Structured Incorporation of Prior Information in Relationship Identification Problems

N.A. SHEEHAN

Departments of Health Sciences and Genetics

University of Leicester

2nd Floor Adrian Building

University Road

Leicester LE1 7RH, UK

Fax: +44 (0) 116 22297250

E-mail: nas11@le.ac.uk

T. EGELAND

Department of Medical Genetics

Ullevål University Hospital

0407 Oslo, Norway

Summary

The objective of this paper is to show how various sources of information can be modelled and integrated to address relationship identification problems. Applications come from areas as diverse as evolution and conservation research, genealogical research in human, plant and animal populations, and forensic problems including paternity cases, identification following disasters, family reunion and immigration issues. We propose assigning a prior probability distribution to the sample space of pedigrees, calculating the likelihood based on DNA data using available software and posterior probabilities using Bayes' Theorem. Our emphasis here is on the modelling of this prior information in a formal and consistent manner. We introduce the distinction between local and global prior information whereby local information usually applies to particular components of the pedigree and global prior information refers to more general features. When it is difficult to decide on a prior distribution, robustness to various choices should be studied. When suitable prior information is not available, a flat prior can be used which will then correspond to a strict likelihood approach. In practice, prior information is often considered for these problems, but in a generally ad hoc manner. This paper offers a consistent alternative. We emphasise that many practical problems

can be addressed using freely available software.

Keywords: Inferring Relationships, Likelihoods, Prior Information, Bayes' Theorem, Pedigrees

1 Introduction

There are many situations which require determination of the true pedigree connecting a given set of individuals from genetic marker data. A small pedigree is usually sufficient to describe the relationship between people involved in a paternity case (Essen-Möller 1938). Similar or slightly more complicated structures are typically required to describe immigration or family reunion cases (Hansen & Morling 1993) and identification problems following disasters could potentially require quite large pedigrees (Gill et al. 1994, Olaisen et al. 1997). Relationship identification is also important in animal and plant applications. Bowers et al. (1999) use parentage analysis to study the origins of wine grapes from Northeastern France, for example. Meagher & Thompson (1987) apply pedigree reconstruction methods to identify maternal and paternal parents of seedlings in a natural population to assess pollen dispersal patterns and evaluate reproductive success. Jones & Ardren (2003) review methods for parentage assignment in natural populations and Blouin (2003) reviews existing methods for relationship estimation from different areas of application and discusses their relevance to wild populations. Clearly, these problems can all be described in terms of determining the most likely pedigree from a set of possible alternatives although this is not usually made explicit.

Misspecification of relationships can severely distort the results of a genetic linkage analysis. Similarly, genetic association studies can be affected if there are close relationships within or between cases and controls that were not specified at recruitment. Much of the literature on inferring relationships for humans concentrates on pairwise relationships and uses genome-scan data, either with view to correcting for errors prior to a linkage/association study or to identify sets of individuals likely to share longer haplotypes around susceptibility alleles (Boehnke & Cox 1997, Göring & Ott 1997, McPeck & Sun 2000, Stankovich et al. 2005). We wish to focus more generally on specifying the relationships amongst several individuals and will usually expect data at no more than 15-20 autosomal microsatellite marker loci. More importantly, we will typically be interested in getting the true relationship for these particular individuals, rather than an overall idea of the graph structure connecting them. Our focus is hence less on large pedigree reconstructions of genetic isolates where the latter emphasis would typically apply and more on forensic problems, wildlife population management applications or, perhaps, construction of pedigrees for linkage analysis from a large population study.

Reconstruction of the pedigree connecting a set of individuals from genetic marker data using likelihood criteria dates back over three decades (Thompson 1974, 1975, 1976a). Besides the marker data, there may sometimes be additional information which is frequently used in practice although it is not always incorporated within a formal modelling framework. Ambiguities due to symmetries in likelihoods for pairwise relationships can often be resolved, for instance, by using age information to distinguish between the parent and offspring when such a relationship is indicated by the genotype data, and sex information

to establish whether it is a maternal or paternal relationship (Thompson 1975). Three pairwise relationships, halfsibs, grandparent-grandchild and aunt-niece, are discussed by Thompson (1986) as indistinguishable on the basis of any amount of data at independently segregating loci. In the absence of relevant non-genetic information, extra DNA data such as that provided by haploid data (i.e. mtDNA or Y-chromosome data) or information on additional relatives (Sieberts, Wijsman & Thompson 2002) or linked markers (Thompson & Meagher 1998) may resolve the problem.

Most interesting identification cases cannot be resolved without some kind of extra information. At least 50 unlinked microsatellite markers are required to distinguish between a pair of half-siblings and unrelated individuals (Blouin 2003) although this number can be approximately halved when maternal profiles are also available (Mayor & Balding 2006). For our purposes, any information that is in addition to the DNA marker data is defined as “prior” information. In practice, such information is not always thought of as prior information, either because it is often incorporated in an informal way, or because of existing prejudices against a Bayesian inferential process. Our approach can be seen as an extension of Egeland et al. (2000), as implemented in the freeware package Familias (<http://www.nr.no/familias>), by allowing for prior knowledge about parts of the structure to be specified and introducing the idea of distinguishing clearly between different sorts of prior information in terms of practical implementation. It comprises four stages: define a sample space of pedigrees, assign a prior probability distribution on this space, calculate the likelihood of the DNA data for each pedigree and compute posterior probabilities using Bayes’ Theorem. The emphasis in this paper is on the first two steps. This four-step procedure allows for a practical implementation whereby each step can be performed separately using whatever tools are available. We show that many of the existing approaches to inclusion of extra information can be viewed as special cases of the simple structured prior function we propose. Of course, a single, albeit general, prior is unlikely to meet the needs of all users so it is important that it can be conveniently tailored to a specific application. The prior function we propose has this facility and can be easily modified.

The main advantage of modelling additional information on an identification problem in a formal way is that all the relevant information is stated “up front” as part of the prior function. Hence, it can be incorporated in the most efficient and appropriate way and not simply be brought in to resolve a particular dilemma when a standard likelihood procedure reaches an impasse. This paper illustrates how many apparently diverse contributions to the literature on estimating relationships can be brought together successfully. As noted by Blouin (2003), the animal genetics literature seems often unaware of relevant developments in human genetics but the reverse argument could also be made. The forensic science literature adds yet another dimension. Since all are working on closely related problems for which many of the key ideas were detailed thirty years ago, it is important to combine the expertise in these areas so that such

issues do not continue to be addressed in parallel. The outline of the paper is as follows. We will begin with a brief review of likelihood-based pedigree reconstruction from DNA data and discuss some approaches to incorporating relevant additional information. We then introduce our prior function and show how many existing approaches from hitherto quite separate sources can be viewed as special cases. We will discuss the advantages and limitations of this modelling approach together with illustrations of how it can be used on real and simulated data.

2 Likelihood-Based Relationship Estimation

In theory, estimating the pedigree for a given set of individuals from genetic marker data is simple: all one has to do is consider all possible joint relationships amongst them and compute the likelihood for each (Cannings & Thompson 1981). However, in a captive breeding programme, for example, precise estimation of the relationships between the existing animals who will be the founders of the future population might require consideration of a vast number of possible alternatives. In practice, brute force enumeration is not always practical and a different approach is required. One such alternative is provided by the sequential algorithm of Thompson (1976*a*) which arrives at a single reconstruction by starting from a position where all individuals are assumed to be unrelated and gradually accepting sibships on the basis of the increase attained in log-likelihood, or *support*. The method is most successful in reconstructing pedigrees connecting a set of closely related individuals and tends to favour large sibships. In particular, it is assumed that the parents of each individual in the sample of interest are either included in the sample themselves, or else are unrelated to any other members of the sample. In general, highly polymorphic loci are better for excluding false parent-offspring links but large numbers of loci are more informative in terms of specifying the relationship since it is the proportion of loci at which individuals have alleles in common that is relevant.

