Using survival data in gene mapping : using survival data in genetic linkage and family-based association analysis

Callegaro, A.

Chapter 1

Introduction and overview

The subject of this thesis is to develop statistical methods to reduce heterogeneity in gene mapping analysis. In fact, many geneticists, epidemiologists and biologists have realized that heterogeneity results in loss of statistical power in studies aiming to identify new genetic factors for complex genetic disorders. To deal with heterogeneity additional risk factors are collected such as age at onset, family-history, genetic and environmental factors. This information can be used to select the most informative cases for the analysis or should be taken into account while searching for new genetic factors by weighting individuals according to their risk profiles.

The main focus of this thesis is to develop new and simple statistical methods to reduce heterogeneity weighting individuals for their age at onset. Further, classical nonparametric linkage analysis methods are extended to include the information given by the family-history. Finally, the family-based association analysis methods are extended to adjust for the number of allele shared identical by descent (IBD) and for gene-covariate interaction.

The thesis is a collection of six articles. The articles are self-contained and they can in principle be read in any order. The objective of this introduction is to present some background notation and information, useful for understanding the articles included in the thesis.

1.1 Frailty models

Survival analysis

The analysis of time-to-event data is called survival analysis. Time-to-event data are encountered in many scientific disciplines including demography, medicine, biology, epidemiology, public health, engineering and economics. One complication in the analysis of these data is the presence of censored observations. There are different types of censoring. The most common type of censoring is the right censoring happening when the study does not span enough time in order to observe the event for all the subjects in the study. If a patient goes through the study without having the event, his time to the event is (right) censored, in the sense that we only know that the event happened after the last time we observed the patient. Whenever the censoring time is less than the event time, the event time is missing. Let T be the life-span, the observed time is given by $y = (t, \delta)$ where t is the observed life-span if $\delta = 1$ and the censoring time if $\delta = 0$. A standard assumption of survival analysis is that event-time and censoring time are independent (noninformative censoring).

Most of the survival analysis methods are based on the hazard function $\lambda(t)$, which is the instantaneous failure rate and is defined as:

$$\lambda(t) = -\frac{\partial \log S(t)}{\partial t}$$

where $S(t) = P(T > t)$ is called survival function. Suppose a patient has survived to time t; then the hazard function is the probability that the patient will have an event in the next instant. Many methods have been proposed in the literature to model the hazard function. Proportional hazard models are widely used in medical statistics, where covariates (X) have a multiplicative effect on the baseline hazard ($\lambda_0(t)$)

$$\lambda(t|X) = \lambda_0(t) \exp(\beta X)$$

For example the famous Cox model (Cox, 1972) is a proportional hazard model with unspecified (nonparametric) baseline hazard function.

Multivariate Frailty models

Multivariate event time data arises when each study subject can potentially experience several types of failures or recurrences of a certain phenomenon, or when failure times are sampled in clusters, such as families, schools, hospitals. During the last two to three decades, a large body of literature on multivariate survival analysis has been developed (Hougaard, 2000). Clustered survival data are encountered in many scientific disciplines including human and veterinary medicine, biology, epidemiology, public health and demography. The statistical analysis of these data is complex, especially when the interest is in the dependence structure. A standard statistical approach to model multivariate failure time data is called frailty model. The hazard rate of the jth individual in the ith cluster is given by

$$\lambda(t_{ij}|X_{ij}, U_{ij}) = \lambda_0(t_{ij}|X_{ij})U_{ij}. \tag{1.1}$$

Note that we assume a general dependence between the baseline hazard (λ_0) and the vector of covariates (X_{ij}). For easy of exposition suppose that clusters are families composed of two siblings. Clayton (1978) and Vaupel et al.

(1979) proposed frailty models where the dependence between the two siblings is modelled by a shared random effect $U_{i1} = U_{i2}$. In order to describe more complex dependency structures the shared model was extended by Yashin et al. (1995). Inspired by the variance components methods for quantitative traits, they decomposed the frailty into the sum of independent effects. The total frailty is given by the sum of an effect which is shared by the two siblings (U_s) and a residual (unshared) effect (U_e). According to the correlated frailty model, the frailties of the two siblings in the ith family can be written as follows

$$\begin{aligned} U_{i1} &= U_{s,i} + U_{e,i1} \\ U_{i2} &= U_{s,i} + U_{e,i2}. \end{aligned}$$

Using this model the dependence between the two siblings is a function of the portion of the total variance explained by the shared effect.

