind such traits is of great value and interest.

Quantitative traits are often studied in experimental populations of e.g. mice. A statistical analysis of an experimental animal population can reveal the locations of the quantitative trait loci, QTL, affecting the trait. A simultaneous search for d QTL involves solving a d -dimensional global optimization problem with many local optima. Using standard software already a two-dimensional search is computationally demanding, and a search in three dimensions is infeasible. This thesis presents efficient numerical methods, several orders of magnitude faster than standard methods, for simultaneous search for up to six QTL in experimental populations. Using these methods, previously impossible analyses can be performed routinely. As a result, analysts can focus future work on issues such as QTL model selection and verification.

Sections 2-5 provide necessary background theory and a description of the computational problems considered in this thesis. Sections 7 and 8 summarize the new methods, including theory and results.

2. Basic genetics

To be able to describe the type of data studied in this thesis, some basic genetics is needed as a background. More detailed theory can be found in standard textbooks such as [2].

2.1 Building blocks of hereditary information

Each cell in an organism contains deoxyribonucleic acid, or DNA for short. The DNA is a long chain of molecular building blocks, nucleotides. There are four different nucleotides, called A, T, G and C, and they make up the alphabet used for the hereditary information. The DNA chain has two strands of nucleotides twisted around each other in a helix, and each nucleotide is paired up with a specific one at the other strand. A is always connected with T, and G with C. Because of this structure, the double helix can always be reconstructed from a single strand by adding the correct complementary nucleotide at each position.

When a cell divides in two, both daughter cells must receive a copy of all the DNA, i.e. the whole genome. During replication the two strands of the mother cell DNA are separated, and new nucleotides are put together to make two double helices identical to the original one, see Figure 2.1.



Figure 2.1: A DNA chain consists of two strands of complementary nucleotides. When DNA is replicated, two double chains identical to the original one are created.

The human genome consists of approximately 3 billion nucleotide pairs. The chain is divided into pieces called chromosomes. A gene is a short segment of a chromosome where the nucleotide sequence gives the blueprint for a particular substance in the body, for example insulin. Only a small fraction of the DNA consists of genes. In between the genes there are long non-coding regions of which the function is largely unknown. A position on a chromosome is called a locus and every gene has its fixed locus, e.g. '2000 nucleotide pairs from the beginning of chromosome 5'.

A human has 23 pairs of chromosomes, i.e. 46 in total. In each pair one chromosome has been inherited from the mother and the other from the father. The chromosomes in a pair are said to be homologous. They have the same genes at the same loci, but they may have different variants, different so called alleles, of the gene. Recall the eye color example from standard high school texts on genetics. We inherit one eye color allele from each parent, either a blue eyes allele or a brown eyes allele. A person who has the same allele on both chromosomes, e.g. blue eyes - blue eyes, is said to be homozygous at that locus. If the two alleles are different the person is heterozygous. Different alleles still have very similar nucleotide sequences, and may differ in as little as only one nucleotide pair. Different allele combinations are referred to as different genotypes.

Genetic polymorphisms is the general term for differences between nucleotide sequences at the same locus. Polymorphisms can occur anywhere in the genome. For some of them there are relatively simple laboratory tests that can reveal which variant of the polymorphism an individual has inherited on each homologous chromosome. Such detectable polymorphisms are examples of so called genetic markers.

2.2 Crossover

Germ cells, i.e. sperm cells and egg cells, differ from all other cells in that they contain only half the genome of the individual, 23 single chromosomes in human germ cells instead of 23 pairs. Germ cells originate from 46-chromosome cells, and a sophisticated process called meiosis ensures that exactly 23 chromosomes, and exactly one from each homologous pair, ends up in each daughter cell. Before the homologous chromosomes are distributed to the daughter cells they are paired up side by side. While they are positioned close together a process called crossover often occurs, see Figure 2.2. The homologous chromosomes randomly exchange large chunks of DNA. As a result, each chromosome that a child has inherited from a parent will most often contain segments from both grandparents.

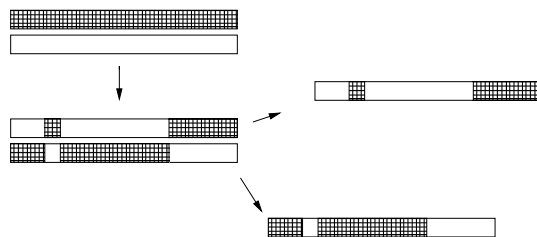


Figure 2.2: When homologous chromosomes are lined up next to each other during the process of germ cell formation, crossing over often occurs. Sections of genetic information are exchanged, and the resulting chromosomes then move to different germ cells.

For example, there are on average 2-3 crossovers per chromosome during human meiosis. Thus the possible number of germ cell variants is much greater than $2^{23} \approx 8.4$ million, which is the number of different ways of picking one chromosome from each of 23 pairs.

2.3 Genetic linkage and distance measures

Two loci are said to be linked if there is a tendency for them to be inherited together. Loci on different chromosomes are always unlinked, since different chromosomes are distributed independently of each other during meiosis. If the loci are located close together on the same chromosome they are linked, i.e. they will often be inherited together, but not always because of occasional recombination in between the loci. If this occurs, the two loci end up on different chromosomes in the homologous pair. The further apart the loci are, the more likely it is that a recombination event occurs, and if they are very far apart the chance of them being separated equals the chance of them staying on the same chromosome, and then the loci are again unlinked.

The likelihood of recombination is not the same everywhere on the chromosomes. For example there is less recombination in the so called centromere region, which plays a crucial role in meiosis and therefore must be kept undisturbed. Thus the number of recombination events is not directly translatable to a certain number of nucleotide pairs.

Using genetic markers, the pattern of inheritance can be tracked through families. For example, by analyzing a marker linked to the eye color gene in several generations, it is possible to determine from which grandparents a child has inherited its eye color alleles. More importantly, finding a marker linked to a disease can lead to location of the faulty gene causing the disease. Finding the gene is very valuable in the search for the cure.

The distance between two loci can be expressed either as physical or genetic distance. The physical distance is the number of nucleotide pairs between the loci. Physical distance is very time-consuming and expensive to measure since it requires sequencing of the DNA. The genetic distance, which is more easy to measure, is the average crossover frequency between the loci, and is expressed in Morgan, M, or centi-Morgan, cM. Note that the observed crossover frequency c_{PQ} between two markers P and Q is not the same as the true crossover frequency, because if there has been two crossover events between P and Q then no recombination will be observed. For the same reason, if we have three markers ordered $A-B-C$, then c_{AC} will be smaller than the sum of c_{AB} and c_{BC} . In order to get an additive distance measure a logarithmic function of the number of crossovers is used, for example Haldane's mapping function. If the observed recombination frequency is c and the genetic distance is d , using Haldane's function we get $d = -(0.5 \ln(1 - 2c)) \cdot 100$ cM. This function only gives an approximation of the true crossover frequency, and

is dependent on some assumptions. More details are given in [42]. Compared to physical distance, genetic distance is easier to measure since it is enough to perform a marker analysis over generations, counting how often two markers located on the same chromosome in the parental generation are separated in the next generation. The fact that the genetic distance is a measure of the crossover frequency is important for the discussion in Section 8.2.

2.4 Phenotypes and genotypes

Phenotype is the visible character of an individual. Hair color, weight, health status and blood group are all phenotypes. The phenotype depends on both the individual's genotype and environment. With environment we mean everything that has happened to and surrounded an individual from time of first cell division. Individuals with the same genotype can have radically different phenotypes. The more similar the environment is, the more similar they will be.

The relative influence of the genotype and the environment varies between different traits. Height is mostly genetically determined, even though e.g. malnutrition can cause arrested growth. In contrast, lung cancer is almost exclusively caused by environmental factors like cigarette smoke and asbestos, although some people are more sensitive than others because of their genotype. The fraction of the phenotypic variation that is caused by genes is called the heritability of the trait.

2.5 Mendelian and quantitative traits

A trait that is governed by a single gene is often called a Mendelian trait, since it follows the rules of inheritance described by the genetics pioneer Gregor Mendel [46]. Variants of Mendelian traits are often clearly separable into discrete classes, such as ABO blood groups or smooth/wrinkled garden pea seeds. Huntington's disease is an example of a recessive Mendelian disorder. Environmental influence may in some cases cause continuous variation also in Mendelian traits.

Quantitative traits are traits that can be measured on a continuous scale. Examples are weight and predisposition to diabetes. Most traits of medical and economic importance have continuous distributions. Many of the quantitative traits show a normal distribution within a population. Often quantitative traits are polygenic, i.e. they depend on many genes acting together as a team.

A single gene with k possible allele pair combinations may be seen as a random variable with k possible states. A trait could be affected by n independent, equally important genes which each could give k different effects on the phenotype. The central limit theorem of statistics gives that if n is

large enough, the overall phenotype will be normally distributed. An alternative perspective is that n genes gives a total number of genotypes of k^n . The phenotypic differences between similar genotypes can be small, and further blurred by environmental variation, giving a continuous distribution overall. If the environment has a very small importance relative to the genotype, i.e. the heritability is large, many genes are needed to give a continuous distribution, while if the variation caused by the environment is large, i.e. the heritability is low, the trait can be continuously distributed even if only a very small number of genes is involved.

The concept of quantitative trait loci is central to this thesis. A quantitative trait locus, or QTL, is a locus where there is a gene or regulatory element that affects a quantitative trait. The term QTL in general refers to a larger region than a gene, and a single QTL may contain more than one gene.

2.6 Genetic effects

Assume a gene has alleles Q and q. For some genes the effect of the alleles on the phenotype is purely additive, i.e. the total effect is the sum of the individual effects. Then an individual with genotype Qq will have a phenotype which is the exact average of phenotypes for genotypes QQ and qq.

Genetic dominance means deviation from additivity, i.e. the phenotype of the heterozygote is different from the mean of the homozygotes. Consider the simple example of a certain flower with a color that is determined by the genotype at a single locus. If we assume that the qq genotype gives white flowers and the QQ genotype red flowers, then if the Qq genotype also gives red flowers, it means that the Q allele is completely dominant.

Dominance concerns alleles at the same locus. When the the total effect of multiple loci is different from the sum of the individual effects it is called epistasis. Epistasis is believed to be very common and important, see e.g. [58, 61, 15]. In [62] it is shown that the vitamin D receptor gene and the collagen I α 1 gene interact and together give a large effect on bone density and osteoporosis in humans, while the individual effects of the genes are small. Even without the experimental evidence available it is reasonable to assume that interactions exist, since genes encode substances that interact in the body.

2.7 Experimental populations

An individual that is homozygous at all loci is said to be inbred. Here we use animals as an example, but the concept is also used for plants. Controlled crosses between inbred lines is a useful tool for understanding the genetics behind a trait. For inbred populations, when crossover occurs during meiosis it does not result in new genetic combinations, since the homologous chromo-

some carry the same alleles. All germ cells will be identical, and the offspring of two inbred animals from the same family, or line, will be genetically identical to the parents. Starting with inbred lines, different kinds of experimental populations can be obtained. Crossing animals from two different inbred lines gives offspring that are heterozygous at all loci. The parent generation is called P₁ and the offspring generation is called F₁. When the F₁ animals form germ cells, crossover will cause chromosomes to contain segments from both P₁ animals. Intercrossing F₁ animals gives the F₂ intercross generation. Sometimes the intercross is also referred to as an F₂ cross. The intercross animals can be homozygous for either grandparental genotype as well as heterozygous, see Figure 2.3.

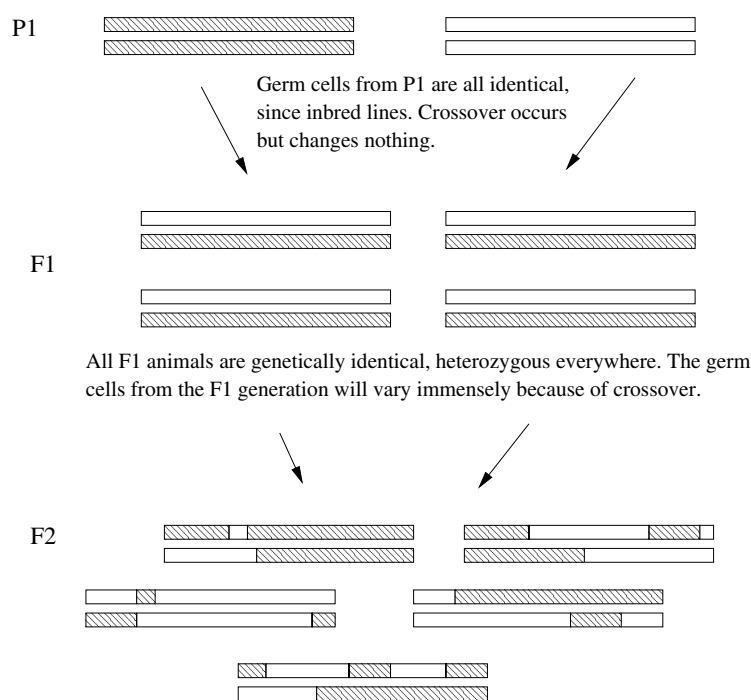


Figure 2.3: A description of how to obtain an intercross, or F₂, experimental population from two inbred lines.