Comparing two pedigrees using likelihood criteria will depend on the allele frequencies assumed for the founders. Rare alleles also tend to give a high likelihood to gene identity by descent and more inbred structures could be heavily favoured in extreme cases. The joint relationship between several individuals determines the probabilities of genetically distinct classes of gene identity states (Thompson 1974). Maximum likelihood estimates of these probabilities, and hence of the relationship, can be obtained using standard methods. Hence, amongst all possible alternative relationships, the true relationship is one of those that maximise the expected log-likelihood “regardless of the number of individuals, the complexities of their relationship, the number of loci for which data are available and the frequencies of alleles and dominance systems exhibited by these loci” (Thompson 1976*b*). Reconstruction methods, such as the sequential procedure outlined above, build up an estimated structure that has high overall likelihood from subunits. The focus here is on those subsets of individuals giving maximum likelihood increase for the acceptance of a given relationship,

relative to the alternative that they are unrelated. For example, which of all non-excluded parent pairs are the most likely parents of A ? The true parents will not necessarily maximise the expected log-likelihood. In fact, a sibling will often have a higher expected log-likelihood for the parent hypothesis than a true parent (Thompson 1986). This will remain the case, however many loci are considered. How important this is will depend on the application. If the main focus of a reconstruction is on the overall shape of the pedigree, with view to gaining anthropological information on age and mating patterns, for example, finding a highly likely pedigree will generally suffice. If the focus, however, is on precise identification of specific relationships, such as in a forensic setting, the set of all possible relationships amongst the given individuals is the correct set of alternatives to consider.

Cannings & Thompson (1981) also suggest that a pedigree can be represented using cluster analysis or multidimensional scaling methods (Thompson 1974) on genetic distance measures, but argue that such approaches rarely provide clarification of the relevant pedigree structure. Cowell & Mostad (2003) combine clustering and likelihood approaches by defining a likelihood-ratio-based distance measure of pairwise relatedness and clustering individuals using an estimate of this measure.

The sequential algorithm of Thompson (1976*a*) exploits age and sex information. Age data, or at least an age ordering, are required to initially sort individuals by descending maternal age. Age data are also used for generation gap restrictions in that mothers are constrained to be between 15 and 50 years older than any of their offspring while fathers must be between 15 and 75 years older. Finally, age data are used to assess the plausibility of selected sibships. Almudevar (2003) considers maximum likelihood pedigree reconstruction via a simulated annealing algorithm that begins with an enumeration of parent-offspring triplets and assembles them into sets of admissible pedigrees on which the likelihood is easily maximised. Age and sex data, although easily incorporated, are not specifically required and the method appears successful on reasonably large pedigrees (one example had 69 individuals) provided there is sufficient kinship structure. As with the sequential approach described above, however, this algorithm assumes what the author calls a “complete” sample in which parents of each individual are either included in the sample themselves or else are unrelated to any other individuals in the sample. Knowledge about mating patterns can also be incorporated in likelihood approaches to pedigree reconstruction. For human populations, Thompson (1976*a*) notes that social prohibitions on polygamy, for example, are usually restricted to concurrent, rather than consecutive matings: age distributions of the relevant offspring groups can be used to distinguish between the two. Prodohl et al. (1998) use natural history information, including lactating status of females and spatial positioning of parents and offspring to refine parentage inferences from genetic likelihoods on a population of armadillos.

The term “prior” immediately introduces the notion of Bayesian reasoning which has had a very mixed reception in legal circles. Indeed, the UK Court of

Appeal appears to have ruled Bayes' Theorem to be inadmissible as evidence. (See Balding (2005) for a recent overview.) The strength of the Bayesian argument lies in its facilitation of discussion via probabilistic statements rather than hypothesis-testing. Despite its obvious relevance, there has not been much discussion of formal use of prior information in forensic applications although attention has been given to the consideration of mutation rates and the deficiencies associated with a single alternative hypothesis (Evet & Weir 1998, Dawid et al. 2001, Vicard & Dawid 2004, Balding 2005). However, despite the reluctance to use Bayesian inference in these settings, Essen-Möller's well-known W or "Wahrscheinlichkeit" statistic has a straightforward Bayesian interpretation. Consider the two standard hypotheses in paternity cases:

- H_1 : The alleged father is the biological father,
- H_2 : A "random" man is the father,

where by "random" we mean that the individual's genes are randomly drawn from the population gene pool. The "W" statistic is defined as

$$W \equiv P(H_1|data) \equiv \frac{LR}{LR + 1} \quad (1)$$

where LR is the relevant likelihood ratio or *paternity index*

$$LR = \frac{P(data|H_1)}{P(data|H_2)}. \quad (2)$$

Although a Bayesian argument was not explicitly used, Essen-Möller (1938) was aware that equality in (1) assumed equally likely competing hypotheses. Practice differs between forensic laboratories as to whether they report the paternity index or W , or both (Egeland & Mostad 2002). It has been argued that although they both contain the same information, the interpretation of W as the (posterior) *probability of paternity* (i.e. posterior probability of H_1) may make it less abstract and less open to misinterpretation than a likelihood ratio (Hummel 1984). Nonetheless, it is often avoided in practice, due to the indirect assumption of equally likely hypotheses (Evet & Weir 1998).

The standard analysis in which only the likelihood ratio, or paternity index, is considered is restricted to the comparison of pairs of alternatives and the "random man" alternative hypothesis, in particular, suffers from well-known deficiencies (Goldgar & Thompson 1988, Balding 2005). Moreover, as has been noted by several authors, these likelihood ratios may differ dramatically for different choices of alternative hypothesis and it is not obvious how to summarise calculations for different such pairs (Egeland et al. 2000). Despite the fact that it is usually the likelihood ratio that is calculated, the questions that are asked in many forensic applications often expect (and mistakenly interpret) an answer in the form of a probability statement. (This is related to discussions in Evet & Weir (1998) on the "transposed conditional" problem.) If paternity probabilities are

required, proper posterior probabilities must be calculated and prior distributions must hence be specified. Consider the completely general case with n competing hypotheses $H_1, \dots, H_n$ having prior probabilities $\pi_1, \dots, \pi_n$, respectively. Note that the original definition in (1) is equivalent to assigning the values $\pi_i = \frac{1}{n}$ corresponding to a flat prior. Let $L_i \equiv P(\text{data}|H_i)$. By Bayes' Theorem, the posterior probability of H_i is

$$P(H_i|\text{data}) = \frac{L_i \pi_i}{\sum_{i=1}^n L_i \pi_i}. \quad (3)$$

Although, posterior probabilities are more meaningful, pairwise comparisons can still be made for standard paternity analyses using posterior probability ratios. Note that the interpretation of such ratios raises issues similar to those discussed for likelihood ratios: not alone are they restricted to pairs of alternatives but they also depend on the choice of prior. For any pair of hypotheses, H_i and H_j we hence have that

$$\frac{P(H_i|\text{data})}{P(H_j|\text{data})} = \frac{L_i}{L_j} \frac{\pi_i}{\pi_j} \text{ for any } i \neq j \quad (4)$$

expressing the posterior probability ratio on the left hand side as the product of the likelihood ratio, L_i/L_j , and the prior ratio, π_i/π_j . When there are only two competing alternatives, (4) is popularly known as the *odds form of Bayes' Theorem* (Evetts & Weir 1998). It is clear from the above representations that the likelihood calculations can be considered separately from the prior probability assignments. From a practical viewpoint, externally assessed prior information can hence be incorporated into an analysis where existing software is used to calculate the relevant likelihood ratios (Mortera, Dawid & Lauritzen 2003).

Many authors have taken a Bayesian approach to the problem of relationship estimation. Gill et al. (1994) address the identification of the remains of the Romanov family where aristocratic origins were implied by indications of gold, platinum and porcelain dental work in some of the bodies. They show how this piece of evidence can be combined with mtDNA using the odds form of Bayes' Theorem, as described above. A likelihood-based approach to the problem of confirming pairwise relationships in sib pairs prior to conducting a linkage analysis is considered by Göring & Ott (1997). The aim here is to increase the power to detect linkage by eliminating false sib pairs. The focus is thus on distinguishing between sibs, half-sibs and unrelated individuals as these are argued to be the most likely alternatives, given the reasons for which they were recruited. Prior probabilities are assigned to the three types of relationship based on knowledge of laboratory error rates and population rates of non-paternity and adoption. Posterior probabilities based on these priors and the likelihoods from the genetic markers are then calculated using Bayes' Theorem. Thompson & Meagher (1987) place prior probabilities on specific relationships between pairs of individuals for inferring parentage. Neff, Repka & Gross (2001) consider a Bayesian approach to calculating "expected" rather than most likely parentage by incorporating additional biological information, such as behavioural observations

during mating, in a prior distribution on parentage vectors. They demonstrate that assuming this prior distribution to be uniform can lead to very misleading results. Goldgar & Thompson (1988) consider the standard paternity-testing problem using a Bayesian interval estimation approach. They reformulate the problem as one of estimating the genetic relationship of the putative father (i.e the tested individual) to the true biological father of the child and so avoid the usual interpretational problems associated with the standard paternity index. A beta prior probability distribution is placed on the coefficient of relationship (Wright 1922) between these two individuals and a posterior interval estimate obtained using numerical integration.

