UPPSALA
UNIVERSITET

IT Licentiate theses
2010-002

# Using Markov Models and a Stochastic Lipschitz condition for Genetic Analyses

CARL NETTELBLAD

UPPSALA UNIVERSITY
Department of Information Technology

# Using Markov Models and a Stochastic Lipschitz condition for Genetic Analyses

*Carl Nettelblad*
carl.nettelblad@it.uu.se

March 2010

*Division of Scientific Computing*
*Department of Information Technology*
*Uppsala University*
*Box 337*
*SE-751 05 Uppsala*
*Sweden*

http://www.it.uu.se/

Dissertation for the degree of Licentiate of Philosophy in Scientific Computing


ISSN 1404-5117
Printed by the Department of Information Technology, Uppsala University, Sweden

**Abstract**

A proper understanding of biological processes requires an understanding of genetics and evolutionary mechanisms. The vast amounts of genetical information that can routinely be extracted with modern technology have so far not been accompanied by an equally extended understanding of the corresponding processes.

The relationship between a single gene and the resulting properties, phenotype of an individual is rarely clear. This thesis addresses several computational challenges regarding identifying and assessing the effects of quantitative trait loci (QTL), genomic positions where variation is affecting a trait. The genetic information available for each individual is rarely complete, meaning that the unknown variable of the genotype in the loci modelled also needs to be addressed. This thesis contains the presentation of new tools for employing the information that is available in a way that maximizes the information used, by using hidden Markov models (HMMs), resulting in a change in algorithm runtime complexity from exponential to log-linear, in terms of the number of markers. It also proposes the introduction of inferred haplotypes to further increase the power to assess these unknown variables for pedigrees of related genetically diverse individuals. Modelling consequences of partial genetic information are also treated.

Furthermore, genes are not directly affecting traits, but are rather expressed in the environment of and in concordance with other genes. Therefore, significant interactions can be expected within genes, where some combination of genetic variation gives a pronounced, or even opposite, effect, compared to when occurring separately. This thesis addresses how to perform efficient scans for multiple interacting loci, as well as how to derive highly accurate empirical significance tests in these settings. This is done by analyzing the mathematical properties of the objective function describing the quality of model fits, and reformulating it through a simple transformation. Combined with the presented prototype of a problem-solving environment, these developments can make multi-dimensional searches for QTL routine, allowing the pursuit of new biological insight.

# List of Papers

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

I C. Nettelblad, S. Holmgren, L. Crooks, Ö. Carlborg (2009) `cnF2freq`: Efficient Determination of Genotype and Haplotype Probabilities in Outbred Populations using Markov Models. Proceedings to BICoB 2009, New Orleans, LNBI 5462, pp. 307-319, Springer Verlag Berlin

II C. Nettelblad, Ö. Carlborg, J. Álvarez-Castro (2010) Assessing Orthogonality and Statistical Properties of Linear Regression Methods for Interval Mapping with Partial Information. Technical Report 2010-005, Department of Information Technology, Uppsala University, 2010.

III K. Harling, C. Nettelblad, S. Holmgren (2010) Efficient Evaluation of the Residual Sum of Squares for Least-Squares Problems in Genetic Mapping of Complex Traits. 2010. Submitted to Computational Statistics.

IV C. Nettelblad, S. Holmgren (2010) Stochastically Guaranteed Global Optima in Multi-Dimensional QTL Searches. Technical Report 2010-006, Department of Information Technology, Uppsala University, 2010.

V M. Jayawardena, C. Nettelblad, S. Toor, P-O. Östberg, E. Elmroth, S. Holmgren (2010) A Grid-Enabled Problem Solving Environment for QTL Analysis in R. Accepted for publication in proceedings to BICoB 2010, Honolulu



Paper IV was also presented in an earlier form at the SIAM conference in Computational Science and Engineering 2009 (SIAM CSE-09) in Miami, Florida. The author to this thesis is the main contributor to papers I, II, IV. In paper III, the author of this thesis formalized some of the mathematical and statistical presentation. An earlier presentation of part of the material was done in the department technical report 2005-033. For paper V, the

authoer of this thesis was the main contribution together with Jayawardena, and implemented the code modifications to make an existing local QTL search code integrated in the R package developed by Jayawardena through flexible serialization. He also designed some of the experiments used to demonstrate the PSE prototype.

*to the future...*

# Contents

**1 Introduction** **9**

**2 Genetic and Biological Background** **13**

2.1 Chromosomes, DNA, Genes . . . . . 13

2.1.1 Haploid and Diploid . . . . . 14

2.1.2 Genetic Distances . . . . . 14

2.1.3 Markers . . . . . 15

2.2 Experimental Populations . . . . . 16

**3 Models for QTL Analysis** **19**

3.1 Normal Mixtures . . . . . 20

3.2 Partial Information . . . . . 20

3.2.1 Imputation Methods . . . . . 21

3.2.2 Linear Models for Partial Information . . . . . 22

3.3 Orthogonal Models . . . . . 23

3.4 Current Models . . . . . 24

3.5 Model Selection . . . . . 25

3.5.1 Model Choice between Parameter Sets . . . . . 25

3.5.2 Orthogonality for Model Selection . . . . . 26

3.5.3 Model selection and significance testing . . . . . 26

3.5.4 Permutation Tests . . . . . 27

**4 Optimization over the Genome Landscape** **29**

4.1 Exhaustive Search . . . . . 29

4.2 Forward Selection and Backward Elimination . . . . . 29

4.3 Genetic Algorithms . . . . . 30

4.4 DIRECT . . . . . 31

4.4.1 General Presentation of DIRECT . . . . . 31

4.4.2 DIRECT for QTL Analysis . . . . . 33

**5 Summary of Attached Papers** **35**

**6 Future Work** **39**

**Acknowledgements** **40**

# Chapter 1

# Introduction

In the popular mind, DNA has become the symbol of the secret of biology. Advances in molecular methods have resulted in that processes and phenomena that previously could only be studied externally, as a "black box", can now be controlled and unravelled in great detail. Efficient sequencing means that the full DNA structure, the genome, for sample individuals within species can almost routinely be determined [50, 23, 34, 49]. An expected development is that sequencing of large numbers of individuals within a species will also become routine. These sequencing results are then not only scientific pinnacles in themselves, but also form the basis for further work. Microarray techniques allow the study and assessment of the presence of almost any possible substance, in any tissue, at any point in time, allowing the study of correlations between genetic data and molecular properties.

Such advances impose a paradigm shift on experimental workflows. While previously, the prudent method was to first decide a candidate substance, a candidate reaction, or a candidate gene, and then studying its biological role and variations, it is now possible to scan or probe complete systems. The scientific challenge is no longer a matter of designing and performing experiments for extracting enough data of enough quality to devise a model. It is rather a matter of how to design a model that can find the networks, reactions and components that are *relevant*, within a cloud of noise.

The field of analysis of quantitative trait loci (QTL) is part of this paradigm shift. A *locus* is a genetic location within the genome. In each locus, different individuals can carry different *alleles* (gene versions). Different alleles can, but will not always, give rise to different observable properties, or traits. Sometimes, the expression (result) of an allele will be highly dependent on the environment. These environmental factors can be external, like chemical factors present or absent during the early development of an organism, or the quality of nutrients available. They can also be internal, in the sense of

what other alleles are present in other loci. Therefore, to study a quantitative trait, i.e. a phenotype variable of some kind that can be measured for each individual, one wants to find a complete set or network of loci where the allelic variations can explain the variation seen in trait values between individuals.

The loci found in such a scan will be dependent on the experimental setting. Loci that potentially have alleles greatly affecting a trait will not be found – if those alleles were absent or infrequent in the population studied. Likewise, external environmental factors might result in a genetic difference never rendering any observable results. For example, an inability or enhanced ability to handle a specific nutrient (such as lactose for humans and bacteria), will only be apparent in an environment where that nutrient is available (i.e. in dairy products).