Crossing an F₁ animal with one of the P₁ lines gives a so called back-cross. The backcross generation will be either heterozygous or homozygous for the P₁ parent genotype at all loci, but never homozygous for the other inbred line. There are additional types of experimental crosses that are not discussed in this thesis, for example recombinant inbred lines and advanced intercrosses.

There is no genetic variation among individuals of an inbred line, whereas so called outbred lines present great genetic variation. An intercross from two outbred lines can still be useful if the two lines are homozygous for different alleles at all the QTL affecting the interesting trait, and the methods presented in this thesis still apply. However, in general outbred populations are genet-

ically very heterogeneous. Humans are outbred, as are livestock and other domestic animals of economic importance.

2.8 QTL mapping in inbred or outbred populations

Most traits of medical or economic importance, including predisposition to the major diseases in the industrialized world and growth rate in crops and farm animals, are quantitative. Therefore it is desirable to find, to map, the underlying QTL so that the mechanisms can be better understood. A recent survey [40] lists genes, connected with e.g. cancer and cardiovascular disease, that have been identified in humans and other mammals using QTL mapping as a first step in the analysis.

In an outbred population, e.g. of humans or livestock, the genetic variation and environmental influence make studies extremely difficult, unless the heritability is unusually large. In general controlled crosses are not available, and QTL mapping must be performed in the population that is available. Introduction to QTL mapping in outbred populations is given in [43], and some techniques are reviewed in [30]. One method is variance components analysis, and in [25] a good background is given along with a proposed two-step strategy.

In experimental populations of e.g. laboratory mice, the genetic variation is better controlled and the heritability can be increased by keeping the environment more constant. Also, the phenotypic differences are often larger, making QTL detection easier. This thesis concerns QTL mapping methods for experimental populations only. Strategies for isolated sub-problems may still be applicable to outbred and natural populations, but this has not been investigated.

2.9 Glossary

Additive effect the effect of n alleles equals the sum of the individual effects.

Allele any of the alternative forms of a gene.

Backcross a cross between animals from F_1 and either of two P1 lines.

Chromosomes Organized DNA structures. Humans have 23 pairs.

Crossover Homologous chromosomes randomly exchange large segments when germ cells are formed.

DNA A double-stranded chain of nucleotides that encode the genetic information.

Dominance effect deviation from additivity at one locus.

Epistasis deviation from additivity between different loci.

F₁ cross between two different inbred lines.

Gene DNA sequence encoding one substance.

Genetic distance An additive measure of crossover frequency.

Genotype (part of) an individual's combination of alleles.

Heritability the proportion of observed variation in a particular trait that can be attributed to the genotype in contrast to environmental factors.

Heterozygosity Different alleles at the same locus on pair of homologous chromosomes.

Homologous chromosomes A pair of chromosomes containing the same linear gene sequences, each derived from one parent.

Inbred line animals homozygous at all loci.

Intercross cross between animals from the same F₁ family.

Linkage Two loci that are close enough on the same chromosome to have a tendency to stay together during meiosis.

Locus (pl. loci) The position on a chromosome of a gene or other chromosome marker; also, the DNA at that position.

Mapping function how observed crossover frequencies are translated to genetic distances.

Marker a sequence of bases at a unique physical location in the genome, which varies in a detectable way between individuals.

Morgan the unit of genetic distance.

Phenotype the visible properties of an organism that are produced by the interaction of the genotype and the environment.

Physical distance The number of nucleotide pairs in between two loci.

Polygenic influenced by multiple genes.

QTL a locus where the genotype affects a quantitative trait.

Quantitative traits traits that vary continuously in a population.

Trait a (partly) inherited characteristic.

3. QTL models for experimental populations

3.1 Choosing a model and quantifying effects

All QTL mapping methods involve a model that describes the relation between genotypes and phenotypes. Two main strategies can be distinguished for choosing a model, finding the QTL and quantifying the effects. Either all three problems are solved simultaneously, or the task is divided into sub-problems that are iteratively investigated until a satisfactory solution has been found.

Bayesian QTL mapping [54, 59] is an example of the simultaneous strategy. The number of QTL, their location, effects and interactions are used together with phenotype and genotype information as parameters in a likelihood function. The probability that e.g. there are two QTL affecting the trait is obtained by integrating the likelihood function over the parameter space where $d = 2$ and normalizing with the integral over the whole space. The integration is computationally involved, and in general Monte Carlo methods are used. There has been some success with the Bayesian approach, although the computational demand limits the applicability of the method. Few initial assumptions about the model are required and therefore this method is useful for model selection.

Partial least squares, PLS, [63, 64] is a method for finding appropriate models for continuous output variables by selecting a subset of descriptors from very large parameter sets. The phenotype is modeled as a linear combination of PLS components obtained from eigenvectors of the genotype and phenotype covariance matrix. All marker genotypes as well as epistatic indicator variables are included in the genotype data matrix. PLS has been tried for QTL mapping [9], but large blocks of correlated data, like genotype information on chromosomes, is difficult to deal with using this technique.

The tasks of choosing a model including the number of QTL and their interactions, handling missing genotype information, finding the QTL locations and estimating the effects can be addressed one at a time to make computations more manageable. In general, many models must be tried and the results compared in order to choose a final model. In this thesis we present methods for the computationally most challenging parts of this strategy.

3.2 QTL models for experimental populations

We focus on QTL mapping in experimental populations which are derived especially for genetic analyses. Standard model classes are described in [45]. Let d be the number of QTL and $x = [x_1 \ x_2 \ \dots \ x_d]$ the vector of d QTL positions. Let n be the number of individuals in the experimental population, y the vector of n phenotype observations, $k \geq d$ the total number of parameters in the model and b the vector of k genetic effects. A general linear QTL model can then be formulated as

$$y_i = \sum_{j=1}^k a_{ij} b_j + \varepsilon_i, \quad (3.1)$$

where y_i is the phenotype value of individual i , a_{ij} is the indicator variable of individual i for the j th parameter, b_j is the corresponding regression coefficient, and ε_i is the error. In matrix form we have

$$y = Ab + \varepsilon, \quad (3.2)$$

where the matrix A is the $n \times k$ design matrix and ε is the error vector.

A common experimental population is the backcross, for which there are two possible genotypes at each locus. Let $g_i^a(x_j)$ denote the indicator variable for the additive effect at the j th QTL for individual i , taking the value 0 or 1 depending on the genotype at x_j . A standard model without epistasis is then

$$y_i = 1 \cdot b_\mu + \sum_{j=1}^d g_i^a(x_j) b_j^a + \varepsilon_i, \quad (3.3)$$

where b_μ is the mean and b_j^a is the additive effect of QTL j . Epistasis can be included in the model by also including terms which are products of indicator variables for the effects at individual loci. A d -QTL backcross model including all pairwise epistatic interactions is

$$y_i = 1 \cdot b_\mu + \sum_{j=1}^d g_i^a(x_j) b_j^a + \sum_{j=1}^{d-1} \sum_{l=j+1}^d g_i^a(x_j) \cdot g_i^a(x_l) b_{jl}^{aa} + \varepsilon_i, \quad (3.4)$$

and a d -QTL model including all pairwise and all three-way epistatic interactions is

$$\begin{aligned} y_i = 1 \cdot b_\mu + \sum_{j=1}^d g_i^a(x_j) b_j^a + \sum_{j=1}^{d-1} \sum_{l=j+1}^d g_i^a(x_j) \cdot g_i^a(x_l) b_{jl}^{aa} \\ + \sum_{j=1}^{d-2} \sum_{l=j+1}^{d-1} \sum_{m=l+1}^d g_i^a(x_j) \cdot g_i^a(x_l) \cdot g_i^a(x_m) b_{jlm}^{aaa} + \varepsilon_i. \end{aligned} \quad (3.5)$$

Higher order interactions than three-way are normally not modeled, since they would be very difficult to verify experimentally.

The intercross is the other most commonly used experimental population. In this case there are three possible genotypes at each locus, and two indicator variables are used for each QTL to model the possibilities. Let $[g_i^a(x_j) \ g_i^d(x_j)] \in \{[1 \ 0], [0 \ 1], [-1 \ 0]\}$ denote the indicator variables of individual i for QTL j , where the values depend on the genotype at x_j . This is the model of [24]. An alternative parameterization is $[g_i^a(x_j) \ g_i^d(x_j)] \in \{[1 \ -0.5], [0 \ 0.5], [-1 \ -0.5]\}$ as in [20], which leads to a different interpretation of the regression parameters. An intercross model without epistasis is

$$y_i = 1 \cdot b_\mu + \sum_{j=1}^d \left(g_i^a(x_j) b_j^a + g_i^d(x_j) b_j^d \right) + \varepsilon_i, \quad (3.6)$$

and one with all pairwise epistatic interactions is

$$\begin{aligned} y_i = 1 \cdot b_\mu + \sum_{j=1}^d \left(g_i^a(x_j) b_j^a + g_i^d(x_j) b_j^d \right) \\ + \sum_{j=1}^{d-1} \sum_{l=j+1}^d \left(g_i^a(x_j) \cdot g_i^a(x_l) b_{jl}^{aa} + g_i^a(x_j) \cdot g_i^d(x_l) b_{jl}^{ad} \right. \\ \left. + g_i^d(x_j) \cdot g_i^a(x_l) b_{jl}^{da} + g_i^d(x_j) \cdot g_i^d(x_l) b_{jl}^{dd} \right) + \varepsilon_i. \end{aligned} \quad (3.7)$$

It is also common to include covariate terms to model the effect of for example sex or birthweight. In the case of a binary sex covariate the indicator variable a_{cov}^i is equal to 1 if animal i is a male and 0 otherwise. The corresponding regression parameter b_{cov}^i will be an estimate of the systematic size difference between males and females. The variables a_{cov} are constant regardless of where the QTL are located. This gives

$$y_i = \sum_{j=1}^{k_{cov}} a_{ij} b_j + \sum_{j=k_{cov}+1}^{k_{cov}+k_{gen}} a_{ij} b_j + \varepsilon_i, \quad (3.8)$$

where k_{cov} is the number of covariates and k_{gen} is the number of genetic parameters. The mean has been grouped with the covariates since the indicator variable is independent of x . The covariate parameters are often discrete, like for males and females, but can also be continuous. Interactions between QTL and discrete covariates can be modeled in the same way as interactions among QTL, i.e. by adding products of the respective indicator variables. The matrix A in (3.2) can be partitioned by columns as $A = [A_{cov} \ A_{gen}]$ where $A_{cov} \in \mathbb{R}^{n \times k_{cov}}$ is constant, and $A_{gen} \in \mathbb{R}^{n \times k_{gen}}$ depends on the QTL positions x . Thus the entries of A_{gen} are functions of x .

As a final example we present a two-QTL intercross model with one binary covariate. No epistasis is included, but all covariate-genotype interactions. We

get

$$\begin{aligned}
y_i = & 1 \cdot b_\mu + g_i^a(x_1)b_1^a + g_i^d(x_1)b_1^d + g_i^a(x_2)b_2^a + g_i^d(x_2)b_2^d + cov \cdot b^{cov} \\
& + g_i^a(x_1) \cdot cov \cdot b_1^{a \cdot cov} + g_i^d(x_1) \cdot cov \cdot b_1^{d \cdot cov} \\
& + g_i^a(x_2) \cdot cov \cdot b_2^{a \cdot cov} + g_i^d(x_2) \cdot cov \cdot b_2^{d \cdot cov} + \varepsilon_i,
\end{aligned} \tag{3.9}$$

which for a specific set of data translates to

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 & 1 & -1 & 0 & 0 & 1 \\ \vdots & & & & & & & & & \end{pmatrix} \tag{3.10}$$

in matrix form.