3 The Sample Space and the Prior

Since, our focus is on finding the most likely pedigree connecting a set of “observed” individuals for whom we will typically have DNA marker data, sequential hill-climbing approaches that could stick at local maxima are not ideal. Moreover, we do not want to be restricted to the assumption that parents of each individual in the sample are either included themselves or else are unrelated to any other individual in the sample. The desire to consider many alternatives is also a complicating feature for our applications.

We would thus wish to consider the correct set of alternative hypotheses by considering all possible pedigrees, but, as has been noted above, this could be a formidable task. A less specific alternative to the “random man” hypothesis for the paternity example above would be: “Some other man is the father”. This, if taken literally, corresponds to an impractically large set of alternatives. In reality, of course, the set of alternative fathers for a specific individual is far from infinite and one important use of prior information is to reduce this set to a manageable size by excluding implausible alternatives. Thompson (1976*a*) recommends exclusions based on particular demographic features provided the aim of the reconstruction is not to investigate any aspects of such features. We will distinguish between *hard* prior information that we will know with certainty and *soft* prior information about which we are only willing to make probabilistic statements. While the latter can favour or downweight particular features, the former can be used to restrict the set of possible alternatives thus making the consideration of “all possible alternatives” a realistic option. Note that the success of the approach presented here depends on being able to generate this sample space of alternatives.

3.1 The sample space

It is also helpful to distinguish between *global* prior information relating to general knowledge about the population or species in question, such as information on breeding patterns, mating behaviour, average numbers of offspring, cultural prejudices etc. and *local* prior information which is particular to the application at hand and relates to specific parts of the pedigree or specified individuals. Whatever algorithm is used to generate the sample space, many structures will

be automatically created which are unlikely candidates for various reasons. For instance, many inbred pedigrees will arise which may not be appropriate for a specific human application and may have a high likelihood if not penalised in some way, especially if rare alleles have been observed. Social and breeding patterns vary widely from one species to another and from one subpopulation to another within a species. Individuals can be categorised as “adults” or “juveniles” either according to their ages or information on whether they have offspring or not. They are deemed to be adults (and hence parental candidates) in the absence of suitable information to the contrary. Sometimes, hard local and global information can combine to significantly reduce the number of possibilities. For example, there are 6720 possible pedigrees comprising two males and three females: declaring one female as a juvenile reduces the number to 2817 whereas if one male is known to be a juvenile, it reduces further to 2128 possibilities. Knowledge on generation gap, such as the bounds on maternal and paternal age differences suggested by Thompson (1976*a*) (Section 2) would give a further reduction.

For many applications, pedigrees involving “unobserved” individuals, i.e. individuals that are not in the original group and for whom we have no genetic data, will be required. For example, for a group of four individuals to comprise a female and her three offspring by the same male, the unknown male must be included in order to describe the relationship, even though there is no other information on him. Several additional individuals might be required to describe more distant relationships. In a sequential approach to a single reconstruction, such additional individuals as described above can be incorporated as unobserved founders and represented simply by the genes they contribute which are assumed to be randomly drawn from the population gene pool. Although we might sometimes wish to hard-wire their relationships with observed individuals, we might not necessarily wish to restrict them to founding positions. Besides, the possibility that these additional individuals could connect the observed set in another way may also be of interest. In this case, they are labelled and added to the observed set and all plausible pedigrees containing the originals plus the extra unobserved ones will be considered, so creating a (considerably) enlarged sample space. An inheritance claim would be an example where this might be relevant. Inheritance laws vary from country to country and will only accept relationship up to a particular degree in order to honour the claim. The number of extra individuals required to cover all possible acceptable relationships between the claimant and the deceased poses an interesting question. Note that there is an upper limit to the degree of relationship that can be detected by any approach based on identity-by-descent (as is implicit in the likelihood calculations here) since the probability that two related individuals do not inherit *any* autosomal DNA from their closest common ancestors increases rapidly with increasing degree of relationship (Donnelly 1983).

All possible pedigrees comprising the final set of individuals can be generated, for example, by listing all possible parent-offspring links and incorporating various consistency checks to ensure that this is valid set of structures. Hard local and global prior constraints should be introduced at this stage to reduce this set.

Realistically, it will not always be sensible to list all reasonable pedigrees at the outset. Enumeration and evaluation of pedigrees could take place sequentially with an algorithm for efficient retrieval of structures with high likelihood and posterior probabilities at the end of the process. For large problems, Markov chain Monte Carlo methods could be considered to sample from this space. Some classification of subspaces of pedigrees might also be appropriate which would enable exploration of the sample space via the separate sub-classes.

3.2 The prior function

Once the sample space for n individuals (some of whom may not be observed) has been determined, a prior probability $Pr(g)$ is assigned to each pedigree g of the following form:

$$Pr(g) = c \prod_{i=1}^s M_i^{b_i(g)} \prod_{\substack{j,k=1 \\ j \neq k}}^n R_{jk}^{o_{jk}(g)} \quad (5)$$

where

$$c = \frac{1}{\sum_g \prod_{i=1}^s M_i^{b_i(g)} \prod_{\substack{j,k=1 \\ j \neq k}}^n R_{jk}^{o_{jk}(g)}}$$

is the normalisation constant. $M_1, \dots, M_s$ are non-negative *global* parameters that allow pedigrees to be weighted according to s specified characteristics. The integer exponent $b_i(g)$ corresponding to M_i provides a particular measure of that characteristic, is internal to pedigree g and thus provides the degree of the relative weightings of different pedigrees for the i^{th} characteristic. For example, if the i^{th} characteristic is inbreeding, $b_i(g)$ would be a measure of the extent to which pedigree g is inbred according to how inbreeding has been defined. Thus, $b_i(g) = 0$ might mean that g has no inbreeding and increasing values of b might correspond to increasing levels of inbreeding. The R-parameters are for local specifications and, as given in (5), allow incorporation of prior information on parent-offspring links. We define $o_{jk}(g) = 1$ if j is the parent of k in pedigree g and $o_{jk}(g) = 0$ if j is *not* the parent of k . For consistency, we must have that $o_{jk}(g) + o_{kj}(g) \leq 1$ for $j \neq k$ in any pedigree g .

If we set

$$M_i = \begin{cases} 0, & \text{if } b_i(g) > 0 \\ 1, & \text{otherwise} \end{cases},$$

only pedigrees, g , with $b_i(g) = 0$ will be allowed. This is equivalent to setting $M_i = 0$ and defining $0^0 \equiv 1$ and, in the example above, would eliminate all inbred pedigrees. If $M_i = 1$, all pedigrees with feature i receive equal weighting from the prior regardless of their respective b -values. Setting all M-parameters to 1 amounts to assigning a flat prior whereby all generated pedigrees have equal probability apriori and there is no penalty associated with any particular feature. A value between 0 and 1 decreases the prior probability of the relevant characteristic while a value exceeding 1 increases it.

The local parameters R_{jk} can be interpreted analogously. Thus $R_{jk} = 0$ rules out the pedigrees featuring j as a parent of k while $R_{jk} > 1$ would give favourable weighting to such structures. For example, if it is believed that individual 1 is the parent of 2, we would set R_{12} to a value larger than 1 to favour those structures, g , with $o_{12}(g) = 1$. Additional information revealing 2 to be a juvenile would allow us to exclude all possibilities where 2 is the parent of 1 so $R_{21} = 0$ disallows all structures g where $o_{21}(g) = 1$. Information on three-way relationships might sometimes be available and appropriate parameters could, in principle, be defined. However, parent-offspring links are probably the most useful in practice as the consistency checks become more complicated with increasing orders since three-way specifications must concur with all the pairwise ones and so on.

