

Population Pedigree Inference from Genomic Data

Due to methodological advances and the phenomenal increase in genetic data from different species, phylogenetic analysis (the inference of evolutionary relationships between species) has risen to prominence and been put on much firmer statistical ground (Felsenstein, 2004; Semple and Steel, 2003). Similarly, considerable progress has been made in characterizing the genealogical relationships between segments of chromosome sampled from individuals from the same species, based on the pattern of mutational and recombinational diversity present in these segments. While such studies are important in their own right, the growing deluge of genome-wide SNP genotype data (and in the near future sequence data) heralds the possibility of inferring the entire pedigree between a set of individuals. For any given contemporary individual, the ancestor(s) of a particular chromosomal segment represent only a fraction of all the ancestors from whom that individual inherited all his or her genetic material. However, when genotypes are obtained from different segments across the genome at a sufficiently dense rate, it becomes at least theoretically possible to reconstruct the genealogical links between contemporary individuals through all ancestors.

Surprisingly little attention has been given to this problem of using genome-wide genetic data to reconstruct the entire pedigree that connects a set of individuals from the same population or species (Hein 2004). Considerable research effort has gone into very restricted applications of this larger task. Thus, for example, multi-locus data are routinely used to test for genealogical relationships between individuals in forensic genetics (Evetts and Weir, 1998) – yielding highly accurate inferences, but for a very limited and shallow set of genealogical links. In addition, such data are used to estimate the proportions of ancestry individuals have from different ancestral gene pools (Rosenberg et al. 2002), providing highly abstracted and summarised information about distant genealogical links between sets of distantly related individuals. Disease gene mapping is a major biomedical activity and there has recently been a switch towards population based (association) mapping at the expense of pedigree mapping. However, genetic data on a massive scale might allow the pedigree to be inferred, which would have major consequences in future approaches to mapping.

Primary aims

The ultimate aim of this project is to devise a statistical approach that can provide efficient and powerful inference of the pedigree relating a set of individuals, where genome-wide genetic data has been obtained. However, in order to achieve this ultimate aim, a number of specific novel theoretical and empirical problems need to be addressed.

1. What are the limits of genealogical inference as function of the number of sampled markers, number of individuals, generations back in time and structure of the pedigree to be inferred?
2. How is pedigree information and certainty rationally represented on a pedigree?
3. What are the most efficient heuristic methods to infer a single pedigree and to traverse the set of pedigrees? And what are the most appropriate algorithms to integrate over the set of pedigrees?
4. What are useful population model parameters that can be used as prior information to restrict the enumeration of possible pedigrees to a relatively small subset of credible pedigrees?

The use of real genealogies and genotype data to validate pedigree inference methods

It would be possible, at least in theory, to develop pedigree inference methods on the basis of simulated genealogies and genotypes. However, remarkably little is known about the parameters required to simulate realistic human genealogies. As a result, the evaluation of pedigree inference

methods on the basis of real genealogies and genotypes is a necessary and essential aspect of the proposed project, providing a firm testing ground and a means of side-stepping potential problems caused by unrealistic model assumptions underlying simulated genealogies. In addition, we will use the real Icelandic genealogies to explore which parameters are most useful for the simulation of credible human genealogies.

The real data used in this project is focussed on a set of 15,000 individuals from the Icelandic population, who have been genotyped for 317 thousand SNPs (and many for about 2000 microsatellites), and for whom the pedigree is comprehensively and accurately known for the past 350 years (10-12 generations). The latter information is from the deCODE Genetics genealogy database (Helgason et al., 2003, 2005), which spans all 300000 extant Icelanders and just over 400000 of their ancestors. The SNPs are from the Illumina HumanHap 300 (HH300) BeadChip, with which one individual is genotyped simultaneously for 317 thousand SNPs, which are randomly distributed across the genome and provide efficient coverage of the genome as they were selected as a highly informative subset of the SNPs included in release 16 of the HapMap project. Added value is provided by the 2000 microsatellite markers, as their high variability makes them extremely informative about identity by descent for the chromosomal segments to which they map. Additional information is available for most individuals in the genealogical database, including geographic location at birth and/or at some time during life. For individuals that are no longer alive, the database also contains information about date of death and in many cases geographic location at death. These variables may be of interest for determining which parameters are most useful for simulating genealogies and restricting the number of enumerated pedigrees.