Different distributions of the frailty have been proposed in the literature like gamma, log-normal, positive-stable etc. The gamma distribution is mathematically convenient because it yields a closed form likelihood which can be readily maximized. Gamma frailty models can also be expressed in terms of observable marginal survival functions. A limitation of the gamma frailty model is that the likelihood becomes too complex for large clusters (families). A possibility to solve this problem is to decompose the likelihood into pairwise contributions (Lindsay, 1998). Another possibility is to use the log-normal frailty model. In contrast to the gamma frailty model the likelihood does not have closed form and numerical approximations are required. However, the log-normal frailty model can be applied to general families and it permits to model complex dependence structures.

1.2 Nonparametric linkage analysis

Linkage analysis is a method to map disease genes along the genome. Using this approach, the rough location of the disease genes are detected by typing DNA sequence called genetic markers of pedigree sets. The method of linkage analysis is based on the concept of biological inheritance. Human chromosomes come in pairs, one is inherited from the father and one from the mother. Before the chromosomes are transmitted to the offspring, the maternal and paternal chromosomes pair up and exchange parts. Such a process is called crossing-over and the exchange of genes is called genetic recombination (figure 1.1).

Since genes close by on the same chromosome tend to be inherited jointly, the frequency of recombination measures the distance between genes. Disease genes are mapped by measuring recombination against a panel of different markers spread over the entire genome. In most cases, recombination

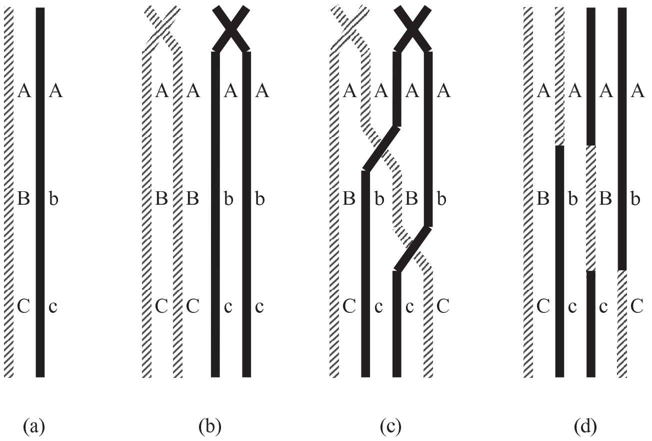

FIGURE 1.1: *Simplification of the meiosis process: (a) maternal chromosomes; (b) the maternal chromosomes duplicate; (c) the four chromosomes crossover in random chromosomal locations; (d) four mixed strands, one of them is randomly transmitted to the offspring.*

will occur frequently, indicating that the disease gene and marker are far apart. Some markers however, due to their proximity, will tend not to recombine with the disease gene and these are said to be linked to it. Genetic linkage analysis test for coinheritance of chromosomal regions with a trait. There are two main classes of linkage analysis, the parametric and the non-parametric methods. Parametric linkage analysis is a powerful approach to localize genes when a genetic model can be approximated, however they can be highly sensitive to misspecification of the linkage parameters (gene frequency, penetrance and degree of dominance) (Clerget-Darpoux et al., 1986). Since for complex traits the mode of inheritance is often unknown, nonparametric methods are usually preferred because they do not make any (explicit) assumptions about the disease model (Kruglyak et al., 1996).

Nonparametric methods are based on allele-sharing between individuals in a pedigree. Two individuals share an allele identical by descent (IBD) if they have both inherited exactly the same allele from a common ancestor. As an example consider the following nuclear family shown on Figure 1.2. At the locus x the two siblings have the same genotype (AA), but they share zero alleles IBD. At the locus y the two siblings share the maternal allele (B) and at the locus z the two siblings share both the maternal (C) and the paternal allele (g), so they share 2 alleles IBD. The biological phenomenon behind the nonparametric linkage analysis is that affected individuals in a family share the same ancestral predisposing DNA segment at a given trait locus. It follows that linkage between a disease locus and marker genotypes can be studied by comparing the observed number of alleles shared IBD to the expected number of alleles in the population. An increase in the number of alleles IBD indicates the presence of a susceptibility gene in the region. For a set of N affected sibling pairs the popular NPL score (Blackwelder and Elston, 1985; Kruglyak et al., 1996) is given by,

$$NPL = \frac{\sum_{i=1}^{N} \gamma_i(\hat{\pi}_i - E\pi)}{\sqrt{\sum_{i=1}^{N} var_0(\hat{\pi}_i)\gamma_i^2}} \tag{1.2}$$