Note that while both global and local parameters, as defined here, can exclude, downweight or favour certain characteristics, they can never assign certainty in the sense that *only* particular structures will be considered. Hard prior information in the form of certainty such as “definitely inbred” or “1 *is* the mother of 2” should always be incorporated at the outset when the sample space is generated. From a practical viewpoint, although information such as age and sex of individuals can naturally be modelled as local specifications, either implicitly in the calculation of $o_{jk}(g)$ or explicitly as indicator variables, such information is often better employed to restrict the set of possible pedigrees down to a manageable size.

3.3 Interpreting the prior function

The general prior function in (5) potentially includes a large number of parameters. However, for many problems, most of these will be set to unity so it is not as cumbersome in practice as it might appear. Choosing non-trivial values for M and R parameters is not straightforward, however. For an animal population with very high levels of inbreeding, an M -parameter greater than 1 would be appropriate although it is not always obvious what value would be most relevant to the application and some experimentation with different values would be required. A straightforward interpretation of the prior is provided by considering a standard situation where comparison between two specific pedigrees, g_1 and g_2 , is of interest via the posterior probability ratio. From Equation (4), the corresponding ratio of the prior probabilities for g_1 and g_2 is the amount by which the likelihood ratio obtained from the DNA data should be adjusted by the prior, or non-DNA, information. (See Section 4.)

The particular characteristics that may be of interest can often be defined in several ways and this will affect the interpretation of the prior. For instance, the software Familias has a built-in prior with three pre-specified global parameters corresponding to inbreeding, promiscuity and generation number. Promiscuity is measured in terms of departure from monogamy and $b_P(g)$ is defined as the number of pairs of offspring in pedigree g with one common parent (i.e. half-siblings) who is also in g . Figure 1 depicts two quite different situations for which this promiscuity b-value would be identical. One might wish to distinguish between the overall number of departures from monogamy and the degree of

polygamy in some applications. The value of $b_I(g)$ provides some measure of

Figure 1: Pedigree (a) is taken from a real human Eskimo pedigree (Gilberg et al. 1978) and would be assigned a value of $b_P = 6$ by the prior of Egeland et al. (2000). This same value would apply to a pedigree containing six different instances of the structure shown in (b).

the degree of inbreeding in pedigree g and is defined as the number of offspring in pedigree g for whom both parents are represented in the pedigree and who themselves have a common ancestor also present in the pedigree (Egeland et al. 2000). Alternatively, one could allow for some inbreeding, while excluding unacceptable levels, by calculating kinship coefficients for each marriage pair and excluding pedigrees with maximum kinship exceeding some pre-specified limit.

While a global M-parameter or a local R-parameter value of greater than 1 will favour pedigrees with a particular characteristic, whether or not it will distinguish between pedigrees exhibiting differing degrees of that characteristic will depend on how the pedigree-specific exponents (b-values) are defined (as shown in Figure 1, for example). For instance, suppose we wish to impose an upper bound on litter or sibship size and we define the relevant index $b(g)$ to be 1 or 0 according to whether the maximum number of offspring of mating pairs exceeds this limit or not. Pedigrees with all sibships of acceptable—but equal—size will receive the same prior weighting as more realistic structures. Extra penalties would have to be imposed to make such distinctions. However, some care has to be taken here due to the multiplicative form of the prior (5). Whether or not a pedigree is inbred, for instance, will typically not be independent of specific parent-child relationships. This would undoubtedly become more problematic if there were several parameters relating to the same feature (e.g. sibship size), as suggested above. However, the dependencies between desirable global and local features of a pedigree are not easy to model.

This prior function is mathematically convenient, straightforward to implement and is particularly tractable in the consideration of ratios for standard forensic analyses, as we noted in Section 3.2. It is also extremely flexible and, although far from perfect, provides a simple approach to the integration of essential non-DNA information. It also includes many current approaches as special cases. The multiplicative prior function for a parentage vector described in Neff et al. (2001) is one such candidate being a product over potential parentage assignments and different biological traits. Likewise, the prior function of Göring & Ott (1997) could be incorporated in the form of three global parameters, one for each of the three relationships considered, whereas that of Thompson & Meagher

(1987) takes the form of a local specification as it relates to specific pairwise relationships. Priors on parameters, such as the beta prior on the relationship parameter of Goldgar & Thompson (1988) do not fall into this framework if they cannot be interpreted as a prior distribution on pedigree structures. Parametric downweighting, however, such as the negative quadratic support on sibship size mentioned by Thompson (1976a) can be easily incorporated. Note that a similar effect can be achieved by assigning a value between 0 and 1 to the relevant global M parameter. Assessment of the number of offspring in the pedigree, as suggested by Thompson (1986) can be achieved, either by relating numbers to an individual, or to a couple. One possibility is to let $b(g)$ be the number of offspring for a marriage couple. Alternatively, let $b(g) = 1$, if the maximum number of offspring for an individual in the pedigree exceeds a specified limit and 0 otherwise. The corresponding M parameter can be used to downweight or favour alternatives in the usual way. Similarly, the restrictions on polygamy and polygyny of Thompson (1976a) can easily be incorporated as global parameters with appropriate definitions of the relevant b values together with a local indicator variable for sex.

4 Examples

Here we consider three examples for which the prior function described in Section 3 is used. In the first example, we consider estimating the relationships between three individuals based on simulated data at one genetic marker locus, for simplicity, and focus on the sensitivity of the results to different prior assumptions. In the second example, we consider a paternity testing problem which was assigned to a set of forensic laboratories as a proficiency test and the third example considers an application in plant genetics taken from Bowers et al. (1999). Both these examples use real data.

4.1 Sensitivity to the prior

Consider three individuals comprising an adult male, an adult female and a juvenile according to their age and sex information. The twelve possible pedigrees involving only these three individuals are shown in Figure 2. In this example, we use the prior of Egeland et al. (2000) which is a special case of (5) and is the prior implemented in the current version of Familias. A prior probability $Pr(g)$ is assigned to each pedigree g of the form

$$Pr(g) = c M_I^{b_I(g)} M_P^{b_P(g)} M_G^{b_G(g)},$$

with M_I , M_P , and M_G allowing pedigrees to be weighted according to inbreeding, promiscuity and generation number, respectively. The b -values, $b_I(g)$ and $b_P(g)$ are described in Section 3.3: $b_G(g)$ is defined as the longest chain of generations beginning and ending with an adult where “adults” are parental candidates, as defined earlier (Section 3.1).

Figure 2: The twelve possible pedigree structures involving an adult male, an adult female and a juvenile. As is consistent with tradition, males are depicted by squares, females by circles and children by diamonds. A parent-offspring relationship is depicted by an arrow directed from the parent to the offspring individual.

Table 1 shows the values of b_I , b_P and b_G for each pedigree and the corresponding value of the prior function for some choices of M_I , M_G and M_P . Note that pedigrees 11 and 12 are both inbred with the male having a child with his own daughter or the female having a child with her own son, respectively. This is reflected in the corresponding values of $b_I = 1$ as there is only one offspring present with related parents, both of whom are also in the pedigree. The value of the promiscuity index, $b_P = 1$, is not immediately obvious for pedigrees 9 and 10 as they are depicted in Figure 2. Figure 3 shows the two possible interpretations of pedigree 9. As we have not declared an additional female to cover the possibility of a common mother, pedigree 9(i) is the assumed default: the unknown mothers are distinct and unrelated to other pedigree members. At first glance it may

Figure 3: Two possible interpretations of pedigree 9 in Figure 2 with b_P taking the value 1 in 9(i) and 0 in 9(ii).

appear strange that $b_G = 1$ for pedigree 7, a family consisting of a grandmother, her son and grand child. Ordinarily, this would be regarded as a two-generational pedigree but, by definition, adult-juvenile links are excluded from the count. This

is to circumvent the apparent inconsistency that would assign $b_G = 2$ to the inbred pedigrees 11 and 12 if such links were allowed. As noted in Section 3, this is just one way of defining b_G .

