Estimation of Mutation Rates from Paternity Cases Using a Bayesian Network

P. Vicard
Università Roma Tre

A. P. Dawid
University College London

J. Mortera
Università Roma Tre

S. L. Lauritzen
University of Oxford

December 6, 2004

Abstract

We present a statistical model and methodology for making inferences about mutation rates from paternity casework. This takes proper account of a number of sources of potential bias, including hidden mutation, incomplete family triplets, uncertain paternity status and differing maternal and paternal mutation rates, while allowing a wide variety of mutation models. A Bayesian network is constructed to facilitate computation of the likelihood function for the mutation parameters. It can process both full and summary genotypic information, from both complete putative father-mother-child triplets and defective cases where only the child and one of its parents are observed. Detailed analysis of a specific dataset is used to illustrate the effects of the various types of biases, and of the assumed mutation model, on inferences about mutation parameters.

1 Introduction

DNA testing is often conducted to resolve a disputed attribution of paternity. However, when this results in an *incompatibility* — values of a child's genotype and those of its presumed parents that are inconsistent with simple genetic segregation — the obvious interpretation of this as due to *non-paternity* is clouded by the alternative possibility that it is in fact due to *mutation*. We cannot properly assess the existence and nature of either of these effects without taking appropriate account of the potential disturbing effect of the other. Dawid *et al.* (2002) constructed Bayesian networks to analyse paternity cases while taking due account of the effect of mutation. Here we emphasise the converse problem: accounting for the effect of possible non-paternity on the estimation of mutation parameters. Once we have good mutation rate estimates, these can be used for

Research Report No.249, Department of Statistical Science, University College London
Date: December 2004

many purposes: in particular, they can be fed back into individual paternity case analyses.

A common estimate of overall mutation at some locus is simply the observed relative frequency of incompatibility at that locus in a sample of families. However, this is subject to a number of confounding biases. In particular, since the principal source of such family data is forensic laboratories conducting analyses to investigate paternity allegations, there will usually be some doubt as to whether the putative father is indeed the true father of the child. Other potential biases are hidden mutation, *i.e.* one not resulting in incompatibility (Chakraborty *et al.* 1996), and differing mutation rates in the male and female lines (Brinkmann *et al.* 1998). Care is needed to ensure that these and other sources of potential bias are handled appropriately. Vicard and Dawid (2004) demonstrated their possible effects and described an appropriate analysis, but this used only summary data and involved some restrictive assumptions. In this paper we relax these restrictions, and show how a Bayesian network can be used to effect a more detailed analysis, using complete data, under a variety of model assumptions.

The paper is organised as follows. First, in § 2, we give a general overview of the genetic background, the paternity casework used for mutation estimation, and the biases that need to be accounted for. In § 3 we describe a variety of models of interallelic mutation transitions. In § 4, and Appendix A, we study the structure of the likelihood function arising from data on several markers for a collection of cases, and show that, to a good approximation, the joint likelihood for all mutation parameters factorises into separate terms, one for each marker. In § 5 we show how to build a Bayesian network model to analyse single-marker data for any single case, and in § 6 we describe how this can be used to compute the likelihood function, either exactly or to an excellent approximation. Detailed analysis of a paternity dataset to estimate the mutation rate of marker VWA is conducted in § 7. We conclude with some cautionary comments on the construction and analysis of such datasets.

2 Background

2.1 Genetic background

2.1.1 Inheritance

A DNA profile, as used for forensic purposes, consists of measurements on a number of *genetic markers*, which we shall also loosely term *genes*. Current practice uses short tandem repeat (STR) markers (Weber and May 1989). These have a finite number (up to 20 or more) of possible values, or *alleles*, generally described by a positive integer (the *repeat number*). For any individual, the measurement on any given marker is represented as a *genotype*, consisting of the unordered allele values for two genes: one, the *paternal gene*, derived from the father, and the other, the *maternal gene*, from the mother — although it is not possible to distinguish which is which. If both genes have the same allele value, we say that the genotype is *homozygous* at that marker; if different, *heterozygous*. The process whereby at each marker a parent transmits to a child a copy of one of his or her two genes, each with probability $\frac{1}{2}$, independently for different transmissions, is called *Mendelian segregation*.

Throughout this paper, for simplicity, we assume both Hardy-Weinberg and linkage equilibrium, *i.e.* independence of an individual's genes both within and across markers. When justifiable, these assumptions greatly simplify the inference task: in particular, each DNA marker in the profile may be handled separately. We further restrict attention to the case that all unrelated "founder" individuals considered can be regarded as having DNA profiles drawn independently from a common randomly-mating population whose allele frequencies are known.

2.1.2 Mutation

The DNA markers used for forensic purposes are particularly prone to mutation, complicating their use. In paternity testing a true father may appear not to be so because a mutation has led to his passing on a seemingly impossible allele to the child. Appropriate assessment of mutation rates is important for correct analysis of paternity cases, as well as for studies in population genetics and for many other biomedical studies (Di Rienzo *et al.* 1998; Wilson *et al.* 2003). A general probabilistic approach to paternity testing accounting for the possibility of mutation has been described by Dawid *et al.* (2001), where it was shown how to compute the likelihood ratio in favour of paternity under some simplifying assumptions. In Dawid *et al.* (2002) a Bayesian network was used to compute this likelihood ratio relaxing these assumptions.

Paternity analyses require estimates both of population allele frequencies and of rates of mutation between specific alleles. Databases have been gathered from which the frequencies of the different alleles, in different populations, can be estimated. In this paper we have used allele frequency estimates for the German/Austrian population, based on data collected at the Institut für Rechtsmedizin, Westfälische Wilhelms-Universität, Münster, Germany. For simplicity, we take these frequencies as correct and known, but in future work the uncertainty attached to allele frequency estimates should be properly accounted for.

Data on inter-allelic mutational transitions are very sparse. For example, marker VWA has a repertory of 11 different alleles, thus requiring specification of 110 distinct allele-specific mutation rates. In the data we use, for this marker there are only 8 out of about 2000 meioses where mutation appears plausible. In order to have any chance of estimating 110 allele-specific mutation rates from such data, we need to construct models expressing them all in terms of a very small number of parameters to be estimated: such models are considered in § 3 below. The main thrust of this work is to estimate such mutational parameters, and thereby the overall mutation rate.

2.2 Paternity casework

Data collected by paternity testing laboratories mainly consist of DNA profiles for mother-putative father-child triplets, as well as some "defective" cases where one of the assumed parents has not been profiled. For each putative family, the data at a given marker can be *either*

Compatible: for example, mother (5, 5), putative father (5, 12), and child (5, 12) — this can be explained by paternity and Mendelian segregation, with the mother handing down either of her 5's, and the father his 12; *or*

Incompatible: for example, mother (5, 5), putative father (5, 12) and child (5, 8) — this can not be explained as above, since there would be no source for the child's 8.

Our data-set is based on paternity cases collected by the Institut für Rechtsmedizin. It consists of a large number of fully compatible triplets, some compatible mother-child and putative father-child pairs, and a few triplets showing a single incompatibility. In all cases maternity is taken as certain, whereas there may be dispute over paternity.

In this laboratory four or five STR markers are routinely used for paternity testing. Fully compatible triplets are regarded as cases of confirmed paternity. When just one incompatibility is detected, another set of five STR markers is profiled: if the incompatibility still remains isolated a "probability of paternity" (Fimmers *et al.* 1992) is computed, and when this is greater than or equal to 0.9997, paternity is regarded as confirmed, the single incompatibility being ascribed to mutation. Otherwise the case is considered as one of uncertain paternity, and is not considered further for purposes of estimating mutation rates. Similar procedures are used in other laboratories (Geada *et al.* 2003).

Mutation rates for the STR markers used in forensic parenthood tests range from around 5×10^{-4} to 7×10^{-3} per generation (Brinkmann *et al.* 1998; Henke and Henke 1999; Sajantila *et al.* 1999). Such a rate is commonly estimated by the observed frequency, s/n, of inferred mutation at that marker in casework triplets, with n the total number of meioses, and s the number of these deemed to be mutations. However, this "naïve estimate" can be very misleading.

2.2.1 Biases

Among the features that should be properly accounted for when modelling and estimating mutation rates are the following:

Hidden mutation. A compatible triplet could nonetheless be the result of a mutation. This effect can lead to a bias of around 15–35% for the naïve estimator (Chakraborty *et al.* 1996).

Differential mutation. Inequality of paternal and maternal mutation rates can affect estimates: Brinkmann *et al.* (1998) estimate the ratio of these rates at about 17:3.

