Improved Maximum Likelihood Reconstruction of Complex Multi-generational Pedigrees

Nuala A Sheehan^{a,*}, Mark Bartlett^b, James Cussens^b

^a*Department of Health Sciences, University of Leicester, UK*

^b*Department of Computer Science, University of York, UK*

Abstract

The reconstruction of pedigrees from genetic marker data is relevant to a wide range of applications. Likelihood-based approaches aim to find the pedigree structure that gives the highest probability to the observed data. Existing methods either entail an exhaustive search, and are hence restricted to small numbers of individuals, or they take a more heuristic approach and deliver a solution that will probably have high likelihood but is not guaranteed to be optimal. By encoding the pedigree learning problem as an integer linear program we can exploit efficient optimisation algorithms to construct pedigrees guaranteed to have maximal likelihood for the standard situation where we have complete marker data at unlinked loci and segregation of genes from parents to offspring is Mendelian. Previous work demonstrated efficient reconstruction of pedigrees of up to about 100 individuals. The modified method that we present here is not so restricted: we demonstrate its applicability with simulated data on a real human pedigree structure of over 1600 individuals. It also compares well with a very competitive approximate approach in terms of solving time and accuracy. In addition to identifying a maximum likelihood pedigree, we can obtain any number of pedigrees in decreasing order of likelihood. This is useful for assessing the uncertainty of a maximum likelihood solution and permits model averaging over high likelihood pedigrees when this would be appropriate. More importantly, when the solution is not unique, as will often be the case for large pedigrees, it enables investigation into the properties of maximum likelihood pedigree estimates

*Corresponding author: Department of Health Sciences, University of Leicester, 2nd Floor Adrian Building, University Road, Leicester LE1 7RH, UK

Email address: nas11@le.ac.uk (Nuala A Sheehan)

which has not been possible up to now. Crucially, we also have a means of assessing the behaviour of other approximate approaches which all aim to find a maximum likelihood solution. Our approach hence allows us to properly address the question of whether a reasonably high likelihood solution that is easy to obtain is practically as useful as a guaranteed maximum likelihood solution. The efficiency of our method on such large problems bodes well for extensions beyond the standard setting where some pedigree members may be latent, genotypes may be measured with error and markers may be linked.

Keywords: Constrained optimisation, integer linear program, Bayesian networks, genetic marker data

1. Introduction

Pedigree information is often required for applications in animal breeding, conservation research and forensic science. Recent trends in human genetics have been towards population-based sampling of unrelated individuals rather than of families for genome-wide association analyses. However, these population samples undoubtedly contain sets of relatives and there is a growing interest in estimating such relationships for various purposes including the study of rare genes and parent-of-origin effects. Analyses that assume all individuals are unrelated can be badly biased unless hidden relationships are excluded or appropriate adjustments made (Stankovich et al., 2005; Choi et al., 2009; Thornton and McPeck, 2010; Jankovic et al., 2010; Staples et al., 2013). Even in linkage studies which require pedigree data, undeclared relationships between pedigree founders and between individuals in different families can bias the results and need to be properly accounted for (Génin and Clerget-Darpoux, 1996; Sheehan and Egeland, 2008; Glazner and Thompson, 2012).

In a recent paper, Cussens et al. (2013) introduced an approach to maximum likelihood pedigree reconstruction using integer linear programming (ILP). Unlike other heuristic likelihood-based approaches (Almudevar, 2003; Riester et al., 2009; Cowell, 2013), the method returns a solution that is guaranteed to have maximal likelihood. Using simulated data, it outperforms other competing approaches for the relatively straightforward situation that is typically considered whereby all individuals have complete data at unlinked markers, founder genotypes are in Hardy-Weinberg equilibrium and transmission of genes from parents to offspring is Mendelian. It also

permits calculation of many high likelihood pedigrees and can hence address uncertainty about the accuracy of a reconstruction. While the alternative guaranteed maximal likelihood approach of Cowell (2009) is only capable of handling pedigrees of up to 30 individuals, simulation studies showed that pedigrees of up to 100 individuals are manageable by our approach and the true pedigree was generally found very quickly unless the pedigree structure was highly complex, as would be the case in animal or plant applications with high levels of inbreeding. Although these would be considered ‘large’ pedigrees in the reconstruction literature, the new software introduced in this paper is much more efficient and can reconstruct much bigger networks. We will illustrate the performance of the method on a large complex human pedigree of an Inuit population, previously published as ‘The Polar Eskimo Genealogy’ comprising 1614 individuals over 7 generations (Gilberg et al., 1978; Edwards, 1992). We argue that this considerable gain in computational efficiency is essential for extensions to more practical reconstruction problems.

The paper will begin with a brief review of the integer linear programming formulation of the reconstruction problem, noting the main differences between the new algorithm and its predecessor. We will omit the technical details here and refer to the machine learning literature for these. We then introduce our test pedigree structure and describe how we will assess a reconstruction of such a complex network. Finally, we present the results of our simulation study and conclude with a discussion of the strengths and weaknesses with indications for future work.

2. Materials and Methods

A pedigree, $\mathcal{G}$, can be described as a directed acyclic graph (DAG) (Lange and Elston, 1975) where each node $v \in V$ (pedigree member) has exactly two parent nodes, either of which may be latent. Graph nodes can be of two possible types, or sexes, and a pair of graph parents must be of different sex. An individual who has no parents in the pedigree is a pedigree *founder* and founders are unrelated to each other and to any non-descendants in the pedigree, by definition. Under the assumptions of complete genetic data at unlinked loci on every individual in the pedigree, Hardy Weinberg proportions of founder genotypes and simple Mendelian inheritance of genes from parents to offspring, the pedigree likelihood for a single marker can be written as a product of ‘local’ conditional probabilities. Specifically, if $\text{Pa}(v, \mathcal{G})$

denotes the parent set of v in $\mathcal{G}$ comprising 0, 1 or 2 members, the likelihood is given by

$$L(\mathcal{G}) = \prod_{v \in V} \tau(v, \text{Pa}(v, \mathcal{G})), \quad (1)$$

where $\tau(v, \text{Pa}(v, \mathcal{G}))$ denotes the conditional probability of the genotype of v given the genotypes of the parent set $\text{Pa}(v, \mathcal{G})$ (Thompson, 2000). Under the assumption of Mendelian segregation, these probabilities assume values of 0, $\frac{1}{4}$, $\frac{1}{2}$ or 1 for individuals with two parents in the pedigree. For parent sets of size 0 or 1, corresponding values can be derived using the population allele frequencies (Cussens et al., 2013).