Ped	(b_I, b_P, b_G)	(M_I, M_P, M_G)				
		(1, 1, 1)	(0, 0, 1)	(0.1, 1, 1)	(10, 1, 1)	(0.1, 0.1, 10)
1	(0, 0, 0)	0.08333333	0.12500000	0.09803922	0.03333333	0.02164502
2	(0, 0, 0)	0.08333333	0.12500000	0.09803922	0.03333333	0.02164502
3	(0, 0, 0)	0.08333333	0.12500000	0.09803922	0.03333333	0.02164502
4	(0, 0, 1)	0.08333333	0.12500000	0.09803922	0.03333333	0.21645022
5	(0, 0, 1)	0.08333333	0.12500000	0.09803922	0.03333333	0.21645022
6	(0, 0, 1)	0.08333333	0.12500000	0.09803922	0.03333333	0.21645022
7	(0, 0, 1)	0.08333333	0.12500000	0.09803922	0.03333333	0.21645022
8	(0, 0, 0)	0.08333333	0.12500000	0.09803922	0.03333333	0.02164502
9	(0, 1, 1)	0.08333333	0.00000000	0.09803922	0.03333333	0.02164502
10	(0, 1, 1)	0.08333333	0.00000000	0.09803922	0.03333333	0.02164502
11	(1, 1, 1)	0.08333333	0.00000000	0.00980392	0.33333333	0.00216450
12	(1, 1, 1)	0.08333333	0.00000000	0.00980392	0.33333333	0.00216450

Table 1: Values of b_I, b_P and b_G for the 12 pedigrees of Figure 2. The last five columns give the prior distribution for the displayed choices of the M-parameters.

The prior probabilities in Table 1 reflect the values of the M-parameters as follows. The heading (1,1,1) for the third column indicates that all M-parameters are 1 and leads to the expected uniform prior of 1/12. The specification (0, 0, 1) completely rules out inbred and promiscuous pedigrees i.e. all pedigrees with $b_I \geq 1$ or $b_P \geq 1$, so pedigrees 9, 10, 11 and 12 are disallowed and the remaining pedigrees are equally likely apriori. In column 5, the values (0.1, 1, 1) downweight inbred pedigrees to 10% of any other pedigree whereas the situation is reversed in the next column with inbred alternatives being 10 times as likely as non-inbred. The last column downweights inbreeding and promiscuity while increasing the prior probability of generation number. This favours pedigrees 4, 5, 6 and 7 apriori as they have more parent-offspring links between adult individuals. Comparing pedigree 11 with pedigree 1, for example, the ratio of the two priors is

$$\frac{c M_I^1 M_P^1 M_G^1}{c M_I^0 M_P^0 M_G^0} = M_I M_P M_G$$

which takes values 1, 0, 0.1, 10 and 0.1, respectively, for the five cases shown in Table 1. If all the M-parameters are unity this ratio is one and both possibilities are equally weighted. Pedigree 11 has one tenth the weighting of the unrelated scenario in pedigree 1 when (a) inbreeding is downweighted by 10% compared to the other characteristics and (b) when inbreeding and promiscuity are both

downweighted but larger numbers of generations are favoured (final column of Table 1).

We now consider genetic marker data at a single locus on these three individuals and assume that the female is (A, A) , the male is (B, B) and the child is (A, B) . The allele frequencies for A and B are both 0.05. We also consider mutation when calculating the likelihoods and assume a model where the probability of a mutation decreases with the distance between the parent and offspring alleles (Dawid et al. 2002). Under this model, an allele with 14 tandem repeats is more likely to mutate to one with 13 or 15 repeats than to one with 12 or 16. For this example, we assumed an overall mutation rate of 0.001 and a factor of 0.1 corresponding to a decrease by one tenth for each additional unit length difference. Genotyping errors are probably more likely than mutations although forensic markers are generally quite well-behaved. However, as genotyping errors and mutations are confounded here (Ewen et al. 2000, for example), one could argue that the inclusion of a mutation model is an albeit crude accommodation for such errors (Sieberts et al. 2002). In the absence of more specific information, Familias uses the entry order of alleles, creating a final pooled category for all unobserved alleles, and assumes that the difference in length between consecutive alleles is 1. We used this default here. Table 2 shows the likelihoods for each of the 12 pedigrees in Figure 2 under this model along with the posterior probabilities corresponding to the prior probabilities of Table 1, for each of the five settings of the M-parameters. Note that the genetic data favour pedigree 8 with the male and female as parents of the child. Without mutation, pedigrees 4-7 and 9-12 would be excluded by the data as they all involve a parent-child relationship between an (A, A) and a (B, B) individual. They are all unlikely assuming mutation, with 4 and 5 being the least likely. The prior specification $(10, 1, 1)$ heavily favours the inbred pedigrees, 11 and 12, while the specification $(0.1, 0.1, 10)$ favours 4, 5, 6, 7. The posterior probabilities favour pedigree 8 for the first three scenarios as the priors also downweight the pedigrees with low likelihood. The posterior distribution corresponding to the $(0.1, 0.1, 10)$ specification also heavily favours pedigree 8: here the genetic evidence outweighs the prior preference for more generations. Only in the $(10, 1, 1)$ scenario do we observe an apparent conflict between the genetic data and the prior: the posterior distribution barely distinguishes between pedigree 8 and the two inbred alternatives, 11 and 12. Comparing pedigrees 8 and 12 we have a likelihood ratio of 10.44: the posterior probability ratio is 1.05 for the $(10, 1, 1)$ scenario but is 104,999 for the $(0.1, 0.1, 10)$ scenario.

4.2 A proficiency test

This example is based on a proficiency test assigned to a large number of forensic laboratories where the main language was Spanish or Portuguese (*Boletín Informativo no 8 GEP-ISFG* 2004). The general theme of the test is the treatment of conflicting information or data. More precisely, three individuals are involved:

Ped	Likelihood	(M_I, M_P, M_G)				
		(1, 1, 1)	(0, 0, 1)	(0.1, 1, 1)	(10, 1, 1)	(0.1, 0.1, 10)
1	0.00000003	0.00385304	0.00454339	0.00442961	0.00167404	0.00408888
2	0.00000031	0.03851204	0.04541226	0.04427504	0.01673246	0.04086936
3	0.00000031	0.03834691	0.04521754	0.04408520	0.01666071	0.04069412
4	0.00000000	0.00036696	0.00043270	0.00042187	0.00015943	0.00389417
5	0.00000000	0.00036696	0.00043270	0.00042187	0.00015943	0.00389417
6	0.00000003	0.00366781	0.00432498	0.00421667	0.00159357	0.03892320
7	0.00000003	0.00365209	0.00430643	0.00419859	0.00158673	0.03875630
8	0.00000616	0.75928801	0.89532999	0.87290905	0.32989049	0.80576396
9	0.00000003	0.00365209	0.00000000	0.00419859	0.00158673	0.00387563
10	0.00000003	0.00366781	0.00000000	0.00421667	0.00159537	0.00389232
11	0.00000059	0.07231314	0.00000000	0.00831342	0.31418141	0.00767394
12	0.00000059	0.07231314	0.00000000	0.00831342	0.31418141	0.00767394

Table 2: Likelihoods and posterior probabilities for the 12 pedigrees of Figure 2 from marker data at one locus and the prior probabilities of Table 1 corresponding to the five different settings of the M-parameter. Note that the actual likelihood for pedigrees 4 and 5 is $2.97619E - 09$.

an adult female, labelled as 1, an adult male 2, and a small boy, or juvenile, 3. There is information indicating that 1 is the mother of 3, but, as we will see, the genetic data are not entirely consistent with this prior assumption. The test asks for paternity probabilities (“La probabilidad de le paternidad”) and so, as we argued in Section 2, posterior probabilities and hence prior distributions are implicitly required although this was not the general interpretation. We discuss this case from the perspective of modelling the prior followed by the likelihood and posterior model calculations.

Prior model The first part of the prior concerns the sample space. There was some guidance given for this task. Most of the possible pedigrees involving these three individuals (see Figure 2) are ruled out by the age information: the two adults cannot feature in a parent-child relationship. It is also highly unlikely that the three individuals in a paternity case are unrelated. We have decided to limit the sample space to the four pedigrees shown in Figure 4, the first three of which correspond to the reasonable possibilities from Figure 2, i.e. pedigrees 8, 2 and 3 respectively. For pedigrees (i) and (ii), 1 is the mother of 3 as is consistent with the given prior information, the mother of 3 is unknown in pedigree (iii) and the sister of 1 is the mother of 3 in the last alternative. More possibilities involving additional individuals (such as (iv)) could be entertained, but we will not do so here. For simplicity, we will assume a flat prior over all global features by setting the relevant parameters to 1 and focus attention on the local parameter R_{13} . Thus by (5), the prior probability is cR_{13} for pedigrees (i) and (ii) and c for pedigrees

(iii) and (iv) where the normalisation constant is given by $c = 1/2(R_{13} + 1)$. Observe that $R_{13} = 1$ corresponds to a flat prior while large values of R_{13} assign priors close to 0.5 for the two alternatives where 1 is the mother of 3.