From a mathematical and statistical standpoint, the problem of QTL analysis is one of model fitting and model selection. The total number of loci included should not exceed those for which statistical significance is plausible. In other words, a repeated experiment, with nominally identical conditions, should result in the same loci being reported. In practice, linear regression models are frequently used. As complete genome sequencing is still not a cost-effective operation to be performed on hundreds of individuals in an experimental population, one resorts to analyzing genomic *markers*. The markers are located at known positions. To analyze loci not coinciding with the markers, the probability for a specific allele being present in the locus needs to be assessed based on the marker value. From such probabilities, the correlation between *genotype* and *phenotype* (trait) can be determined. For models of equal size, that correlation should be maximized, the result being considered the true QTL. When choosing between models of different size, adjustment is needed to correct for the fact that a model with increased degrees of freedom will always be able to provide a better fit, compared to a limited model. One way to empirically determine significance is through random modifications or permutations of the experimental data, in one sense a form of cross-validation.

In this thesis, advances are presented in the areas of preprocessing marker data to compute genotype probabilities, the structure of the linear regression model used, as well as new developments for efficiently performing permutation testing for determining significance. The latter advances are related to the practical utilization of high-performance computing (HPC) infrastructure, as well as a mathematical reformulation of the optimization problem, which accelerates both actual QTL searches and permutations, but especially the latter. In all, these advances allow for routinely including a larger number of loci in models, allowing a better understanding of the genetical architecture underlying complex traits in experimental populations.

The remainder of this thesis is organized as follows. In Section 2 the structure of input data is reviewed. This defines the scope of populations for which our methods are applicable, as well as what pre-processing is required to efficiently handle that data. In Section 3 an overview is given of linear regression models, and some comments on how those relate to other existing models for QTL analysis. Section 4 discusses the actual optimization problem of finding one or more QTL in an optimization search, given preprocessed input. The papers forming the basis for the thesis are summarized in Section 5, followed by a short presentation of possible future extensions and developments.

# Chapter 2

# Genetic and Biological Background

## 2.1 Chromosomes, DNA, Genes

The *genome* of an individual is primarily structured in *chromosomes*. Each chromosome contains two interlinked complementary molecules of deoxyribonucleic acid (DNA). These can be used to replicate each other, through the process of base-pairing allowed by the structure first described in [51]. Within a chromosome, the genetic code is a sequence of nucleotides, where each nucleotide is represented by one of the letters A, C, G, T. Pairing is restricted in such a way that A can only pair to T in the complementary strand, while C pairs to G. A single chromosome can be millions of such base pairs in length. A mutation in the genetic code can include the deletion of a stretch of base pairs, an insertion of base pairs originating from another region, a duplication of a short genetic segment, or the modification of a single base pair (insertion-deletion, indel). Therefore, different individuals of a species tend to share the same genomic structure on a macroscopic level, but there is not always a one-to-one correspondence for individual base pairs.

A gene is a specific region encoding some product, affecting the function of the cell. The gene product is frequently a protein, which is produced through transcription into messenger RNA, followed by translation into a protein in a ribosome external to the cell nucleus, where DNA is preserved. However, the definition of a gene is not completely straightforward. In eukaryotic organisms (organisms that do have a cell nucleus), the gene might be "interrupted" by introns, base pair sequences that are not included in the actual mRNA and hence are not translated into a protein sequence.

The structure of the introns, as well as surrounding genetic material, both upstream and downstream relative to the protein-coding sequence, can also influence the practical expression of the gene in different ways. There are also many pseudo-genes in most genomes, i.e. genetic material that is clearly similar to true genes, but where transcription or translation never takes place, possibly due to missing critical regulatory sequences, or due to some kind of suppression.

For these reasons, a quantitative trait locus will frequently coincide with a gene, but it cannot be said beforehand that the genetic cause of trait variation will actually be found within the coding sequence. A gene can also sometimes be moved, *translocated* within the genome, so that it can appear at different loci. Such events will cause problems for the methods described in this thesis. Furthermore, a locus determined using these methods will frequently be wide, so that multiple closely linked genes are indicated as associated with the trait, making other methods necessary to discriminate between them.

### 2.1.1 Haploid and Diploid

Most of the genetic material in an organism from a species that practices sexual reproduction is duplicated, with one copy originating from each parent. This means that every chromosome exists in pairs. The genome in such a cell is called diploid. During the life cycle of an individual, every cell will preserve that separation, with one copy of maternal origin and another of paternal origin, replicated through each cell division (mitosis). Only in meiosis, when new gametocytes (oocytes and spermatocytes) are created, these two copies will intermingle. From the pairs of chromosomes, a new, recombined, genome is created for each gametocyte, half of which is actually included. This genome, half the size of the genome in a conventional somatic cell, is called haploid, with one copy of each homologous chromosome.

This mixing of genetic material for each chromosome is called recombination. An individual instance where the source is switched is known as a cross-over. The frequency of cross-overs is not constant over the genome. It has been shown that some sequences are related to triggering them. There are also macroscopic variations over the genome, recombination being more frequent in some regions compared to others.

### 2.1.2 Genetic Distances

From the perspective of this thesis, and QTL analysis in general, the chromosome can be seen as a continuous object. Distances are not measured

in discrete base pairs, but rather in the unit centimorgan, cM. A one centimorgan interval corresponds to that the probability of crossing over taking place between two endpoints is 1%. The probability for some crossing to occur over a distance of 2 cM is then naturally 2%. However, there is a definite probability of double recombination, so that the detected origin will be identical to the source. Hence, the probability for a difference in sequence origin for a very large distance will approach 50%.

In the literature, several mapping functions have been suggested for the translation between recombination frequencies and genetic distances. The Morgan map function, is completely linear, assuming no double recombinations. The Kosambi map function [30], on the other hand, includes modelling of the biologically known fact that a single cross-over tends to shadow the neighboring regions, making another recombination event nearby more unlikely than otherwise expected (interference). Kosambi distances are not additive, though.

The Haldane mapping function is more complex than the Morgan function, in that there is no linear correlation between the net parity of recombination events and distance, but on the other hand simpler than Kosambi, ensuring that the distance between $A$ and $C$, traversing $B$, is equal to the distances $AB + BC$, a much needed convenience for modelling purposes. The Haldane function, defined for non-negative distances in cM, is given by

$$p = 0.5 + 0.5e^{-0.02x}, \tag{2.1}$$

where $p$ is the probability of identical genetic phase (chromosome origin) at 0 and $x$. This relationship can be derived from realizing that the recombination process itself, assuming no interference, is a Poisson process, and then summing over the probability of an even state (i.e. a total of $0, 2, 4, 6...2k$ recombination events between the two points considered) [19].

### 2.1.3 Markers

It is generally not feasible to study each haploid genome separately. Rather, the resulting genome in diploid cells in the offspring is studied. Full sequences are generally also not detected, instead specific markers are used. Markers are known location of variability, which can be based on genetic repeats, including so-called microsatellites. A repeat is a rather short pattern that is repeated in multiple instances adjacent to each other in the genome sequence. Due to molecular details in the process of genome replication, the exact number of such repeats can change due to "slippage". Therefore, the number of microsatellites can vary even between closely related individuals. For this reason, this class of marker has been used in forensic studies. It was also the first class that could be routinely detected in the 1980s.

Another class of markers is the so-called single-nucleotide polymorphisms (SNPs). These are point-mutations, where a single base-pair (letter) in the genetic code has been altered. An SNP, when it has occurred, tends to be preserved. As there are billions of base-pairs in an ordinary genome, the general probability of another mutation in the same position is low. Therefore, all individuals with a specific allele are assumed to share a common ancestor or have a direct ancestor-descendent relationship. The standard situation is therefore that only one wild-type allele (the most common) and one mutant version exist, despite the fact that there are theoretically four possible bases in each position.