where $\hat{\pi}_i$ is the estimated proportion of alleles shared IBD between the ith sib-pair; $E\pi$ is the expected proportion of IBD under the null hypothesis and it is equal to 0.5 for sibling pair; $var_0(\hat{\pi}_i)$ is the variance estimated under the null hypothesis of no linkage and γ_i is the weight assigned to the *i*th sib pair. The NPL score is computed at a grid of marker positions. The final statistic is the maximum value of NPL scores over all marker loci. Significant linkage is detected whenever the maximum is larger than a threshold. The value of the threshold to control false detection rate has been discussed by Feingold et al. (1993), Lander and Schork (1994) and Lander and Kruglyak (1995). Kruglyak et al. (1996) noted that, under the null hypothesis, the variance of the NPL score is generally smaller than one. This effect is due to the fact that in general, the information on descendent is incomplete. Assuming that the variance of NPL score is equal to one lead to conservative p-value estimates in the case imperfect data. A possibility to solve the problem is to estimate the variance by simulations, or to use a particular likelihood-ratio approach (Kong and Cox, 1997).

For quantitative traits, the phenotypes of the family-members are usually modelled by multivariate normal distribution, with variance-covariance matrix depending on linkage, on residual genetic and/or on environmental effects. Since families are typically chosen based on their trait values, the retrospective likelihood of the marker data conditioned on the trait is adequate to account for the ascertainment process. From the retrospective likelihood different score statistics have been proposed in the literature (Lebrec et al., 2004; Sham and Purcell, 2001; Tang and Siegmund, 2001; Tritchler et al., 2003) which are similar to the NPL score statistic (1.2). In this case, the weight function γ_i is derived from the retrospective likelihood and it is a function of the family trait values standardized against known population parameters.

For age at onset traits, different frailty models have been proposed in the literature (Commenges, 1994; Jonker et al., 2009; Li and Zhong, 2002; Pankratz et al., 2005). However, no NPL score statistics are available. For this reason a major part of the thesis (Chapter 2-5) is devoted to the derivation of NPL score

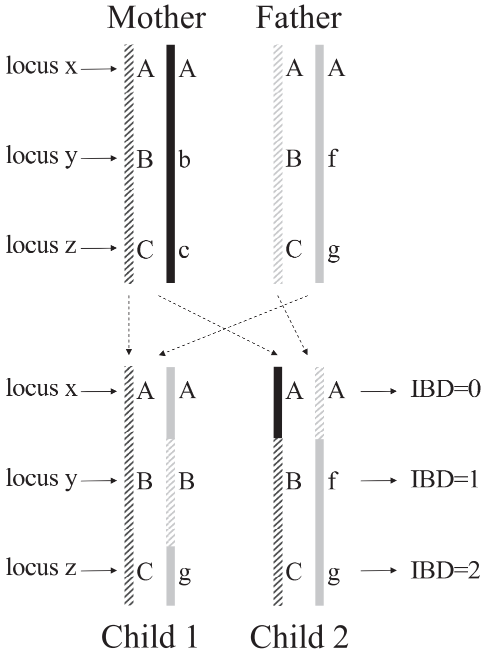

FIGURE 1.2: *IBD sharing in a nuclear family.*

statistics for age at onset data.

NPL score adjusting for age at onset

Many complex disease studies have suggested that casual genes can influence the age at which the event occurs (age at onset). In the literature methods have been proposed for linkage with age at onset data (Commenges, 1994; Jonker et al., 2009; Li and Zhong, 2002; Pankratz et al., 2005) where the linkage effect is modeled as a random effect of frailty models (1.1). To simplify the exposition suppose that families are composed of two siblings. For notational simplicity, we will suppress the family index. Let's represent the four paternal alleles of a particular family with the integers $k = 1, 2, 3, 4$. The segregation of these alleles to the two siblings can be represented by the sequence $s = (s_{11}, s_{12}, s_{21}, s_{22})$ where s_{j1} and s_{j2} label the paternal and the maternal alleles of the jth sibling. The number of distinct integers represents the number of genetically distinct alleles among the two siblings. Let's represent the effect of the four paternal alleles by four independent random variables $U_{p,k}$, $k = 1, ...4$. Then the frailties of the two siblings explained by the inherited paternal alleles can be written as follows

$$\begin{aligned} U_{a,1} &= U_{p,s_{11}} + U_{p,s_{12}} \\ U_{a,2} &= U_{p,s_{21}} + U_{p,s_{22}}. \end{aligned}$$