Unknown paternity status. Particularly in paternity casework, one must allow for uncertainty as to whether or not the putative father is indeed the true father.

Previous work accounting for all these features (Vicard and Dawid 2004) required various simplifying assumptions and approximations. In this paper we relax or remove these. In particular:

(i). We can analyse detailed genotype data, with case-specific prior probabilities.

(ii). Defective cases, *i.e.* mother-child and putative father-child pairs, can be included.

(iii). General mutation models can be used.

Our approach is based on the construction and numerical analysis of a *Bayesian Network* model (Cowell *et al.* 1999) of the problem.

3 Mutation models

For a given marker, let $q_{i\to j}$ denote the probability that allele i mutates to allele j in a parent-child transmission. Then the probability that allele i is transmitted unmutated is $q_{i\to i} = 1 - \sum_{j\neq i} q_{i\to j}$, and the overall mutation rate is

$$\begin{aligned} \mu &= \sum_{(i,j):i\neq j} p_i\, q_{i\to j} \\ &= 1 - \sum_i p_i\, q_{i\to i}\,, \end{aligned}$$

where p_i denotes the population frequency of allele i.

Various models describing the allele-specific mutation rates $q_{i\to j}$ in terms of a small set of adjustable parameters have been proposed (Ohta and Kimura 1973; Valdes *et al.* 1993; Durret and Kruglyak 1999; Egeland and Mostad 2002; Dawid *et al.* 2001; Dawid *et al.* 2002). Assessment of the overall mutation rate μ may be sensitive to the particular mutation model assumed for the interallelic transitions.

3.1 Scalar models

A very flexible class of mutation models comprises the *scalar mutation models*, having the general form:

$$q_{i\to j} = \lambda\, s_{i\to j} \quad (j \neq i),$$

where $\lambda \geq 0$ is an unknown scale parameter to be estimated. Define $s_{i\to i} := 1 - \sum_{j:j\neq i} s_{i\to j}$. By rescaling the $s_{i\to j}$ $(j \neq i)$ and reciprocally rescaling λ, if necessary, we can assume without loss of generality that $s_{i\to i} \geq 0$, so that $S := (s_{i\to j})$ is a transition matrix.

The probability of no mutation out of state i is $q_{i\to i} = 1 - \lambda\,(1 - s_{i\to i})$, and the overall mutation rate per transmission is

$$\mu = \kappa\,\lambda \tag{1}$$

where

$$\kappa := 1 - \sum_i p_i s_{i\to i}. \tag{2}$$

3.2 Construction

We can regard a general scalar mutation process as operating as follows. From an original (unmutated) gene `og` we first generate a "possibly mutated gene" `mutg` according to the transition matrix S; and then take the actual (mutated) gene `ag` to be *either* `mutg`, *or else* identical with `og`, according to the outcome, 1 or 0, of a binary mutation indicator `mut`, where the probability that `mut` $= 1$ is λ.

3.3 Specific models

In this work we consider in particular the following scalar models (Vicard and Dawid 2004):

Proportional model

This has $s^{\text{prop}}_{i\to j} = p_j \ (i \neq j)$, $\kappa^{\text{prop}} = 1 - \sum p_i^2$.

Single-step mutation model (SMM)

This has

$$s^{\text{SMM}}_{i\to j} = \begin{cases} \frac{1}{2} & \text{if } |i-j| = 1, i \neq i_{min} \text{ or } i_{max} \\ 1 & \text{if } |i-j| = 1, i = i_{min} \text{ or } i_{max} \\ 0 & \text{other } i \neq j, \end{cases}$$

and $\kappa^{\text{SMM}} = 1$. Here i_{min} and i_{max} denote the boundary values for the repertory of alleles.

Mixed model

With probability h a transition follows the SMM, and with probability $(1-h)$ the proportional model, independently for each potential mutation event. This yields

$$\begin{aligned} s^{\text{mixed}}_{i\to j} &= h s^{\text{SMM}}_{i\to j} + (1-h) s^{\text{prop}}_{i\to j} \quad (i \neq j), \\ \kappa^{\text{mixed}} &= h + (1-h)\left(1 - \sum p_i^2\right). \end{aligned}$$

We can either fix the mixing parameter h (a high value such as 0.9 is biologically plausible), or leave it to be estimated.

3.4 Parametrisations

Assume a common scalar mutation model in both the maternal and paternal lines, but with possibly differing line-specific scale parameters, λ_M and λ_P. The line-specific mutation rates are thus $\mu_M = \kappa\,\lambda_M$, $\mu_P = \kappa\,\lambda_P$, with κ given by (2).

We introduce: $\tau := \mu_M + \mu_P$, the *total mutation rate*; and $\rho := \mu_P/\tau$, the *paternal fraction* of the total mutation rate. We thus have

$$\mu_P = \rho\,\tau \tag{3}$$
$$\mu_M = (1-\rho)\,\tau. \tag{4}$$

We further introduce the *total mutation parameter* $\xi := \lambda_P + \lambda_M$. Then

$$\tau = \kappa\,\xi \tag{5}$$

and

$$\rho = \lambda_P/\xi. \tag{6}$$

Thus

$$\lambda_P = \rho\,\xi \tag{7}$$
$$\lambda_M = (1-\rho)\,\xi. \tag{8}$$

In this paper we take the total mutation rate τ, or equivalently ξ, as the principal parameter of interest (Dawid (2003) used $\lambda := \xi/2$.) We treat ρ as fixed, but the analysis is easily extended to allow estimation of ρ along with τ.

4 Likelihood function

We consider mother–putative father–child triplets, on each of which we have full or partial information ("findings") about their DNA profiles. We suppose (somewhat unrealistically in view of the description in § 2.2) that we have a collection of such cases randomly drawn from the relevant population.

For any case, full information at any given marker would comprise the genotypes of all three individuals for that marker. Partial information comprises both defective cases, in which data on one of the "parents" are missing; and/or summary data, for example a simple report as to compatibility.

4.1 General structure

For each marker m, we suppose we have a model — not necessarily a scalar model — describing its mutation process in terms of a parameter ξ_m. We wish to use all the data to make inferences about the vector $\boldsymbol{\xi} = (\xi_m : i = 1, \ldots, M)$ of mutation rate parameters of the various markers, allowing in particular for the possibility of non-paternity.

The structure of the joint likelihood function for $\boldsymbol{\xi}$ is analysed in the Appendix. There it is shown that, under reasonable assumptions, we can approximate this joint likelihood by a product of terms, one for each ξ_m. Any such term is itself computed as a product over the individual cases. Considering a single such case, with findings f_m for marker m, its contribution to the likelihood for ξ_m has the form:

$$\ell(\xi_m) \propto \pi^*_{\setminus m}\,\mathrm{pr}(f_m \mid \xi_m, P) + (1-\pi^*_{\setminus m})\,\mathrm{pr}(f_m \mid \xi_m, N), \tag{9}$$

where P [resp. N] denotes paternity [resp. non-paternity], and $\pi^*_{\setminus m}$ is the probability of paternity for the case, calculated from the findings on all markers other than m, assuming no mutation — this may be calculated from standard formulae (Essen-Möller 1938).

Finally we multiply all the likelihood contributions across the different families to obtain the overall likelihood for ξ_m.

4.2 Algebraic form

One way of proceeding is to develop algebraic formulae for the terms $\mathrm{pr}(f \mid \xi, P)$ and $\mathrm{pr}(f \mid \xi, N)$ in (9) (for a single marker m, here omitted from the notation).

Suppose that we have full genotype data on a certain triple: mother's genotype AB, putative father genotype CD, and child's genotype EF (here A, B, C, D, E, F are arbitrary and any of them could be identical). Let $q^M_{i\to j}$ denote the mutation transition rate from allele i to allele j for the maternal line, and $q^P_{i\to j}$ that for the paternal line. Then under paternity, P, we have

$$\begin{aligned} \mathrm{pr}(f \mid \xi, P) &\propto \mathrm{pr}(EF \mid AB, CD; P) \\ &\propto (q^M_{A\to E} + q^M_{B\to E})(q^P_{C\to F} + q^P_{D\to F}) \\ &\quad + (q^M_{A\to F} + q^M_{B\to F})(q^P_{C\to E} + q^P_{D\to E}). \end{aligned} \tag{10}$$

Whittaker *et al.* (2003) use formula (10) to develop a likelihood analysis and comparison of various mutation models, assuming paternity in all cases.