Given the genetic marker data on a set of individuals, our aim is to find a pedigree graph that maximises the likelihood or, equivalently, the log-likelihood of these data. The above factorisation of the likelihood is the *directed local Markov property* that defines a Bayesian network (BN) (Lauritzen, 1996) and is the core of the ILP formulation. Maximising the pedigree likelihood is thus a BN optimisation problem using the fitted pedigree likelihood as the *score* (Cowell, 2009). The trick is to impose appropriate constraints that restrict the search to valid pedigrees.

2.1. Likelihood Optimisation using ILP

In the interests of readability, we provide a brief description of how to encode a maximum likelihood pedigree reconstruction as an ILP problem here. A more detailed discussion can be found in Cussens et al. (2013).

An integer linear program requires an objective function that is linear in a set of variables (at least some of which are integer-valued) representing unknown quantities of interest and aims to find a set of values for these variables that maximises the objective function subject to constraints. Since our aim is to find a pedigree that maximises the likelihood for observed marker data, the ILP variables must be able to specify any possible pedigree. Binary variables determining parentage assignments do precisely this since a pedigree is fully specified by the enumeration of all parent-offspring sets. Thus, for each individual $v \in V$, we define

$$I(W \rightarrow v)(\mathcal{G}) = \begin{cases} 1 & \text{if } v \text{ has parents } W \text{ in } \mathcal{G} \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $W \subseteq V \setminus \{v\}$ and $|W| \leq 2$. For any particular pedigree $\mathcal{G}$, the variable $I(W \rightarrow v)(\mathcal{G})$ only takes the value 1 when W is the selected parent set for

v (i.e. $W = \text{Pa}(v, \mathcal{G})$). Hence, the pedigree log-likelihood for a single marker can be expressed as a linear function of these binary variables with known coefficients derived from the segregation probabilities, $\tau()$. Specifically,

$$\begin{aligned} \log L(\mathcal{G}) &\equiv \sum_{v \in V} \log \tau(v, \text{Pa}(v, \mathcal{G})) \\ &= \sum_{v, W} (\log \tau(v, W)) [I(W \rightarrow v)(\mathcal{G})]. \end{aligned} \quad (3)$$

The corresponding log-likelihood for a set of independently segregating markers is just a sum of the relevant expressions given in Equation (3) so we will only discuss the method in terms of a single marker for simplicity.

A valid pedigree determines a joint instantiation of the parentage assignment variables by setting exactly $|V| \equiv n$ of them to one. However, there are many possible such assignments that do not constitute a pedigree. Optimising the pedigree likelihood is thus a problem of finding a set of values $I(W \rightarrow v)$ to maximise the objective function in Equation (3) subject to constraints that restrict the search to valid pedigrees.

In order to exploit efficient ILP solvers, the required constraints must be expressed as linear equations or inequalities. As shown previously (Cussens, 2010; Cussens et al., 2013), the first constraint ensuring that each individual can only have one parent set (of size 0, 1 or 2) is given by:

$$\forall v : \sum_W I(W \rightarrow v) = 1. \quad (4)$$

Further linear inequalities are defined so that sex can be assigned consistently since members of a parent pair should be of opposite sex. Note that we do not assume that the sex of each individual is known. We merely ensure that the pedigree structure is such that sex can be consistently assigned. However, if an individual's sex is known, it is easy to incorporate this information and so restrict the search to pedigrees where these known sexes are fixed.

Finally, constraints are required to restrict the search to acyclic graphs. In Cussens et al. (2013), this was achieved by adding $n(n-1)$ auxiliary variables for 'generation numbers' assigning a value of zero to each pedigree founder and a value greater than that of each parent to every non-founder. These constraints eliminate *directed cycles* since they ensure that no individual can be his own ancestor or descendant in the pedigree with maximal likelihood.

The main difference in the algorithm we use here is with respect to how cycles are excluded.

In this paper, cycles are ruled out using *cluster* constraints (Jaakkola et al., 2010) which do not use auxiliary variables. These are linear inequalities which are based on the observation that any subset (‘cluster’) of individuals $C \subseteq V$ must contain at least one individual who has no parents in C :

$$\forall C \subseteq V : \sum_{v \in C} \sum_{W: W \cap C = \emptyset} I(W \rightarrow v) \geq 1. \quad (5)$$

An important step in solving our ILP problem is to solve the *linear relaxation* of the problem where the restriction on the $I(W \rightarrow v)$ variables to integer values is removed. The linear relaxation can be solved quickly and provides a crucial upper bound for the maximum likelihood. This upper bound is far tighter when using cluster constraints than it is when using our earlier ‘generation number’ approach and leads to much faster solving of the ILP problem.

However, there are clearly exponentially many cluster constraints so it is impractical to add them all to the ILP. Instead they are added as *cutting planes* during the course of the solving process. The key to success with such an approach is (i) to only add those cluster constraints that are required (not all potential cycles need be ruled out) and (ii) to implement an efficient way of finding good cutting planes. Although cluster constraints are sufficient to ensure acyclicity, there are further constraints which can be added to improve the upper bound provided by the linear relaxation. Details of the various issues involved in the cutting plane approach as applied to general Bayesian network learning are provided in Cussens (2011) and Bartlett and Cussens (2013). The novelty in this paper is the adaptation of this approach to learning much larger networks and, specifically, to pedigree learning.