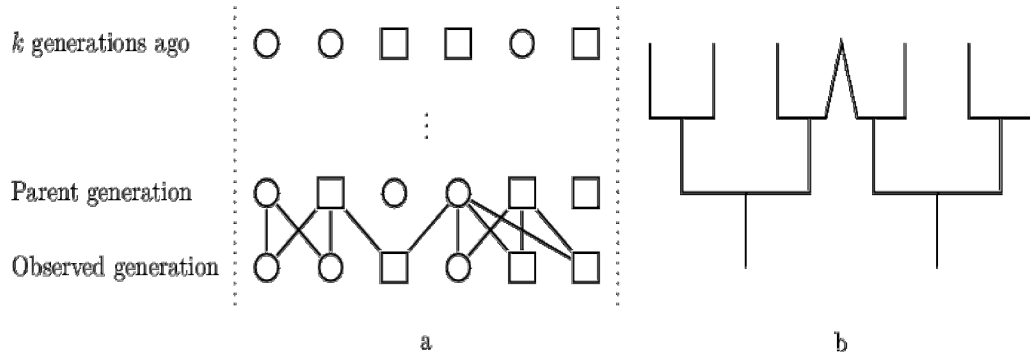
The combination of the deCODE Genetics genealogy database and the availability of extensive genetic data from the Icelandic population present a unique opportunity to develop population pedigree inference methods. In principle, other genealogy databases could be used for such research, for example large genealogy databases from Quebec (Heyer et al., 1999) and Utah (et al., 1996). However, these databases do not provide as complete coverage of a single and clearly defined population as the Icelandic genealogy database and the availability of genotypes for a large proportion of the individuals in the genealogy database is a situation that is wholly unique to Iceland.

Related Work

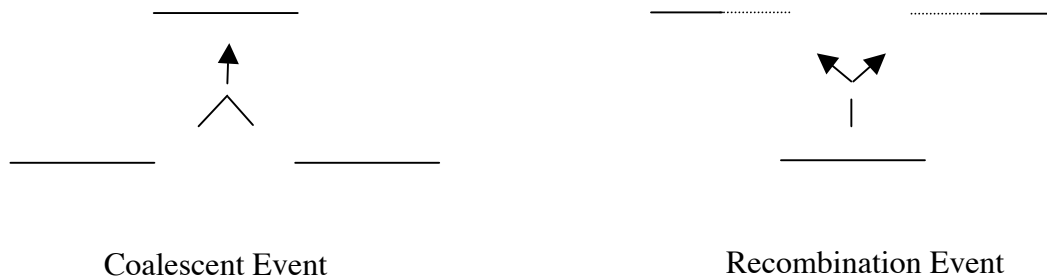
The work closest in spirit to this proposal is found in Cowell and Mostad (2002) and their program FAMILIAS. This infers small pedigrees from genetic data, but does not scale or address population pedigrees or investigates reconstructability as far down in time as possible. Programs like FASTLINK (Cottingham, Idury, Schaffer, 1993), VITESSE (O'Connell JR, Weeks DE, 1995) and SUPERLINK (Fishelson and Geiger, 2003) are concerned with algorithmic speed, but from a different perspective – namely disease gene mapping, that has conceptual similarities. The mapping methods involve summing over the unknown genetic configurations in the known pedigree, while simultaneously investigating the likelihood of data including phenotypes explained by a causative unknown position. Steel and Hein (2005) derives a series of results on combinatorial properties of pedigrees under idealized models related to reconstructability.

Background and basic concepts

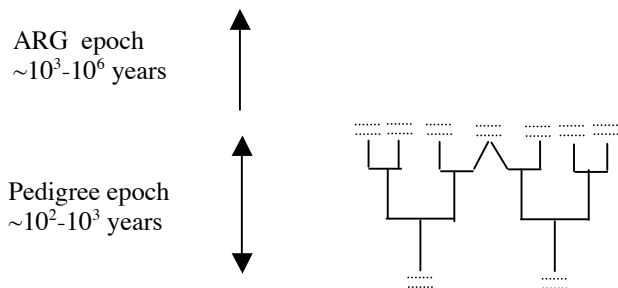
In the current setting, a *Pedigree* (see illustration below) refers to a graph with extant individuals labelled with distinct names and typically unlabelled ancestors. The individuals are nodes. Each individual has two edges, pointing to the parents (father and mother) in the previous generation.