Some SNP testing methods also only test for the known variants. The tests are frequently done by hybridization against oligo-nucleotides in a microarray. When the genetic material binds to a short sequence of either allele, the binding will have greater affinity if the match is perfect. A new mutant might therefore not be detected, as it will match neither probe. With a proper preparation, binding can be detected by luminescence. Hybridization arrays for $10,000$ or more SNPs can be manufactured industrially, allowing a very high number of markers to be analyzed simultaneously.

Due to the nature of marker determination, the ordering of markers in the actual genome may now be known. Specific tools exist to devise plausible marker maps, including mapping distances between markers, from unordered marker lists [33, 48]. The algorithms used are based on different simplifying assumptions, as it is computationally prohibitive to test for the exponential number of possible marker orderings. From the perspective of this thesis, we consider the determined orderings of markers, as well as their interspersing distances, to be accurate. However, it should be noted that existing methods for setting up marker maps are conceptually similar to our new algorithms for determining genotype probabilities presented in this thesis. The problem of determining proper marker maps is also being influenced by the continuing change in total marker counts, and the fact that markers now frequently have a known genome location in the genomes of fully sequenced species, which makes the ordering known.

## 2.2 Experimental Populations

Theoretically, QTL analysis can be performed in natural as well as experimental populations. In this context, a natural population is a population that would exist, with similar characteristics, even if no QTL experiment would have taken place. A wild population of some mammal species would be one example of this. In a natural population, the total genetic variability might be great and not completely characterized. Furthermore,

environmental conditions are not well-defined, and may even be unknown.

The alternative to a natural population is an experimental population. In many cases, these are based on different lines within a species. A line can be based on inbreeding, where all individuals within the line are identical. A line can also be defined as a specific breed or variant, or based on geographic origin, and similar factors. In these outbred conditions, individuals are still expected to be more similar within each line than between lines. Inbred lines can be preferable, since variations in genetic background between the different founder individuals will not affect the analysis, which is the case in outbred conditions. On the other hand, an inbred line might have genetic deficiencies in addition to the intended trait under study, making the experiment results highly significant, but at the same time resulting in a limited opportunity for generalization.

When the lines demonstrate differences in some trait, such as body weight or length of flowering time, crosses between the lines can help elucidate the genetic structure. By convention, the founder individuals of such a cross are called the $F_0$ generation. Individuals from the two lines can be mated, in such a way that each resulting individual has one parent from line 1, and one from line 2. Ideally, parents of both sexes should be represented for both line origins, to avoid confusing genetic effects with e.g. epigenetic effects due to imprinting, or womb effects. The resulting individuals form the so called $F_1$ generation. All individuals in the $F_1$ generation show the same genetic background, with one allele from each line in every gene. The phenotype distribution for the $F_1$ individuals is therefore expected to be identical for all individuals, any differences arising from environmental factors.

From an $F_1$ generation, two general crossing schemes are common. One is the backcross, where $F_1$ individuals are crossed with individuals from either founder line. In such a structure, the backcross offspring will have 1 or 2 alleles from the founder line chosen in the backcross, and 0 or 1 alleles from the other line, in every locus. No effects will be seen in e.g. loci where the founder line used is dominant, as dominance can obscure the presence of another allele. The average genetic contribution of the line used in the backcross is then 75%.

The other common cross structure is the intercross. In an intercross, $F_1$ individuals are crossed once again, to form an $F_2$ generation. The average genetic contribution is the same as in the $F_1$, 50% from each original line in each resulting individual. However, at the locus level, greater variability can be observed, with 0, 1 or 2 alleles from both lines being possible, with the relative proportions $1 : 2 : 1$. An intercross can detect dominant alleles, but the additional degrees of freedom might reduce the exact assessment of effects.

More elaborate crossings are directly conceivable from these basic structures, like repeated crossings of individuals from successive generations, going on from the $F_2$. A multi-generational design of this type will introduce more recombination events between the markers, allowing a finer mapping of QTL, assuming the marker map is dense and informative enough to properly trace inheritance.

# Chapter 3

# Models for QTL Analysis

In Section 2 of this thesis, we have briefly covered the critical aspects of the genetics involved. When probabilities of state or line origin are defined in any putative set of candidate loci, a proper model is needed to analyze the quality of a fit. Several approaches are available in the literature. Many are related to interval mapping (IM), which is in fact a specific approach for handling non-definite genotypes. Earlier approaches only analyzed single markers, discarding those samples where genotype information was not conclusive. Such an available case analysis is well-known in the statistical literature, but known to be flawed, with limited power as well as distorted results as possible consequences [35].

What constitutes a model is not always evident, especially as different methods and frameworks in some cases give identical or directly equivalent results, while the outcome can vary in more complex cases. One modeling aspect is what assumptions are made on the phenotype distribution for single individuals, given their genetic information. This results in a statistical framework for the analysis. Another is what phenotype values are expected for different genotypes, defining the parametrization of the model. The framework will primarily affect the detection of a proper QTL location, while the parametrization is critical for making the estimated genetic effects possible to interpret and analyze from a biological standpoint. These two aspects tend to be intertwined, something that becomes even more clear when model selection is considered. Beforehand, it is in general not known what number of loci and genetic mechanisms to expect, so different parametrizations need to be tested in a common framework.

## 3.1 Normal Mixtures

The main hypothesis of all QTL models described here is that the total variance in a trait can be decomposed into multiple independent components, $V_{tot} = V_g + V_e$. Here, the component $V_g$ is the genetic variance, while $V_e$ is an environmental component. These components are assumed to be separate, or statistically independent, i.e. with zero covariance. General analysis of similarity between relatives can indicate the portion of genetic variance, the broad-sense heritability $h^2$, without identifying the specific loci causing that variance [37]. If a proper separation of the genetic variance is constructed, the with-in class variance for each genotype identified would be $V_e$.

If a genetic property is in fact interacting with the environment, or with another locus not covered by the model, the within-class variance will increase. The actual explainable variance from the model will probably be smaller than $V_g$, as multiple loci are exerting a limited influence on the trait and not included in the model. An effect counteracting this is the problem of over-fitting, where, by chance, the recombinations between an analyzed position and the true position can redistribute individuals between classes in such a way that the with-in class variances are reduced. Even at the correct locus, a model with a high number of parameters can give a better-than-true fit, as the parameter estimates are fitted against the specific individuals sampled and can describe the random variations within that sample as part of the model.

Assume that the genotype class for each individual $i$ is described by a vector $z_i = (00\cdots010\cdots0)$, where the single non-zero element is 1 represents the true genotype. The distribution for observed trait values for $i$ can then be expressed as:

$$\mathcal{L}_i = \sum_j z_{ij} N\left(\mu_j, V_e\sigma^2)\right), \tag{3.1}$$

where $\sigma^2$ is ideally expected to approach $V_e$. This interpretation is common to all the models presented here, when $z$ indeed is a binary vector. The maximum likelihood (ML) can in this case be performed using linear regression, or using the PERF algorithm presented in paper III. The variables $\mu_j$ are fitted directly, and $\sigma$ computed from those variables.

## 3.2 Partial Information

In the case of partial information, several interpretations are possible. The original presentation of IM reuses (3.1), but defines $z$ to be a general

probability vector, i.e. only imposing the conditions $\sum_j z_{ij} = 1, 0 \leq z_{ij} \leq 1$. The result is a normal mixture, a linear combination of multiple normal distributions describing the expected phenotype for the individual $i$ [31]. It should be noted that this distribution is not equivalent to the distribution of the sum of two normally distributed variables. Rather, the shape of the single distribution is the sum of multiple identical normal distributions, weighted by probability, and with identical standard deviations, determined by $\sigma$.