Since $\sum_j q_{i\to j} = 1$, (10) can always be written in terms of the (typically very small) transition probabilities $(q^M_{i\to j},\, q^P_{i\to j} : i \neq j)$, and will have the form, in self-explanatory notation, *constant* + *linear* + *bilinear*. Apart from exceptional cases where the "child" shares no alleles at all with either parent, at least one of the q's in (10) will be of the form $q_{i\to i}$, and then *linear* will not vanish; to a very good approximation we can then ignore the term *bilinear* and treat (10) as a linear function of the small q's.

When the findings are incompatible, (10) must reduce to 0 when all the $(q^M_{i\to j},\, q^P_{i\to j} : i \neq j)$ are 0, so that *constant* will be 0. The remaining term *linear* then agrees with the formulae for L_P given in Table 1 of Dawid *et al.* (2001).

Under non-paternity, N,

$$\begin{aligned} pr(f \mid \xi, N) &\propto \mathrm{pr}(EF \mid AB, CD; N) \\ &\propto 2\left\{(q^M_{A\to E} + q^M_{B\to E})\,p_F + (q^M_{A\to F} + q^M_{B\to F})\,p_E\right\}. \end{aligned} \tag{11}$$

(In both (10) and (11) the omitted constants of proportionality are the same: in the first line, the product of the probabilities for genotypes AB and CD; in the second, 1/4 if $E \neq F$, 1/8 if $E = F$.)

Expression (11) is a linear function of $(q^M_{i\to j} : i \neq j)$, the constant term being non-zero except for a case of maternal incompatibility. This constant supplies the formulae for $L_{\overline{P}}$ given in Table 1 of Dawid *et al.* (2001). Note however that, in constrast to that paper, for current purposes we can not ignore the linear term.

The likelihood contribution $\ell(\xi)$ for this case is now calculated from (9).

4.2.1 Scalar model

For a scalar model we have $q^M_{i\to j} = \lambda_M s_{i\to j}$ $(i \neq j)$, $q^M_{i\to i} = 1 - \lambda_M(1 - s_{i\to i})$, and similarly for the paternal line. Expressing $\lambda_P = \rho\,\xi$, $\lambda_M = (1-\rho)\,\xi$, we see that, as a function of ξ for fixed ρ, $\ell(\xi)$ is exactly quadratic, and very near to linear. The coefficients of this quadratic function can readily be expressed in terms of the $(s_{i\to j})$ and ρ.

The above implies:

Lemma 4.1 *For any findings f on a case, $\ell(\xi)$ is a quadratic function of ξ.*

However for incomplete findings f, for example a simple report that the triplet is incompatible, expression of the coefficients of this quadratic function in algebraic form is likely to be impracticable.

The quadratic term in $\ell(\xi)$ is entirely due to considering the possibility of two simultaneous mutations, under paternity — a very rare event; hence $\ell(\xi)$ will be linear to a very close approximation. We remark that the approximate factorization of (A.2), on which we rely, is itself valid only up to linear terms, so that in any case we should not regard the quadratic term as very meaningful.

For an incompatible case $\mathrm{pr}(f \mid \xi)$ will typically[1] be increasing in ξ. Using (5) and rescaling, we then have an expression for the likelihood contribution, in terms of the total mutation rate τ, of the form

$$\ell(\tau) \propto a + \tau + O(\tau^2). \tag{12}$$

The intercept a determines the behaviour of the likelihood contribution for this case.

4.2.2 Influence of the constant

Plausible values of τ typically range from 0.002 to 0.01. When the intercept $a > 0.02$, the likelihood contribution (12) is effectively constant in this region, and the case can thus be discarded as essentially uninformative about τ. On the other hand, when $a < 0.001$, $\ell(\tau)$ can be well approximated by τ, equivalent to taking paternity as confirmed. These are the only two options that have typically been considered in previous analyses. However, in the intermediate

[1] See footnote 10 for an exception to this.

range, $0.001 < a < 0.02$, the function $\ell(\tau)$ is not well approximated either by a constant or by τ. Then, rather than either discard the case or assume paternity, we need to compute and use its correct likelihood contribution, $\ell(\tau) \approx a + \tau$.

5 Bayesian network

The above algebraic approach is unsuited to the analysis of incomplete findings, such as summary information that a triplet is compatible. To handle such cases, in addition to full triplet data, we have explored an alternative route, involving the construction of a computer software system, built as a *Bayesian network* (Cowell *et al.* 1999), to calculate the appropriate likelihood expression numerically. The advantages of this are that the same system can be used to analyse any findings that one might wish to enter. In addition (Dawid *et al.* 2002) it is straightforward to extend such a system to allow for additional complexities, such as profiles measured on relatives of the putative father when he himself is unavailable, laboratory errors, silent alleles, *etc.*

A *Bayesian network* (BN), or *Probabilistic Expert System* (PES), is a joint graphical and numerical representation that allows the application of fast general-purpose computer software such as HUGIN[2]. For this paper we have used HUGIN Version 5.7: Dawid (2003) describes the use of the object-oriented Version 6 to simplify the tasks of network specification and construction.

In a Bayesian network, complex interrelationships are broken down into simple local dependencies, from which a full graphical representation can be built in a modular fashion. A genetic pedigree fits naturally into this framework: the family relationships constitute the local graphical modules, with the required conditional probability tables being simply specified, for example by Mendel's laws of inheritance, or by the logical relationship between a genotype and its constituent genes. The conditional independence properties encoded in a Bayesian network hold naturally for genetic inheritance, since a child's genes must be conditionally independent of all other (non-descendant) individuals' genes, given its parents' genes. We refer the reader to Dawid *et al.* (2002) for an introduction to and detailed description of the construction and use of Bayesian networks to analyse problems of forensic identification inference, and for further examples to Dawid and Evett (1997); Mortera (2003); Mortera *et al.* (2003).

5.1 Bayesian network representation of a paternity case

The Bayesian network in Figure 1, whose notation is summarised in Table 1, will be used to compute the contribution to the likelihood of the mutation parameters for a given marker, based on a case consisting of full or partial information on the genotypes of a specific mother-putative father-child triple. Each node in the BN represents a random variable relevant to such a case.

[2] `http://www.hugin.com`

We further need to specify, for each node, its probabilistic or deterministic dependence on its 'parents'.[3] We now describe these ingredients in detail.

5.1.1 Genotypes

For each genotype we introduce (unobservable) nodes representing its paternally and maternally derived genes: `mapg` and `mamg` for the mother, `pfapg` and `pfamg` for the putative father, and `capg` and `camg` for the child. Each of the "founder" nodes `mapg`, `mamg`, `pfapg` and `pfamg` is supplied with the population allele frequency distribution: we use the values in Table 2. The maternal genotype is now represented by a pair of nodes `mming: = min(mapg, mamg)`, `mmaxg := max(mapg, mamg)`, selecting the minimum and maximum values of the mother's parental genes. Similarly for the putative father and child genotypes, (`pfming`, `pfmaxg`) and (`cming`, `cmaxg`).

5.1.2 Inheritance

To model Mendelian inheritance, the child's original maternal gene `comg` is generated by drawing one of the mother's two genes, `mamg` and `mapg`, according to the value of the segregation indicator `segr`, which is a Bernoulli($\frac{1}{2}$) variable: `segr := Bin(1, 0.5)`. Thus `comg:= if(segr = 0, mapg, mamg)`[4], and similarly for the child's original (putative) paternal gene `copg`.

5.1.3 Non-paternity

To take account of possible non-paternity, we introduce a "query node" `tf=pf?`, a Boolean (*true/false*) indicator of whether or not the true father, `tf`, is identical with the putative father `pf`. This is assigned a uniform prior distribution; however this is merely a mathematical device, and should not be interpreted as implying an actual prior paternity probability of 50%.

5.1.4 Mutation

Maternal mutation can be incorporated as a process whereby the child's "original" maternal gene `comg` is probabilistically transformed into a new "actual" maternal gene `camg`, according to some specified mutation process. Similarly a mutation process for the (putative) paternal line is applied to `copg`, resulting in `capg1`.

The child's actual paternal gene `capg` is now taken as identical with `capg1` if the putative father is the true father; otherwise it is identical with `otherpg`, a new founder node, generated from the population gene-frequency distribution,

[3] The term 'parent' — in single quotes — is used in its graph-theoretic sense, meaning a node in the BN out of which an arrow into the given 'child' node originates. This must not be confused with biological parenthood!

[4] The HUGIN expression `x:= if(a = y, b, c)`, means: *if* `a = y` *then* `x := b`; *else* `x := c`.

derived from a "random man" unrelated to the putative father.[5] Thus `capg := if(tf=pf?, capg1, otherpg)`.