The *Ancestral Recombination Graph* (ARG) is the graph that describes the relationship of a set of sequences (Hein, Schierup and Wiuf, 2005). See illustration below for the two basic events in an ARG – the coalescent and the recombination event. An evolutionary history can be translated into an ARG by starting in the present and going backwards in time until all positions of the sequences have found one single ancestor. Going back in time, sequences encounter coalescences and recombinations. Coalescent events will merge sequences that are identical, reducing the sample size by one. Recombinations will redistribute a single sequence to two sequences, where one sequence will carry the material to the left of the recombination point and the other the material to the right of that point. In most analysis the ARG ignores that sequences are in individuals and thus describes a population of sequences, not of individuals with sequences.



The sequences can be modified by *mutations* (backwards in time) that will change a single position in a single sequence. Pedigree inference will typically span a period of 50 years to maximally a thousand years, while the ARG will go back in the range 10^5 - 10^6 years. There are two consequences of this for the proposed project. Firstly, most of the mutations observed as differences in the genetic material of extant humans, will not have occurred in the Pedigree Epoch, but will enter as differences already present in the founders of the pedigrees (about 10.000 in Iceland). Secondly, stochastic models describing the ARG can be used to describe the relationship of the sequences that appear in the founders of the pedigree.



Genealogies are combinatorial objects, but there are natural probability measures induced by the evolutionary process of reproduction, recombination and mutation. The fundamental models defining the probability distributions underlying pedigrees are almost inherent in the models underlying coalescent theory. The major missing components are couple formation (possibly including an infidelity parameter), progeny distribution for a given couple and distribution of birth times. These have been explored and can easily be incorporated in population reproduction models. This naturally defines a prior distribution on pedigrees. Given a pedigree, and the genomes at the founders (nodes where both parents are not known) the recombination and mutational process defines the probability of the genomes of the individuals.

Methodological Workplan:

Probabilistic pedigree inference naturally falls in 3 stages: Firstly, defining a prior distribution on pedigrees. Secondly finding the likelihood of the data for a fixed pedigree. Thirdly, the set of pedigrees must be investigated: A reasonable initial guess of a pedigree must be obtained and a neighbourhood of pedigrees must be explored.

1. Priors on Pedigrees.

For the present purpose, we must be able to sample pedigrees and assign probability/density to a pedigree efficiently. Probability measures on pedigrees have been studied for decades (Moehle, 1994; Kammerle, 1989, 91), but have experienced a recent revival (Chang, 1999; Derrida et al., 2000 and Rhode et al., 2004). Clearly, given genotypes from a set of individuals, the space of possible pedigrees connecting them is immense. However, only a smaller subset of all possible pedigrees are actually reasonable given prior knowledge about human life-history traits and, where it is available, information about the demographic history of the population to which the individuals belong. Thus, the use of priors that take such factors into account are likely to prove extremely valuable for pedigree inference. Among the potentially useful factors are those relating to fecundity (offspring numbers, interval between offspring, age at first and last offspring, etc.) and mating (geographic location, number of different mates, avoidance of relatives). Most existing models are based on discrete generations and have unrealistic treatment of life-history factors. One of the aims of the proposed project is to evaluate to what extent the inclusion of such information aids pedigree inference procedures. One recent simulator developed by Gasbarra et al. (2005) used coalescent-based theory to simulate relationships -- and hence a pedigree -- backwards in time, according to fidelity (average number of 'marrying' partners) and fertility (average number of children) parameters. Priors have also been designed for pedigrees without reference to an explicit population process. Egeland et al. (2002) is an example of this where a prior is a product of factors that can penalize unlikely properties of a pedigree.

Most of the priors have the property that the change in probability of a pedigree due to a local change can be calculated much faster than a total recalculation. Since an estimation of population parameters is an aim in this proposal, population based priors will be pursued.

2. Likelihood of data given a pedigree

Classic methods for this problem were developed by Elston and Stewart (1971), Lander and Green (1987) and Cannings, Skolnick and Thompson (1978), but recent developments have created flexible frameworks allowing more efficient algorithms that may be further developed to scale to the present problem using probabilistic graphical models (Lauritzen, 1996). They form a natural general framework to express and manipulate a number of important aspects of computation in statistical genetics, in particular problems involving pedigree analysis (Lauritzen and Sheehan, 2003).