Optimizing this likelihood as a function of the vector $\theta = \{\sigma, \mu_j\}$ is a non-linear problem which is usually solved by the expectation-maximization (EM) algorithm. This algorithm, first presented in its general-purpose form with that name in [14], is relatively robust, but sometimes converges slowly. The algorithm consists of two main steps, explaining the name:

1. Expectation: Compute the expected value of the full (log-)likelihood function for the population, given a current parameter vector $\theta$.
2. Maximization: Compute the optimal $\theta$, given the log-likelihood function.

When the optimization is complete, marginalization can be done to determine a posterior probability vector where the prior genotype probabilities $p_{ij}$ are replaced by posterior probabilities $\pi_{ij}$.

### 3.2.1 Imputation Methods

In the IM methodology, a single mixture realization of the population is considered. The prior probabilities for the individual genotypes are used over all iterations. There are competing schemes, called imputation methods, that use multiple realizations of the population. One form of this is the multi-QTL model (MQM) method [25], which encompasses e.g. the inclusion of covariates for markers at non-modelled positions to handle genetic background, and as the name indicates, the modelling of multiple loci within a single model.

However, in the context of treatment of partial information, MQM differs from IM by its use of repeated weighted linear regression. The individuals are not realized as a single phenotype distribution as a normal mixture, but rather multiple separate distributions, where the total influence of each onto the composite likelihood for the population is defined as $L_j^{\pi_{ij}}$. The distributions resulting from the different states for a single individual are multiplied by each other, rather than merged into a single distribution by addition. In the likelihood space, a choice between an arithmetic or a geometric mean must be made.

Furthermore, MQM is adjusting the probabilities $\pi_{ij}$ at each step, using these within the model. The genotype probability is then based on a combination of prior information and the resulting phenotype probabilities. This two-sided scheme is in fact a variation of the general location model, where an excellent description of the general case and different specializations found in [35]. It should be noted that the general location model is not equivalent to generalized linear models, although specific realizations of one can fall within the other.

The linear regression case for full information, the partial information model in IM, and the partial information model in MQM share a central property of posing a single realization of the population. MQM will introduce multiple rows for each individual, but they are all part of a common parameter fitting step, which is iterated until the effects, as well as the $\pi_i$ have converged, together.

There are several methods employing Bayesian and Monte Carlo methodology for QTL analysis. Someare "true" Markov Chain Monte Carlo approaches, like the ones described in [47], but there are also other methods employing the methodology towards the analysis of a single set of candidate loci, i.e. a comparable setting to the cases for the models already discussed. The input for these algorithms consists of genotype probabilities and phenotype values, the output a likelihood of a fit, and a set of model parameters focusing on the genetic effects and the residual variance. In [45] one such approach is described, where multiple full realizations of the genotype at evenly spaced "pseudomarkers" are constructed.

In the context of a single position, creating a set of realizations through pseudomarkers is equivalent to sampling from the individual genotype probabilities already seen in the other methods. Each realization will be a case of full information, which can be handled through the first method described above. These realizations are then weighted together, based on the residual variance determined. The residual variance has an immediate algebraic relationship to the likelihood for the model. In effect, the method is approximating sampling from the posterior distribution by taking samples from the (much simpler) prior distribution, and then applying this weighting. The authors of [45] note that there is a close similarity to general IM, the critical difference being that IM finds a single optimum parameter vector, whereas imputation methods marginalize over all vectors.

### 3.2.2 Linear Models for Partial Information

IM using the EM algorithm is computationally expensive. The evaluation of each iteration requires evaluation of the Gaussian probability distribution

function in different points for each individual. In addition, the EM method has only linear convergence, normally resulting in a large number of iterations. For these reasons, the authors of [38, 20] independently suggested the use of linear regression, also for partial information, where the $z_i$ vectors are not binary. The result is a probability distribution with $\sigma^2$ variance, centered not on either genotype mean, but on a point between them, as if the continuum of probabilities directly corresponded to a continuity of genetic effects on phenotype.

The simplicity, both in computation time and regarding simple use of existing software tools, has made this form of linear regression highly popular [24, 9, 43]. However, multiple limitations have been pointed out, including a confusion of residual variance with unexplainable variance due to lack of information [52], and other aberrations found in systematic simulation studies [28]. Suggested remedies have included iteratively re-weighted regression [53], and estimating equations (EE) [15]. The latter of these can also be implemented using Fisher scoring [22].

## 3.3 Orthogonal Models

From a biological perspective, handling arbitrary mappings of full genotype realizations for all loci modelled into mean phenotype values is cumbersome. The total number of parameters also becomes very large for multiple loci, as the total set of phenotype effects would have size $n^d$, where $n$ is the number of genotypes per locus (generally 3 for $F_2$ populations), and $d$ is the number of loci. In addition, a number of covariates can be included in the model, e.g. sex and other factors not captured directly by the markers, while these factors are still expected to have a significant influence on the trait under study. By correcting for such effects, the actual correlation between the genotypes and the trait can be made more clear.

In basic Mendelian genetics, the traits are binary. The inheritance pattern for the trait can introduce an assymmetry between alleles, with one dominant allele, meaning that the presence of that specific allele results in the trait being "active". Other alleles are recessive, meaning that the trait connected to the allele will only be seen if the dominant allele is not present, i.e. that both the copies in a diploid genome are copies of the recessive allele. The dominant and recessive effects can frequently be understood as a matter of working versus silent alleles. If one copy of the allele is damaged in some way, it might never be used, with the working copy being used instead. The damaged, or silent, allele is then recessive, as no phenotype change is seen unless both copies are silent. Against a background of silent alleles, a single working allele will be considered dominant, as its presence will modify the

phenotype. Two copies of the dominant allele will not affect the trait further than a single copy, as the working piece of genetic code is already present.

Based on Mendelian thinking, a continuous trait with two alleles can be dissected into an additive and a dominant portion. The additive portion is linear to the allele count, while the dominant portion is contrasting heterozygotes against homozygotes. For a $F_2$ population, a design matrix $\mathbf{S}$ can be defined as

$$\mathbf{S}_{F_2} = \begin{pmatrix} 1 & -1 & -0.5 \\ 1 & 0 & 0.5 \\ 1 & 1 & -0.5 \end{pmatrix}, \tag{3.2}$$

where the columns represent base vectors for the variables $\mu, \alpha, \delta$. This design matrix can be multiplied with a $\mathbf{Z}$ matrix with full-information individuals to build the model for linear regression. Similar approaches can also be devised for the partial information methods, using the parameters to compute the per-genotype class means. The original $F_2$ model was due to Fisher, in [17]. Further extensions were made in the 1950s for more properly handling epistatic interactions in loci showing $F_2$ frequency proportions.

## 3.4 Current Models

The renewed interest in quantitative genetics due to the molecular developments presented in Section 2 has also spurred new efforts regarding on the basic modelling and parametrization issues. These efforts have been focused on achieving *orthogonality* in more cases. In mathematical terms, the property of an orthogonal model is equivalent to the full model design matrix $\mathbf{X} = \mathbf{ZS}$ being orthogonal, for which $\mathbf{X^T X}$ being diagonal is a sufficient and necessary condition.

Orthogonality, in this context, is generally analyzed in the case of individuals of full information, independently of the regression model and parametrization. The intent is to ensure that estimates of the different variables (model parameters) are independent. This independence ensures that the remaining parameter estimates are unaffected if a parameter is removed. The explained variance from the model can also be clearly and uniquely attributed to the parameters. The original $F_2$ model and other models were conceived to be orthogonal for populations exactly matching that structure. No population of finite size will be a perfect $F_2$, as the allele and genotype frequencies will not match the expectation values exactly. Therefore, no matter the problem structure, a slight deviation from orthogonality will be present for actual data. Subsequent developments implement support for other frequency distributions, first handling deviations where Hardy-Weinberg equilibrium (genotype frequencies being products of allele frequencies)

still holds [56] and finally handling any genotype frequencies for a single locus [3]. The parameter estimates for one specific population can also be translated into corresponding functional estimates (estimates describing gene function). Functional estimates derived from different populations can then be accurately compared.