5.1.5 Scalar model

Suppose we use a scalar mutation model, as in § 3.4, with parameters ξ and ρ, represented in the network by nodes `xi` and `rho`. We can then generate `camg` from `comg` using the construction of § 3.2: we first generate a "possibly mutated gene" `mutmg` from `comg` using transition matrix S, and then choose between `mutmg` and `comg` according to the mutation indicator `mut_m`, where the probability that `mut_m` $= 1$ is $\lambda_M = (1 - \rho)\,\xi$. That is, `mut_m:= Bin(1, (1 - rho) * xi)`.[6] We proceed similarly for the paternal line, taking `mut_pf := Bin(1, rho * xi)`.

5.1.6 Mixed model

In the network described here we use the mixed model of § 3. For the allele distribution of Table 2, this has $\kappa = 0.193\,h + 0.807$. (In particular, for the SMM $\tau = \xi$, while for the proportional model $\tau = 0.807\,\xi$.)

First we generate a founder gene `newmg` from the population gene frequency distribution. This represents the outcome of a potential mutation under the proportional model. Next we generate a mutated maternal gene `1step` according to the SMM: `1step` := `comg` + `fcoin`, if `comg` is not equal to the minimum or maximum allele in its repertory. Here `fcoin` represents a "fair coin" toss, having probability $\frac{1}{2}$ for each of the two states $+1$ and -1. When `comg` is the minimum [maximum] allele value, we take `1step` := `comg` + [−]1.

To form `mutmg` we select `newmg` or `1step`, according as the outcome of a biased coin toss, `bcoin := Bin(1, h)`, is 0 or 1: thus `mutmg := if(bcoin = 1, 1step, newmg)`. The mixing parameter `h` is itself represented as a node in the network, so allowing us to set its value: we give it possible values 0, 0.5, 0.7, 0.9, and 1, and use the same values for setting the paternal fraction node `rho`.

We apply the method described in § 5.1.4 above to transform `mutmg` into the child's actual maternal gene, `camg`. A similar transformation of `copg` yields `capg1`.

For node `xi` we have chosen 43 states in the range 0 to 0.0198, assigning a fictitious uniform prior distribution over these states.

5.1.7 Pairs

Sometimes we need to deal with "defective" or "pair" cases, consisting of the child and just one of the parents. Care is then needed to ensure appropriate

[5] The network constructed in Dawid (2003) allowed for possible mutation of an original gene inherited from the "random man", leading, for a non-stationary model (*e.g.* the mixed model with $h \neq 0$), to a different distribution for the child's actual paternal gene under non-paternity. While there is no obvious logical reason to prefer one of these assumptions to the other, our current model is simpler.

[6] The rôle of the additional 'parent' node `switch` of `mut_m` in Figure 1 is explained in § 5.1.7.

modelling of the unobserved genotype. For example, if we just have data on the child and putative father, we would want to model the child's unobserved maternal gene `camg` as generated from the population gene frequency distribution. However as it stands the network would produce an actual gene that had undergone a mutation process, and when that process is non-stationary this would change the distribution. In order to avoid this we include an additional Boolean node `switch`, a 'parent' of `mut_m`. For analysing triplet data, this must be set to *false*, in which case the dependence of the binary mutation indicator `mut_m` on the mutation parameters `xi` and `rho` is exactly as described in § 5.1.4. However, for cases where the maternal genotype is missing we set `switch` to *true*, which has the effect of turning off the maternal mutation process by forcing `mut_m` to be 0. The full expression is `mut_m := Bin(1, if(switch, 0, (1 - rho) * xi))`.

We could similarly insert a switch node to turn off paternal mutation when the data on the putative father are absent. However this is unnecessary, since in this case we can bring about the same effect by setting `tf=pf?` to *false*, thus ensuring that `capg` is identical with `otherpg`, which is drawn, as required, direct from the population gene-frequency distribution.

When $h = 0$, the mutation process is stationary; in this special case the above switches have no effect and may be ignored.

5.2 Compatibility

The network as described so far is sufficient to analyse full genotype data on the individuals constituting a triple or pair. However, we wish also to be able to analyse summary information: for example whether or not the configuration of genotypes is compatible. In cases of incompatibility, this might be further classified by whether or not it could be explained by a single mutation on the paternal, or maternal, line. To allow entry of such summary evidence we need to create further nodes in the network.

5.2.1 Simple compatibility

To represent simple compatibility we introduce a Boolean node `comp`: this takes the value *true* when `camg` is either `mapg` or `mamg` and `capg` is either `pfapg` or `pfamg`; or *vice versa*.

As a first attempt, we added a single new node `comp`, having 'parents' `mapg`, `mamg`, `pfapg`, `pfamg`, `capg`, `camg`, and defined by the expression: `comp := or(and(or(camg = mapg, camg = mamg), or(capg = pfapg, capg = pfamg)), and(or(camg = pfapg, camg = pfamg), or(capg = mapg, capg = mamg)))`. However, this representation is computationally very inefficient. A measure of the complexity of a Bayesian network (Cowell *et al.* 1999) is its total clique-table size. This initial approach created a 7-node clique, containing the `comp` node with 2 states, and its 6 'parent' nodes which (for marker VWA) each have 11 states; the associated table thus had $11^6 \times 2 = 3{,}543{,}122$ entries. Optimal triangulation required a further 8-node clique of size 7,086,244, and the total network clique-table size was 10,793,106. To eliminate this complexity we split the single logical expression for `comp` into simpler logical constituents. Although this creates more cliques, they are much smaller, and the network becomes more computationally tractable: for marker VWA the largest clique-table size is now 234,256 (this occurring twice), and the total clique-table size becomes 640,552 — a 17-fold reduction.

As shown in Figure 1, simple compatibility is now described by means of seven Boolean nodes, representing the following three steps:

a) We first check the compatibility of the child's maternal gene with his/her mother's and putative father's genes. This is defined by the logical disjunctions: `cmg_m := or(camg = mamg, camg = mapg)`, and `cmg_pf := or(camg = pfamg, camg = pfapg)`. These expressions are *true* if either 'parent' node is *true* and *false* otherwise. Similarly, nodes `cpg_m` and `cpg_pf` check the compatibility of the child's paternal gene with his/her mother's and putative father's genes, respectively.

b) In the second step, we define `c1comp` and `c2comp` as the logical conjunctions: `c1comp := and(cmg_m, cpg_pf)`, and `c2comp := and(cpg_m, cmg_pf)`, which are *true* if both 'parent' nodes are *true*.

c) Finally, we define the overall compatibility by the Boolean node `comp := or(c1comp, c2comp)`, which is *true* if at least one 'parent' node is *true*.

We can now enter evidence of simple compatibility or incompatibility at node `comp`.

5.2.2 Pair compatibility

We can further elaborate the network to enable us to enter evidence on compatibility status for "pair" cases. Compatibility between mother [resp. putative father] and child is represented in Figure 1, by the Boolean node `comp_m` [resp. `comp_pf`], with 'parents' `cmg_m`, `cpg_m` [resp. `cmg_pf`, `cpg_pf`]. Each of these is *true* if either of its 'parent' nodes is *true*, else *false*. Thus `comp_m := or(cmg_m, cpg_m)`.

When entering such summary evidence on pairs, we also need to set appropriate switches, as described in § 5.1.7.

We can also use nodes `comp_m` and `comp_pf` to specify incompatibility status in more detail. Thus to represent a case of "paternal exclusion" we would enter `comp_m` = *true*, `comp_pf` = *false* (and possibly, but redundantly, `comp` = *false* also); while a case of ambiguous exclusion would be given by `comp` = *false*, `comp_m` = *true*, `comp_pf` = *true*.

6 Using the network

6.1 Computation of the likelihood

For any case, a single propagation in the Bayesian network for marker m allows us to calculate numerically the likelihood function $\ell(\xi)$, based on the findings in that case, for the mutation parameter $\xi \equiv \xi_m$.

We first enter the desired values of ρ and h at node `rho` and `h`, and then the findings on the case. In addition (see § 5.1.7) we set appropriate values of `switch` and (if appropriate) `tf=pf?`, according as we have triplet or pair data. When we have data on the putative father we also need to incorporate the prior probability π of paternity. This can be done by inserting evidence at node `tf=pf?` in the form of a likelihood vector proportional to $(1 - \pi, \pi)$ over the states (*false*, *true*). We now propagate, using the software, and interrogate the node ξ to obtain the posterior probabilities for its various states. Because the initial probabilities at these states were constant, these posterior probabilities are proportional to the desired likelihood function $\ell(\xi)$ for ξ, based on this case, evaluated at these points. The overall likelihood of ξ at these points is now obtained by multiplication across all cases.