The flexibility of graphical models makes it particularly straightforward, in principle, to represent pedigree uncertainty, exploited for example by Hansen and Pedersen (1994) in a problem concerning fur colour of foxes (Skjøth et al., 1994) and in a smaller scale concerning forensic identification problems (Dawid et al. 2002), which formally can be seen as problems of pedigree determination

(Egeland et al., 2000). Such problems can be solved with currently known computational methods, which are essentially exact. However, the pedigree determination problems to be addressed and tackled in this project have a scale and scope expected to go beyond the limits of known exact algorithms, hence new methods need to be developed to represent and manipulate the associated problems in graphical models.

One promising way ahead is the use of advanced Monte-Carlo methods, known as *blocking Gibbs sampling* (Jensen et al., 1995; Jensen and Kong, 1999; Lund and Jensen, 1999). Although this has been tried successfully in many cases, there is a need to refine the algorithms and the understanding of their properties. In particular, efficient algorithms should be developed to identify blocking sets, as well as verifiable conditions for correctness and strong mixing of associated MCMC algorithms.

Another promising approach uses iterative loopy local propagation (Heskes 2004), essentially ignoring cycles in the associated graphical model. Such approximate methods do not yield the correct likelihood, but as they tend to over-inflate high likelihood points they may be very efficient in identifying potential candidates for pedigrees with high likelihood.

To calculate likelihood ratios between similar pedigrees, methods need to be developed which do not traverse the entire pedigree but are based on the idea that the relevant effect of changes only affect small subsets of the pedigree. This can either be exactly or approximately true. A systematic approach to exploit this leaves great scope for efficiency improvement in the calculations, for example as indicated in Olesen and Madsen (2002) and Flores and Olesen (2003). Extending this with probability computation based on aggregation of probabilities in a balanced representation of the junction tree resulting from the graphical model can potentially lead to exponentially faster likelihood computations.

3a. Initial Pedigree

By modifying a method first proposed by Thompson (1975), we may infer pairwise relationships by a maximum likelihood technique that computes a set of three kinship coefficients. This and derived methods (for instance Cowell and Mostad, 2002) increase in power if each allele has many states. This can be obtained by combining neighboring SNPs. Heuristic reconstruction can be done using methods with analogues such as reconstructing more and more of the global pedigree further and further back in time. A given pedigree will satisfy all pairwise relationships to a quantifiable degree and local rearrangements of the pedigree will define a local search algorithm. Pairwise relationships going several generations back in time are usually much harder to pinpoint than single generation relationships like siblings and parent-child relationships. Hence we intend to investigate to what extent imputation of parental genetic material is helpful and necessary for reconstructing deep pedigrees.

3b. Pedigree Space Traversal and Pedigree Metrics/Similarity Functions

Neighborhoods can be defined in terms of natural edit operations on pedigrees. Natural edit operations would include addition and deletion of an individual and changing a parent of an individual, but also operations like merging two individuals, i.e. postulating that two individuals in the pedigree represent the same real person, or the reverse operation of splitting one individual into two should be explored. Given such operations, Markov Chain Monte Carlo (MCMC) methods are ideally suited for exploring the space of pedigrees with high likelihood. A main aim of the project will be to develop efficient MCMC procedures for sampling pedigrees according to their posterior probability. Efficient will here necessarily have to be defined in the context of methods developed for incremental computation of pedigree likelihood, as discussed below. Hence, not only the efficiency of sampling the space of likely pedigrees but also the efficiency with which likelihoods of pedigrees can be computed need to be taken into account.

Data Analysis :

Analysis of data will have two components – simulations and analysis of the Icelandic data. The analysis will be accompanied by making the user-friendly and generally available software.

i. Simulations

The analysis of simulated data serves two functions – rigorous methodology testing and mapping quantities of interest – such as reconstructability – for different scenarios.

Rigorous testing implies continuous generation of data and simultaneous analysis, to test the code and map computational requirements as a function of number of individuals and markers. This is crucial to assess the stability and scalability of the code.

It is of great interest to analyze the reliability of reconstructions as a function of markers, chosen individuals and time.