Figure 4: The four alternative pedigrees considered for the proficiency test. Pedigrees (i) and (ii) comply with the prior information that 1 is the mother of 3. The mother of 3 is held to be unknown in pedigree (iii) while a sister of 1 is the true mother in (iv).

Likelihood The data for a standard forensic set of nine markers (Butler 2006) are shown in Table 3. The first eight are consistent with 1 and 2 being the parents of 3. However, for marker D7S820, 1 and 3 share no alleles. Although genotyping error is also a plausible explanation, we will follow the instructions for the proficiency test and consider mutation to be the reason for this consistency. A reasonable approach would be to choose mutation models for each marker based on all available knowledge and general data but we assigned a mutation probability of 0.0015 to marker D7S820 and 0 to the others, as recommended to the laboratories.

The recommended model assumes that the probability of mutating to any allele i is proportional to the allele frequency p_i with proportionality constant $k = 0.0015 / \sum_{i=1}^{13} p_i(1 - p_i)$, but other reasonable mutation models give essentially the same posterior probabilities. (Details of this mutation model can be found in the Familias manual, for example.) Likelihoods for all markers on all pedigrees are given in Table 4. From this table, conventional likelihood ratios corresponding to a flat prior, can be computed. For instance, the LR comparing pedigrees i and ii is $3.02\text{E-}32 / 2.15\text{E-}37 = 139989$.

Posterior model The line corresponding to “Flat” in Table 4 gives the posterior probabilities for a flat prior. These correspond to scaled likelihoods and clearly give the same information as the total likelihoods on the preceeding line. The last alternative where the sister of 1 is the true mother, is the most likely corresponding to this model with a posterior probability of 0.96. Postulating apriori that it is 100

Marker	1	2	3
THO1	6,9	6,9	6,9
TPOX	8,11	8,8	8,8
CSF1P0	10,11	10,12	10,12
D3S1358	17,18	12,15	12,18
VWA	16,18	16,16	16,16
FGA	24,24	21,24	21,24
D5S818	11,11	11,14	11,11
D13S317	8,8	8,8	8,8
D7S820	7,9	11,13	10,13

Table 3: The marker data for the proficiency test. There are two adults (a female 1 and a male 2) and a child 3. The man is not excluded for any marker whereas there is an inconsistency at D7S820 for 1 being the mother of 3.

times more likely that 1 is the mother of 3 than not, implies that $R_{13} = 100$: the corresponding results are depicted in the last line of Table 4. Now the posterior for pedigree (i) is increased to 0.56 while that for (iv) is decreased to 0.43. It is always difficult to quantify prior information. Hence, as indicated in the first example, it is good practice to vary the assumptions. Figure 5 shows the posterior probability for pedigree (i) with 1 and 2 being parents of 3, as a function of $\log_{10}(R_{13})$. For a value of $\log_{10}(R_{13})$ in excess of 3, i.e. R_{13} exceeds 1000, pedigree (i) has a posterior probability close to 1.

Our treatment of this example is not so much presented in the form of a proposed solution but rather more as a broad discussion of the issues brought up by an important case that are bound to occur often enough to be of practical importance. In practice, one would endeavour to obtain more information or more marker data, or re-analyse the existing data with a view to possibly correcting genotyping errors, but these were not viable options for the proficiency test. The consensus amongst the participating laboratories was that the data for marker D7S820 should be disregarded and the paternity index calculated as usual. The defending argument was that maternal mutations are ignorable since the woman is assumed to be the true biological mother of the boy and, in any case, the maternal terms cancel out in the ratio (2) for the standard “random man” alternative hypothesis. While the identity of the father might often be the sole focus of a paternity case, we do not believe this to be a reasonable approach in general, especially when there is genetic evidence against a biological mother-child relationship and given that the probability of an error in forensic marker data is usually assumed to be small. The fact that the only discrepancy was due to a single offspring allele differing by one repeat unit from one of the two maternal alleles supports the need to consider a mutation model as STR mutations are predominantly one-step (Mayor & Balding 2006). However, genotyping errors can also show a similar preference for neighbouring alleles due to stutter pattern in microsatellite markers (Ewen et al. 2000). We have already commented on

System	Ped i	Ped ii	Ped iii	Ped iv
TH01	3.98E-03	1.69E-03	1.69E-03	2.84E-03
Tpox	3.57E-02	1.80E-02	3.60E-02	3.58E-02
CSF1PO	7.38E-03	4.88E-03	8.95E-03	8.16E-03
D3S1358	4.52E-05	4.89E-07	1.63E-05	3.08E-05
vWA	1.77E-03	3.99E-04	7.98E-04	1.28E-03
FGA	4.92E-04	1.63E-04	0.000152	3.22E-04
D5S818	3.45E-04	2.33E-04	1.16E-04	2.31E-04
D13S317	4.38E-04	6.34E-05	6.34E-05	2.51E-04
D7S820	4.82E-09	3.10E-09	7.50E-06	3.80E-06
Total	3.02E-32	2.15E-37	5.96E-32	2.32E-30
Flat ($R_{13} = 1$)	1.25E-02	8.93E-08	2.47E-02	9.63E-01
$R_{13} = 100$	5.59E-01	3.99E-06	1.10E-02	4.30E-01

Table 4: Results for the proficiency test. The likelihoods for 9 markers are given for the four pedigrees considered. The total likelihood is obtained by multiplying. The line corresponding to “Flat ($R_{13} = 1$)” gives posterior probabilities for the four alternatives ignoring prior information while the last line assumes that, apriori, pedigrees where 1 is the mother of 3 are 100 times as likely as those where this is not the case.

the importance of considering other alternative hypotheses. Including a mutation rate (which is a crude adjustment for typing errors) together with a high prior probability on the mother-child relationship is surely a preferable approach.

4.3 The Parentage of Wine Grapes

We now consider the study of Bowers et al. (1999) concerning the parentage of wine grapes from northeastern France. Most of the current cultivars are centuries old and many only exist in core collections. Based on an examination of 322 cultivars and genetic data at 32 microsatellite loci, 16 cultivars, including the great Chardonnay, were consistent with being the progeny of a single cross between the Pinot and Gouais blanc strains. As the Pinot grape has been in Burgundy for longer than any other variety, finding that it is a parent of several other varieties was not surprising. What *was* surprising was that Gouais blanc, very popular in that region in the Middle Ages but banned at various times for its mediocrity and no longer planted anywhere in France, was the other parent. The authors argue that, despite its poor reputation, Gouais blanc clearly has great genetic potential as a parent and that its genetic dissimilarity with the Pinot variety (sharing only 20 of the 62 alleles considered) contributed to the success of this unlikely parental combination. The message for modern grape breeding programs is to retain sufficient diversity in the core collections to enable consideration of comparably different parents.

Table 5 shows some of the published data from Bowers et al. (1999) at 4 of the 32 loci for Chardonnay (C) and its putative parents, Pinot (P) and Gouais blanc (G), along with the frequencies of the progeny (i.e. C) alleles based on the

Figure 5: Posterior probability of pedigree (i) (i.e. 1 and 2 are the parents of 3) plotted as a function of $\log_{10}(R_{13})$.

Locus	Genotype			Frequency
	P	G	C	
VVMD28	221	231	221	0.057637
	239	249	231	0.115274
VVS2	137	133	137	0.040346
	151	143	143	0.17147
VVMD31	216	212	214	0.086455
	216	214	216	0.214697
VrZAG79	239	237	243	0.094697
	245	243	245	0.108696

Table 5: Data from Bowers et al. (1999) giving the genotypes of Chardonnay, C, and its assumed parents Pinot, P, and Gouais blanc, G, at four loci. The estimated allele frequencies relate to the progeny alleles and so the frequency of 221 at locus VVMD28 is 0.057637 (or 0.06 as reported by Bowers et al. (1999)).

genotype data from all 322 cultivars considered. In the original paper, likelihood ratios were calculated for pairwise comparison of the pedigree with P and G as parents of C versus the five alternatives: both parents are unknown, either one of the two is a parent and the other is unknown, either P and a close relative of G is a parent or vice versa. A close relative was defined to be a parent, an offspring or a sibling. Figure 6 depicts these alternatives with pedigrees 5 and 6 distinguishing between the “parent of G” and “sibling of G” possibilities, for example. (We have omitted the alternatives where the close relative was an offspring here but clearly these could have been included.)