Unlike the analysis of Dawid *et al.* (2001), this analysis is "exact" in that it allows for simultaneous mutation in both the paternal and the maternal germline. However, we do not directly obtain the continuous likelihood function: instead, we evaluate it at the specific values (43 in our implementation) chosen as states for node `xi` in the network. We can then fit a curve to these points to obtain, either exactly or approximately, the continuous likelihood contribution $\ell(\xi)$. This is particularly straightforward for a scalar model, when, according to § 4.2.1, $\ell(\xi)$ is exactly quadratic, and very close to linear.

6.2 Incompatible cases

The number of incompatible cases will typically be small (just 4 for marker VWA in the data-set we use): we enter the full genotype evidence on each of these and, using the network as described above, extract the corresponding likelihood contribution. Alternatively, we can bypass the network and use the algebraic approach of § 4.2.

6.3 Compatible cases

For any marker m the vast majority of cases considered will be compatible. In principle, given the full genotype data, we could analyse each one of these in detail, just as we do for the incompatible cases, finally multiplying all their contributions together. However, it may be convenient on account of the large number of such cases, or necessary because of unavailability of detailed genotype data, to enter for each case just the summary "compatibility" property, as described in § 5.2.1, and compute the corresponding likelihood contribution.

6.3.1 Fully compatible triplets

Many of the compatible cases will typically be triplets that are *fully compatible*, *i.e.* compatible on all markers. For these the appropriate prior probability $\pi^* := \pi^*_{\setminus m}$, although varying from case to case in the light of the findings on the other markers (and other evidence), will always be near 1. For a general scalar model, by equation (42) of Vicard and Dawid (2004) the likelihood contribution $\ell(\xi) = \mathrm{pr}(\mathrm{Comp} \mid \xi)$ from such a case will have the form

$$\ell(\xi) = 1 - \alpha\xi + O(\xi\epsilon) + O(\xi^2),$$

where $\epsilon := 1 - \pi^*$; so that, so long as π^* is very close to 1, we have

$$\ell(\xi) = 1 - \alpha\xi + o(\xi), \tag{13}$$

insensitive to the precise value of π^*. To the same order of approximation we may thus assume certain paternity:

$$\ell(\xi) = \mathrm{pr}(\mathrm{Comp} \mid \xi, P) + o(\xi).$$

Now $\mathrm{pr}(\mathrm{Comp} \mid \xi, P)$ can readily be computed from the network, as described in § 5.2 and § 6.1, simultaneously for each of the discrete set of states of ξ. We then evaluate α in (13) by fitting a linear relationship to these values. We note (Vicard and Dawid 2004, §5) that the probability of compatibility does not depend on ρ when paternity is assumed, and hence neither does α.

Let there be n_{FC} fully compatible cases. The overall likelihood contribution from these cases is

$$L_{FC}(\xi) \approx (1 - \alpha\xi)^{n_{FC}}. \tag{14}$$

Although we do not use it for computation, we note the further approximation to expression (14):

$$L_{FC}(\xi) \approx \exp\left(-\alpha\, n_{FC}\xi\right).$$

Although ξ is small, n_{FC} will typically be large enough that $n_{FC}\xi$ is appreciably non-zero — and hence non-constant — over the range of plausible values for ξ. Thus the likelihood contribution from fully compatible cases can not be ignored.

6.3.2 Locally compatible triplets

A further, typically small, number n_{LC} of the cases compatible on this marker m will be only *locally compatible*, *i.e.* incompatible on one or more of the other markers. Then $\pi^*_{\setminus m}$ is essentially 0, and the likelihood contribution from each such case is

$$\ell(\xi) = \mathrm{pr}(\mathrm{Comp} \mid \xi, N) + o(\xi). \tag{15}$$

Clearly $\mathrm{pr}(\mathrm{Comp} \mid \xi, N)$ only depends on the maternal mutation rate, $\mu_M = (1-\rho)\,\xi$, and will be of the form $\mathrm{pr}(\mathrm{Comp} \mid \xi, N) = 1 - \beta(1-\rho)\,\xi + o(\xi)$. We can calculate $\mathrm{pr}(\mathrm{Comp} \mid \xi, N)$ from the network, and β can thus be obtained in similar fashion to α above.[7] The total likelihood contribution from locally compatible cases is then

$$L_{LC}(\xi) \approx \{1 - \beta(1-\rho)\,\xi\}^{n_{LC}}. \tag{16}$$

There is information about ξ in locally compatible cases, although if, as is typical, $n_{LC} \ll n_{FC}$, this will be of a smaller order of magnitude than that in fully compatible cases.

6.3.3 Compatible pairs

For cases of mother-child or putative father-child compatible pairs, we can conduct an analysis essentially as above, but entering the appropriate findings at `comp_m` or `comp_pf`, and setting switches as described in § 5.1.7.

For a mother-child case we can without loss of generality always assume non-paternity, generating a likelihood contribution (accounting for possible hidden mutation) that we can compute exactly from the network, and that is expressible in the form

$$\ell(\xi) \approx 1 - (1-\rho)\,\gamma\,\xi + o(\xi), \tag{17}$$

this linear approximation being again excellent, with γ computable in similar fashion to α and β above.

For a putative father-child case, if there is an incompatibility at some other marker we can assume non-paternity — but that now leaves only the data on the child, which is clearly uninformative about mutation. Hence such cases can be discarded. Otherwise, there being no other incompatibility, it will again usually be an acceptable approximation to assume paternity. This case is then formally identical to a mother-child pair except for the interchange of maternal and paternal mutation rates. It will therefore generate a likelihood contribution expressible as

$$\ell(\xi) \propto 1 - \rho\,\gamma\,\xi + o(\xi), \tag{18}$$

with the same γ as in (17).

7 Analysis of data

We illustrate our methodology by applying it to the estimation of the total mutation rate τ for the STR marker VWA. The allele frequencies for this marker in the reference population are as given in Table 2. We use the mixed mutation model as described in § 3.3 and § 5.1.6, considering values 0 (proportional model), 0.5, 0.9 and 1 (SMM) for the mixing parameter h.

We analyse case data, comprising 2013 meioses, collected at the Institut für Rechtsmedizin. Information about these cases is largely taken from Brinkmann

[7] Note that the algebraic analysis of Vicard and Dawid (2004) for this case was incorrect, and its subsequent results are thus potentially inaccurate, since it applied an approximation valid only for small values of the parameter λ_P to the case $\lambda_P = 1$. Numerical propagation using HUGIN avoids this problem.

et al. (1998), with some variations and additional information for incompatible cases supplied by Brinkmann's laboratory. The cases are not a random selection from general casework, but were selected as those for which paternity was regarded as established with very high (≥ 0.9997) probability. We comment below on the likely biasing effects of such selection.

The total number of meioses is odd because some of the compatible cases are deficient, *i.e.*, pairs. Each triplet contributes 2 meioses, whereas a pair contributes only one.

There are four triplets exhibiting incompatibility at marker VWA. We analyse the detailed genotype data for these incompatible triplets. We could in principle similarly analyse complete case data for the compatible cases, using a batch processing interface to HUGIN to handle the large number of cases. However we did not have access to the detailed compatible case data. We have instead conducted an analysis using summary information only, as described in § 6.3.

7.1 Data assumptions

Traditional analysis of data such as these involves two stages. At the first stage, only those cases where paternity is regarded as firmly established are retained for further analysis. At stage 2, any incompatible case is counted as contributing one mutation event, further classified as a paternal or a maternal mutation, while compatibility is taken as equivalent to no mutation. If we assume paternal and maternal mutation rates equal, a naïve estimate of their common value is then given by the ratio of the number of incompatibilities to the total number of meioses. Because the way in which the meioses are divided between triples and pairs does not affect this analysis, detailed information on this division is frequently not provided; and this was the case for the data as supplied to us. However, the information is relevant for our own analysis, so that we were obliged to make some assumptions about it. These are described below. In § 7.5 we consider the sensitivity of our conclusions to variations in these assumptions.

We were informed that roughly 4% of meioses in the data are from deficient cases. For our illustrative analysis, we assume that we have 8 meioses from the 4 incompatible triplets, 1924 meioses from 962 compatible triplets, and a further 81 meioses from 81 compatible pairs. Among the 962 triplets compatible on VWA, 943 are fully compatible, and 19 only locally compatible.