Finally, in order to take account of model uncertainty, we can find the k most likely pedigrees, for some arbitrary value k , by repeatedly solving the ILP problem and adding the following constraint

$$\sum_{v, W: I(W \rightarrow v)(G)=1} I(W \rightarrow v) \leq n - 1 \quad (6)$$

after each solution, G , has been found (Cussens et al., 2013). By ensuring that the next reconstruction differs in at least one parent set from all others

previously found, this prevents a previous solution from being returned again and will provide valid pedigrees in decreasing likelihood order.

2.2. The Pedigree and the Data

The *Polar Eskimo Population* has lived for centuries in the Thule district of North West Greenland and is believed to be the most northern human native population on Earth (Gilberg et al., 1978). The first record of contact with the outside world was in 1818 so the population prior to that can be regarded as a *breeding isolate* being genetically (and geographically) isolated from external gene sources. The pedigree in Figure 1 depicts 1614 descendants from this breeding isolate and ranges over 7 generations—excluding the baseline founder generation—from a birth in 1805 to one in 1975. There are 225 founders in total. The pedigree is not completely connected and comprises 11 components or sub-pedigrees. The largest of these contains most of the population with 1581 individuals, the next one has 11 members and the others all have fewer than 6.

The structure of this population has long been of interest for demographic and population genetic research, in particular with regard to studies of survival in the high Arctic (Gilberg et al., 1978; Edwards, 1992). Notably, despite the size and isolation of the population, there are remarkably low levels of inbreeding: the highest inbreeding coefficient is $\frac{1}{8}$ corresponding to a marriage between one individual and his niece and there are only 5 first cousin marriages, one of which is actually via half-siblings rather than full siblings. What makes the structure complex, is the existence of multiple marriages and marriage chains, whereby individuals are inter-related by marriage in many different ways. An example of this complexity is presented in Figure 2 which shows two marriage chains comprising 5 and 8 individuals, respectively, who are involved in 11 marriages amongst themselves and 12 additional marriages with individuals who are not in either chain. Three brother-sister pairs are involved and the two chains are connected via two of these pairings (Edwards, 1992). In all, there are 563 individuals with a single marriage, 129 with 2 marriages, 21 with 3 marriages and 8 individuals with four marriages in the pedigree.

2.3. Assessing the Reconstruction

In order to assess how well the method performs on a large complex structure, we simulated *complete* data on the 1614 individuals in Figure 1 using the original Caucasian allele frequencies for 15 autosomal short tandem

Figure 1: A marriage node graph of the ‘Polar Eskimo Genealogy’, taken from Sheehan (1992), depicting 7 generations of 1614 individuals descended from a breeding isolate in North West Greenland.

repeat loci (the 13 forensic CODIS core loci plus D19S433 and D2S1338) corresponding to Table 1 of Butler et al. (2003). The number of alleles at each marker ranges from 7 to 15. Unrealistic allele frequencies greatly simplified the solving problem in our earlier work (Cussens et al., 2013), so we prefer to focus on realistic frequencies here. Specifically, for each simulated dataset, we assigned genetic profiles to the pedigree founders from these allele

Figure 2: Two marriage chains, adapted from Edwards (1992), comprising 13 individuals (shaded) involved in 11 marriages between themselves and 12 other marriages with individuals who are not in either chain. Only one child from each marriage is depicted.

frequencies assuming random union of gametes (and hence Hardy-Weinberg equilibrium) at each marker independently and Mendelian transmission of genes from parents to offspring. We note that the assumption of Mendelian segregation is crucial to the required decomposition of the pedigree likelihood in (1) (Lauritzen and Sheehan, 2003).

Given the size of this pedigree, accurate recovery of every single edge (parent-child relationship) is often an unreasonable expectation and capturing particular features of the structure may be sufficient. We will categorise such features into two groups: *local* features, which include sibship sizes, numbers of marriages for specific individuals and particular relationships amongst a set of individuals, and *global* features which include the number of pedigree components, number of founders (pedigree *width*), number of generations (pedigree *depth*) and number of pedigree descendants for a given individual. We define the generation of an individual to be the longest distance (in terms of directed edges) to an ancestor who is a founder. As examples of local features we used the known uncle-niece marriage mentioned earlier, the linked marriage chain structure in Figure 2, an individual whom we have labelled as 141 who had 10 children from 2 marriages, and the overall distribution of marriages (i.e. 0 to 4) in the pedigree. For global features, we used the number of components and numbers of individuals within the main component, the number of founders, the number of generations and an individual (labelled 49) known to have 4 marriages and 452 descendants in the pedigree (Edwards, 1992). For completeness, and also to be consistent with the BN learning literature, we look at the number of edges but instead

of simply counting the number we get right (there are 2778 directed edges in the true pedigree), we consider the *recall* and *precision*. Specifically, *recall* is defined as the proportion of the 2778 real edges that were found in the reconstruction. *Precision* is defined as the proportion of all edges in the reconstruction that are actually real ones. High precision implies that there are more correct than incorrect edges in the reconstructed graph whereas high recall means that most of the correct edges have been found. Incorrect edges are defined as edges in the reconstruction that do not exist in the real pedigree, edges in the real pedigree that are missing in the reconstruction, or edges in the reconstruction that are in the right place but have the direction reversed. We note that recall and precision are equivalent to the notions of sensitivity and positive predictive value in diagnostic testing. We also note that our algorithm’s task is to obtain a maximum likelihood pedigree. This may, or may not resemble the true pedigree since genetic inheritance is probabilistic. Thus, the closeness of a reconstruction to the true pedigree is not necessarily an appropriate evaluation of the success or efficiency of a likelihood maximisation approach and the two issues should not be confused.

All results were produced on a single core of a normal desktop computer (2.7GHz dual-core processor with 6GB of memory running Linux) with version 1.3.1 of the software (freely available at <http://www.cs.york.ac.uk/aig/sw/gobnilp>) which uses the SCIP 3.01 framework for constraint integer programming and CPLEX 12.5 as the underlying linear programming solver. All reported times omit the preprocessing time taken to produce the local log likelihood scores ($\log \tau()$). The time for computation of these scores was of the order of 3 seconds per dataset.