Figure 6: Eight alternative pedigrees for the relationship of Chardonnay with Pinot and Gouais blanc where we make no distinction between male and female plants.

Bowers et al. (1999) did not take a formal Bayesian approach in their analysis of these data but what we have referred to as “hard” prior information was incorporated to restrict the set of alternatives. Grape cultivars are generally hermaphroditic and thus self-pollinating. However, it would appear that grape is “intolerant of inbreeding” as only one self-pollinated cultivar was found in all 322 examined and so inbred alternatives were not considered. Historical information ruled out the options where C could be the parent of either P or G. We assume, as was implicit in the original paper, that loci are unlinked and in Hardy-Weinberg equilibrium. Our focus here is purely on priors so we will not consider the possibility of mutations and null-alleles although Bowers et al. (1999) do note that these should not be disregarded.

Our prior function (5) is also applicable to this situation. Table 6 gives the likelihoods, based on the published data at 4 loci for the eight pedigrees of Figure 6. “Posterior 1” corresponds to the flat prior (“prior 1”) assigning equal probabilities of $\frac{1}{8}$ to all pedigrees and agrees with the likelihood analysis that the pedigree with P and G both as parents is by far the most likely. “Prior 2” (with corresponding posterior probabilities) is a reflection of the pervading disbelief that Gouais blanc could possibly have had such a successful offspring as Chardonnay. In particular,

Pedigree	Prior 1	Prior 2	Likelihood	Posterior 1	Posterior 2
1	0.125	0.085470085	1.00994E-17	0.000012	0.000101
2	0.125	0.008547009	5.62162E-13	0.660551	0.564966
3	0.125	0.854700855	1.45349E-15	0.001708	0.146074
4	0.125	0.008547009	2.44133E-16	0.000287	0.000245
5	0.125	0.008547009	8.09692E-14	0.095140	0.081373
6	0.125	0.008547009	8.09692E-14	0.095140	0.081373
7	0.125	0.008547009	6.26209E-14	0.073581	0.062933
8	0.125	0.008547009	6.26209E-14	0.073581	0.062933

Table 6: Likelihoods and posterior probabilities for the eight pedigrees of Figure 6 for two choices of the prior (5).

the pedigree where both parents are unknown is 10 times more likely than one featuring Gouais blanc, or one of its relatives, as a parent under this prior, while any pedigree with Pinot, or one of its relatives, as a parent with another unknown species is 100 times more likely apriori. This corresponds to the prior of (5) with all global parameters set to 1 and appropriate settings of the local parameters for the relationships in question. As it turns out in this instance, the evidence from the marker data is so overwhelmingly in favour of Gouais blanc as a parent that it overrides the prior. In fact, when all 32 markers are considered (data not shown), the posterior probability of pedigree 2 is 0.99897. Pedigrees 5 and 6 (and 7 and 8) cannot be distinguished on the basis of their likelihoods and were combined in the original paper. Nonetheless, it is conceivable that pertinent prior information could separate them and this can be easily incorporated in our framework.

5 Discussion

For most relationship identification applications, we are rarely in a situation in which we know absolutely nothing besides the DNA marker data so it makes sense to consider a Bayesian approach to the problem. For forensic applications, the necessity for such an approach is implicit for the results that are often required in practice, such as paternity probabilities (Section 2). Prior information is often used in practice but is frequently incorporated informally at an interim stage of an analysis, such as when a likelihood approach produces what is clearly an unfavourable answer. This paper stresses the importance of stating all relevant information at the outset so that it can be integrated as efficiently as possible in a formal and transparent way.

As noted by Thompson (1975), finding the most likely relationship amongst a set of individuals is not the same statistical problem as identifying the most likely individuals for a specific relationship. No matter how much information is available, the latter does not necessarily assign individuals to the true relationship as the true relationship may not be among the alternatives considered. However, consideration of all possible pedigrees connecting the individuals of interest, is a formidable (sometimes impossible) task in general. The approach of Egeland et

al. (2000) attempts this by brute force enumeration of all possible alternatives but is restricted to very small problems and has been used only for forensic science applications. Besides extending the prior function to incorporate any number of global features and local parent-offspring relationships, we have shown that hard prior information to which we can attach certainty (e.g. there is *no* inbreeding or *A* is the mother of *B*) can play a vital role in reducing the set of alternatives to a manageable size, thus making such an approach tractable. Alternatively, efficient ways of generating and exploring the search space could be investigated.

The prior function (5) of Section 3 defines a prior distribution on pedigrees, rather than on model parameters and has obvious limitations: the multiplicative form may not always appear reasonable and there are no general guidelines for selecting values for the M and R parameters, besides the simple options of 0 and 1. The effects of any prior will be diluted in the presence of a lot of data but priors can potentially have heavy influence in our applications. Sensitivity to the choice of prior parameters should always be investigated for any particular application and a flat prior used if there is no other information. However, all priors have limitations. This prior has an advantage in being simple to extend and interpret. Moreover, many existing methods for incorporating prior information into relationship identification problems can be shown to be special cases or straightforward adaptations of this prior.

Unlike other approaches in the human genetics literature, we will typically not have genome-scan data for the applications we wish to focus on, we may not wish to assume that all parents are either in the sample or else are unrelated to other individuals in the sample, we are not purely interested in pairwise relationships and our interest lies in the true relationship rather than the best from a limited set of alternatives or a reasonable approximation. This approach is hence relevant to wildlife applications where researchers have traditionally been slow to adopt existing methods based on genome scan data because wildlife biologists usually work with small numbers of loci (Blouin 2003). For instance, it should be routine to calculate the relationships among founders in captive breeding programs but there are very few published examples where this has been done. Likewise, there have been few attempts to use reconstructed sibships as a means of estimating the effective number of founders that contributed to a particular population. Estimation of relatedness amongst individuals in a case-control study and estimation of subgroups of related individuals from a large population-based biobank genetic association study, either to identify those likely to share longer haplotype blocks around disease susceptibility genes of interest, or to construct pedigrees for a subsequent linkage analysis, are other potential applications of this approach.

The fact that pedigree applications can be expressed as *Bayesian networks* (Lauritzen & Sheehan (2003)) permits an interpretation of relationship estimation as a Bayesian network (BN) learning problem with a lot of structural constraints. So far, the existing BN learning algorithms are not appropriate for these problems but one has to suspect that they might be adaptable. A first step in this direction is being made by Angelopoulos & Cussens (2005) in extending their work on

defining probability tree-based priors on model structures using stochastic logic programs to pedigrees and then sampling from the posterior distribution via Markov chain Monte Carlo.

Many human applications of relationship estimation are concerned with error detection which tends to be viewed as a separate problem despite the important overlap. There are two main types of error that can occur: pedigree errors which are systematic and affect all loci, and genotyping errors which are sporadic and arise for various reasons including data entry errors, gel misreading or mutation. Distinguishing between them is difficult when limited data are available: Mendelian inconsistencies can be a symptom of either and Mendelian compatibilities do not necessarily imply that both are absent. Despite various claims to the contrary, Mendelian consistency checking of pedigree information can be shown to be an NP-complete problem and thus it is highly unlikely that popular existing algorithms such as those of O’Connell & Weeks (1999) and Abecasis et al. (2001), for example, are of polynomial complexity at worst (Aceto et al. 2004). This has obvious implications for the analogous problem of inferring pedigrees. Although forensic markers are generally well chosen, typing errors can be very common in other areas of application. It is possible to model genotyping errors in relationship estimation problems (Boehnke & Cox 1997, McPeck & Sun 2000, Sieberts et al. 2002, Sobel et al. 2002) and, as noted in Section 4, a mutation model can be interpreted in this light. This of course highlights the importance of the prior information in reducing the set of alternatives as the genetic data will never eliminate an implausible option when an error model is included. However, the computational issues still have to be addressed. Combining the expertise in all the diverse areas of application is surely a first step.