## 3.5 Model Selection

Selecting the appropriate model is a matter of finding the proper set of loci, with the proper parameter set best describing the external factors. There is an inherent trade-off between specifying multiple loci and several parameters, and the risk for fitting the specific sample, rather than capturing the true biological properties of the underlying population.

One common approach is to separate the selection of model size (or rather a specific parameter set), and the selection of the optimum locus set for this model configuration. This has also been the view used in the papers in this thesis. The other option would be a more general Markov chain Monte Carlo walk, with transitions including variations in model configuration as well as loci [47].

### 3.5.1 Model Choice between Parameter Sets

The traditional likelihood for a model is, in essence, the likelihood for the observed data, given the model, i.e. $P(y|M)$. If we want to choose the single most likely model, we should maximize $P(M|y)$. However, assuming the prior probability for any model $M$ to be a constant $C = P(M)$, we can see that, according to Bayes' law, $P(M|y) \propto P(y|M)$.

The assumption of a constant $P(M)$ will break down if models of varying sizes are analyzed. A larger number of parameters increases the model space, although not all models are relevant. Several criteria have been suggested for introducing a correction to the likelihood to account for model size differences. Two general approaches (not restricted to QTL analysis) are the Akaike information criterion [2], and the Bayesian information criterion [44]. These have been studied for QTL applications [6], including simple modifications of the BIC. In recent work, a more context-sensitive approach is proposed, handling main effect and interaction parameters with separate weights [5, 54]. A review of these methods can be found in [40]. For reference, the expressions for AIC, BIC are given below:

$$AIC = n \ln\left(\frac{RSS}{n}\right) + 2K \tag{3.3}$$

$$BIC = n \ln\left(\frac{RSS}{n}\right) + K \ln n \tag{3.4}$$

$$mBIC = n \ln\left(\frac{RSS}{n}\right) + (a+b)\ln n + 2a\ln(l-1) + 2bln(u-1) \tag{3.5}$$

Here, $n$ is the number of individuals, $K$ is the number of parameters, and $RSS$ is the residual sum of squares (which relates to the model likelihood in the linear regression case). In mBIC, $a$ is the number of main effects, $b$ is the number of interaction effects. The values of $l$ and $u$ are based on the number of markers and marker pairs, respectively, somewhat arbitrarily chosen by default as $l = N_m/2.2, u = N_p/2.2$.

### 3.5.2 Orthogonality for Model Selection

If the modelling approach is not orthogonal, all models in the model selection set must be evaluated to select the most suitable model. The parameter estimates and variance components per parameter cannot directly be used to estimate the total explainable variance in a model where that parameter would be excluded. For $K$ parameters, $2^K$ evaluations need to be made. This is prohibitive even for limited values of $K$.

If the total modelling approach is orthogonal, only the full model has to be fitted. The variance components for specific parameters will be unchanged when the total parameter set is modified. Therefore, an optimal combination of variance components can be selected with no new model fitting. This optimum can generally be found using a simple dynamic programming algorithm, depending on what constraints are imposed on the parameter set, e.g. requiring main effects to be present before allowing corresponding interaction effects, as frequently advocated [45].

### 3.5.3 Model selection and significance testing

The concept of model selection is closely related to significance testing. For model selection, as treated here, the interest is to determine which model is the most likely to be true for the underlying population. Significance testing is assessing if the effects found are significant from the null hypothesis, i.e. if the model including the specific set of loci chosen has more than, for example, 95% support. In a way, this is a very specific case of model selection, with the only models specifying 0 and $n$ loci, respectively.

### 3.5.4 Permutation Tests

Deriving theoretical expressions for the distribution of the likelihood function is a complex task. Some attempts have been made [31, 42], since the QTL analysis setting exhibits non-symmetric population structures, non-normal phenotype distributions and non-homogeneous marker maps and marker information patterns, the efforts based on conventional statistical theory by necessity become crude approximations.

Instead, the standard approach is to perform permutation testing [12]. From the original dataset, permuted versions are created. Here, the individuals and their genetic makeup are kept unchanged, while the phenotypes are permuted. Several of the sources for specific deviations in the objective function will be kept by this method. However, if covariates of different kinds are included in the model, the permutation needs to take these into account, possibly by doing permutes within separate classes defined by the covariates [13]. The intent is to do permutes between individuals considered to be identical, except for the actual genotypes modelled, but this is exceedingly hard when more background factors ate taken into account.

Assuming suitable permutes can be created, the objective function distribution for the null hypothesis can be found empirically. If 99 % of the values are inferior to the determined objective function value for the real model, the probability of a Type-I error is 1 %, resulting in a practical implementation of a significance test.

# Chapter 4

# Optimization over the Genome Landscape

In order to find a proper set of QTL, using the data described in Chapter 2, and the models described in Chapter 3, the genome needs to be scanned in some manner, to find the actual optimum. This optimum is then considered to be the true QTL, although the literature is frequently reporting multiple "peaks" of independent models, assuming independence between them [29, 9]. However, for such cases, a multi-dimensional model with only main effects would be preferable, to avoid confounding effects from other QTL to be included in the respective main effect estimates.

## 4.1 Exhaustive Search

The most straightforward, almost naïve, approach to explore the model fit landscape over the genome (in whatever dimensionality $d$), is to loop over all genome positions in a closely spaced grid, say every $1cM$, evaluating the model at each of these positions. The position resulting in the highest likelihood will be stored and reported in the end.

## 4.2 Forward Selection and Backward Elimination

Exhaustive search is more or less computationally intractable for any dimensionality $d > 2$. A common approach is then forward selection [7, 5, 54, 55]. The basis for forward selection is to find a single QTL with the highest possible likelihood, for inclusion of that QTL in the model. When that locus has been selected, a new scan is performed, keeping the first

locus fixed and adding a second locus. This process can continue to an arbitrary dimensionality $d$. Unless full independence of parameters per locus can be guaranteed, this approach will not necessarily find the optimal set of loci, unless full independence of parameters per locus would be guaranteed. If interaction effects are included, the forward selection approach is also inappropriate, since the selection of the first locus will not take into account what level of explanative power might arise when interactions are taken into account. One possible approach is then to select pairs of loci as the basic forward selection unit. However, this solution will not handle networks of gene interaction, where the same locus is part of multiple pairs. The best first locus to include would be a "hub" in such a network, but the hub will not be the first selected locus in a pair-based forward selection method.

The suboptimality of forward selection is also acknowledged by authors suggesting use of the method [7]. This can be compensated by a later step of backward elimination. A model of a size greater than the optimal final model is developed through forward selection. A backward iteration is then initiated, systematically testing the removal of the included loci. In each iteration, a new model is created, removing the locus where the removal proved to have the lowest reduction of the likelihood. The resulting models of size $[1, d]$ will most likely have a higher likelihood than the original models of equivalent size during the forward selection phase, since the removal step has a form of hindsight with a larger total set of relevant loci, where the non-independence and interaction contributions that make forward selection sub-optimal can better be taken into account.

The results are, despite the use of backward elimination, not guaranteed to be optimal. The total set of models to test per iteration is much smaller for backward elimination, since the space only consists of those QTL already identified in the forward selection phase, which can be contrasted to the dense grid of all locations in the genome, or even all pairs of such locations, that form the basis for the exhaustive search in forward selection. Therefore, using forward selection without backward elimination can rarely be warranted. It is also possible to extend the single "cycle" of forward selection and backward elimination by adding additional and more flexible paths of extension and elimination from a model. Optimality of the solution given still relies on the problem structure being overly simple, or the signals from individual loci strong.