Markers can be SNPs or microsatellites with an increasing emphasis on the former. These should be range from 1-2 percent of complete knowledge of genomic variation to complete knowledge. Since pedigree reconstruction poses theoretical problems of reconstructability it would also be of interest to have simulations corresponding to 10 and 100 times the human genome to monitor how good pedigree reconstruction could be in idealized cases. Additionally, markers might be phased or not. Finally, markers can be chosen to represent common alleles according to criteria of different strengths.

Chosen individuals: Different number of individuals can be chosen from a small percentage to the complete population. The individuals can be chosen randomly, be gathered in families or selected to minimize relatedness.

Time from present: Clearly more recent relatedness is easier to detect than more ancient on average. However, this could fluctuate dependent on random events and occasionally reconstructions could be possible far back in time in certain areas of the pedigree.

Mapping based on pedigrees or population samples have been the two major ways to locate disease genes. Presently there is a major tilt towards association mapping, but this approach might make suboptimal use of valuable pedigree information. Tabulating the size of this effect is of great interest and could be done with the developed methods by assuming standard models of disease inheritance: single or several causative SNPs, whose phenotypic effect is determined by recessive/dominance with a certain penetrance. Using the knowledge of the true pedigree is most likely an important factor, but using mapping techniques based on models explicitly taking the pedigree into account in the modeling could also increase power.

ii. The Icelandic Data

The Icelandic Data is the optimal data set in the world for testing the proposed methods and to analyze for biological questions. The Icelandic Pedigree should be

- Annotated with posterior probability of each asserted relationship. The Icelandic Pedigree has about 700.000 child-parent relationships (pedigree edges), so these will be investigated by exploratory statistical methods. What is the distribution of posterior probability in general? If ranked in according to birth of child? Are there surprising gender correlations? Are there areas of certainty and uncertainty in the pedigree?
- Is the pedigree stable or would local re-arrangements lead to a more likely pedigree?
- Both with and without pedigree knowledge, the distribution of histories of ancestral genetic material can be sampled and questions like the amount of ancestral material to the present individuals at a given time and the amount of ancestral material in the proposed founder population to Iceland can be answered.
- The parameters defining the prior distribution on pedigrees is not the main focus, but key parameters relating to fecundity and fidelity should be tested. In the Icelandic case the

pedigree is almost known, but in many anticipated applications the pedigree will be a hidden state that must be integrated out according to its probability. In both cases the central parameters such a population growth, infidelity rates etc. can be estimated.

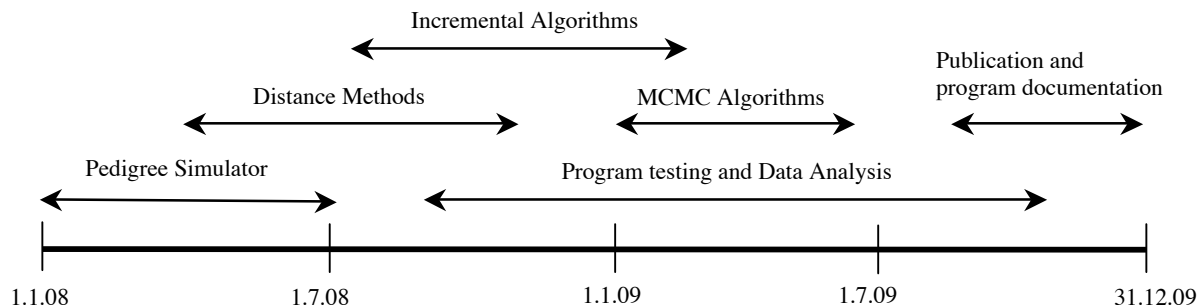
Software Development

The main focus in this project will be on developing methods for large-scale reconstruction of pedigrees from genomic data, and application of these methods on real data sets. The aim is not to develop a polished and easy-to-use package for pedigree inference. Still, to apply methods to large data sets these will necessarily have to be efficiently and robustly implemented. It will remain a priority in implementation phases to achieve a high degree of flexibility in the software to allow its easy use and application to a wide range of related problems. All software will be made available to the research community for download and, where feasible, as web servers.

Future Work

With this project we also hope to stimulate interest in working on pedigree related problems, both locally by formulation and supervision of relevant student projects and in the broader bioinformatics research community. All three applicants have the possibility of recruiting PhDs in their current situations.

Work Schedule and Milestones.



Dependencies between the different parts of the proposed investigation to a large extent dictates the time ordering of these parts. The resulting work schedule in terms of major deliverables and tasks is shown graphically in the diagram above.