According to Brinkmann *et al.* (1998), the overall ratio of maternal to paternal meioses is approximately 1. Where necessary, we assume there are 40 mother-child pairs and 41 putative father-child pairs, all these cases being fully compatible.

7.2 Compatible cases

Compatible cases carry information on hidden mutation. Uncertainty about paternity is irrelevant: as shown in § 6.3.1 and § 6.3.2, to the level of approximation at which we are working we can assume paternity for a fully compatible case, or non-paternity for a locally compatible case.

The approximate likelihood contribution from the summary information on compatible triplets is computed using formulae (14) for the fully compatible cases and (15) for the locally compatible cases, while that from the compatible pairs is obtained from (17) for mother-child cases and (18) for fully compatible putative father-child cases. The overall approximate likelihood term for all compatible cases is thus

$$(1-\alpha'\,\tau)^{943}\,\{1-\beta'\,(1-\rho)\,\tau\}^{19}\,\{1-\gamma'\,(1-\rho)\,\tau\}^{40}\,(1-\gamma'\,\rho\,\tau)^{41}\,. \qquad (19)$$

In (19) the linear approximations are expressed in terms of the parameter of interest τ rather than ξ. Thus $\alpha' = \alpha/\kappa$, $\beta' = \beta/\kappa$ and $\gamma' = \gamma/\kappa$, where α, β, γ are calculated from the network, as described in § 6.3. The coefficients α', β' and γ' are shown in Table 3 as functions of the mixing parameter h.[8] We see that the dependence on the specific mutation model, determined by h, is essentially ignorable.

For all our further analyses here we set the parameter ρ at 0.5 (in this case we do not need to know how the 81 compatible pairs cases divide into mother-child and putative father-child types). The overall likelihood function from compatible cases becomes

$$(1-\alpha'\tau)^{943}\,(1-0.5\,\beta'\tau)^{19}\,(1-0.5\,\gamma'\tau)^{81}\,.$$

7.3 Incompatible cases

The findings on VWA for the four incompatible cases are given in Table 4.

In contrast to the compatible cases, the assumed probability of paternity, π, does affect the information about τ contained in an incompatible case. These four triplets are compatible on all other markers. As shown in the Appendix, § A.4, it is thus appropriate to use, as the probability of paternity, $\pi^* := \pi^*_{\setminus VWA}$, based on the findings on the other markers. These values, reported in Table 5, are computed by means of the formula of Essen-Möller (1938), assuming that the probability of paternity before taking account of any of the DNA data was 0.5.[9]

For Case 4 we use $\pi^* = 0.9961$ as the probability of paternity. This is smaller than the true value, because it is based on only a subset of the other observed markers. We have done this for illustration, to investigate the actual amount of information in a case that would have been rejected on account of uncertain paternity.

[8]The "naïve" approach, which simply counts each compatible triplet as exhibiting no mutation, is equivalent to taking $\alpha' = \beta' = \gamma' = 1$; departure of these coefficients from 1 is due to taking proper account of hidden mutation. It is clear from Table 3 that this effect is important and would lead to the naïve estimates being strongly biased (Chakraborty *et al.* 1996; Vicard and Dawid 2004).

[9]This is a common default in paternity analyses, but will rarely be entirely appropriate. There is no problem in principle in using other starting values, to take account of the non-DNA evidence in each case. However, so long as such a starting value is not too close to 0 or 1 the effect on the value of π^* should be relatively small.

7.3.1 Individual cases

We first investigate, for the incompatible cases, the sensitivity of the associated likelihood contribution $\ell(\tau)$ to the assumed paternity probability π. From (12) we have $\ell(\tau) \propto a + \tau + o(\tau)$. For each incompatible case we evaluated the constant a for various values of π and h, using both the algebraic approach of § 4.2 and, as a check, numerical computation in the network. The value of a is to be interpreted according to the guidelines in § 4.2.2. Table 6 and Table 7 report the results for Cases 1 and 4; Cases 2 and 3 were very similar to Case 1.

Case 1 is a case of paternal exclusion, where either a 1-step or a 6-step paternal mutation could explain the incompatibility. From Table 6 we see that, for the proportional model, we have $a \leq 0.001$, effectively confirming paternity, so long as $\pi \geq 0.9995$; however $a \leq 0.02$, and thus the case is already informative about τ, as soon as π exceeds about 0.995. For the SMM and mixed models, the case is informative even for π around 0.9, although paternity is confirmed only for $\pi \geq 0.9995$. Thus Case 1 contains non-negligible information about τ for values of π that would normally lead to rejection of this case. Since the likelihood contribution $\ell(\tau) \approx a + \tau$ increases with τ, such a rejection would result in a downward bias in the mutation rate estimate.

Case 4 is a case of ambiguous exclusion, which can be explained under paternity either by a 3-step maternal mutation or by a 2-step paternal mutation. Under the SMM neither of these can happen and non-paternity becomes the only possible explanation, leading to an essentially uninformative likelihood contribution $\ell(\tau) \propto 1 - \frac{1}{2}\tau$.[10] Otherwise (see Table 7), the case is already informative for $\pi \geq 0.995$ (for the proportional model, and mixed model with $h = 0.5$) or $\pi \geq 0.9995$ (for $h = 0.9$); while paternity can be taken as established for $\pi \geq 0.9995$ (proportional model) or $\pi \geq 0.9997$ (mixed model, $h = 0.5$). However, under the mixed model with $h = 0.9$ we can not safely proceed as if paternity were established even when $\pi = 0.9997$. Intuitively this is because this model, which is close to the SMM, gives very low probability to the observed data under paternity, so favouring the alternative explanation of non-paternity — which is much less informative about mutation.

7.4 Overall likelihood

We now examine the sensitivity of the overall likelihood of τ (incorporating all cases, both compatible and incompatible) to the mutation model and the probabilities of paternity used for the four incompatible cases.

We consider the proportional and mixed ($h = 0.9$) models, together with the following choices for π :

(i). $\pi = 1$. This ignores any effect of non-paternity, although hidden mutation is correctly accounted for.

[10] This results from the implied finding that maternal mutation, which has rate $\frac{1}{2}\tau$, has not occurred in this case.

(ii). $\pi = 0.9997$. At this value paternity is normally taken as established, further analysis proceeding as for (i) above. We explore the accuracy of this approximation.

(iii). $\pi = \pi^*$. This makes correct use of the data on the other markers to assess the relevant probability of paternity.

(iv). $\pi = 0.995$. Such cases would normally be discarded. We investigate how much information on the mutation rate they contain.

The corresponding overall likelihoods for τ are shown in Figure 2 for the proportional and in Figure 3 for the mixed model. We see that, on introducing even a very small probability of non-paternity, smaller values of τ become more likely; the smaller π is, the more the curve shifts to the left. Intuitively this is because, as we decrease π, we can more easily explain away an incompatibility as due to non-paternity rather than mutation. The effect of this on the maximum likelihood estimate is shown in Table 8.

The value of π affects the overall likelihood differently for the proportional model and the mixed model. This is essentially due to the differences between the columns for $h = 0$ and $h = 0.9$ in Table 6 for Case 1 (with similar behaviour for Cases 2 and 3); and in Table 7 for Case 4.

Under either model, when $\pi = 0.9997$ the contribution from the first three cases was well approximated by τ, just as if paternity were established. However (see § 7.3.1), this approximation is poor for Case 4 under the mixed model with $h = 0.9$, on account of its intercept $a = 0.00393$. Allowing a very small chance that the observed incompatibility might be due to non-paternity has led to a flatter likelihood contribution. It is essentially just this effect for this one case that has caused the overall likelihood to move noticeably to the left under the mixed model.

Very similar considerations apply when we take $\pi = \pi^*$. For Case 4, the flattening effect of its low value $\pi^* = 0.9961$ is now seen even under the proportional model; while under the mixed model this case is effectively discarded as a non-paternity (paternity would require a mutation of at least 2 steps, which is very unlikely for this model).

When $\pi = 0.995$, under the proportional model all four incompatible cases have non-negligible intercept a. These flatter contributions result in a likelihood favouring very small values of τ. The behaviour is quite different under the mixed model. For Cases 1–3, the observed incompatibilities could be explained by one-step mutations, which are quite probable under this mutation model, to the extent that it becomes safe to assume paternity, together with mutation, for these cases; conversely Case 4, which can not now be easily explained by mutation, can be discarded as a non-paternity.