## 4.3 Genetic Algorithms

Genetic algorithms (GA) is a class of methods for general global optimization, frequently used when details about the analytical properties

of the problem at hand are scarce. Inspired by natural evolution, a population of "individuals" are evolved. Each individual carries some form of "genetic code" determining the properties of the individual. Each individual represents an object that can fulfil a defined purpose or objective, i.e. a solution to the optimization problem. Between generations, reproduction is taking place, mixing the genes between the individuals, followed by selection, removing individuals with inferior performance relative to the objective. Many different schemes exist with different specific methods for reproduction, representation of genes, mixing of genetic material, but the main ideas are shared by all concepts. A summary of GA approaches in general can be found in [4].

For QTL problems, genetic algorithms were pioneered in [10], where each locus was treated as a separate GA gene. As the effects from the loci are frequently almost independent, separate searches would tend to find good candidate loci, or small networks. By mixing the GA genes, these individual genes or networks are merged into a total optimum.

## 4.4 DIRECT

In global optimization theory, different mathematical properties of the objective function in the model space can be exploited. One example of such a property is that of Lipschitz continuity, which can be defined as

$$|f(x+v) - f(x)| \leq rK, |v| \leq r, \tag{4.1}$$

where $v$ is an arbitrary vector of maximum length $r$, $r, K \in \mathbf{R}, > 0$, and $x$ is some vector within the model space. In other words, within any environment, there is a limit on the total variation relative to its center point. When we seek for an optimum of $f(x)$, this property can be used to reduce the space explored. If there is a known $f_{min}$ candidate, and $f$ has been evaluated in another point $x'$, where $f_{min} < f(x')$, no point $x''$ where $|x' - x''| \leq \frac{f_{min} - f(x')}{K}$ can result in a value $f(x'') < f_{min}$. Exploiting (4.1) for Lipschitz optimization has been frequently described in the literature [46, 41, 18, 39], and even more frequently as an implicit assumption in different heuristics within other methods.

### 4.4.1 General Presentation of DIRECT

A straight-forward application of Lipschitz optimization requires $K$ to be known. However, if $K$ is unknown, the algorithm DIRECT, for DIviding RECTangles, can be used [27]. In DIRECT, concurrent hypotheses are

examined for $K$. This is done by dividing the search space in so-called boxes. For each box, the objective function is evaluated in the center. The first box will cover the entire search space, and this box is split into three boxes along some dimension. The centroid of the center box will coincide with the centroid of the original box, while the objective function is evaluated in the two new boxes adjacent to the center box. From this point on, a convex hull is taken in a space consisting of box radius on one axis, and function value on the other. All boxes represented by vertices included in this hull are split in one iteration of the DIRECT algorithm.

A box will be part of the convex hull if there is no box with higher radius that presents a smaller value for $f$ in the centroid. Among the active boxes (the original box in a splitting is always removed, replaced by the three children), one, out of potentially many, with maximum radius will always be split. This is due to the fact that the optimal minimum of the function might actually be within that box, no matter what the value of $f$ is in the centroid, assuming some arbitrarily high value of $K$. Likewise, for all boxes in the hull, there will be some value of $K$ which would allow a value of $f$ superior to the currently known $f_{min}$ to exist within the box radius from the centroid. If a larger box $b_1$ has a smaller value of $f$ than some smaller box $b_2$, the minimum possible $f$ value within $b_1$ would always be lower than that of $b_2$, no matter the value of $K$. Hence, $b_2$ is not a suitable candidate for splitting.

During the iterations of DIRECT, a possible result is that the environment of $b_1$ is shown not to contain any value close to the theoretical Lipschitz bounds possible with free assignment of $K$, and thus $b_2$ might be considered at a later iteration, at the very least when the descendent boxes after splitting of $b_1$ all have a smaller radius than $b_2$. As the largest box of the active set will always be split, there is no natural termination condition for DIRECT. The objective function can be studied in ever-increasing detail, but assuming no upper bound on $K$, there is in theory always a possibility that there will be a new optimum to be found within the vicinity of the centroid of some box.

Standard termination criteria for the iteration process can be based on simply running a fixed number of iterations. Other options include termination after a certain number of iterations with no new global optimum found, or a certain minimum maximum radius (i.e. a maximum size on the largest remaining box), or similar heuristic variations shown to be very effective in practice [16]. If a maximum radius is chosen equivalent to that of the lattice spacing in an exhaustive search, a DIRECT search will require at least the same number of function evaluations as the corresponding exhaustive configuration.

### 4.4.2 DIRECT for QTL Analysis

The QTL objective function, whether it is based on the linear regression residual or the log-likelihood from some other model, is a candidate for optimization using DIRECT. There is no known Lipschitz constant, but on the other hand there is clearly a continuous quality to the shape of the function in the limit of an infinite population with recombination taking place in a continuous manner. DIRECT has also been successfully used for QTL analysis, with the minor modification of introducing so-called chromosome boxes [36].

When DIRECT is used for QTL analysis, the initial state used represents any combination of QTL (the total number determined by a chosen dimensionality $d$) located on the chromosomes included, rather than a single box representing the full search space. This is done since even though an arbitrary concatenation of the chromosomes of the genome could be constructed, no continuity in the objective function is expected over chromosome boundaries. Although DIRECT supports values of $K$ arbitrarily large in theory, the method also becomes arbitrarily inefficient in actually finding the optimum under such conditions. The division into chromosome boxes also presents an obvious choice for loosely connected parallelization of the DIRECT search, executing a search for part of the search space, as defined on a subset of the initial chromosome boxes, on each node in e.g. a computational grid environment [26].

# Chapter 5

# Summary of Attached Papers

The papers included in this thesis cover many aspects needed for QTL analysis for line crosses. The material covered follows the workflow of QTL analysis, starting with the issues around pre-processing of marker and pedigree data. This is followed by a treatment of the methods for evaluating linear models, and an analysis of their consequence, followed by analysing the objective function in the space of different loci within the genome, for allowing more efficient searches. The final paper starts treating the infrastructure issues hampering the implementation of high-dimensional models for actual QTL scientists interested in biology, with limited experience in high-performance computing.

The results in this thesis form the basis for actual experiments unravelling the secrets of biology, as well as realizing possible economical gains for important crops and animal breeds. The thesis is not focused directly on the genetic applications, but it has been done with an awareness of the needs arising, including the need to describe gene networks and epistasis, aspects critical for more complex traits [11]. Another limitation in scope is that the results are directly applicable only to applications with a clear, known, population structure.

## Paper I

In this conference paper, Hidden Markov Models are used to determine genotype probabilities. The use of HMMs for genetics was first introduced over 20 years ago [33, 32]. The approach has also been used in specific cases for QTL analysis [8, 45]. In paper I, we present a general tool replicating the results for outbred populations in [21], but now exploiting HMMs, reducing algorithm complexity from exponential to log-linear. This development is

critical in modern applications, as the total number of markers increases. Furthermore, while it is simple to inspect a larger number of markers, the general data quality will decrease. Values will be missing for some individuals, or there will be ambiguities. These properties were handled in [21], but with an algorithm requiring an exponential increase in time and memory use, making the problem intractable in practice.

In paper I, we also explore using the HMMs for haplotyping, inferring the phasing for linked markers in especially the $F_0$ and $F_1$ generations. This is a critical development for making the genotype probabilities in the $F_2$ generation independent. If the probabilities are not independent, a proper analysis of the population should take the dependence structure into account, something current models will not do. Hence, the probability adjustments due to haplotyping will not only decrease the uncertainty of the $F_2$ probabilities, but also remove a bias in the estimates.