3. Results

The cumulative distribution of solving times for 100 simulated data sets, generated as described in Section 2.3 is shown in Figure 3. Reconstructions took about 10 minutes on average and ranged between 3 and 42 minutes. We should note that the earlier version of our algorithm, using generation numbers instead of cluster-based constraints to rule out directed cycles, had solving times of between 3 and 100 *hours* for the same data so this is a big gain in efficiency.

Figure 3: Cumulative distribution of solving times for a single maximum likelihood reconstruction from 100 simulated data sets on the 1614-member Eskimo pedigree of Figure 1. The dotted lines depict the lower quartile, median and upper quartile, respectively, and indicate that three quarters the reconstructions were found within 12 minutes, for example.

3.1. How Good is the Reconstruction?

The variability amongst these reconstructions from the 100 different data sets can be assessed by considering the extent to which the various local and global features discussed in Section 2.3 are captured. It should be stressed that measuring the quality of these reconstructions assesses the performance of maximum likelihood based pedigree reconstruction in general and not just the specific software used. As the software finds a guaranteed maximum likelihood pedigree, one would observe similar findings using any other technique that was capable of obtaining a maximum likelihood pedigree.

Starting with the local features, 87% of the reconstructions found the known uncle-niece marriage. The marriage chain structure in Figure 2 was also recovered with reasonable accuracy: 78% of the maximum likelihood pedigrees correctly identified the three pairs of siblings and 64% identified all 23 marriages. The number of marriages ranged from 19 to 23 with an aver-

	Number of Marriages				
	1	2	3	4	> 4
True Pedigree	563	129	21	8	0
Mean Number	554.4	159.9	35.7	10.5	0.7
Percent Exact Match	3	0	0	9	51

Table 1: Distribution of the number of marriages, 1, 2, 3, 4 or > 4 averaged across the 100 maximum likelihood reconstructions from different simulated data sets. The first row gives the distribution in the true pedigree. The bottom row indicates the percentage of reconstructions that had the same number as the true pedigree for each category.

age of 22.6 so no reconstruction fared too badly. 65% of the time, individual 141 indeed had 10 children with an average of 9.7 across the hundred reconstructions. The distribution of marriage numbers, averaged over the 100 optimal pedigrees, is shown in Table 1 along with the proportion of pedigrees with exactly the right number of marriages in each category. There is a tendency to over-estimate the number of multiple marriages but, although very few reconstructions had exactly the right numbers in each category, the overall distribution seems quite reasonable. We note that the median values were very similar to the means here and so are not reported. Finally, the 4 marriages of individual 49 were correctly detected 87% of the time and ranged from 3 to 5 with a median of 4.

With regard to the global features, the number of components was always underestimated to be about 7 on average, with a single maximum value equal to the true value of 11. The size of the largest component was over-estimated to be about 1600 on average, with a single minimum at the true value of 1581. The number of founders was also consistently under-estimated to be about 170 on average: however, the identified founders usually were real founders in the true pedigree. The number of generations was badly over-estimated, ranging from 24 to 70 and so was the number of descendants of individual 49 with a mean of 1006 and median of 1059, by comparison with the true value of 452. The number of reconstructed edges was also over-estimated ranging from 2810 to 2844 but the recall was 96.8 % and the precision was 95.2 %, which would seem to be reasonably high values in the absence of any specific knowledge about which edges are wrong.

Essentially, in the absence of all information other than the marker data, a maximum likelihood approach will endeavour to create as many relatives

as possible in order to optimise the likelihood, hence the lower numbers of founders, more generations and more multiple marriages. For local features — especially those ‘within’ the pedigree — however, maximum likelihood appears to be doing surprisingly well without any additional information.

3.2. Using Prior Information

The pedigree file on which we simulated data was formatted as a standard ‘triplet file’ with a row for each individual in the pedigree and three columns giving the individual’s identifier followed by the identifiers for both parents. Founders have no parents so the labels in the second and third columns are both 0 to reflect this. When the pedigree data were originally collected, a relatively small number of individuals with only one known parent were assigned a newly created founder as second parent (Edwards, 1992). It was acknowledged that this could lead to a slightly conservative estimate of the overall inter-relatedness. However, the implication for our purposes is that we happen to ‘know’ some extra information about the possible parent sets. Specifically, we can ‘inform’ the ILP solver that the parent sets W in Equation 2 are always of size 0 or 2 and so parent sets of size 1 do not have to be considered. In practice, there will always be some sort of ‘prior’ information available. The nature of this information will depend on the application. While we would not always expect to be able to rule out single parent sets, it serves as an example of how additional information can affect the reconstruction process.

The practical result of using this information is that the number of potential parent sets that must be considered for each individual is vastly reduced. When single parents are allowed, there is an average of 7.3 parent sets per individual, with the majority of individuals permitted 6 or more parent sets. The restriction to parent sets of size 0 or 2 reduces this greatly to 2.1 on average, with 95% of individuals having 3 or fewer potential parent sets. Given that all individuals may have the empty parent set and that non-founders also have a real non-empty parent set, this means the number of incorrect potential parent sets is vastly pruned.

Incorporating this information led to a big reduction in solving times which are now in the order of seconds rather than minutes. Specifically, the solving times for 100 simulated data sets averaged at about 39 seconds and ranged from 7 seconds to about 4 minutes. With this gain in efficiency, the k most likely pedigrees can be obtained in very reasonable time so we generated the first 100 pedigrees for each of these 100 simulated data sets.

Figure 4: Mean and median solving times (in seconds) for the 100 most likely pedigrees from 100 sets of simulated data on the 1614-member pedigree of Figure 1. The vertical bars depict inter-quartile ranges.

Mean and median solving times for finding the k^{th} pedigree across all 100 data sets are shown in Figure 4. Interquartile ranges are shown at every 10^{th} pedigree to indicate the range of solving times. We can see from the figure that the maximum value of 4 minutes was an exceptionally high solving time as most pedigrees are found in less than a minute. The first pedigree takes the longest to find and the median solving times for subsequent pedigrees are generally around 2 seconds with occasional departures. We do not see an obvious trend of increasing solving times with each new pedigree found as reported previously (Cussens et al., 2013). This is because our new solver retains information about sub-optimal pedigrees found while seeking the first one and hence has better starting points from which to search for subsequent pedigrees.