6 Acknowledgements

The authors acknowledge research support from Leverhulme Research Interchange Grant F/07134/K, which instigated our collaboration. We are grateful to Petter Mostad, Vanessa Didelez, Robert Cowell and Elizabeth Thompson for helpful comments and discussions. We also thank John Bowers for providing additional data for the plant genetics example.

References

- Abecasis, G., Cherny, S., Cookson, W. & Cardon, L. (2001), GRR: graphical representation of relationship errors, *Bioinformatics* **17**, 742–7439.
- Aceto, L., Hansen, J., Ingólfssdóttir, A., Johnsen, J. & Knudsen, J. (2004), The complexity of checking consistency of pedigree information and related problems, *Journal of Computer Science and Technology* **19**, 42–59.
- Almudevar, A. (2003), A simulated annealing algorithm for maximum likelihood pedigree reconstruction, *Theoretical Population Biology* **63**, 63–75.

- Angelopoulos, N. & Cussens, J. (2005), Exploiting informative priors for Bayesian classification and regression trees, *in* Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05), Edinburgh, pp. 641–646.
- Balding, D. J., ed. (2005), *Weight-of-evidence for Forensic Profiles*, Statistics in Practice, John Wiley & Sons, Ltd.
- Blouin, M. S. (2003), DNA-based methods for pedigree reconstruction and kinship analysis in natural populations, *Trends in Ecology and Evolution* **18**, 503–511.
- Boehnke, M. & Cox, N. (1997), Accurate inference of relationships in sib-pair studies, *American Journal of Human Genetics* **61**, 423–429.
- Boletín Informativo no 8 GEP-ISFG* (2004). IX Jornadas de Genética Forense Manóas, 2-4 de Junio de 2004.
- Bowers, J., Boursiquet, J.-M., This, P., Chu, K., Johansson, H. & Meredith, C. (1999), Historical genetics: the parentage of Chardonnay, Gamay and other wine grapes of northeastern France, *Science* **285**, 1562–1565.
- Butler, J. M. (2006), Genetics and genomics of core short tandem repeat loci used in human identity testing, *Journal of Forensic Sciences* **51**, 253–265.
- Cannings, C. & Thompson, E. A. (1981), *Genealogical and Genetic Structure*, Cambridge University Press.
- Cowell, R. G. & Mostad, P. (2003), A clustering algorithm using DNA marker information for sub-pedigree reconstruction, *Journal of Forensic Sciences* **48**, 1239–1248.
- Dawid, A. P., Mortera, J. & Pascali, V. L. (2001), Non-fatherhood or mutation? A probabilistic approach to parental exclusion in paternity testing, *Forensic Science International* **124**, 55–61.
- 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.
- Donnelly, K. (1983), The probability that related individuals share some section of the genome identical by descent, *Theoretical Population Biology* **23**, 34–63.
- Egeland, T. & Mostad, P. F. (2002), Statistical genetics and genetical statistics: a forensic perspective, *Scandinavian Journal of Statistics* **29**, 297–307.
- Egeland, T., Mostad, P. F., Mevåg, B. & Stenersen, M. (2000), Beyond traditional paternity and identification cases. Selecting the most probable pedigree, *Forensic Science International* **110**, 47–59.

- Essen-Möller, E. (1938), Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis. Theoretische Grundlagen, *Mitteilungen der Anthropologische Gesellschaft (Wien)* **68**, 9–53.
- Evett, I. W. & Weir, B. S. (1998), *Interpreting DNA Evidence*, Sinauer, Sunderland MA.
- Ewen, K. R., Bahlo, M., Treloar, S. A., Levinson, D. F., Mowry, B., Barlow, J. W. & Foote, S. J. (2000), Identification and analysis of error types in high-throughput genotyping, *American Journal of Human Genetics* **67**, 727–736.
- 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.
- Gill, P. E., Ivanov, P. L., Kimpton, C., Piercy, R., Benson, N., Tully, G., Evett, I., Hagelberg, E. & Sullivan, K. (1994), Identification of the remains of the Romanov family by DNA analysis, *Nature Genetics* **6**, 130–135.
- Goldgar, D. E. & Thompson, E. A. (1988), Bayesian interval estimation of genetic relationships: application to paternity testing, *American Journal of Human Genetics* **42**, 135–142.
- Göring, H. H. & Ott, J. (1997), Relationship estimation in affected sib pair analysis of late onset diseases, *European Journal of Human Genetics* **5**, 69–77.
- Hansen, H. E. & Morling, N. (1993), Genetic investigations in immigration cases and frequencies of DNA fragments of the VNTR systems D2S44, D5S43, D7S21, D7S22 and D12S11 in Turks, *Forensic Science International* **60**, 23–35.
- Hummel, K. (1984), On the theory and practice of Essen-Möller’s W value and Gürtler’s paternity index (PI), *Forensic Science International* **25**, 1–17.
- Jones, A. G. & Ardren, W. R. (2003), Methods of parentage analysis in natural populations, *Molecular Ecology* **12**, 2511–2522.
- Lauritzen, S. L. & Sheehan, N. (2003), Graphical models for genetic analyses, *Statistical Science* **18**, 489–514.
- Mayor, L. R. & Balding, D. J. (2006), Discrimination of half-siblings when maternal genotypes are known, *Forensic Science International* **159**, 141–147.
- McPeck, M. & L. Sun (2000), Statistical test for detection of misspecified relationships by use of genome-screen data, *American Journal of Human Genetics* **66**, 1076–1094.
- Meagher, T. & Thompson, E. (1987), Analysis of parentage for naturally established seedlings of *chamaelirium luteum* (*liliaceae*), *Ecology* **68**, 803–812.

- Mortera, J., Dawid, A. P. & Lauritzen, S. L. (2003), Probabilistic expert systems for DNA mixture profiling, *Theoretical Population Biology* **63**, 191–205.
- Neff, B. D., Repka, J. & Gross, M. R. (2001), A Bayesian framework for parentage analysis: the value of genetic and other biological data, *Theoretical Population Biology* **59**, 315–331.
- O’Connell, J. R. & Weeks, D. (1999), An optimal algorithm for automatic genotype elimination, *American Journal of Human Genetics* **65**, 1733–1740.
- Olaisen, B., Stenersen, M. & Mevåg, B. (1997), Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster, *Nature Genetics* **15**, 402–405.
- Prodohl, P. A., Loughry, W. J., McDonough, C. M., Nelson, W. S., Thompson, E. & Avise, J. C. (1998), Genetic maternity and paternity in a local population of armadillos assessed by microsatellite DNA markers and field data, *The American Naturalist* **151**, 7–19.
- Sieberts, S. K., Wijsman, E. M. & Thompson, E. A. (2002), Relationship inference from trios of individuals in the presence of typing error, *American Journal of Human Genetics* **70**, 170–180.
- Sobel, E., Papp, J. C. & Lange, K. (2002), Detection and integration of genotyping errors in statistical genetics, *American Journal of Human Genetics* **70**, 496–508.
- Stankovich, J., Bahlo, M., Rubio, J., Wilkinson, C., Thomson, R., Banks, A., Ring, M., Foote, S. & Speed, T. (2005), Identifying nineteenth century links from genotypes, *Human Genetics* **117**, 188–199.
- Thompson, E. A. (1974), Gene identities and multiple relationships, *Biometrics* **30**, 667–680.
- Thompson, E. A. (1975), The estimation of pairwise relationships, *Annals of Human Genetics* **39**, 173–188.
- Thompson, E. A. (1976a), Inference of genealogical structure, *Social Science Information* **15**, 477–526.
- Thompson, E. A. (1976b), A paradox of genealogical inference, *Advances in Applied Probability* **8**, 648–650.
- Thompson, E. A. (1986), *Pedigree Analysis in Human Genetics*, The Johns Hopkins University Press, Baltimore.
- Thompson, E. A. & Meagher, T. R. (1987), Parent and sib likelihoods in genealogy reconstruction, *Biometrics* **43**, 585–600.
- Thompson, E. A. & Meagher, T. R. (1998), Genetic linkage in the estimation of pairwise relationships, *Theoretical and Applied Genetics* **97**, 857–864.

Vicard, P. & Dawid, A. P. (2004), A statistical treatment of biases affecting the estimation of mutation rates, *Mutation Research* **547**, 19–33.

Wright, S. (1922), Coefficients of inbreeding and relationship, *The American Naturalist* **56**, 330–338.