## Paper II

This report deals with the question of whether orthogonal estimates are feasible in a relevant form in spite of partial genotype information. As partial information, with varying information patterns, is currently becoming the norm, this is an issue of increased relevance. Also, recent contributions to the field [3, 56] have attempted to achieve orthogonality in more general configurations, but only in the case of full information. We describe how the residual sum of squares is modified in the presence of partial information, and also suggest an imputation method where the parameter estimates are independent. This method will select QTL candidates with a preference for marker locations with high information content in a way that might be problematic. In the report, we show that this approach, IRIM, in practice performs well when compared to Haley-Knott regression in a chicken intercross.

## Paper III

This paper explores more efficient methods of evaluating linear regression models, compared to traditional methods such as QR factorization. The presented PERF algorithm can be used with any model parametrization, and significant performance improvements are shown in different settings. One limitation of PERF is that full information cases are needed. However, from an algorithm perspective, full information is also present in individual imputations. Therefore, PERF is highly applicable for imputation methods.

## Paper IV

Paper IV improves upon the previous use of the DIRECT Lipschitz optimization method. The results are based on a novel transformation of the residual sum of squares, which is the objective function previously used for DIRECT. The transformed objective function is shown to be Lipschitz continuous, with a well-defined Lipschitz constant in many settings, given a theoretical infinite-size population and perfect recombination frequencies.

Actual real populations are not infinite-size and recombinations never have an ideal distribution. Therefore, the original DIRECT iteration approach is retained, but an additional pruning condition based on the stochastic Lipschitz constant is introduced. Using simulations, we show how this condition gives rise to a termination condition resulting in that the correct optimum is found with very high probability. Using this condition, extremely high performance can be achieved in permutation tests. When starting from a set of candidate models, with a specific objective function value $f_{min}$, DIRECT searches on permuted sets can be performed, with the goal of finding an optimum superior to $f_{min}$. The termination condition will allow most such searches to terminate early, if $f_{min}$ is significant.

## Paper V

The larger datasets and the interest in epistasis and accompanying multidimensional models increases the computational needs for QTL analysis radically. Therefore, there is a clear motivation to utilize modern high-performance computing resources, such as computational grids. In this paper, we extend on previous work by integrating the R statistical environment [1] to a DIRECT-based implementation for QTL searches in multiple dimensions. Efficient random permutation tests will also require substantial computational resources, especially in order to determine relevant thresholds for high significance levels.

# Chapter 6

# Future Work

This thesis covers many aspects of QTL analysis. In the future, this work will need to be disseminated further into the biological community. The prototype described in paper V is a first step in this direction. One avenue for further work would be to use our tools to analyze new and coming datasets in the search for significant three-way interactions. The motivation for choosing 3D searches is that the conventional approach of forward selection would be inadequate in that case, and that the general issue of the importance of high-level interactions is contested.

From a modelling perspective, the issue of orthogonality carries great interest, as an orthogonal definition of main effects would change the possible methodology for model selection within larger parameter sets in a drastic manner.

The PERF algorithm presented in paper III changes the relative performance benefits of Haley-Knott regression, and methods based on multiple imputations with full-information genotype realizations. The work in papers III, IV could be combined, with an implementation of multiple imputation schemes similar to the one presented in [45]. Each imputation gives one candidate value for the likelihood. These could be used for deriving bounds on the possible objective function value in each locus tested, with the higher and lower parts of the bound used in separate parts of the DIRECT algorithm pruning condition, to err on the side of caution. An approach of this type could limit the upper limit constant found in paper IV.

The results in this thesis form the basis for actual experiments ravelling the secrets of biology, as well as realizing possible economical gains for important crops and animal breeds. Results from experimental crosses in other organisms can also further the understanding of homolog structures in man.

## Acknowledgements

Numerous people have, in different ways, made this work possible, some in formal ways, some informally. First, I would like to thank Jessica Flodin. You have introduced a new perspective, a new dimension, in my life. You have also shown a tremendous support for my work. You have listened and given feedback on research-related issues where my mind was too confused to get things straight.

I would also like to thank Robert Rosén. You have been a good friend for almost ten years now, and our discussions regarding the practical conditions of life as a PhD student have been very valuable. It has also been fun to give and receive advice on mathematical, textual and other issues where you and I have been able to appreciate the specific field of the other.

Sverker Holmgren, José M. Álvarez-Castro, Örjan Carlborg: all of you have been critical for my project. Sverker has always believed in my ideas. Papers I and IV would never have come to fruition unless you would have shown interest in new aspects of the field. My insight and understanding of genetic models would have been very different if I would not have been so lucky to have José as my assistant advisor. This has been a basis for all my work. The ever-lasting energy and devotion to QTL analysis shown by Örjan has been a continuous inspiration. If it were not for him, I would never have entered this specific field.

Finally, I would like to thank Karin Nettelblad and Folke A. Nettelblad, my parents. Without you, I would not only most likely not have entered this field, but I would have been far worse off in innumerable ways. Your attitude to life, science, and truth is something I admire. You have also been a practical support in reading different texts and manuscripts. Old habits die hard, and after my long cooperation with both of you as a technical translator, sending off some texts for an external perspective has been very useful. Naturally, I have also used you, like Robert and Jessica, to test off the strangest ideas and as talking partners to set the thought fibers in my mind straight.

This work was supported by the Graduate School in Mathematics and Computing (FMB).

# Bibliography


[1] The R project for statistical computing. http://www.r-project.org.

[2] H. Akaike. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723, 1974.

[3] J. M. Alvarez-Castro and O. Carlborg. A Unified Model for Functional and Statistical Epistasis and Its Application in Quantitative Trait Loci Analysis. *Genetics*, 176(2):1151–1167, 2007.

[4] W. Banzhaf, F. D. Francone, R. E. Keller, and P. Nordin. *Genetic programming: an introduction: on the automatic evolution of computer programs and its applications*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1998.

[5] M. Bogdan, J. K. Ghosh, and R. W. Doerge. Modifying the Schwarz Bayesian Information Criterion to Locate Multiple Interacting Quantitative Trait Loci. *Genetics*, 167(2):989–999, 2004.

[6] K. W. Broman. *Identifying quantitative trait loci in experimental crosses*. PhD thesis, Department of Statistics, University of California, Berkeley, 1997.

[7] K. W. Broman and T. P. Speed. A model selection approach for the identification of quantitative trait loci in experimental crosses. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 64(4):641–656, 2002.

[8] K. W. Broman, H. Wu, S. Sen, and G. Churchill. R/qtl: QTL mapping in experimental crosses. *Bioinformatics*, 19(7):889–890, 2003.

[9] G. R. Brown, D. L. Bassoni, G. P. Gill, J. R. Fontana, N. C. Wheeler, R. A. Megraw, M. F. Davis, M. M. Sewell, G. A. Tuskan, and D. B. Neale. Identification of Quantitative Trait Loci Influencing Wood Property Traits in Loblolly Pine (Pinus taeda L.). III. QTL Verification and Candidate Gene Mapping. *Genetics*, 164(4):1537–1546, 2003.

[10] O. 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.

[11] O. Carlborg and C. Haley. Epistasis: too often neglected in complex trait studies? *Nature Reviews Genetics*, 5:618–625, 2004.

[12] G. Churchill and R. Doerge. Empirical threshold values for quantitative trait mapping. *Genetics*, 138:963–971, 1994.

[13] G. A. Churchill and R. W. Doerge. Naive Application of Permutation Testing Leads to Inflated Type I Error Rates. *Genetics*, 178(1):609–610, 2008.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[15] B. Feenstra, I. Skovgaard, and K. W. Broman. Mapping quantitative trait loci by an extension of the Haley-Knott regression method using estimation equations. *Genetics*, 173:2269–2282, 2006.