7.5 Pairs and triplets

Assuming full compatibility, the likelihood contribution from a pair of meioses is approximately $(1 - \alpha'\tau)$ if they belong to a single triplet, whereas it is $(1 -$

$\frac{1}{2}\gamma'\tau)^2 \approx (1-\gamma'\tau)$ if we are dealing with two pairs. Since (see Table 3) $\alpha' \neq \gamma'$, our conclusions will be sensitive to the way the total number of fully compatible meioses splits between triplets and pairs.

To explore this sensitivity, we assume that we have our 4 incompatible triplets and 19 locally compatible triplets, and compare various assumptions as to how the 1967 meioses from fully compatible cases divide among triplets and pairs:

(a) All triplets except for one pair. The associated likelihood contribution is

$$(1-\alpha'\tau)^{983}\,(1-0.5\,\gamma'\tau).$$

(b) All pairs, with likelihood contribution

$$(1-0.5\,\gamma'\tau)^{1967} \approx (1-\ \gamma'\tau)^{983}(1-0.5\,\gamma'\tau).$$

(c) 943 triplets and 81 pairs, as in § 7.1, with likelihood contribution

$$(1-\alpha'\tau)^{943}\,(1-0.5\,\gamma'\tau)^{81}.$$

For cases (a) and (b), Table 9 shows the maximum likelihood estimates of τ for the proportional model (with $\alpha' = 0.768$, $\gamma' = 0.548$) and the mixed model with $h = 0.9$ ($\alpha' = 0.788$, $\gamma' = 0.561$), for various choices of the probability of paternity π assumed for the 4 incompatible cases. Figure 4 compares the three full likelihood functions, using the mixed model and $\pi = 0.9997$.

Since $\gamma' < \alpha'$, when we take all the fully compatible meioses as resulting from pairs higher values of τ become more likely. From Table 9 we see that the estimates for this case are about 45% greater than when we take almost all of the compatible meioses to be from triplets; furthermore the likelihood (b) in Figure 4 shows a higher dispersion.

From Figure 4, and comparison of Table 9 with Table 8, we see that there is little difference between cases (a) and (c): so long as the number of pairs is only a small proportion of total meioses, it is not important to know it with precision. However, this will not be the case when that proportion is relatively large.

7.6 Using summary incompatibility data

Our analysis of compatible triplets and pairs was based on summary data only, without access to the detailed DNA profiles in these cases. Even when these are available, owing to the typically very large number of such cases it is tempting to base the analysis on compatibility alone, instead of exploiting the detailed genotype information. But one might nevertheless be concerned that important information, or at least precision, would thereby be lost.

In the absence of the data that would allow us to investigate this issue directly, we examine instead the effect of still further reduction, using now only summary information on the 4 incompatible cases. We shall however take into account the specific type of incompatibility, as in Vicard and Dawid (2004). Thus Cases 1 and 2 are paternal exclusions, *i.e.* (assuming paternity) incompatibility is unambiguously due to the paternal line; while Cases 3 and 4 are ambiguous exclusions, since incompatibility could be explained either by a paternal or by a maternal mutation.

Once again, the likelihood contribution from taking into account only the line-specific incompatibility type can be expressed in the form $\ell(\tau) \approx a+\tau+o(\tau)$. The corresponding intercept term a is given in Table 10 for the proportional and mixed ($h = 0.9$) models, and $\pi = 0.9997$ or π^*. These terms were calculated using HUGIN, on inserting evidence about the specific incompatibility as described in § 5.2.2. Cases 1 and 2 are now treated identically, as are Cases 3 and 4, except for their differing values for π^*.

For the proportional model, we could also apply equations (43) and (47) of Vicard and Dawid (2004), taking $\rho = 0.5$, $\tau = 0.807\xi$. This yields:

$$a = 0.807 \times 2 \times (1-\pi)/\pi \quad \text{for a paternal exclusion,} \tag{20}$$

$$a = 0.807 \times 2 \times (1-\pi)/(1+\pi) \quad \text{for an ambiguous exclusion.} \tag{21}$$

Even though (see footnote 7) these formulae are based on a potentially inaccurate approximation, the values obtained agree with those in Table 10 to the precision displayed.

We see from Table 10 that the mutation model assumed makes virtually no difference to inferences about mutation based on summary information. This is in contrast to what appears in Tables 6 and 7, and could be regarded as a pleasing robustness property of using summary information only. A contrary interpretation is that, if we do take seriously a specific mutation process, it could be misleading not to use the full genotype data.

Comparing the values for Case 1 with the corresponding entries of Table 6, using only the summary information makes no discernible difference if we can assume the proportional model, and this holds for Cases 2 and 3 also. There are more noticeable differences under the mixed model, but the value of a in all these cases nevertheless remains small enough that we can regard the likelihood as proportional to τ — equivalent to taking paternity as certain.

For Case 4 the difference becomes noticeable for the proportional model, and significant for the mixed model, especially for $\pi = \pi^*$ where we now obtain $a = 0.0030$ as against the value 0.0523 that resulted from full analysis. In this case it is necessary to see the full data to discover that the exclusion could only be explained by a 2-step mutation, which, as we saw, has the effect of downweighting the possibility of mutation to the extent that the case can be discarded as uninformative. However, this does not apply when only summary information is used: misleadingly, this yields an informative likelihood.

Figure 5 shows the overall likelihoods for τ computed with either (a) full or (b) summary data, using the mixed mutation model with $h = 0.9$ and paternity probability π^*. The maximum likelihood estimate for (b) is 0.00422, as compared with 0.00375 for (a). The difference between the curves is essentially due entirely to the treatment of Case 4. Compared with the full analysis, which effectively discards this case, the summary analysis yields a likelihood contribution that increases with τ, thus shifting the curve towards higher values.

Finally, we could consider the effect of summarising still further, noting merely the presence of an incompatibility, but not its type. In fact this makes no discernible difference to the value of the intercept a for the ambiguous Cases 3 and 4.[11] For the paternal exclusions, Cases 1 and 2, the effect of noting merely that there is an exclusion of some sort would (to a very close approximation) be to divide the value of a by 2.[12] Since the a values were already small enough to proceed as if paternity were established in these cases, the overall likelihood is not noticeably affected by such further reduction. However in general, because it has half the value for a, a nonspecific exclusion favours larger values of τ more than a paternal exclusion does, so that not using this fuller information could bias the inferences upwards.[13]

8 Discussion, extensions, recommendations

We have described how a Bayesian network can be used to make inferences about mutation rates from paternity casework, taking full account of a variety of logical subtleties and computational complexities, including in particular hidden mutation and uncertain paternity. This has been applied to an illustrative dataset.

We have not considered here such complications as more complex pedigrees, allelic dropout, measurement error or uncertainty in the population gene frequencies. In principle these could be incorporated in a Bayesian network description of the problem in a straightforward manner. We hope to pursue these and similar extensions in the future.

For simplicity, in our data analysis we fixed ρ, the paternal fraction of the total mutation rate, at 0.5 (so taking $\mu_M = \mu_P$). One could easily explore the sensitivity of inferences to the value assumed for ρ.[14] Our basic framework can also be used (see Appendix, § A.5) to make inferences about ρ, treated as an unknown parameter along with τ. Methodologically, this would require inserting an additional node into the network to represent the pair (ξ, ρ); alternatively an algebraic analysis could be undertaken. We have not conducted inference for ρ here: many more incompatible triplets would be needed to obtain a reliable estimate.[15] However if we could assume a common value of ρ across all markers

[11] For the proportional model this would follow directly on noting that, according to equations (45) and (46) of Vicard and Dawid (2004), ambiguous exclusion and nonspecific exclusion yield proportional likelihoods; and even though those equations have not been fully justified they appear numerically accurate.

[12] Again, this can be justified for the proportional model to the extent that equations (20) and (21) are accurate.

[13] Conversely, a maternal exclusion is certain evidence of mutation, and ignoring this specific information would bias inferences downwards.

[14] Brinkmann *et al.* (1998) interpret their data as showing a higher mutation rate in the paternal than in the maternal germline in a ratio about 17:3, corresponding to $\rho = 0.85$.

[15] For the analysis of Vicard and Dawid (2004) the information about ρ in the data was

the combined information could become appreciable. Evidence could also be combined across different studies.

We have seen that mutation estimates can be sensitive to the mutation process assumed, expressed in our case by the value of the mixing parameter h. With a little further elaboration we could treat h as another unknown parameter, and estimate it, too, from the same data. In particular we could then compare the relative support for the SMM ($h = 1$) as against the proportional model ($h = 0$). Similarly we could compare other distinct mutation models.