The first 100 reconstructed pedigrees for each simulated data set had exactly the same likelihood — albeit a different likelihood for each data set

— so they are all optimal pedigrees. This is because the graph is so large that there are many possible small changes that do not affect the overall likelihood. The true pedigree was never found as the true pedigree was not a maximum likelihood pedigree in any of the datasets. We considered all 10,000 reconstructed pedigrees to evaluate performance in terms of recapturing the local and global features of Section 2.3. Locally, the maximum likelihood pedigrees were quite similar to those in the previous analysis which also allowed for single parent sets. They were slightly better at recovering the marriage chain structure in Figure 2 with 80.6% of the pedigrees featuring all three pairs of siblings and 72.2% correctly identified all 23 marriages. The distribution of marriage numbers was also very similar to that in Table 1. They were quite a bit more successful in detecting the number of marriages for individual 49 with about 97% of the pedigrees getting it exactly right. The global features were all much better, however. On average, the right number of pedigree components was found, the number ranged between 10 and 12 and 92% of the reconstructions found exactly 11. The number of individuals in the main component was 1581.1 on average and 92% had the exact number of 1581. The mean (and median) number of founders was the true value of 225: 99% of the pedigrees had the right number and 92% had correctly identified all 225 real founders. The number of descendants of 49 now averaged at 456.8 with a median of 453 and ranged from 335 to 549. Although only 23% of the optimal pedigrees had exactly the true value of 452, this is a big improvement on the performance where single parent sets could also be considered. The number of generations was still over-estimated taking its minimum at the true value of 7 and getting as high as 18 but with a median of 9 and mean of 9.8, it is much more reasonable. 99% of the reconstructed pedigrees had the correct number of 2778 edges. They were not always the true edges, however, as the recall and precision were both 97.3%. Since most of the pedigrees had the same number of edges as the real pedigree, the existence of an incorrect edge implies that a correct one was omitted and so the two measures are equal.

3.3. Adding Extra Marker Data

Although 15 microsatellite markers are generally sufficient to solve most forensic identification problems for ‘non-deficient’ pedigrees, this is a far larger network to reconstruct. Hence, the extent to which additional marker data affects the reconstruction is of interest, particularly by comparison with extra non-genetic data such as that presented in the previous section. To

this end, we performed the simulations again to obtain a single reconstruction from 100 datasets using twice as many loci and without the limitation that individuals must be founders or have both parents. To avoid introducing unnecessary differences between settings, the 15 additional markers introduced had the same allele frequencies as those used previously.

The increased number of markers served to rule out some potential parent-child relationships, reducing the average number of parent sets per individual from 7.3 to 5.4. Reconstruction times also decreased significantly, taking less than a minute in all cases. The improved timing behaviour is expected, as the runtime of GOBNILP is strongly correlated with the number of possible parent sets in the problem (Malone et al., 2014).

The similarity of the reconstructed pedigree to the true one was also markedly improved. For 77 of the 100 simulated datasets, the reconstructed maximum likelihood pedigree was the true pedigree. In a further 18 cases, a single individual had a true parent removed and an incorrect one added, while in the final 5 cases, two individuals had an incorrect parent assigned instead of their true one. Thus, none of these incorrect edges corresponded to a reversal of a true edge. The local and global features were all correct in every case except for the number of marriages, which was affected by the incorrect parentage assignments, and there was a single dataset for which an additional generation was reported.

The high degree of accuracy achieved in reconstructing these datasets stems from the fact that 30 markers were sufficient to rule out almost all possible parent sets except those corresponding to (possibly reversed) edges in the true pedigree. In fact, each of these datasets permitted an average of only 228 parent sets (0.14 per individual) that were not formed from true parents, reversed true parent-child edges, or the empty set. The reconstruction task was thus almost reduced to simply choosing orientations of these true edges. The method proved very successful at this.

3.4. Comparing Different Maximum Likelihood Pedigrees

Returning now to the case of using only 15 markers and the results of section 3.2, we find that, in contrast to our previous findings for smaller pedigrees (Cussens et al., 2013), the true pedigree did not feature at any point among the first 100 equally optimal reconstructions and always had a likelihood lower than the maximum value. Indeed, longer runs (data not shown) revealed that the likelihood did not decrease after finding up to 400

Figure 5: Median values of F-scores comparing each of the 100 most likely pedigrees with the first pedigree found across the 100 simulated data sets. The vertical bars depict inter-quartile ranges.

pedigrees. We note that it is not only possible to have a non-unique maximum likelihood solution, but the number of such solutions will depend on various factors. In particular, it will decrease with increasing amounts of information available and increase with increasing size of the structure. When the parent sets W in equation (2) were allowed to range over all three values 0, 1 and 2, the ‘best’ pedigrees always had higher likelihood than those obtained when single-parent sets were excluded. However, the ‘best’ pedigrees found by adding this constraint, or by adding extra marker data, were far more accurate reconstructions of the true pedigree reflecting the importance of providing additional information that allowed us to exclude incorrect parent sets (Thompson, 1986; Cussens et al., 2013).

We saw in Section 3.1 that there is quite a lot of variability among the maximum likelihood pedigrees estimated from different sets of data. It is also of interest to consider how the 100 equally likely pedigrees reconstructed from

Figure 6: Median values of F-scores comparing each of the 100 most likely pedigrees with (a) the real pedigree and (b) the previous pedigree found across the 100 simulated data sets. The vertical bars depict inter-quartile ranges.

the *same* set of data differ from each other. A single measure of performance in terms of precision and recall is given by the *F-score* defined as:

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \quad (7)$$

The F-score is the harmonic average of recall and precision. It assumes values between 0 and 1 and is equal to p when precision and recall are both equal to p . Using 15 markers and allowing parents sets of size 0 or 2, there was only one simulated dataset which did not yield 100 maximum likelihood reconstructions all with the same number of edges as the true pedigree. For each simulated dataset the precision and recall never differed by more than 0.1% for those 100 maximum likelihood reconstructions. Hence, the F-score here is more or less equal to the precision and to the recall. In order to assess how subsequent solutions differ from the first one found, the F-score comparing each pedigree with the first pedigree is calculated. Figure 5 shows medians (and interquartile ranges) for these scores across the 100 different data sets and indicates that later maximum likelihood reconstructions are not more distant from the initial solution than earlier ones by this measure. There is a definite oscillating pattern in the graph which could be due to the fact that the solutions all differ by varying a few edges at a time. Figure 6 shows analogous plots comparing each pedigree with (a) the true pedigree and (b) the previous (i.e. $(k - 1)^{th}$) pedigree found. These show that (a) the reconstructions do not deteriorate in accuracy as k increases, which is reassuring as they are all equally likely, and (b) each reconstruction is not more similar to the one found previously than it is to the first one, for example.

