Reconstructing Pedigrees from Genetic Marker Data
Nuala A. Sheehan, James Cussens & Elinor M. Jones

Likelihood-based approaches
In theory, estimating the pedigree connecting a given set of individuals from genetic marker data is simple: just consider all possible pedigrees and compute the likelihood for each [5].

Due to the potentially enormous number of possibilities, brute force enumeration is only practical for small numbers of individuals [4].

‘Sequential’ (i.e. greedy) algorithms can be used which efficiently produce a single high likelihood reconstruction but not necessarily a maximum likelihood (ML) pedigree [6, 1].

For maximum likelihood reconstruction of reasonably large pedigrees, we propose:
• complete search using integer linear programming (ILP) [2] and state-of-the-art ILP optimisation solvers [3],
• for convenience, we view the problem as one of Bayesian network (BN) learning,
• as for other current approaches, we will assume:
  — a complete sample i.e. a parent of a given individual is either observed or is a pedigree founder
  — genotype marker data at unlinked loci
  — Hardy-Weinberg proportions for founder genotypes and Mendelian segregation from parents to offspring.

Learning Pedigrees
A pedigree is a directed acyclic graph (DAG) with well-defined structural constraints: each node has exactly two parent nodes (although either can be latent) and parents must be of opposite sexes.

A pedigree DAG where the nodes represent the genotypes of the pedigree members, rather than the individuals themselves is a Bayesian network (BN).

Under the above assumptions, the pedigree likelihood decomposes into a product of conditional probabilities each of which is a function of an individual node and its parent nodes (e.g. the usual Mendelian transmission probabilities of $\frac{3}{4}$ or 1 for an individual with two observed parents):

$$L(G) = \prod_{v \in V} \tau(v, Pa(v, G)).$$

(1)

Integer Linear Programming (ILP)
An integer linear program (ILP) is defined by:
1. a set of variables $X$, representing unknown quantities, some of which are restricted to be integer values;
2. an objective function of the form $\sum_{x \in X} c_x x$ where the coefficients $c_x$ are fixed constants, and
3. linear equations and inequalities putting joint constraints on the values the variables can take.

ILP optimisation problem: find an assignment of values to $X$ which maximises the objective function while respecting all constraints. (NP-hard when some variables are constrained to be integers.)

CHALLENGE: Find a suitable ILP encoding of the pedigree learning problem.
This can be done for the above assumptions.

References

Encoding the Pedigree Learning Problem
• For a sample of $n$ individuals, create $O(n^2)$ binary variables to represent all possible parentages:

$$I(W \rightarrow v)(G) = \begin{cases} 1 & \text{if } v \text{ has parents } W \text{ in } G \\ 0 & \text{otherwise} \end{cases}$$

where $W \subseteq V \setminus \{v\}$ and $|W| \leq 2$

subject to the constraint $\forall v : \sum_{W} I(W \rightarrow v)(G) = 1$ i.e. each individual has exactly one parent set.

• Extra constraints are required to rule out cycles and to ensure that sex can be consistently assigned.

• From (1), the optimisation problem is to maximise the log-likelihood which is now in the desired form:

$$\text{Maximise } \sum_{v,W} \log \tau(v, W) I(W \rightarrow v)(G) \text{ subject to all constraints}.$$

Reconstructing a 59-member pedigree

Conclusions and Future Work
• Simulated 1000 datasets using 10 marker loci each with 8 equifrequent alleles.
• 75% found the maximum likelihood pedigree within 0.4386 seconds: the longest run was 15.26 seconds.
• MUCH faster than the simulated annealing results of $> 6$ minutes for the same application in [1] and which are not guaranteed to find a maximum likelihood solution.
• Model uncertainty: ML estimation does not guarantee finding the ‘true’ pedigree but a maximally probable structure given the data.
• Prior information can be incorporated.
• We can find multiple distinct high (posterior) probability pedigrees and use Bayesian model averaging to account for model uncertainty.

ILP provides a hopeful improvement on alternative approaches to reconstructing pedigrees. The simple decomposition in (1) breaks down when markers are linked and some pedigree members are unobserved — extra variables required for unknown phase and missing information.

Possible — provided we can keep the log-likelihood function linear.