The number of parameters in the different models can be calculated using a general formula. Let p be the number of parameters for one locus. With an intercross population $p = 2$ (g^a and g^d), and with a backcross population $p = 1$ (g^a only). The mean gives one parameter and the individual QTL gives $d \cdot p$ parameters. Including all pairwise epistatic interactions requires $p^2 \binom{d}{2}$ parameters, and including all 3-way epistatic interactions requires $p^3 \binom{d}{3}$ parameters. This means that, in total, we get at most $\sum_{j=0}^3 (p^j \binom{d}{j})$ parameters.

4. The computational problem

After selecting a model of a form described in Section 3.2, the next step in QTL mapping is to find the QTL positions. This constitutes the main computational problem in QTL mapping. Given a model, any set of d hypothetical QTL positions x can be used as input when building the design matrix $A(x)$. To distinguish between hypothetical and true positions, the latter will henceforth be denoted x_{QTL} . The goal is to optimize the model fit over all possible positions x and to compute the corresponding residual sum of squares, RSS_{opt} , according to

$$RSS_{opt} = \min_{b,x} (A(x)b - y)^T (A(x)b - y), \quad (4.1)$$

where the system of equations is sometimes weighted by a diagonal matrix W . The position vector x that minimizes (4.1) is denoted x_{opt} , and is the most probable set of QTL positions for the given model. The expression (4.1) is a separable non-linear least-squares problem where the model is a linear combination of non-linear functions. Following [27], the solution of (4.1) can be separated into two parts: The inner, linear problem,

$$RSS(x) = \min_b (A(x)b - y)^T (A(x)b - y), \quad (4.2)$$

is referred to as evaluation of the objective, or kernel, function. Methods for solving (4.2) are presented in Papers I and IV in this thesis. The outer, non-linear problem,

$$\min_{x \in G^d} RSS(x), \quad (4.3)$$

is referred to as the global search problem and is the topic of Papers II, III and V.

There will always exist a set of hypothetical QTL positions x_{opt} that optimizes (4.3), but the question remains whether it can be established that $x_{opt} = x_{QTL}$. It is also necessary to compare models with different number of QTL and different sets of interaction parameters. To evaluate the result, a statistical test is performed. RSS_{opt} is compared to a significance threshold, which may be derived theoretically. However, a theoretical derivation depends on a large number of assumptions that may not be valid. An alternative method is to compute empirical significance thresholds for each experimental population and model using permutation tests [19, 23]. In this case, at least 1000 QTL analyses of randomly permuted data are explicitly performed and the optimal RSS-values are sorted, giving an empirical, as opposed to theoretical, distribution of random test results. This method is robust but time-consuming.

In summary, the minimization problem (4.1) arises in two settings. When searching for a set of QTL in a real data set, both the optimal set of loci x_{opt} and the corresponding RSS_{opt} are interesting, while when performing permutation tests for determining a significance threshold, only RSS_{opt} is needed.

5. Search space

The outer computational problem (4.3) is a d -dimensional global optimization problem, which normally has a large number of local minima. The QTL search should in principle be performed over all x in a d -dimensional hypercube where the side is given by the size of the genome (which is measured in the unit *Morgan*).

There are two levels of structure in the search space G^d , which are illustrated in Figure 5. The genome is divided into C chromosomes, and this gives the first level of structure. The search space hypercube consists of a set of C^d d -dimensional unequally sized *chromosome combination boxes*, cc-boxes. A cc-box is identified by a vector of chromosome numbers $c = [c_1 \ c_2 \ \dots \ c_d]$, and consists of all x for which x_j is a point on chromosome c_j . The ordering of the loci does not affect the value of the objective function (4.2). Therefore, we can restrict the search space G^d to cc-boxes identified by non-decreasing sequences of chromosomes. In addition, in cc-boxes where two or more edges span the same chromosome, e.g. $c = [1 \ 8 \ 8]$, we need only consider values of x such that $x_k < (x_{k+1} - S)$ for k for which $c_k = c_{k+1}$. The distance S between two hypothetical QTL on the same chromosome must be chosen large enough for some recombination to have occurred between x_k and x_{k+1} . Otherwise the corresponding columns of $A(x)$ would be identical. The function (4.2) is continuous within cc-boxes but discontinuous between them, since the individual chromosomes are completely unlinked and the entries of the design matrix may change in any manner when moving from one chromosome to another.

On each chromosome, a set of marker positions defines the locations where the genetic information is completely determined by the experimental procedure (for perfect data sets). This gives the second level of structure. Each cc-box consists of a set of d -dimensional unequally sized *marker boxes*, m-boxes, defined by the marker positions and the endpoints of the chromosome. The function (4.2) is smooth within m-boxes but has discontinuous derivatives across markers. Two QTL cannot be resolved within the same marker interval, since without any marker in between, there are no verified recombination events that confirm that the genotypes at the two loci are different.

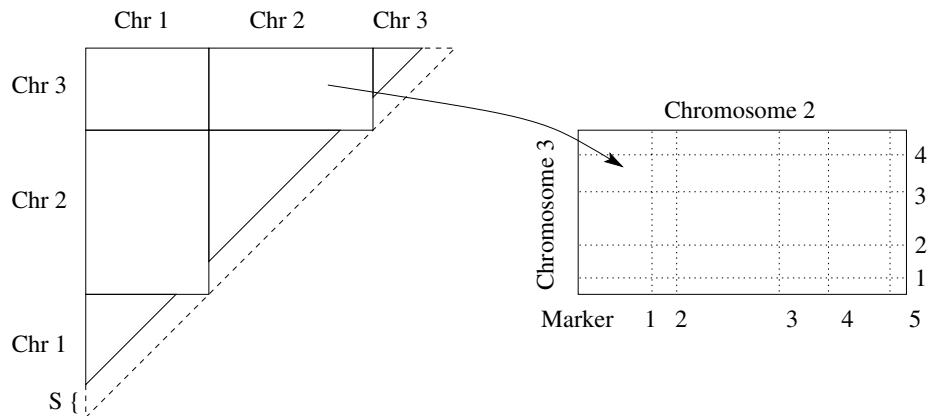


Figure 5.1: The structure of part of the search space when $d=2$. Each chromosome combination box can in turn be divided into marker boxes. The symmetry of the search space leads to that only chromosome combination boxes with non-decreasing sequences of chromosomes need to be considered. S is the minimum distance between two QTL on the same chromosome.

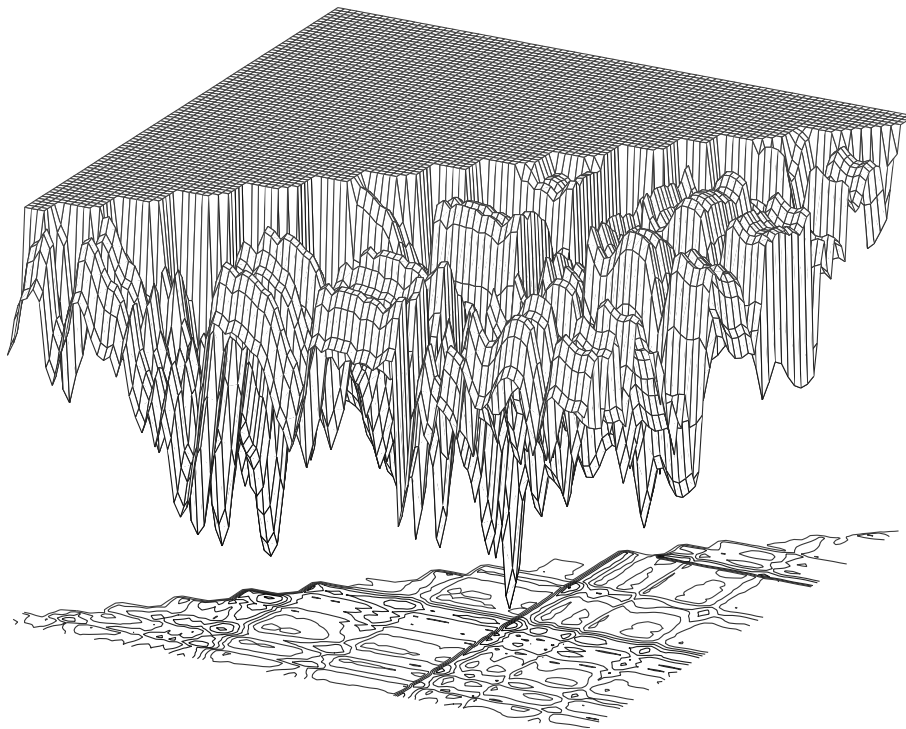


Figure 5.2: An example of part of the function surface for $d=2$. The function has a large number of widely spaced optima. The discontinuities at chromosome boundaries can be seen as straight lines in the contour plot below the surface.

6. Summary of contributions

Sections 2-5 provided background theory and a description of the computational problem considered in this project. Sections 7 and 8 summarize the theory and results of the papers which this thesis is based on:

Paper I A method for solving (4.2) using updated QR factorizations is presented. The more covariates are included in the model, the larger the gain from using the method. Applied on a real data set from an intercross between wild boar and domestic pigs [3] the method reduces the CPU time by one order of magnitude compared to the standard method.

Paper II An adaptation of the global optimization algorithm DIRECT [36] is used for solving (4.3) with real data [3, 55] in two and three dimensions. The algorithm is compared with two other solution methods, and is found to be equally accurate and at least one order of magnitude faster than the previously fastest method.

Paper III A two-phase algorithm where global DIRECT is coupled to a local optimization algorithm is applied on (4.3) in two to six dimensions with simulated data. Using a local algorithm accelerates convergence and the optimum is found up to eight times faster than when performing a run of DIRECT only.

Paper IV Problem (4.2) is efficiently solved for the case when complete genotype information at x is available. The gain in CPU time is one order of magnitude or more compared to the standard method. The method is applicable to any data set if an appropriate statistical method is used.

Paper V A parallel grid computing implementation of exhaustive grid search for (4.3) in two and three dimensions is presented. The tool can be used for verifying the location of the global optimum for real data sets, which in turn is valuable as a reference result when developing high-dimensional search methods as for example in Paper III.

7. Evaluating the objective function

This section concerns evaluation of the objective function (4.2), also called the kernel function, and summarizes background, theory and results of Papers I and IV.

7.1 Missing genotype data

In order to evaluate (4.2), the design matrix $A(x)$ must be constructed, and this requires genotype information at x . The genotypes are only known at the set of genetic marker positions $M = \{m_{ij}\}$, where m_{ij} is the coordinate of the j th marker on chromosome i . The chromosomes vary in length and have different numbers of irregularly spaced markers. Often some marker genotypes are missing for some individuals, and in between markers no genotypes are known. Normally it is interesting to evaluate the kernel function (4.2) also for $x \notin M$, which makes it necessary to deal with the problem that part of the genotype information required to build the design matrix $A(x)$ using the rules presented in Section 3.2 is missing. The naïve way to deal with missing information is to only evaluate the kernel function for $x \in M$ and to discard individuals whose marker genotypes are unknown. As the density of marker sets increases and the quality of data acquisition improves, this strategy becomes increasingly attractive, but other methods are still more widely used.

Not all genotypes are equally probable at a locus with missing information. The process of crossover, see Section 2.2, follows certain rules, and therefore the genotypes at informative markers close to the locus of interest greatly influence the likelihood of a certain genotype. Crossover is often modeled as a Poisson process. The Haldane and Kosambi mapping functions provide standard formulas for estimating the a priori probability that a crossover has occurred between two loci [43]. This structure of the uncertainties in the data make for example total least squares, described e.g in [8], an unsuitable method for QTL mapping.

There are several ways to use the mapping functions for incorporating incomplete genotype information in the kernel function. The traditional approach is interval mapping, IM, for mapping of single QTL [41]. The extension to multiple loci is multiple interval mapping, MIM [38]. Given a model of the type described in Section 3.2, a likelihood function is formulated based on the phenotypes and known genotypes together with the a priori recombination probabilities. The likelihood is then maximized as a function of QTL loca-

tions, effects and unknown genotypes. The relation to Bayesian QTL mapping could be noted. There a likelihood function of similar type is formulated, but it is integrated instead of maximized. Maximizing the likelihood is a nonlinear problem, which is different from (4.2) and must be solved using an iterative method.