[16] D. E. Finkel. *DIRECT Optimization Algorithm User Guide.* North Carolina State University, March 2003.

[17] R. A. Fisher. The correlation between relatives on the supposition of mendelian inheritance. *Trans. R. Soc. Edinburgh*, (52):399–433, 1918.

[18] E. Galerpin. The cubic algorithm. *J of Mathematical Analysis and Applications*, pages 635–640, 1985.

[19] J. B. S. Haldane. The combination of linkage values, and the calculation of distance between the loci of linked factors. *J Genet*, 8:299–309, 1919.

[20] C. S. Haley and S. A. Knott. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, 69(4):315–24, 1992.

[21] C. S. Haley, S. A. Knott, and J. M. Elsen. Mapping Quantitative Trait Loci in Crosses Between Outbred Lines Using Least Squares. *Genetics*, 136(3):1195–1207, 1994.

[22] L. Han and S. Xu. A Fisher scoring algorithm for the weighted regression method of QTL mapping. *Heredity*, 101:453–464, Nov 2008.

[23] L. D. W. Hillier et al. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature*, 432(7018):695–716, December 2004.

[24] D. D. Houston, C. S. Haley, A. L. Archibald, and K. A. Rance. A qtl affecting daily feed intake maps to chromosome 2 in pigs. *Mammalian Genome*, 16:464–470, 2005.

[25] R. C. Jansen. Interval Mapping of Multiple Quantitative Trait Loci. *Genetics*, 135(1):205–211, 1993.

[26] M. Jayawardena and S. Holmgren. Grid-enabling an efficient algorithm for demanding global optimization problems in genetic analysis. In *3rd IEEE International Conference on e-Science and Grid Computing, IEEE Conference Proceedings 10.1109*, pages 205–212. IEEE, 2007.

[27] D. Jones, C. Perttunen, and B. Stuckman. Lipschitzian optimization without the lipschitz constant. *J. Optimization Theory App*, 79:157–181, 1993.

[28] 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.

[29] S. Kerje, O. Carlborg, K. Schütz, C. Hartmann, P. Jensen, and L. Andersson. The twofold difference in adult size between the red junglefowl and white leghorn chickens is largely explained by a limited number of QTLs. *Animal Genetics*, 34(4):264–274, 2003.

[30] D. Kosambi. The estimation of map distances from recombination values. *Ann. Eugen.*, pages 172–175, 1944.

[31] E. S. Lander and D. Botstein. Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps. *Genetics*, 121(1):185–199, 1989.

[32] E. S. Lander and P. Green. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America*, 84(8):2363–2367, 1987.

[33] E. S. Lander, P. Green, J. Abrahamson, A. Barlow, M. J. Daly, S. E. Lincoln, and L. Newburg. Mapmaker: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, 1(2):174 – 181, 1987.

[34] K. Lindblah-Toh et al. Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature*, 438:803–819, Dec 2005.

[35] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley Series in Probability and Statistics. Wiley, New York, 2nd edition, 1987.

[36] K. Ljungberg, S. Holmgren, and O. Carlborg. Simultaneous search for multiple QTL using the global optimization algorithm DIRECT. *Bioinformatics*, 20(12):1887–1895, 2004.

[37] M. Lynch and B. Walsh. *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, 1 edition, January 1998.

[38] O. Martinez and R. Curnow. Estimating the location and the sizes of effects of quantitative trait loci flanking markers. *Theor Appli Genet*, 85:480–488, 1992.

[39] R. Mladineo. An algorithm for finding the global maximum of a multimodal, multivariate function. *Mathematical Programming*, pages 253–271, 1987.

[40] C. Nettelblad. Model selection criteria in the NOIA framework for gene interaction. Master's thesis, School of Engineering, Uppsala University, Sweden, Oct. 2007.

[41] J. Pinter. Globally convergent methods for n-dimensional multiextremal optimization. *Optimization*, 17:187–202, 1986.

[42] A. Rebai, B. Goffinet, and B. Mangin. Approximate Thresholds of Interval Mapping Tests for QTL Detection. *Genetics*, 138(1):235–240, 1994.

[43] K. E. Schütz, S. Kerje, L. Jacobsson, B. Forkman, O. Carlborg, L. Andersson, and P. Jensen. Major growth QTLs in fowl are related to fearful behavior: possible genetic links between fear responses and production traits in a red junglefowl x white leghorn intercross. *Behav Genetics*, 34:121–130, 2004.

[44] G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

[45] S. Sen and G. A. Churchill. A Statistical Framework for Quantitative Trait Mapping. *Genetics*, 159:371–387, 2001.

[46] B. Shubert. A sequential method seeking the global maximum of a function. *SIAM J. on Numerical Analysis*, pages 379–388, 1972.

[47] M. J. Sillanpää and J. Corander. Model choice in gene mapping: what and why. *Trends in Genetics*, 18(6):301 – 307, 2002.

[48] J. Slate. Quantitative trait locus mapping in natural populations: progress, caveats and future directions. *Molecular Ecology*, 14:363–379(17), February 2005.

[49] C. M. Wade et al. Genome sequence, comparative analysis, and population genetics of the domestic horse. *Science*, 326:865–867, Nov 2009.

[50] R. H. Waterston et al. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, Dec 2002.

[51] J. D. Watson and F. H. C. Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953.

[52] S. Xu. A comment on the simple regression method for interval mapping. *Genetics*, 141:1657–1659, 1995.

[53] S. Xu. Further investigation on the regression method of mapping quantitative trait loci. *Heredity*, 80:364–373, 1998.

[54] M. Zak, A. Baierl, M. Bogdan, and A. Futschik. Locating Multiple Interacting Quantitative Trait Loci Using Rank-Based Model Selection. *Genetics*, 176(3):1845–1854, 2007.

[55] Z.-B. Zeng, J. Liu, L. F. Stam, C.-H. Kao, J. M. Mercer, and C. C. Laurie. Genetic Architecture of a Morphological Shape Difference Between Two Drosophila Species. *Genetics*, 154(1):299–310, 2000.

[56] Z.-B. Zeng, T. Wang, and W. Zou. Modeling Quantitative Trait Loci and Interpretation of Models. *Genetics*, 169(3):1711–1725, 2005.

**Recent licentiate theses from the Department of Information Technology**

**2010-001** Anna Nissen: *Absorbing Boundary Techniques for the Time-dependent Schrödinger Equation*

**2009-005** Martin Kronbichler: *Numerical Methods for the Navier-Stokes Equations Applied to Turbulent Flow and to Multi-Phase Flow*

**2009-004** Katharina Kormann: *Numerical Methods for Quantum Molecular Dynamics*

**2009-003** Marta Lárusdóttir: *Listen to Your Users - The Effect of Usability Evaluation on Software Development Practice*

**2009-002** Elina Eriksson: *Making Sense of Usability - Organizational Change and Sense-making when Introducing User-Centred Systems Design in Public Authorities*

**2009-001** Joakim Eriksson: *Detailed Simulation of Heterogeneous Wireless Sensor Networks*

**2008-003** Andreas Hellander: *Numerical Simulation of Well Stirred Biochemical Reaction Networks Governed by the Master Equation*

**2008-002** Ioana Rodhe: *Query Authentication and Data Confidentiality in Wireless Sensor Networks*

**2008-001** Mattias Wiggberg: *Unwinding Processes in Computer Science Student Projects*

**2007-006** Björn Halvarsson: *Interaction Analysis and Control of Bioreactors for Nitrogen Removal*

**2007-005** Mahen Jayawardena: *Parallel Algorithms and Implementations for Genetic Analysis of Quantitative Traits*

**2007-004** Olof Rensfelt: *Tools and Methods for Evaluation of Overlay Networks*

UPPSALA
UNIVERSITET

Department of Information Technology, Uppsala University, Sweden