With regard to the local and global features of Section 2.3, the number of edges and the number of founders did not vary among the 100 maximum likelihood pedigrees for any given data set. However, the distribution of marriages within the pedigree, the number of generations and the number of descendants of individual 49 all tended to vary within the different reconstructions. Local features such as the number of children of individual 141, the number of partners for 49, the three sets of siblings and number of marriages in the marriage chain structure of Figure 2 appeared to be more stable and, when they did differ, they tended to oscillate between close values.

3.5. Comparison With An Approximate Approach

FRANz is a software package for maximum likelihood pedigree reconstruction that can handle very large pedigrees (Riester et al., 2009) and is a strong competitor for our ILP approach. In contrast to our approach, FRANz can deal with missing population members. In the complete data case, such as we have here, the reconstruction is performed using the simulated annealing approach of Almudevar (2003) so it cannot be guaranteed to find the actual maximum likelihood pedigree but it is very quick. For

comparison with our software (GOBNILP), we tested FRANz on the Eskimo pedigree reconstruction using the same 100 simulated data sets of Section 3.1 (i.e. with 15 markers and *no* additional information) using default values for all parameters except for the mutation rate which was set to 0. Solving times, together with those obtained previously using GOBNILP, are shown in Figure 7. As previously, reported times do not include time initially spent calculating the local likelihoods.

Figure 7: Solving times using FRANz and GOBNILP for a single maximum likelihood reconstruction from 100 simulated data sets on the 1614-member Eskimo pedigree of Figure 1.

The median solving time is slightly faster (by 28 seconds) for GOBNILP but the mean solving time is faster (by 65 seconds) for FRANz due to the substantially lower variability in run times in the latter case. The FRANz reconstructions usually had higher (log)likelihood than the true pedigree although they were actually worse on three occasions. This is at variance with the reported results in the original paper (Riester et al., 2009) and is presumably due to the fact that our pedigree is not a simple ‘out-breeding’ pedigree. They always had lower likelihood than the GOBNILP reconstructions indicating that, on this occasion, FRANz never found any of the maximum likelihood pedigrees. In terms of solving times, the ILP approach thus competes well with simulated annealing. Additionally, it guarantees a maximum

likelihood reconstruction and can deliver the k ‘best’ pedigrees.

Assessment of how well the 100 reconstructed pedigrees found by FRANz resemble the true one was conducted using the measures given in Section 2.3 and compared with the GOBNILP performance in Section 3.1 where no additional information was supplied. Specifically, a single optimal reconstruction was obtained from each method on the same 100 simulated data sets, with 15 markers and all parents sets of 0, 1 and 2 were considered. These results are summarised in Table 2.

In terms of recovering the global features, FRANz tended to be slightly worse in general, but on only one aspect was the difference statistically significant. The number of founders, and the number of those which were truly founders, was notably lower and even further from the true value of 225. The average pedigree depth, average number of components, size of the largest component and number of descendants of individual 49 were equally poor for both approaches.

Performance of FRANz with regard to the local features was also slightly inferior to that of GOBNILP. The proportion of reconstructions picking up the uncle-niece marriage was lower at 80%. There was a slight, but significant, fall in the average number of children assigned to individual 141, from 9.7 for the maximum likelihood approach to 9.3, but the number of reconstructions in which the correct number occurred fell notably, from 65% of reconstructions to 54%.

The average number of sibships and marriages detected in the marriage chain was virtually unchanged on average, but the three sibships in the marriage chain were only recovered 74% of the time by FRANz with only 58% of reconstructions featuring all 23 marriages. The number of marriages involving individual 49 was almost exactly the same on average and the overall distribution of the number of marriages was not particularly different from that found for the GOBNILP reconstructions. At the overall pedigree level, FRANz finds fewer edges and has both lower recall and precision.

In this case at least it appears that, aside from being theoretically interesting, finding an actual maximum likelihood pedigree is also of practical importance. This is particularly so when specific features rather than simply the overall ‘shape’ of the pedigree may be of interest.

Feature	GOBNILP	FRANz	p
Number of founders	171.0	159.1	< 0.001
Number of correct founders	171.0	158.3	< 0.001
Number of components	7.1	7.4	0.238
Size of largest component	1599.5	1597.5	0.066
Number of mates of 49	3.9	3.8	0.245
Number of descendants of 49	1005.9	962.0	0.194
Number of generations	47.8	50.8	0.164
Individuals with 1 marriage	554.4	554.8	0.994
Individuals with 2 marriages	159.7	157.2	0.055
Individuals with 3 marriages	35.7	35.4	0.684
Individuals with 4 marriages	10.6	10.3	0.372
Individuals with over 4 marriages	0.7	0.8	0.597
Uncle and niece marriage	87%	80%	0.037
Number of children of 141	9.7	9.3	0.008
Sibships found in marriage chains	2.8	2.7	0.454
Marriages in marriage chains	22.6	22.5	0.304
Number of edges	2826.4	2815.2	< 0.001
Recall	96.8%	95.4%	< 0.001
Precision	95.2%	94.1%	< 0.001

Table 2: Comparison of most likely pedigrees found by GOBNILP and FRANz. See Section 2.3 for a full description of each of the criteria. All quantities are averages over 100 datasets except that for ‘Uncle and Niece Marriage’ which is the proportion of the datasets in which the relationship was found. p values are calculated using a 2-sided Mann-Whitney U test (with continuity correction where values are integral) except for ‘Uncle and Niece Marriage’ which uses a chi-squared test. No correction for multiple comparisons is made.