Another nonlinear approach, which however retains the basic form of the kernel function (4.2), is the weighted linear model method of [34]. Here, the design matrix rows are replicated for all individuals with uncertain genotypes, and one row for each possible genotype is included in an augmented design matrix A_{aug} . Each row is weighted by the a posteriori probability of the corresponding genotype, where the probability is a function of the known genotypes, the mapping function, the regression parameters and the phenotypes. The weights are iteratively refined until convergence is achieved.

Multiple imputation [57, 4] is a strategy giving only linear subproblems, thus avoiding the need for iterative methods. In both [57] and [4] imputation is used to generate a set of complete genotype data realizations, consistent with the known genotypes, using a hidden Markov model following the rules for genetic crossover. Then one alternative is to combine the realizations in a single augmented design matrix and give each replication a weight $1/r$, where r is the number of data realizations [4]. This is related to the weighted linear model method described above, but no iterative scheme is required, and some allowed genotypes may not be sampled at all. A second alternative is to compute the RSS separately for each realization and use an average of the results as the kernel function value [57].

A computationally attractive method to handle missing information, without excluding individuals as with the naïve method, is the linear regression approximation to interval mapping [44, 28, 39]. Here the discrete indicator variables $g^a \in \{1, 0, -1\}$ and $g^d \in \{1, 0\}$, described in Section 3.2, are replaced with continuous functions $g^a = P_{AA} - P_{BB}$ and $g^d = P_{AB} + P_{BA}$, where P_{GG} is the a priori probability, as computed using the known genotype information and a mapping function, of genotype GG . If the genotype is exactly known, the probabilities are either 0 or 1 and the resulting indicator variables equal to the known genotype case of Section 3.2. With this method, model parameterizations that are equivalent in the full information case can result in different RSS for the same data when genotypes are uncertain.

The results when using IM, the linear regression approximation or multiple imputation have been compared by several authors [44, 28, 37, 57]. The smaller the proportion of missing data, the smaller the differences between the methods, and when the genotypes are exactly known the methods are equivalent. In Papers II, III and V, which all investigate the global problem (4.3), the linear regression method is used. Paper I presents methods for IM and linear regression, and in Paper IV algorithms for multiple imputation and the weighted linear model method are presented.

7.2 Numerical methods for linear regression

7.2.1 Continuous indicator variables

When using the linear regression approximation to interval mapping [44, 28, 39], evaluating the objective function amounts to solving the standard least squares problem

$$\min_b RSS(x) = (A(x)b - y)^T (A(x)b - y), \quad (7.1)$$

where the discrete genotype indicator variables making up the matrix $A(x)$ have been replaced with continuous, exponential genotype probability functions of x .

In Paper I we studied (7.1). This problem can be solved using any standard algorithm for least squares problems. In [29] the NAG software library routine G02DAF is used, in [17] the normal equations are solved using LU factorization of $A^T A$, and in the QTL analysis package QTL Cartographer [6] the LINPACK routines SQRDC and SQRSL are used. In Paper I we derived a more efficient scheme where we exploit the specific structure of the QTL mapping objective function. First we note that, as described in e.g. [8], it is not necessary to compute the parameter vector b , which is done in all three examples above, to obtain RSS. Instead we compute the residual sum of squares $RSS = (Ab - y)^T (Ab - y)$ using the QR factorization of A . Furthermore, in Paper I we point out that since A_{cov} does not depend on x , much computational work can be saved by updating the QR factorization for each new point x , instead of solving the complete problem for every new point as in the software mentioned above. Updating techniques for QR factorization of matrices with a number of constant columns are described in [8]. Details of the updating for the QTL mapping problem, using Householder and Givens transformations, are given in Paper I.

In Paper I we implemented the updating scheme in a routine of our own, called UQRSL, but the updating can also be performed using the LAPACK library routines DGEQRF, or DGEQP3 if column pivoting is used, and DORMQR. In the initiation step $Q_{cov}R_{cov} = A_{cov}$ is computed using either DGEQRF or DGEQP3, and $Q_{cov}^T y$ is computed using DORMQR. These results are stored. For each RSS evaluation A_{QTL} is built using the genotype information at x , and $Q_{cov}^T A_{QTL}$ is computed. Then DGEQRF or DGEQP3 is used to compute $\tilde{Q}\tilde{R} = (Q_{cov}^T A_{QTL})_{k_{cov}+1:n, 1:k_{QTL}}$, the QR factorization of the $k_{cov} + 1$ to n last rows of $Q_{cov}^T A_{QTL}$. Finally the RSS is computed as $\|\tilde{Q}^T (Q_{cov}^T y)_{k_{cov}+1:n}\|_2^2$. Performing computations on only the lower blocks of $Q_{cov}^T A_{QTL}$ and $Q_{cov}^T y$ is trivially achieved by sending pointers to the appropriate array elements to the library functions.

If A is (almost) rank deficient, i.e. there are columns in A which are (almost) linearly dependent, the Householder QR factorization procedure may break down. QR with column pivoting can be used in order to be able to guarantee

an accurate solution in the (rare) case of near rank-deficiency. User-friendly software for QTL mapping should notify the user when A is rank deficient, as this implies that there is a problem with either the model or the data. Linearly dependent columns in A imply that two or more QTL genotype coefficient columns are almost the same. This can occur for example when two putative QTL too close together are considered. Then, there will be no (or very few) recombination events between the putative QTL positions, and the corresponding columns in A_{QTL} will be identical (or very similar). In this case there is not enough information in the data to accurately model the effects of two putative QTL. Rank deficiency can also occur when the marker genotype data is insufficient and the uncertainty leads to equal a priori genotype probabilities for many individuals.

The gain in arithmetic operations from using updated QR factorizations depends on the number of individuals m and the number of parameters $k = k_{cov} + k_{gen}$. The QR factorization of A_{cov} requires about $2k_{cov}^2(n - k_{cov}/3)$ arithmetic operations. The complete factorization of A without updating requires $\sim 2k^2(n - k/3)$ arithmetic operations. Updating with standard library functions, as described above, requires extra function calls and slightly increased memory traffic, but experience shows that the effect of this is marginal. The relative gain from employing the updating algorithms is roughly proportional to $(k_{cov}/k)^2$. When $k_{gen} = 2$ and $k_{cov} = 31$, a real example examined in Paper I, the number of arithmetic operations is reduced by $\sim 88\%$. For a small ratio (k_{cov}/k) the gain is only about 1%, and then it can in practice be faster to perform a regular factorization without updating. Empirical evidence suggest that updating is beneficial whenever there is at least one cofactor in addition to the mean.

Figure 7.1 shows the total CPU time required for solving the least squares kernel problem in an exhaustive grid search over the entire genome for two one-QTL and two two-QTL models. The data dimensions and models are copied from real data examples [3, 55]. The timings for G02DAF are extrapolated from smaller numbers of function evaluations. The gain in using UQRLS, the updating algorithm presented in Paper I, instead of G02DAF, used in [29], is dramatic. The difference is 2 – 3 orders of magnitude for all problem sizes tested. UQRLS requires 20 minutes and 2 h 40 minutes respectively for the two-dimensional scans, while G02DAF would need 45 hours and 54 days, respectively. The differences depend both on the reduction in work introduced by the updating procedure and the extra calculations performed by G02DAF. Comparing SQRDC/SQRSL, used in [6], and G02DAF shows that the unnecessarily detailed analysis of G02DAF is responsible for a 1 – 2 orders of magnitude increase in computational effort, compared to a routine that computes the QR factorization using the standard algorithm. Hence, a comparison of SQRDC/SQRSL and UQRLS demonstrates the gain of our updating algorithm. Updating reduces the CPU time by approximately one order of magnitude for the one-QTL models and slightly less for the two-QTL models.

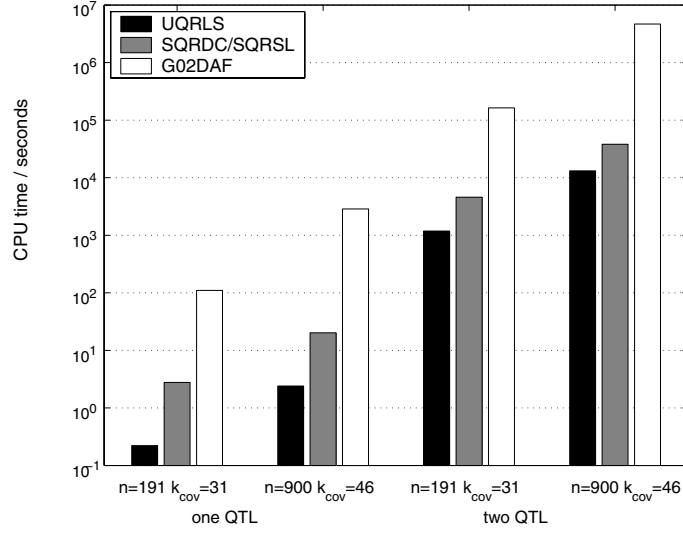


Figure 7.1: Computing times for the kernel problem in an exhaustive grid search in one and two dimensions, using our method UQRLS and two algorithms from standard software.

7.2.2 Polynomials for g^a and g^d

In Paper I we use the standard procedure of evaluating the genotype probability functions g^a and g^d on a regular grid and storing the results in a look-up table. This table is then referred to when building A for different x . In practice this means that the continuous g^a and g^d are replaced by step functions. In Paper III the exponential functions g^a and g^d are replaced with third degree polynomial approximations which are fast and easy to evaluate. This opens the possibility of applying gradient-based optimization methods to the global problem (4.3). Also, storing the polynomial coefficients requires less memory space than the grid values.

For a backcross the analytical expression for $g^a = P_{AA}$ between markers k and $k+1$ is

$$g^a(z) = K(1 + c_1 e^{-2z})(1 + c_2 e^{2z}) = K(1 + c_1 c_2 + c_1 e^{-2z} + c_2 e^{2z}) \quad (7.2)$$

where the local coordinate z on the marker interval replaces x and satisfies $0 \leq z \leq D_k$, D_k is the distance between markers k and $k+1$, $K = 0.5/(1 + e^{-2(k_1+k_2)})$, $k_1 \geq 0$ is the distance from marker k to the closest informative marker to the left, $k_2 \geq D_k$ is the distance from marker k to the closest informative marker to the right, $c_1 = \pm e^{-2k_1}$ with sign depending on the genotype at the left informative marker and $c_2 = \pm e^{-2k_2}$ with sign depending on the genotype at the right informative marker. The analytical expression for g^a and g^d for an intercross have a similar form. The approximating polynomials must give the exact function value at the markers, for continuity reasons. Thus two degrees of freedom remain, and the two free parameters are chosen so that

the squared integral of the error is minimized. In the case D_k is very small only a first degree polynomial is used, in order to avoid oscillations of the approximating function. An unpublished investigation of the polynomial approximation, where interval analysis [48] was used to compare it to the step function traditionally used, revealed that already second degree polynomials give sufficiently accurate approximations of the exact functions.

7.2.3 Complete genotype information

When complete genotype information is available at x , the objective function evaluation can be simplified in a way that significantly reduces the computational load. This is described in Paper IV. The imputation method [57, 4] and the weighted linear model method of [34] also result in problems with the same structure, which makes the method applicable also when genotype information is missing in the original data set. As a consequence, Paper IV presents a solution to the weighted linear model problem mentioned, but not studied, in Paper I.

The objective function (4.2) can, when including a diagonal weight matrix W , be rewritten according to

$$\begin{aligned}
 RSS &= y^T W y - y^T W A (A^T W A)^{-1} A^T W y \\
 &= y^T W y - y^T W U P (P^T U^T W U P)^{-1} P^T U^T W y \\
 &= y^T W y - y^T W U (L D L^T)^{-1} U^T W y \\
 &= y^T W y - y^T W U L^{-T} D^{-1} L^{-1} U^T W y \\
 &= y^T W y - z D^{-1} z
 \end{aligned} \tag{7.3}$$

where $U \in \mathbb{R}^{n \times k}$, $P \in \mathbb{R}^{k \times k}$, $\text{rank}(P) = k$, $U^T W U = L D L^T$ is the factorization of $U^T W U$ into a unit lower triangular matrix L , a diagonal matrix D , and the transpose of L , and z is the solution vector to the unit triangular system $Lz = U^T W y$. The matrix A can also be an augmented design matrix A_{aug} , in which case the number of rows n is increased. The computationally most expensive step when implementing (7.3) is performing the matrix-matrix and matrix-vector multiplications to form $U^T W U$ and $U^T W y$. In Paper IV it is shown how to easily build $U^T W y$ and $U^T W U$ from the genotype class counts and phenotype sums and thus avoid performing the costly matrix multiplications. This gives an at least ten times reduction in computing time compared to computing the QR factorization of A . Additional time can be saved by exploiting the sparsity pattern of the matrix $U^T W U$ during the $L D L^T$ factorization, and of the equation system $Lz = U^T W y$ when solving for z . In the paper, the method is referred to as PERF, Pseudomarker Evaluation of the RSS Function.