Using this frailty model the dependence between the two siblings increases with the number of allele shared IBD. Commenges (1994) proposed a frailty model where the random effect of the sibling j is given by $U_j = U_{a,j}$. The random effect comes from an unspecified distribution with $E(U_a) = 0$ and $E(U_a^2) = 1$. This model assumes that all the dependence between siblings depends on a single causal gene, in fact, under the null hypothesis of no linkage the two siblings are independent. In order to take into account residual genetic or environmental correlation, Li and Zhong (2002) proposed an additive gamma frailty model $U_j = U_{a,j} + U_s$ where $U_{a,j}$ models the linkage effect and U_s is an independent random variable. Jonker et al. (2009) further extended the Li and Zhong model (1.3) by adding an unshared random effect ($U_{e,ij}$). Instead of using gamma distributed random effects, Pankratz et al. (2005) proposed a log-normal frailty model for age at onset linkage analysis.

In order to test for linkage most of the authors proposed likelihood-ratio tests (Jonker et al., 2009; Li and Zhong, 2002; Pankratz et al., 2005). However, the score test is a computationally faster, locally most powerful, and robust alternative to the likelihood ratio test. For these reasons, in Chapter 2-5 we derived new score tests for age at onset linkage analysis. The derived score statistics are classical NPL scores (1.2) with weight functions γ_i depending on the age at onset times standardized against known population parameters. In the simple case that the siblings are independent under the null hypothesis, the weight is the product of the martingale residuals of the two siblings $\gamma_i = (\delta_{i1} - \Lambda_1(t_{i1})) \times (\delta_{i2} - \Lambda_2(t_{i2}))$, where Λ is the known marginal cumulative hazard function.

NPL score adjusting for family-history

Wallace and Clayton (2006) showed that selecting cases with positive family history of disease generally increases the power to detect linkage for common complex diseases. The increase in power is observed particularly in the presence of environmental effect and with rare disease risk alleles. Instead of selecting families with positive family-history, an alternative strategy may be to recruit unselected families, but collect information on family history. Then a variable summarising the family history may be included in the analysis.

A standard approach to take into account ungenotyped affected individuals is by sampling the distribution of the missing marker data given the observed marker throught MCMC algorithms. Multiple imputation methods has been proposed for linkage analysis with a two step procedure: first, inferring missing genotypes from the genotype of the observed relatives, and, second applying standard linkage analysis (Almasy and Blangero, 1998; Sobel and Sengul, 2001). Skrivanek et al. (2003) proposed a sequential imputation method to approximate the allele sharing statistics. These MCMC methods are complex, computationally intensive and they are not powerful when there is a large number of samples completely untyped (Sobel and Sengul, 2001).

In Chapter 6 we derive a simpler method where the family-history is included in the weight of the NPL score. The weight depends on known information so no complex sampling methods are necessary. Suppose that affected sibling pairs (ASPs) have been genotyped for the analysis and suppose that a portion of them has one untyped affected siblings (or one untyped parent). In this simple case the proposed method is the NPL score (1.2) with weight $\gamma = 1$ for the ASPs without positive family history and weight $\gamma = 1.5$ for the ASPs with positive family history.

1.3 Family based association analysis

Genome-wide linkage analysis are often followed by association studies of candidate genes located under the linkage peak. With these genetic association studies one hopes to identify candidate genes whose variation causes the excess IBD sharing of marker alleles in the linkage study. Genetic association studies compares alleles or genotype frequencies in affected individuals with those in unaffected individuals. A marker may be associated with the disease because it is in linkage disequilibrium with a causal variant at the disease locus. Linkage disequilibrium is the condition in which the haplotype frequencies in a population deviate from the values they would have if the genes at each locus were combined at random. In contrast with linkage, the linkage-disequilibrium is a result from ancestral recombination events and it is a measure of co-segregation in the population, instead of a measure of co-segregation in a pedigree.

Disease-marker association may also be due to population stratification. Population stratification is the presence of a systematic difference in allele frequencies between subpopulations. To eliminate false positive results, family-based designs are used. The unified approach to Family-Based Tests of association (FBAT) have been proposed by Rabinowitz and Laird (2000) and Laird et al. (2000), builds on the original TDT method (Spielman et al., 1993) in which alleles transmitted to affected offsprings are compared with the expected distribution of alleles among offsprings. Let $X_{g,ij}$ denotes some function of the jth offspring's marker genotype in the ith family. Usually the association effect is modeled as a covariate, so for age at onset trait a natural choice is

$$\lambda(t_{ij}|X_{g,ij}, U_{ij}) = \lambda_{0ij}(t_{ij}) \exp(\beta_g X_{g,ij}) U_{ij}, \tag{1.3}$$

where the random effect U_{ij} models the dependence between siblings and it may have components that are attributable to the linkage effect, to shared environmental and polygenic effect.