4. Discussion

For the first time, we have a method that can deliver guaranteed, rather than approximate, maximum likelihood reconstructions of large complex pedigrees efficiently for the standard situation where we have a complete sample, unlinked markers and Mendelian transmission of genes from parents to offspring. As a result, the properties of such estimates can now be thoroughly investigated. Furthermore, the performance of approximate approaches which aim to maximize the likelihood can be properly evaluated and we can begin to address the often debated issue of whether we actually need

a maximal likelihood solution to any particular problem. To our knowledge, this is the first time a guaranteed maximum likelihood reconstruction from genetic marker data has been obtained for such a large Bayesian network and, in particular, a pedigree. It would seem in this case that there is an advantage in being able to obtain a truly optimal solution rather than a high likelihood one. Unlike previous work for smaller pedigrees (Cussens et al., 2013; Cowell, 2013), the true pedigree is not found quickly with 15 microsatellite markers and does not typically feature in the top k pedigrees. This is because in this case the true pedigree will generally not have maximum likelihood.

When problems of the size we have presented here are being considered, it is really important to know the purpose of the reconstruction: there could be an enormous number of maximum likelihood solutions and with a modest number of markers some additional or ‘prior’ information to guide the search towards the true structure will be required. As an example, we showed that information restricting parent sets to be of size 0 or 2 yielded a marked improvement but other application specific information such as the number of generations, mating patterns, partial age orderings, sex information and known parent-offspring edges, which are usually available, can and should be incorporated. When all pedigree members are observed, we have shown that 30 microsatellite markers are adequately informative for a maximum likelihood reconstruction. This will not be the case when we have missing individuals and more markers will definitely be required.

In order to be of practical use, a reconstruction algorithm will need to be able to handle missing data, at the very least, and ultimately linked markers. Some individuals may only have partial information and, more importantly, certain individuals who are essential for defining relationships amongst some of the observed individuals may be completely missing. A big pedigree like the Polar Eskimo genealogy will typically only have data on the living members of the population so it is perhaps not realistic to think about reconstructing a vast pedigree from marker data observed on a small fraction of the population. However, even for more reasonably sized pedigrees, there will be many possible solutions to explore and the question of how many generations back we should go has to be addressed. Moreover, the pedigree likelihood calculations become intractable when there are too many missing individuals: the expression in (1) now has to be summed over all possible genotypes of the unobserved individuals in any particular reconstruction. The complex inter-relationships by marriage in the Polar Eskimo pedigree create large undirected cycles or *loops* in the graph which require

conditioning on very large ‘cutsets’ of individuals in the ‘peeling’ algorithms making exact likelihood calculation impossible even for a single diallelic locus when only the living individuals are observed. Approximations to the likelihood have to be considered and it is no longer possible to guarantee that the returned reconstructions have maximal likelihood. Furthermore single-site updating MCMC methods may not define an irreducible Markov chain in the above case where we have no mutation or genotyping error (Sheehan, 2000). In principle, our ILP approach can accommodate missing individuals by adding additional variables during the solving process, using a technique (‘column generation’) similar to that used to add additional cluster constraints (5) during solving. However, the extension from the fixed sample size considered here is non-trivial and is an important issue for future work. The consideration of missing data — even on relatively small problems with tight restrictions on what patterns of missing data are allowed — greatly increases the size and computational complexity of the reconstruction problem.

Our method is not restricted to unlinked markers, such as we have used here. This is important, given the dense sets of single nucleotide polymorphisms (SNPs) that are now routinely available. It has already been noted that the simple decomposition of the likelihood in (1) breaks down when inheritance of genes in offspring from parents is not Mendelian since offspring genotypes are no longer conditionally independent given the parental types (Lauritzen and Sheehan, 2003). The same is true, and by a very similar argument, for linked markers. In both cases, knowing the genotype of one offspring provides information on either the segregation law or the parental phase thus affecting the genotype probabilities for subsequent offspring (Cussens et al., 2013). Additional variables for parental phase and meiosis indicators can be added to the solver, just as above. This is equivalent to using a more complete BN description of a pedigree (Lauritzen and Sheehan, 2003). Likelihood computations with linked markers are more complicated as we cannot perform the calculations for one locus at a time, as we did here, and exact calculation now requires the Lander-Green algorithm (Lander and Green, 1987) which is restricted to quite small pedigree structures. Again, various likelihood approximations need to be considered and compared for sensible application to linked marker data.

Since we have been dealing with simulated data throughout in order to test the efficiency of the ILP approach to pedigree reconstruction, we have ignored mutation and genotyping error. This will of course not be reasonable when considering real data. Microsatellite markers have quite a high

mutation rate (of about 10^{-3}) so ignoring this over a real multi-generational pedigree would be incorrect. Genotyping error is not so much an issue for these data. There is no problem at all in incorporating a (germline) mutation model, such as the 1-step mutation model described in Dawid et al. (2002). The effect will be to greatly increase the number of candidate parent sets to be considered and hence increase the solving times. Somatic mutations are not inherited and could look more like genotyping errors. SNPs have a much lower mutation rate (of about 10^{-9}) so it is not unreasonable to ignore mutation here. However, the genotyping error is not ignorable and, being sporadic, will affect all pedigree members and not just descendants of a germline mutation carrier. The model is hence easier as it can be simply a matter of adding noise to the data but the number of candidate parent sets to be considered will be even larger.

Finally, we used the ‘known’ population allele frequencies throughout for our likelihood calculations here so the only uncertainty is the pedigree structure. This will not be the case in practice and frequencies often have to be estimated from the observed individuals. Depending on how many individuals are observed, and how inter-related they are, we can expect such frequency estimates to have an adverse effect on performance. We note that this is a common problem for analyses on population isolates, in general, but we now have a method that could allow us to investigate these effects.