Figure 7.2 shows the PERF and QR factorization computing time as a function of the number of model parameters k for 19 different 1-4 QTL models using a backcross data set with 999 mice. The times are normalized with the computing time for PERF applied to the two parameter model. The updating

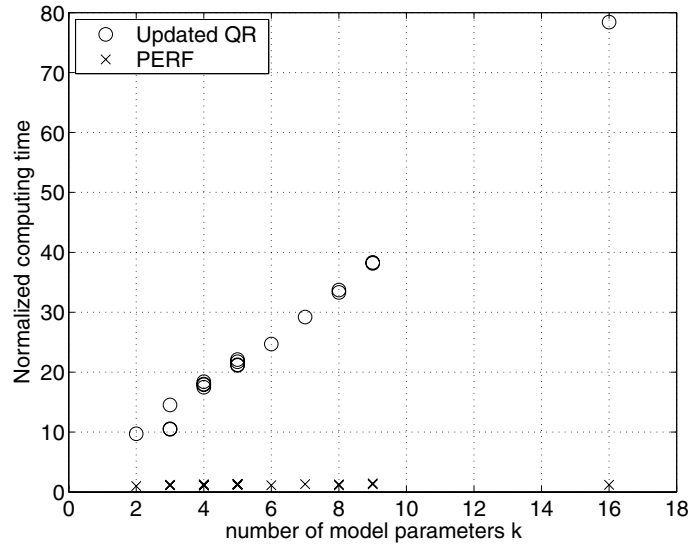


Figure 7.2: The computing time for updated QR factorization and PERF in a back-cross data example. All results are normalized with the computing time for the $k = 2$ model using PERF. The computing time for PERF is close to constant for the different models, while it increases greatly for the standard method QR factorization.

algorithm presented in Section 7.2.1 was used for the QR factorization in the cases when it improved the efficiency, resulting in different computing times for models with the same k but different numbers of covariates. The time for QR factorization increases rapidly with the number of model parameters k , while the computing time for PERF is close to constant. In all cases QR factorization takes 10 times longer than PERF or more.

Table 7.1 shows how many RSS evaluations with PERF that can be performed during the time required for a single QR factorization in the case of 1-4 QTL models without covariates. The results could be interpreted such that if the number of imputations chosen for a particular data set equals the number reported in Table 7.1, the imputation method is as fast as the regression method, see Section 7.2.1. The number of imputations required varies with the type of data and the amount of missing information [57].

7.3 Numerical methods for interval mapping

Traditional interval mapping involves maximizing a likelihood function, resulting in a nonlinear problem that is commonly solved via the ECM algorithm [47], a Gauss-Seidel type method which approximates the EM algorithm of [22].

Model	Gain backcross	(k)	Gain intercross	(k)
1 QTL	10	(2)	13	(3)
2 QTL, -	12	(3)	20	(5)
3 QTL, -	14	(4)	25	(7)
4 QTL, -	16	(5)	29	(9)
2 QTL, pairwise	16	(4)	39	(9)
3 QTL, pairwise	24	(7)	72	(19)
3 QTL, pairwise and 3-way	29	(8)	168	(27)

Table 7.1: The gain from using *PERF* instead of *QR* factorization for the same model in a dataset of 999 individuals. The gain is computed as (CPU for *QR*) divided by (CPU for *PERF*). The models include no covariates. '-' indicates no interactions, 'pairwise' denotes pairwise epistatic interactions and '3-way' three-way interactions.

In Paper I we show that maximizing the likelihood function is equivalent to solving a generalized least squares problem of the form

$$\min_{b_{cov}} RSS = (A_{cov}b_{cov} - y)^T V^{-1} (A_{cov}b_{cov} - y), \quad (7.4)$$

where the matrix V^{-1} is a function of the regression parameters and RSS, which makes the problem nonlinear. In [8] general linear models resulting in problems of the form (7.4) are described.

When the problem is formulated as (7.4) we can use the EM algorithm to find the value of the objective function instead of the approximating ECM algorithm. Solving (7.4) for a fixed matrix V^{-1} corresponds to the M-step in the algorithm, and we present a fast method for this computation using the fact that V^{-1} is only a small rank modification of the identity. Computing a new matrix V^{-1} given the solution to the M-step represents the E-step.

We also compare the performance of the ECM algorithm applied to the formulation in [66] with the EM algorithm applied to (7.4), using a model where $n = 190$, $k_{cov} = 31$ and $k_{gen} = 2$ as a test example. Three different ways of choosing the initial values are examined, namely using the null hypothesis parameter estimates, the least squares estimates and the estimates from the previous position. This is explained further in Paper I.

Figure 7.3 shows the normalized numbers of iterations required for a full one-dimensional genomic scan for the two algorithms. 'Null' denotes using parameter estimates from the null hypothesis model as initial values, 'prev' denotes using the estimates from the previous genome positions x , and 'LS' denotes using the least squares estimates. The computational work needed for one iteration is very similar for the EM and ECM algorithms. The results show that the EM algorithm on average converges in less than half the number of iterations. An improvement was expected since the ECM algorithm is an approximation of the EM algorithm. There are many other algorithms that could

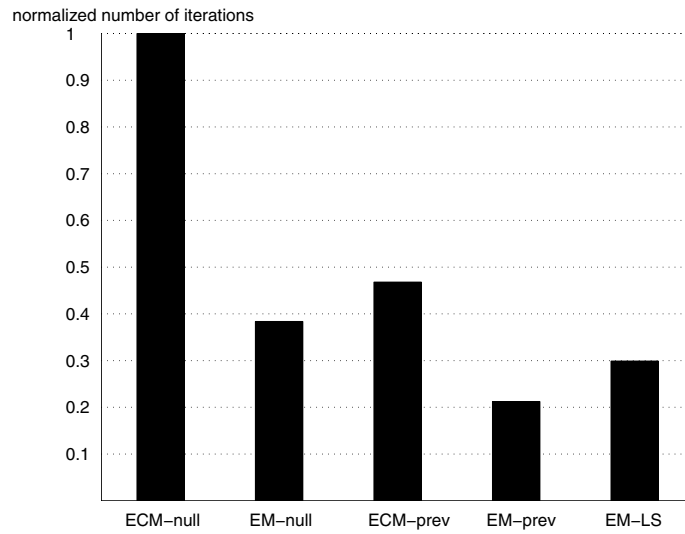


Figure 7.3: Normalized numbers of iterations needed for a full genomic scan using the EM or ECM algorithm with different starting values. A single iteration takes approximately the same time for the two methods, implying that using the EM algorithm is twice as fast as the ECM algorithm. Using parameter estimates from the previous coordinates as initial values is more efficient than using the null hypothesis estimates.

be applied for solving the nonlinear IM problems, but we have not investigated this issue further.

8. Finding the global optimum

As described in Section 4, the global task in QTL mapping is to find the point x that minimizes (4.3), which is a multidimensional global optimization problem with a large number of local optima. This problem is treated in Paper II, III and V. The optimization algorithms considered in the articles can be used together with any variant of the objective function, but in our three papers we have chosen the linear regression method.

A challenge in global optimization is determining when the optimum has been found, i.e. it is difficult to know when to stop when performing a computation on real data. It is in general impossible to know when a potential solution is sufficiently close to the exact global optimum of a continuous function, unless some additional information of the function is available. Interval analysis [48] can give rigorous bounds on the remaining error, but often it is very computationally demanding to obtain a sufficiently sharp enclosure of the optimum. The inherent “unsolvability” of the global optimization problem implies that there is no single best way to search for the optimum. A wide variety of algorithms have been developed that are suitable for different types of functions. There are stochastic and deterministic algorithms, as well as algorithms including both stochastic and deterministic steps. A collection of current algorithms is found in [51].

8.1 Global optimization algorithms

Before going into details about the methods used in this thesis, a brief overview of some different types of global optimization algorithms is given.

8.1.1 Stochastic algorithms

The most simple example of a purely stochastic algorithm is Monte Carlo optimization, where a large number of randomly selected points are evaluated, and the one with the lowest function value is taken as the optimum. More advanced stochastic methods include Simulated Annealing [53]. Here a collection of points in the search space are selected randomly and the function values evaluated. Then a small, random displacement is generated for each point and the new function values are computed. A new position is accepted with probability 1 if the displacement results in a smaller function value, or, if the new function value is greater, with probability $e^{-\Delta f/kT}$, where T is the

“temperature”, k is the Boltzmann constant and Δf is the change in function value. The system is started at a high temperature and then gradually cooled until only displacements giving a reduction in function value are accepted. Tabu search [26] is related to simulated annealing, since in both methods trajectories that can go both downhill and uphill in a landscape are generated. In Tabu search all points in a neighborhood of the current point are considered for the next step. The point giving the largest decrease, or smallest increase, in function value is selected, although points recently visited are impermissible, tabu, unless they give a decrease in the function value.

Genetic optimization algorithms is another stochastic approach [31]. A population of solution vectors x is created randomly. The corresponding function values are evaluated and used for assigning a fitness value to each solution vector x . Vectors with high fitness are randomly combined and modified to create new solutions, and this process is iterated. A genetic optimization algorithm has been used for simultaneous mapping of two QTL [13]. In Paper II we used the same algorithm on (4.3) for comparison with a new method, see Section 8.1.4. The advantage of the genetic algorithm is that it is a general purpose method, which is easy to implement for widely different optimization problems and problem dimensions. Standard software packages are available. The drawbacks include slow local convergence and a large number of parameters that need to be well-tuned for good performance. The parameters must often be set differently for similar problems. Furthermore, in the computations it can be difficult to take advantage of special knowledge about the objective function.

8.1.2 Deterministic methods

A deterministic, brute-force method to find the global optimum of (4.3) is to perform an exhaustive grid search in small steps, e.g. $1cM$. This is referred to as the enumerative strategy in [51]. Given that the grid of search points resolves the fastest variation in the objective function, this method is guaranteed to locate the optimum, however it will often be very computationally expensive. An exhaustive n -dimensional grid search in a $L cM$ genome requires $\binom{L}{n}$ function evaluations. In a typical experiment $L \approx 2500$. Then a three-dimensional search amounts to $\sim 3 \cdot 10^9$ evaluations, and a four-dimensional search $\sim 2 \cdot 10^{12}$ function evaluations. Exhaustive three-dimensional grid search is possible using the linear regression method, but it is very computationally demanding and a parallel computer is needed. Four-dimensional searches are in practice impossible using today’s computational methods without access to many hundreds of processors, at least if permutation tests are required. In Papers II and V we perform a number of exhaustive two- and three-dimensional searches in real data sets using an efficient implementation on parallel computers, in order to know the global optimum when later testing more advanced methods. For these

computations we extended the updating methods from Paper I and used them also for parts of A_{QTL} . When x_1 and x_2 are fixed and x_3 varies, the factorization of columns affected by only x_1 and x_2 does not change and can be saved between function evaluations.

Forward selection is a popular method in practical QTL analysis, and is based on the simple idea of exhaustive search while avoiding the problem of slow multidimensional optimization. Here, a series of one-dimensional exhaustive searches is performed. The scalar coordinate x_{opt} in each search is taken as the position of a QTL, which is included in the model as a known covariate in the next one-dimensional search. The search sequence is terminated when the inclusion of an extra parameter does not give a sufficiently large decrease in the objective function. For general QTL models, it is not clear how accurate forward selection is. It could be anticipated that the scheme can fail to detect QTL that only affect the phenotype through interactions with other QTL. Several analyses of real data sets have revealed such interactions between pairs of QTL, some of which were only detectable by solving the full two-dimensional optimization problem [58, 61, 16]. Such results motivate our interest for developing efficient algorithms also for high-dimensional QTL mapping problems. Furthermore, in Paper III it is shown that forward selection can fail also when no epistatic interactions are present.