The score statistic to test $H_0 : \beta_g = 0$ (from a retrospective likelihood) gives the so called FBAT statistic

$$FBAT = \frac{\sum_{i=1}^{N}(X_{gi} - EX_{gi})'\gamma_i}{\sqrt{\sum_{i=1}^{N}\gamma_i' var_0(X_{gi})\gamma_i}}, \tag{1.4}$$

which is a linear combination of offspring genotypes (X_{gi}) and weights (γ_i). EX_{gi} denotes the expectation of the offspring's marker genotype conditioned on the parental genotypes. The weight is a function of the trait values. For example, if we assume that the frailty is constant $U_{ij} = 1$, the weight is the martingale residual $\gamma_{ij} = \delta_{ij} - \Lambda_0(t_{ij})$. In particular, using a gamma distributed random effect Zhong and Li (2004) derived a particular weight function to test for association in the presence of linkage.

In Chapter 7 we derive a new class of weights from a generalized linear mixed model. For survival (age at onset) data we use the Poisson model. Further, we extended the FBAT statistics in order to adjust for gene-covariate interaction.

1.4 Outline of the thesis

This thesis consists of three parts. The first part consists of Chapter 2 up to 5. It develops new NPL score tests for age at onset linkage analysis. The second part consists of Chapter 6, which presents a new method to test for linkage analysis taking into account the family-history. Finally, the third part, Chapter 7, addresses the subject of family-based association analysis adjusting for linkage effect and/or gene-environmental interaction.

Chapter 2 deals with the age at onset linkage analysis of selected sibling pairs. We derive a NPL score statistic from the retrospective likelihood of a gamma-frailty model. We use the model proposed by Jonker et al. (2009) but with a different parametrization which permits to use information known from twin studies such as the sib-sib correlation. Simulation studies show that the proposed method is robust and more powerful than standard nonparametric linkage methods. As illustration we apply the new score statistic to data from a breast cancer study.

Chapter 3 extends the score test derived in Chapter 2 to include the parental age at onset. NPL score statistics are derived from a gamma frailty model and from a log-normal frailty model, respectively. In order to investigate how age at onset of sibs and their parents affect the information for linkage analysis the weight functions were studied for rare and common disease models, realistic models for breast cancer and human lifespan. We studied the performance of the methods by simulations. As illustration, the score statistics were applied to the GAW12 data. The results show that it is useful to include parental age at onset information in genetic linkage analysis.

Chapter 4 addresses the issue of testing for age at onset linkage analysis in general pedigrees. The score test derived in Chapter 2 is extended to general pedigrees using a pairwise likelihood approach (Lindsay, 1998). Further, this method is compared by simulations with the approximated log-normal frailty model derived in Chapter 3. The two methods are applied to the GAW16 Framingham data.

Chapter 5 is concerned with robust score tests for aggregation and linkage analysis of human longevity. We propose a new statistic for aggregation analysis, which tests for a relationship between a family history of excessive survival of the sibships of the long-lived pairs and the survival of their parents and their offspring. For linkage analysis, we derive a new NPL score statistic from a shared gamma frailty model, which is similar in spirit to the score test derived in Chapter 2. We apply the methods to data from the Leiden Longevity Study (Schoenmaker et al., 2006).

Chapter 6 is concerned with a new class of allele-sharing statistics which takes into account the information given by the family history. Such an information is included into the scoring functions of classical allele-sharing statistics. We consider pedigrees of affected sibling pairs with positive family-history given by one ungenotyped affected relative. By simulating using models for complex diseases we showed that taking into account family-history generally increases the power to detect linkage. Allele-sharing methods were applied to the symptomatic osteoarthritis GARP study where taking into account the family-history increased considerably the power to detect linkage in the surrounding of the $DIO2$ susceptibility locus.

In Chapter 7, we develop a score test for family-based association analysis. In order to study family based association in the presence of linkage we extend a generalized linear mixed model proposed for genetic linkage analysis (Lebrec and van Houwelingen, 2007) by adding a genotypic effect to the mean. The corresponding score test is a weighted statistic, where the weight depends on the linkage effect and on other genetic and shared environmental effects. To test for genetic association in the presence of interaction, we propose a linear regression method where the family-specific score statistic is regressed on family-specific covariates.

In the last chapter the results presented in chapters 2 to 7 are summarized. Finally, the appendix illustrates the use of `arthur` package, a software which was built to apply most of the methods discussed in this thesis.