Overall, the fact that our algorithm can tackle such a large problem in the straightforward case with very acceptable solving times bodes well for extensions to problems where the complexity is not simply due to size and structure but to different patterns of missing data and correlated observations.

Acknowledgements

We are grateful to three anonymous reviewers and the editor for useful criticisms of an earlier version of this paper. The authors acknowledge support from the Medical Research Council (Project Grant G1002312).

Almudevar, A., 2003. A simulated annealing algorithm for maximum likelihood pedigree reconstruction. *Theoretical Population Biology* 63, 63–75.

Bartlett, M., Cussens, J., 2013. Advances in Bayesian network learning using Integer Programming. In: Nicholson, A., Smyth, P. (Eds.), *Proceedings of*

- the 29th Conference on Uncertainty in Artificial Intelligence (UAI 2013). AUAI Press, pp. 182–191.
- Butler, J. M., Schoske, R., Vallone, P. M., Redman, J. W., Kline, M. C., 2003. Allele frequencies for 15 autosomal STR loci on U.S. Caucasian, African American, and Hispanic populations. *Journal of Forensic Sciences* 48 (4).
- Choi, Y., Wijsman, E. M., Weir, B. S., 2009. Case-control testing in the presence of unknown relationships. *Genetic Epidemiology* 33, 668–678.
- Cowell, R. G., 2009. Efficient maximum likelihood pedigree reconstruction. *Theoretical Population Biology* 76 (4), 285–291.
- Cowell, R. G., 2013. A simple greedy algorithm for reconstructing pedigrees. *Theoretical Population Biology* 83, 55 – 63.
- Cussens, J., 2010. Maximum likelihood pedigree reconstruction using integer programming. In: *Workshop on Constraint Based Methods for Bioinformatics (WCB-10)*, Edinburgh.
- Cussens, J., 2011. Bayesian network learning with cutting planes. In: *Cozman, F. G., Pfeffer, A. (Eds.), Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI 2011)*. AUAI Press, pp. 153–160.
- Cussens, J., Bartlett, M., Jones, E. M., Sheehan, N. A., 2013. Maximum likelihood pedigree reconstruction using integer linear programming. *Genetic Epidemiology* 37, 69–83.
- Dawid, A. P., Mortera, J., Pascali, V. L., van Boxel, D., 2002. Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics* 29, 577–595.
- Edwards, A. W. F., 1992. The structure of the Polar Eskimo genealogy. *Human Heredity* 42, 242–252.
- Génin, E., Clerget-Darpoux, F., 1996. Consanguinity and the sib-pair method: an approach using identity by descent between and within individuals. *American Journal of Human Genetics* 59, 1149–1162.

- Gilberg, A., Gilberg, L., Gilberg, R., Holm, M., 1978. Polar Eskimo Genealogy. Vol. 203 Nr.4 of *Meddelelser om Gronland*. Nyt Nordisk Forlag Arnold Busck, Kobenhavn.
- Glazner, C., Thompson, E. A., 2012. Improving pedigree-based linkage analysis by estimating co-ancestry among families. *Statistical Applications in Genetics and Molecular Biology* 11, issue 2, Article 11.
- Jaakkola, T., Sontag, D., Globerson, A., Meila, M., 2010. Learning Bayesian network structures using lp relaxations. In: *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS 2010)*. Vol. 9. *Journal of Machine Learning Research Workshop and Conference Proceedings*, pp. 358–365.
- Jankovic, A., vonHoldt, B. M., Rosenberg, N. A., 2010. Heterozygosity of the Yellowstone wolves. *Molecular Ecology* 19, 3246–3249.
- Lander, E. S., Green, P., 1987. Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences (USA)* 84, 2363–2367.
- Lange, K., Elston, R. C., 1975. Extensions to pedigree analysis. I. Likelihood calculations for simple and complex pedigrees. *Human Heredity* 25, 95–105.
- Lauritzen, S. L., 1996. *Graphical Models*. Clarendon Press, Oxford, United Kingdom.
- Lauritzen, S. L., Sheehan, N., 2003. Graphical models for genetic analyses. *Statistical Science* 18, 489–514.
- Malone, B., Kangas, K., Järvisalo, M., Koivisto, M., Myllymäki, P., 2014. Predicting the hardness of learning Bayesian networks. In: *Proceedings of the 28th AAAI Conference on Artificial Intelligence (AAAI 2014)*. AAAI Press.
- Riester, M., Stadler, P. F., Klemm, K., 2009. FRANz: reconstruction of wild multi-generation pedigrees. *Bioinformatics* 25, 2134–2139.
- Sheehan, N., 1992. Sampling genotypes on complex pedigrees with phenotypic constraints: the origin of the B allele among the Polar Eskimos. *IMA Journal of Mathematics Applied in Medicine and Biology* 9, 1–18.

- Sheehan, N. A., 2000. On the application of Markov chain Monte Carlo methods to genetic analyses on complex pedigrees. *International Statistical Review* 68, 83–110.
- Sheehan, N. A., Egeland, T., 2008. Adjusting for founder relatedness in a linkage analysis using prior information. *Human Heredity* 65 (4), 221–231.
- Stankovich, J., Bahlo, M., Rubio, J. P., Wilkinson, C. R., Thomson, R., Banks, A., Ring, M., Foote, S. J., Speed, T. P., 2005. Identifying nineteenth century links from genotypes. *Human Genetics* 117, 188–199.
- Staples, J., Nickerson, D. A., Below, J. E., 2013. Utilizing graph theory to select the largest set of unrelated individuals for genetic analyses. *Genetic Epidemiology* 37, 136–141.
- Thompson, E. A., 1986. *Pedigree Analysis in Human Genetics*. The Johns Hopkins University Press, Baltimore.
- Thompson, E. A., 2000. *Statistical Inference from Genetic Data on Pedigrees*. Vol. 6 of NSF-CBMS Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics and the American Statistical Association, Beachwood, Ohio, USA.
- Thornton, T., McPeck, M. S., 2010. ROADTRIPS: Case-control association testing with partially or completely unknown population and pedigree structure. *American Journal of Human Genetics* 86, 172–184.