Branch and bound is an important class of deterministic global methods where the search space is divided into smaller and smaller regions. In each iteration the choice of which regions to divide further is made based on computed limits on the function values in each region. Regions are discarded whenever the limits exclude the possibility of improving the currently best actual function value. So called greedy algorithms always choose the region with the currently best actual value for division, while other algorithms also consider the potential of improvement in each region. In Lipschitz optimization the search space is divided into regions, and a Lipschitz constant K , a global maximum on the rate of change of the objective function, is used to eliminate regions which are guaranteed not to contain the optimum. A strict lower limit on the potential function value in a region is computed using K and known function values nearby. Regions where the lower limit is higher than the best known function value are discarded, and the most promising regions are explored further. Lipschitz optimization is further described in [32]. Interval analysis [48] is another method that can be used to compute rigorous bounds on the function values.

DIRECT, presented in [36], is a deterministic optimization algorithm which uses the ideas of Lipschitz optimization but does not require the knowledge of a Lipschitz constant. In Paper II we have adapted DIRECT for problem (4.3), and in Paper III the method is developed further. DIRECT can be viewed as a grid search algorithm where the evaluations have been ordered so that regions with potentially good function values are explored early while searching in less promising regions is delayed. An heuristic criterion is used to determine

when to stop the computations. DIRECT does well in comparison with other methods in the case of a fixed budget of function evaluations [36]. A more detailed description of the algorithm is given in Section 8.1.4.

8.1.3 Hybrid methods

There is a large collection of hybrid optimization methods, combining stochastic, deterministic and local methods. A search can be initiated by selecting a number of starting points, either randomly or according to a defined pattern. Then a local search is performed from each point. It is possible to use the information from the starting points to guide the continued search. A simple way is to cluster points believed to lie close to the same local optimum and only perform a local search from one point in each cluster [7]. Starting points can also be obtained by running a global algorithm for a number of iterations, and switching algorithm when approaching a local optimum in order to accelerate convergence. A more elaborate approach is to use selected points to build a model of the landscape and then concentrate the search to the “valleys” of the model landscape, c.f. the CGU method [52]. In Paper III we examine the use of a number of local algorithms in combination with DIRECT. More details are given in Section 8.2.

8.1.4 The DIRECT algorithm

The original DIRECT algorithm [36] searches for the global minimum of a Lipschitz continuous function $f(x)$. The search is performed within a d -dimensional box defined by $l_i \leq x_i \leq u_i$, where l_i and u_i are the lower and upper bounds on variable x_i . The search space is divided into gradually smaller hyper-boxes, and in each iteration a new set of boxes is chosen for subdivision. The function value at the center of a box is denoted f_c , and the distance between the center and the vertices is d_{cv} . If the maximal rate of change of $f(x)$ were K , the minimal value that $f(x)$ could possibly attain in a box would be $f_c - K \cdot d_{cv}$. For a given value of K , the box with the lowest limit on the function values is potentially optimal. In DIRECT all boxes that are potentially optimal for any value of K from zero to infinity are selected for subdivision. This means that K does not need to be known. The box selection step is illustrated in Figure 8.1. In a scatter plot of f_c versus d_{cv} , the minimal potential function value in a box is obtained as the intercept of a line with slope K drawn through the box-dot. Selecting all potentially optimal boxes for any K corresponds to determining the lower convex hull of the cloud of box-dots in Figure 8.1, which is a computationally easy task. In the figure lines for $K = 0.5$ and $K = 0.2$ are shown, drawn through the dots representing the boxes giving the smallest y -intercepts for these two K -values.

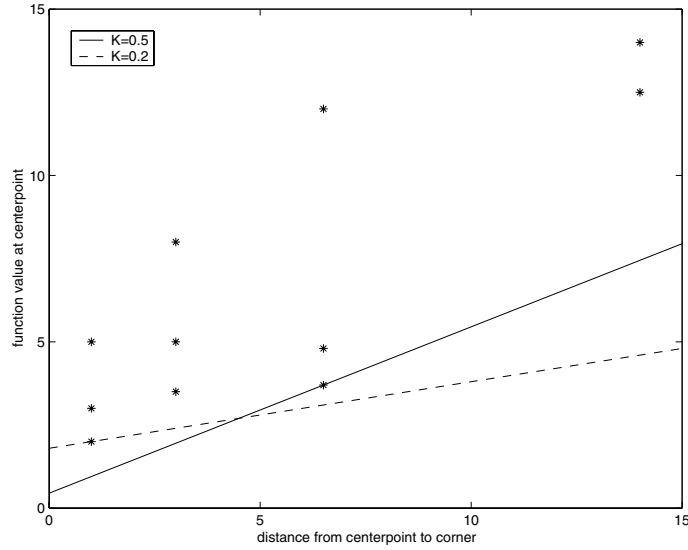


Figure 8.1: Visualization of DIRECT box selection. Each box is represented by a dot in a plot of the function value at the box center versus the distance between the box center and the box corners. A line with slope K through a dot will intercept the y-axis at the minimum function value that could possibly be attained in the box if the Lipschitz constant were equal to K .

8.2 DIRECT applied to QTL mapping

In Paper II we apply DIRECT to the QTL mapping problem (4.3). To adapt the algorithm to the special structure of the search space, which is described in Section 5, the original algorithm is modified. The function (4.2) is discontinuous between cc-boxes, and therefore the search space is divided into cc-boxes already at initiation, and the center of each box is sampled. This is sufficient to fulfill the Lipschitz continuity condition of DIRECT, since the Lipschitz method is used for bounding (4.3) within, not across, hyper-boxes. Symmetric cc-boxes, i.e. boxes where two or more edges span the same chromosome, are divided according to a pattern that ensures that $x_k < (x_{k+1} - S)$ for x_k and x_{k+1} on the same chromosome.

Also, in contrast to the original method, the box sizes are not normalized in order to retain the relation between the distance measure centi-Morgan and change in the genotypes. This relation gives a motivation for using a Lipschitz type algorithm as DIRECT for QTL mapping. A short distance corresponds to few crossover events and little change, while a large distance corresponds to many crossover events. The distance can obviously not be directly translated to difference in objective function value since the change in the objective function depends also on the phenotypes of the individuals in the marker genotype classes.

There is no well-defined convergence criterion for DIRECT, and in Paper II, as well as in [21] and [5], it is observed that the local convergence of DIRECT

is rather slow. We have chosen to run DIRECT for a fixed number of function evaluations and then perform a local exhaustive search as a refinement step. In the three-dimensional searches an intermediate DIRECT phase is performed on the best chromosome combination after the initial set of iterations. Paper III describes improvements achieved by implementing a more efficient local search.

In Paper II we have tested DIRECT on real data sets from two intercrosses between outbred lines. The first is a wild boar \times domestic pig intercross [3] with 191 individuals and a genome size of approximately 2300 cM , and the second is a white leghorn \times red jungle-fowl chicken intercross [55], with 852 animals and a genome size of 2500 cM . All the studied traits were growth related, for example bodyweight at eight days of age in the chicken, and ham weight (including meat, fat and bone) in the pig. The results show that DIRECT accurately finds the global optimum in all real data test cases, and that the computations are performed about one order of magnitude faster than when using the genetic optimization algorithm.

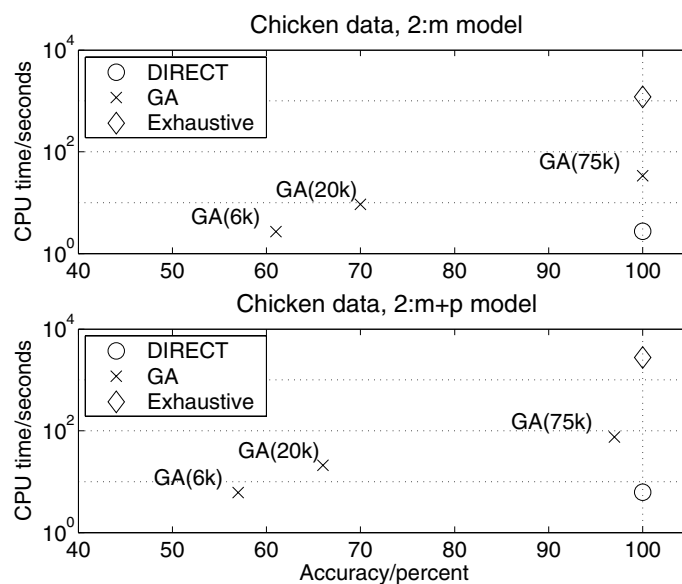


Figure 8.2: CPU time for two-dimensional searches as a function of the percentage of successful localizations of the global optimum for the chicken data set. DIRECT and exhaustive grid search always find the global optimum, while the genetic algorithm sometimes fails. The fewer evaluations are allowed in the genetic algorithm, the more often it fails.

Figure 8.2 shows the average CPU times and accuracy over 9 phenotypes of the chicken data. The models used are a two-QTL model denoted 2:m, including only marginal (additive and dominance) effects, and a model denoted 2:m+p, where also all pairwise interaction effects were included in addition to the 2:m effects. An exhaustive search with the 2:m model requires about

20 minutes, and 46 minutes with the 2:m+p model. DIRECT finds the global optimum in less than 3 and 7 seconds respectively. GA(75k), GA(20k) and GA(6k) denote the genetic optimization algorithm with parameter settings using 75.000, 20.000 and 6.000 function evaluations, respectively. GA(75k) finds the global optimum in close to 100% of the runs, with CPU time 34 and 76 seconds. Using GA(6k), the genetic algorithm with the same number of function evaluations and thus practically the same CPU time as DIRECT, reduces the accuracy from close to 100% to around 60%. GA(20k), which corresponds to the setting of [13], gives intermediate results. The genetic algorithm has more difficulties finding the global optimum when epistasis is included in the model. It was observed already in [13] that the genetic algorithm sometimes failed when a QTL pair lacked significant marginal effects. This can be explained by the forward selection property of the algorithm as discussed in Paper II.

As stated in Section 4.1, the goal when analyzing real data is to find the optimal position x_{opt} and RSS_{opt} , while when analyzing randomized data the goal is only to determine the value RSS_{opt} accurately enough to give usable empirical significance thresholds. In Paper II we find that DIRECT does not always locate the exact optimum when applied to randomized data, which is not surprising since the randomization will in general “smear” the features of the function landscape resulting in many local optima with almost the same function value. Still, we show that the computed significance thresholds are accurate enough for practical use.

In Paper III we developed a two-phase method where DIRECT is coupled to a local optimization algorithm in order to accelerate the final convergence. Searches in up to 6 dimensions were performed on high-noise simulated data. We used the efficient, polynomial based continuous evaluation method described in Section 7.2.2 for evaluation of (4.2), and computed gradients via numerical differentiation when needed.

When a hyper-box smaller than a certain limit is chosen for subdivision, it is sent to the local phase. There it is first investigated whether the box lies completely within an m-box (see Section 5). If not, the box is divided along the marker interval boundaries. This ensures that the local algorithm is only applied within a region where the objective function is smooth, allowing for the application of gradient-based optimization methods. Three local methods were implemented, and the performance compared with a single global DIRECT run, denoted **D**. More details are given in Paper III.

- **DIRECT (D-D)**: We used DIRECT as a local algorithm. In our experiments, the local iteration is stopped when there is no function value improvement for the last two iterations. Note that this two-phase algorithm is not equivalent to a single global DIRECT run with more iterations.
- **Steepest Descent (D-SD)**: The most straight-forward gradient based scheme steepest descent was tested. We exploit an Armijo line search along the negative gradient, where the maximum step length is defined by

Algorithm	D	D-D	D-SD	D-QN
$p_{\text{alg}} \cdot G$	41	32	25	22

Table 8.1: *Stopping rule parameters. The maximum number of function evaluations allowed without function value improvement is $N_f = (p_{\text{alg}} \cdot G)^d / d!$, where d is the number of dimensions.*

the box boundary. The bound constraints are accounted for by a simplified barrier method, where a component of the negative gradient pointing out of the box is set to zero if the current point is close to a box boundary.

- **Quasi-Newton (D-QN)**: We also tried a quasi-Newton scheme, including approximative second derivative information together with the gradient in the local optimization algorithm. We implemented the same line search and barrier method as with steepest descent, but choose the search direction using the BFGS method where an approximate inverse of the Hessian is repeatedly updated during the iterations using the gradients, see e.g. [50].

