

Population Pedigree Inference from Genomic Data

7.1.08

Supervisors: Jotun Hein and Steffen Lauritzen

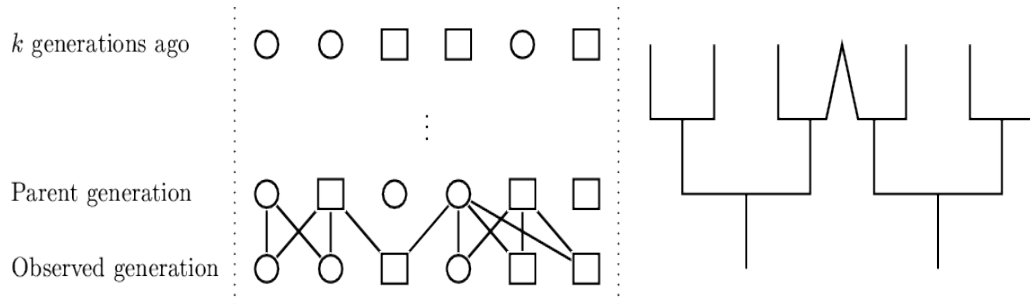
Goal: To investigate the possibility of inference of the pedigrees in a population in very simple cases. To this end, three sets of programs must be made/run:

- A simulation tool to generate sequence data and pedigrees in a small population
- Algorithms calculating the likelihood of data for a given pedigree.
- Pedigree modification. This is necessary to traverse the set of possible pedigrees.

It is then possible both to simulate and infer pedigree. Due to the hardness of the problem, the size of the population, number of sampled extant individuals and pedigree depth will be kept very small, but how to extend methodologies to real data sets should be considered.

Motivation: 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.

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. 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.

Proposed Research:

1. Simulating Pedigrees and Genomes. For the present purpose, we must be able to sample pedigrees and assign probability/density to a pedigree efficiently. Pedigrees will be generated by explicit mating schemes and genomes will be traced within a pedigree by obeying Mendelian inheritance. We will assume discrete and constant generation time. Above the chosen pedigree depth, sequence relationship will be described by standard coalescent theory. 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. Yun Song has also written a simulator that is available.

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).

3. Pedigree Space Traversal. 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.

However, given the time limitations, we propose only to use random sampling of pedigrees and exhaustive search. If time permits more advanced approaches should be investigated.

Proposed Investigations and Simulations:

The analysis of simulated data serves two functions – 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.

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

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.

Work Plan

- Read key articles and expand these 2 pages into a 5 page detailed work plan
- Generate data with Yun Song's or Gasbarra's programs
- Try standard packages for pedigree likelihood calculations
- Implement random and exhaustive pedigree generator
- Infer pedigrees for tiny populations (<8), sampled extant individuals (<5) and pedigree depth (<4).

Comments: i. This project is demanding and would need good programming skills, ability to read literature from diverse backgrounds and good understanding of basic combinatorics and probability theory. ii. This project could be expanded to a full PhD devoted to pedigree inference based on real re-sequencing data. This field is expected to be of increasing importance in coming years. iii. Besides Hein and Lauritzen, Lyngsø, Thatté and Hellenthal are also interested in this project, so there should be ample opportunity for discussion.

References:

- Cannings, Skolnick and Thompson (1978) "Probability Functions on Complex Pedigrees" Adv. Appl. Prob. 10:26-61.
Elston, R.C. and Stewart, J. (1971) "A general model for the genetic analysis of pedigree data" Human Heredity 21: 523-543.
Felsenstein (2004) "Inferring Phylogenies" Sinauer
Gasbarra, Sillanpää and Arjas (2005) "Backward simulation of ancestors of sampled individuals". Theor. Pop. Biol. 67:75-83
Hein, J.J. (2004) Human evolution: Pedigrees for all humanity. Nature 431:518-519
Hein, J.J., Schierup, M.H. and Wiuf, C.H. (2005) Gene Genealogies, Variation and Evolution. Oxford University Press, 296 pages.
Lauritzen, S. and Sheehan, N. (2003) "Graphical Models for Genetic Analyses" Statist. Sci. 18, no. 4. 489-514
Lander, E.S. and Green, P. (1987) "Construction of multi-locus genetic linkage maps in humans" PNAS 84: 2363-2367
Semple, C. and Steel, M (2003) "Phylogenetics" Oxford University Press