The search is terminated after a certain number of function evaluations have been performed without any improvement. The number is chosen in a manner that is generalizable to any number of dimensions. For a d QTL model, the size of the search space is $G^d / d!$, where G is the length of the genome in centi-Morgan. This motivates us to set $N_f = (p_{\text{alg}} \cdot G)^d / d!$, where the parameters p_{alg} are determined by performing a large number of numerical experiments for each algorithm, adjusting p_{alg} so that the global optimum is found in all data sets. In Table 8.1, the values of $p_{\text{alg}} \cdot G$ are shown.

Some results are given in Figure 8.3, which shows the largest number of function evaluations performed before termination for each group of simulated data sets with a fixed number of QTL. It is clear that using a two-phase algorithm significantly reduces the number of function evaluations required, even when the DIRECT algorithm is used also for the local optimization. It is also clear that if the gradient based methods are employed, this gives a considerable further improvement. The difference between the **D-SD** and **D-QN** schemes is not very large for the current data sets.

Using numerical experiments it was found that the objective function is often non-convex around the global optimum, and therefore the **D-SD** and **D-QN** methods are not guaranteed to converge, even though in all cases tested here the optima were indeed found. The **D-D** method is not affected by the convexity properties, and will always find the optimum if run for sufficiently many iterations. Therefore a suggested strategy is to exploit the two-phase **D-D** algorithm with a generous number of iterations for local search for determining the best model fit x_{opt} for the genetic data. Then the **D-QN** scheme can be employed for optimization during the permutation test used for determining the significance of the result. In the significance testing, the effect of an eventual small error in x_{opt} is not important, c.f. Section 4.

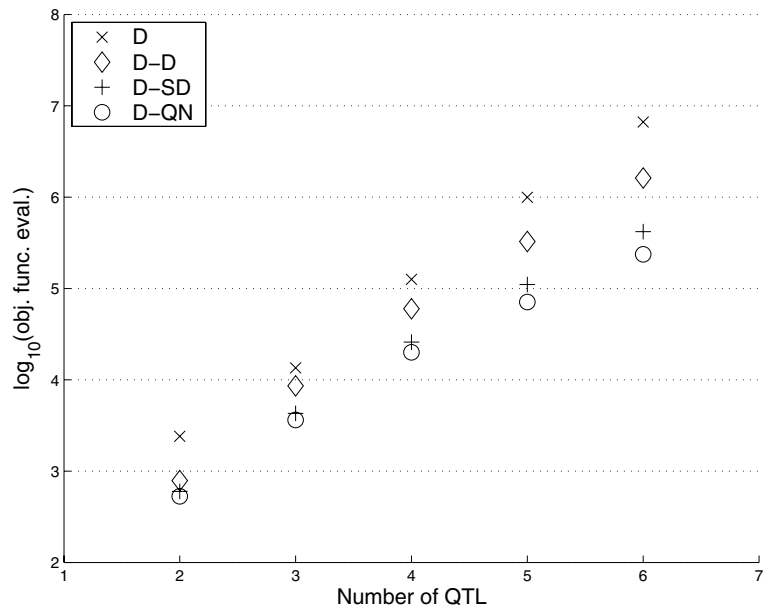


Figure 8.3: The maximal number of function evaluations required for convergence for different values of d , using DIRECT without local search and DIRECT coupled to local DIRECT, Steepest descent and Quasi-Newton. For a fixed d the number of function evaluations is directly proportional to the computing time. When increasing d the time for a single function evaluation also increases. The data is from a simulated backcross population with epistatic interactions between pairs of QTL.

9. Conclusions and outlook

In this thesis we have presented efficient algorithms for simultaneous search for multiple QTL in experimental populations. Algorithms for both the objective function and global optimization problem have been developed. Compared with standard software, the new methods give a two to three orders of magnitude speed-up for typical problems in two dimensions, and the additional improvement for searches in higher dimensions is also measured in orders of magnitude. From this follows that computations which were previously infeasible can now be performed routinely on a desktop computer, including simultaneous search for up to at least six QTL.

Genetic researchers are now able to focus on, for example, the important issue of QTL model selection. Permutation testing [19, 23] is a robust tool when evaluating the significance of a result, and requires extensive computations which now can be easily performed. Large numbers of different parametric models can also be compared, using e.g. the Akaike information criterion [1] or Schwarz Bayesian information criterion [56], see e.g. [4, 11, 10].

It is important to ensure that the methods are readily available for interested geneticists. The two-dimensional optimization method has been widely spread, and is popular among the users. The algorithm has been implemented in Web-QTL [18] and Pseudomarker [65], which are freely available to the public, and in a still unreleased version of R/qrtl [12]. Incorporation is underway of the two-dimensional search and the QR factorization updating method into Grid-QTL [33]. In addition, the method is being integrated into a very large collaborative genetic mapping project [60], where it will be used for detecting epistatic interactions.

It would probably be possible to further improve the methods presented in this thesis:

- A deeper analysis of the objective function in both one and several dimensions could be performed, investigating the convexity properties in the marker boxes. The results could be used for more efficient local search.
- A Lipschitz constant would be useful in the optimization. It may be possible to derive two types of such a constant, one that describes the maximum variation within a chromosome combination box between marker loci, and one that concerns the variation within a marker box.
- A completely different global optimization algorithm could be tried, like for example convex global underestimation.

The algorithms presented in this thesis are efficient, and would be beneficial for more researchers than are currently using them. Therefore another valuable future project could be to:

- Create a grid portal with all algorithms neatly packaged and easy to apply on standard format data sets.

It may be difficult to achieve additional order-of-magnitude reductions for the computational problems discussed in this thesis. However, there is a need for algorithm development in other related fields:

- QTL mapping in natural populations is an important topic. An introduction to the field is given in [43], and some statistical methods are reviewed in [30]. The related computational problems should be investigated.
- So called heterogeneous stocks [49] are derived from inbred lines but are much more genetically diverse than backcross or intercross populations. The statistical methods used to analyze these populations present new computational challenges.
- It is common to collect data from multiple phenotypes for a single population. High-throughput, parallel methods for traditional QTL analysis of each individual phenotype, preferably reusing computations that depend on the genotype data only, are needed. This issue is discussed in [14].
- The expression levels of single genes can be measured using microarrays. The levels vary among individuals, and are examples of quantitative traits. Thousands of expression phenotypes can be measured for a single population. In the field of so called genetical genomics [35], QTL analysis is used to find the QTL that govern these expression phenotypes. Statistical methods for detecting interaction networks and dependencies between the expression levels of different genes are being developed, giving new computational problems that require efficient solutions.

10. Acknowledgments

This work was funded by The Graduate School in Mathematics and Computing (FMB) and the Linneaus Centre for Bioinformatics, both at Uppsala University.

My sincere thanks are due to

- Sverker Holmgren, for outstanding leadership. You have provided guidance when needed and trusted me with freedom otherwise. You have listened in a way that made my thoughts clear and ideas clever, and have provided pivotal input. Your impressive way of discriminating between the essential and the superfluous, only giving specific directions in crucial matters, has made me trust your judgment like no one else's. I would be very fortunate to have a boss like you at my next job.
- Mum and Dad, for instilling in me the unshakable conviction that I can master any intellectual challenge provided I devote some effort to it.
- Örjan Carlborg, for identifying an open problem in an interesting field of research. I have enjoyed much extra attention thanks to the great need for solutions in this area. You also set an excellent example in careful scientific analysis, and your feedback on my work is always most valuable.
- Leif Andersson, for playing a crucial role during the initiation of the project, sharing your valuable data sets and offering your time and expertise in support of my work.
- Anders Sjöberg, for persistent invitations to come and work at TDB. I would not have ended up here if it were not for you.
- Martina Hägglund, Katya Mishchenko and Mahen Jayawardena, for fruitful collaboration and discussions.
- Friends at TDB, for making me look forward to going in to work, and to friends and loved ones outside the workplace, for making my free time enjoyable and relaxing.
- Providence, for Staffan.

11. Summary in Swedish

Algoritmer för genetisk kartläggning av kvantitativa egenskaper i experimentella djurkorsningar

I grupper av djur och växter kan man se en kontinuerlig variation i de flesta egenskaper som är medicinskt eller ekonomiskt betydelsefulla. Egenskaperna kan alltså mätas kvantitativt, och exempel är mjölkproduktion, tillväxthastighet, blodtryck och kolesterolnivåer. Dessa egenskaper påverkas av både miljön och flera genetiska faktorer i samverkan. I detta sammanhang inkluderas individuella livsstilsfaktorer, till exempel mat- och motionsvanor, i miljön. Områden i arvsmassan där det finns gener som påverkar kvantitativa egenskaper kallas QTL, vilket är en förkortning av engelskans *quantitative trait loci*.

Eftersom kvantitativa egenskaper är så viktiga är det intressant att studera deras genetiska bakgrund. Dock är det ofta mycket komplicerat, eftersom påverkan från varje enskild QTL kan vara liten och därför svår att urskilja från effekterna av miljön och andra QTL. Ett sätt att underlätta analysen är att undersöka djur som hållits i en konstant miljö och som har enkla och väldefinierade släktskapsförhållanden. En statistisk analys av sambanden mellan djurens genetiska skillnader och skillnader i den studerade egenskapen kan avslöja var i arvsmassan det finns QTL som påverkar egenskapen.

En central del i den statistiska analysen är att minimera en komplicerad matematisk funktion. Antag att man tror att det är fem QTL som påverkar egenskapen, och har formulerat en funktion utifrån det. Det går då att räkna ut ett funktionsvärde för varje kombination av fem hypotetiska QTL-positioner i arvsmassan. Funktionsvärdet är ett mått på hur väl det går att matematiskt beskriva variationen i egenskapen hos djuren med den genetiska variationen på de fem positionerna. Ett litet funktionsvärde betyder att skillnaden är liten mellan den matematiska modellen och de nivåer på egenskapen man har mätt upp på riktigt. De positioner som ger den minsta skillnaden indikerar var i arvsmassan det är mest troligt att hitta de sökta QTL. Detta minimeringsproblem har två delar ur beräkningssynpunkt.

Den första delen består i att så snabbt som möjligt räkna ut värdet på den matematiska funktionen, för en enda kombination av QTL-positioner. I den här avhandlingen presenteras tre olika algoritmer för detta som kan användas för fyra av de vanligaste sätten att definiera skillnadsfunktionen.

- Den första algoritmen kan användas på skillnadsfunktion A , och utnyttjar att vissa delberäkningar i funktionen ibland kan vara desamma för alla

kombinationer av positioner. Det sker när man i funktionen har inkluderat så kallade fixa effekter, en matematisk beskrivning av miljöpåverkan och annat som kan ändra egenskapen utan att det har något med den genetiska information vid olika positioner att göra. Till exempel är det vanligt att inkludera kön i modellen, eftersom hanar och honor ofta är olika utan att det har något samband med intressanta QTL. Genom att återanvända beräkningarna som är lika för alla positioner kan den totala beräkningstiden minskas med upp till 90%, som i ett verkligt fall som använts som exempel i en artikel i denna avhandling.

- Algoritm nummer två baseras på en omformulering av skillnadsfunktion B. Genom att ta en ny beräkningsväg till samma resultat blir det möjligt att använda befintliga beräkningsmetoder som är mer effektiva än de som används i standardmjukvara. Det leder till att beräkningsarbetet kan reduceras med ungefär hälften.
- Den tredje algoritmen kan användas på skillnadsfunktion C och D. Den utnyttjar att det i vissa fall ingår nästan bara ettor och nollor i beräkningarna. Antag att man ska multiplicera ett stort antal talpar och summera resultatet. Om man samtidigt vet att bara ettor och nollor ingår, till exempel $(0 \cdot 1) + (1 \cdot 1) + (0 \cdot 0) + (0 \cdot 1) + \dots = ?$, så räcker det att räkna i hur många av paren det ingår två ettor, eftersom alla andra produkter blir noll. I en dator tar det tid att utföra multiplikationer med noll och räkna ut $1 \cdot 1$, och det går att spara tid om man kan undvika alla sådana beräkningar i ett problem. Genom att formulera om funktion C och D går det att styra så att den största delen av beräkningarna består av just $1 \cdot 1$ eller multiplikationer med noll. Den tredje algoritmen använder den formuleringen och undviker sedan i princip alla dessa multiplikationer, vilket leder till att beräkningarna går mellan 10 och 100 gånger snabbare än med en standardmetod.

Tillsammans täcker de tre algoritmerna de mest populära varianterna av skillnadsfunktionen.

Den andra delen av minimeringsproblemet är att söka igenom möjliga kombinationer av QTL-positioner och hitta den kombination som ger det minsta värdet på skillnadsfunktionen. Detta kan liknas vid att hitta koordinaterna för botten av den djupaste gropen i ett stort landskap med en mängd hål av olika storlek. I avhandlingen används DIRECT, en algoritm som presenterats tidigare av andra forskare. DIRECT jämförs med metoder som tidigare används för QTL-sökningsproblemet, och visas vara mer än 10 gånger snabbare än den tidigare snabbaste metoden, och dessutom mer pålitlig när det gäller att hitta det allra minsta värdet på funktionen. För att ytterligare förbättra resultaten kopplas DIRECT ihop med algoritmer som är effektiva när man har hittat kanten av en "grop" i funktionen och snabbt vill bestämma de koordinater som motsvarar dess botten. Denna strategi snabbar upp beräkningarna ytterligare några gånger.

Algoritmerna som presenteras i denna avhandling gör att beräkningsproblemet som tidigare utgjorde oöverstigliga hinder i QTL-analys av kontrollerade

djurpopulationer nu kan genomföras rutinmässigt. Delar av programkoden har redan införts i ett antal olika mjukvarupaket som används av genetiker i flera länder. De nya, snabba metoderna gör det möjligt att utföra en mer noggrann statistisk utvärdering av resultaten och på ett säkrare sätt bestämma antalet viktiga QTL och var de finns. Detta ger en bättre utgångspunkt i laboratoriet för nästa steg i analysen, då de centrala genernas exakta position och funktion ska ringas in.

Bibliography

- [1] H. Akaike. Fitting autoregressive models for prediction. *Annals of the Institute of Statistical Mathematics*, 21:243–247, 1969.
- [2] Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, and James Watson. *Molecular Biology of the Cell*. Garland Publishing, Inc., third edition, 1994.
- [3] L. Andersson, C. Haley, H. Ellegren, S. Knott, M. Johansson, K. Andersson, L. Andersson-Eklund, I. Edfors-Lilja, M. Fredholm, and I. Hansson. Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science*, 263:1771–1774, 1994.
- [4] R. Ball. Bayesian methods for quantitative trait loci mapping based on model selection: Approximate analysis using the Bayesian information criterion. *Genetics*, 159:1351–1364, 2001.
- [5] M. Bartholomew-Biggs, S. Parkhurst, and S. Wilson. Using DIRECT to solve an aircraft routing problem. *Computational Optimization and Applications*, 21:311–323, 2002.
- [6] C. Basten, B. Weir, and Z.-B. Zeng. *QTL Cartographer, Version 1.15*. Department of Statistics, North Carolina State University, Raleigh, NC, 2001.
- [7] R W Becker and G V Lago. A global optimization algorithm. In *Proceedings of the 8th Allerton Conference on Circuits and Systems Theory*, pages 3–12, 1970.
- [8] Å. Björck. *Numerical Methods for Least Squares Problems*. Society for Industrial and Applied Mathematics, Philadelphia, 1996.
- [9] Å. Bjørnstad, F. Westad, and H. Martens. Analysis of genetic marker-phenotype relationships by jack-knifed partial least squares regression (pls-r). *Hereditas*, 141:149–165, 2004.
- [10] M. Bogdan, J. Ghosh, and R. Doerge. Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics*, 167:989–999, 2004.
- [11] K. Broman and T. Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society: Series B*, 64:641–656, 2002.

- [12] K. Broman, H. Wu, S. Sen, and G. Churchill. R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, 19:889–890, 2003.
- [13] Ö. Carlborg, L. Andersson, and B. Kinghorn. The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Genetics*, 155:2003–2010, 2000.
- [14] Ö. Carlborg, D.-J. de Koning, K. Manly, E. Chesler, R. Williams, and C. Haley. Methodological aspects of the genetic dissection of gene expression. *Bioinformatics*, 21:2383–2393, 2005.
- [15] Ö. Carlborg and C.S. Haley. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, 5:618–625, 2004.
- [16] Ö. Carlborg, S. Kerje, K. Schütz, L. Jacobsson, P. Jensen, and L. Andersson. A global search reveals epistatic interaction between QTL for early growth in the chicken. *Genome Research*, 13:413–421, 2003.
- [17] Örjan Carlborg. *New methods for mapping quantitative trait loci*. PhD thesis, Swedish University of Agricultural Sciences, 2002. Acta Universitatis Agriculturae Sueciae, Veterinaria 121.
- [18] E. Chesler, L. Lu, J. Wang, R. Williams, and K. Manly. WebQTL: rapid exploratory analysis of gene expression and genetic networks for brain and behavior. *Nature Neuroscience*, 7:485–486, 2004.
- [19] G. Churchill and R. Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138:963–971, 1994.
- [20] C. Cockerham. An extension of the concept of partitioning hereditary variance for analysis of covariance among relatives when epistasis is present. *Genetics*, 39:859–882, 1954.
- [21] S. Cox, R. Haftka, C. Baker, B. Grossman, W. Mason, and L. Watson. A comparison of global optimization methods for the design of a high-speed civil transport. *Journal of Global Optimization*, 21:415–433, 2001.
- [22] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society*, 39:1–38, 1977.
- [23] R. Doerge and G. Churchill. Permutation tests for multiple loci affecting a quantitative character. *Genetics*, 142:285–294, 1996.
- [24] R. Fisher. The correlation between relatives on the supposition of Mendelian inheritance. *Proceedings of the Royal Society Edinburgh*, 52:399–433, 1918.

- [25] A. George, P. Visscher, and C. Haley. Mapping quantitative trait loci in complex pedigrees: a two-step variance component approach. *Genetics*, 156:2081–2092, 2000.
- [26] F. Glover. Tabu search - part i. *ORSA Journal on Computing*, 1:190–206, 1989.
- [27] G.H. Golub and V. Pereyra. The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate. *SIAM Journal on Numerical Analysis*, 10:413–432, 1973.
- [28] C. Haley and S. Knott. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69:315–324, 1992.
- [29] C. Haley, S. Knott, and J.-M. Elsen. Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics*, 136:1195–1207, 1994.
- [30] I. Hoeschele, P. Uimari, F. Grignola, Q. Zhang, and K. Gage. Advances in statistical methods to map quantitative trait loci in outbred populations. *Genetics*, 147:1445–1457, 1997.
- [31] J. Holland. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, 1975.
- [32] R. Horst, P.M. Pardalos, and N.V. Thoai. *Introduction to Global Optimization*. Kluwer Academic Publishers, second edition, 2000.
- [33] Institute of Evolutionary Biology, Roslin Institute and National e-Science Centre, Edinburgh. *GridQTL*. <http://www.gridqtl.org.uk/>.
- [34] R. Jansen. Interval mapping of multiple quantitative trait loci. *Genetics*, 135:205–211, 1993.
- [35] R. Jansen and J.-P. Nap. Genetical genomics: the added value from segregation. *Trends in Genetics*, 17:388–391, 2001.
- [36] D. Jones, C. Perttunen, and B. Stuckman. Lipschitzian optimization without the Lipschitz constant. *Journal of Optimization Theory and Application*, 79:157–181, 1993.
- [37] C.-H. Kao. On the differences between maximum likelihood and regression interval mapping in the analysis of quantitative trait loci. *Genetics*, 156:855–865, 2000.
- [38] C.-H. Kao, Z.-B. Zeng, and R. Teasdale. Multiple interval mapping for quantitative trait loci. *Genetics*, 152:1203–1216, 1999.
- [39] S. Knapp, W. Bridges, and D. Birkes. Mapping quantitative trait loci using molecular marker linkage maps. *Theoretical and Applied Genetics*, 79:583–592, 1990.

- [40] R. Korstanje and B. Paigen. From QTL to gene: the harvest begins. *Nature Genetics*, 31:235–236, 2002.
- [41] E. Lander and D. Botstein. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, 121:185–199, 1989.
- [42] B.-H. Liu. *Statistical Genomics*. CRC Press, 1998.
- [43] M. Lynch and B. Walsh. *Genetics and analysis of Quantitative Traits*. Sinauer Associates, Inc., 1998.
- [44] O. Martinez and R. Curnow. Estimating the locations and the sizes of effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics*, 85:480–488, 1992.
- [45] K. Mather and J. Jinks. *Biometrical Genetics*. Chapman and Hall, 1982.
- [46] G. Mendel. Versuche über Pflanzen-Hybriden. *Verhandlungen des Naturforschenden Vereins*, 1866.
- [47] X.-L. Meng and D. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80:267–278, 1993.
- [48] R.E. Moore. *Interval Analysis*. Prentice Hall, Englewood Cliffs, New Jersey, 1966.
- [49] R. Mott, C. Talbot, M. Turri, A. Collins, and J. Flint. A method for fine-mapping quantitative trait loci in outbred animal stocks. *Proc. Natl Acad. Sci. USA*, 97:12649–12654, 2000.
- [50] J. Nocedal and S. Wright. *Numerical Optimization*. Springer-Verlag, New York, 1999.
- [51] P.M. Pardalos and H.E. Romeijn, editors. *Handbook of Global Optimization Volume 2*. Kluwer Academic Publishers, 2002.
- [52] A. Phillips, J. Rosen, and V. Walke. Molecular structure determination by convex global underestimation of local energy minima. *DIMACS Series in Discrete Math & Theoretical Computer Science*, 23:181–198, 1995.
- [53] C. Gelatt S. Kirkpatrick and M. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.
- [54] J. Satagopan, B. Yandell, M. Newton, and T. Osborn. A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics*, 144:805–816, 1996.

- [55] K. Schütz, S. Kerje, Ö. Carlborg, P. Jensen, and L. Andersson. Analysis of a red junglefowl \times white leghorn intercross reveals trade-off in resource allocation between behavior and production traits. *Behavioural Genetics*, 32:423–433, 2002.
- [56] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.
- [57] S. Sen and G. Churchill. A statistical framework for quantitative trait mapping. *Genetics*, 159:371–387, 2001.
- [58] K. Shimomura, S. Low-Zeddies, D. King, T. Steeves, A. Whiteley, J. Kushla, P. Zemenides, A. Lin, M. Vitaterna, G. Churchill, and J. Takahashi. Genome-wide epistatic interaction analysis reveals complex genetic determinants of circadian behavior in mice. *Genome Research*, 11:959–980, 2001.
- [59] M. Sillanpää and E. Arjas. Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics*, 148:1373–1388, 1998.
- [60] L. Solberg, W. Valdar, D. Gauguier, G. Nunez, A. Taylor, P. Hernandez, S. Davidson, P. Burns, W. Cookson, R. Deacon, J. Rawlins, R. Mott, and J Flint. A protocol for high throughput phenotyping, suitable for quantitative trait analysis in mice. *Mammalian Genome*, 2005. Accepted.
- [61] F. Sugiyama, G. Churchill, D. Higgins, C. Johns, K. Makaritsis, H. Gavras, and B. Paigen. Concordance of murine quantitative trait loci for salt-induced hypertension with rat and human loci. *Genomics*, 71:70–77, 2001.
- [62] A. Uitterlinden, A. Weel, H. Burger, Y. Fang, C. Van Duijn, A. Hofman, J. Van Leeuwen, and H. Pols. Interaction between the vitamin D receptor gene and collagen type I α 1 gene in susceptibility for fracture. *Journal of Bone Mineral Research*, 16:379–385, 2001.
- [63] H. Wold. Soft modelling: The basic design and some extensions. In *Systems under indirect observation II*, pages 1–54. Amsterdam: North Holland Publishing Co., 1982.
- [64] S. Wold, C. Albano, W. Dunn, K. Esbensen, S. Hellberg, E. Johansson, and M. Sjöström. Pattern recognition: Finding and using regularities in multivariate data. In *Food research and data analysis*, pages 147–188. London: Applied Science Publishers, 1983.
- [65] H. Wu, S. Sen, K. Ljungberg, K. Broman, and G. Churchill. *Pseudomarker, Version 2.01*, 2005. <http://www.jax.org/staff/churchill/labsite/software/pseudomarker>.
- [66] Z.-B. Zeng. Precision mapping of quantitative trait loci. *Genetics*, 136:1457–1468, 1994.

Acta Universitatis Upsaliensis

*Digital Comprehensive Summaries of Uppsala Dissertations
from the Faculty of Science and Technology 133*

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-6248



ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2005