An issue deserving fuller investigation than we were able to give it in § 7.6 is how to quantify the loss of information when analysing only summary data, rather than full genotype information, for compatible cases. Although this does not lead to systematic biases, it can reduce the precision of parameter estimates, as well as affecting their values. When full data are available, the large number of cases could be analysed either numerically, using a suitable batch-processing interface to the Bayesian network, or by constructing a program to handle the algebra and related calculations as described in § 4.2.

We can not emphasise too strongly that our focus in this paper has been entirely methodological. Our data analysis is intended purely to illustrate this methodology, and its results must not be taken as substantively meaningful. This is because the paternity data-set we have analysed, like most of those used for mutation estimation, was not randomly selected from casework, but instead comprised only cases where the assessed probability of paternity was very close to 1. As we have observed in our treatment of Case 4 in § 7, there can be valuable information about mutation even in cases that would be rejected by this criterion, and ignoring this is likely to lead to biases in the estimates obtained. We therefore recommend retaining for analysis a much more complete collection of cases than usual. If it is not possible to analyse all, or a large random sample, of cases, then the choice of those to exclude should be based on an assessment (for a variety of realistic assumptions about the probability of paternity and the mutation model) of the constant a in (12), according to the criteria of § 4.2.2. Taking proper account of cases where the value of the probability of paternity is below the threshold commonly considered as confirming paternity should lead to more precise estimates of both τ and ρ — which could differ notably from those based on the biased sets of cases usually analysed.

In addition, the information we were able to use about compatible cases was very incomplete: we did not know exactly how many cases were triplets or maternal or paternal pairs, nor did we have detailed genotype information.

We hope that future work based on a complete analysis of full paternity casework will provide better estimates of mutation rates on specific markers. Analysis of casework data from various reference populations could enable valuable assessment of the variation of mutation rates across populations. But, prior to any analysis, however simple or sophisticated, an appropriate data collection,

shown to be on a par with the information about the bias of a coin that would be obtained from tossing it k times, where k is the number of unambiguous incompatibilities observed in the data — just 2 for VWA in our current data-set.

selection and recording process is essential if we are to obtain credible estimates of mutation parameters.

Acknowledgments

This research was supported by a Research Interchange Grant from the Leverhulme Trust. We are indebted to our collaborators in this research programme for helpful discussions. We are also grateful to Bernd Brinkmann and colleagues for information about their data.

A Appendix: The joint likelihood

In this Appendix we consider the joint likelihood for all mutation parameters, and show that, to a good approximation, it can be regarded as factorising into a number of terms, one for each marker, each term being readily evaluated using a suitable Bayesian network.

A.1 The likelihood

Suppose we have data on markers $1, \ldots, M$. Let ξ_m be a mutation parameter for marker m. We assume that $\xi_m = 0$ corresponds to the case of no mutation, and that ξ_m is small enough that we can neglect higher order terms. We initially assume that any other parameters (e.g. h, ρ) that might be present in the mutation model, as well as the population allele frequencies, are fixed, so that we can regard the totality of unknown parameters as $\boldsymbol{\xi} := (\xi_m : m = 1, \ldots, M)$. We consider the incorporation of additional parameters in § A.5 below.

Consider a single case, *e.g.* a putative family triplet or pair. We assume that no doubt attaches to the issue of maternity, but wish to take account of the possibility that the putative father may not be the true father. Denote by S the (unobserved) binary paternity indicator: $S = P$ if the putative father is the true father, N if not. We denote the prior probability of paternity before taking account of any of the DNA information, $\mathrm{pr}(P)$, by π_0, and assume $0 < \pi_0 < 1$.

Let the finding for this case on marker m be f_m: this might give details of the genotypes of each party, or only cruder information such as compatibility/non-compatibility. We write the totality of findings on the case as $\boldsymbol{f} := (f_m : m = 1, \ldots, M)$. We shall denote $(f_i : i \neq m)$ by $\boldsymbol{f}_{\setminus m}$, *etc.*

We make the following reasonable assumptions, expressed in the symbolism of conditional independence (Dawid 1979):

(a). By itself, the paternity status S is uninformative about mutation:

$$S \perp\!\!\!\perp \boldsymbol{\xi}.$$

(b). Given $\boldsymbol{\xi}$ and S, the markers behave independently:

$$\perp\!\!\!\perp \{f_m : m = 1, \ldots, M\} \mid \boldsymbol{\xi}, S.$$

(c). Each mutation parameter only affects its associated marker:

$$f_m \perp\!\!\!\perp \boldsymbol{\xi}_{\setminus m} \mid \xi_m, S.$$

It follows readily from assumptions (a)–(c) that the joint likelihood function $l(\boldsymbol{\xi}) = \mathrm{pr}(\boldsymbol{f} \mid \boldsymbol{\xi})$, based on this case, for the full parameter $\boldsymbol{\xi}$ is given by:

$$l(\boldsymbol{\xi}) = \pi_0 \prod_{m=1}^{M} \mathrm{pr}(f_m \mid \xi_m, P) + (1 - \pi_0) \prod_{m=1}^{M} \mathrm{pr}(f_m \mid \xi_m, N). \tag{A.1}$$

A.2 Factorisation

We seek an approximation for $l(\boldsymbol{\xi})$ of the following form:

$$l(\boldsymbol{\xi}) \propto \prod_{m=1}^{M} a_m(\xi_m) \, \{1 + o(\boldsymbol{\xi})\}. \tag{A.2}$$

We assume that $\mathrm{pr}(f_m \mid \xi_m, S)$ is a differentiable function of ξ_m for both $S = P$ and $S = N$: this is certainly so for a scalar mutation model, since then, by Lemma 4.1, this expression is quadratic. We further make the very mild additional assumption that the dth derivative from the right, $\mathrm{pr}^{(d)}(f_m \mid \xi_m, S)$, is non-zero at $\xi_m = 0$ for some $d \geq 0$ (again this clearly holds for any scalar mutation model). Since $\mathrm{pr}(f_m \mid \xi_m, S) > 0$ for $\xi_m > 0$, this derivative must be positive.

Definition A.1 Let $g(\cdot)$ be a non-negative real-valued function of $\boldsymbol{\xi} \in [0,1]^M$. We say that g possesses the *approximate factorisation property* if we can express it in the form

$$g(\boldsymbol{\xi}) = \prod_{m=1}^{M} a_m(\xi_m) \, \{1 + o(\boldsymbol{\xi})\}. \tag{A.3}$$

The following simple lemma gives a sufficient condition for the approximate factorisation property to hold.

Lemma A.1 *Let $g(\cdot)$ be differentiable at $\mathbf{0}$, with $g(\mathbf{0}) > 0$. Then g possesses the approximate factorisation property.*

Proof. We can expand $\log g$:

$$\log g(\boldsymbol{\xi}) = \log g(\mathbf{0}) + \sum_m b_m \xi_m + o(\boldsymbol{\xi}). \tag{A.4}$$

We thus have

$$g(\boldsymbol{\xi}) \propto \prod_m \left(e^{b_m \xi_m}\right) \{1 + o(\boldsymbol{\xi})\}, \tag{A.5}$$

displaying a factorisation of the desired form. $\square$

A.3 Factorisation of the likelihood

A.3.1 Compatibility with maternity

In a typical paternity case, for each m the finding f_m is compatible with the assumed mother-child relationship (assuming non-paternity), in the absence of mutation. Then $\mathrm{pr}(f_m \mid \xi_m = 0, N) \neq 0$, and so, since $\pi_0 < 1$, (A.1) then gives $l(\mathbf{0}) \neq 0$. We can thus apply Lemma A.1 to deduce that $l(\boldsymbol{\xi})$ possesses the approximate factorisation property.

A.3.2 Incompatibility with maternity

Now consider the contrary case that some finding f_j is *prima facie* incompatible with maternity, even assuming non-paternity. Such cases are rare — there are none in our dataset — but it is nonetheless important to be able to analyse them. In such a case $\mathrm{pr}(f_j \mid \xi_j = 0, S) = 0$ both for $S = P$ and for $S = N$. We still have $\mathrm{pr}(f_j \mid \xi_j, S) > 0$ for $\xi_j > 0$ and the first non-zero derivative (the d_S^jth, say) of this probability at $\xi_j = 0$ is positive. (For a scalar model, with $\mathrm{pr}(f_j \mid \xi_j, S)$ quadratic, d_S^j is 1, except for cases where the child is incompatible with each parent, when $d_P^j = 2$).

Let Q_m denote the probability that a random man has the same genotype as the putative father at marker m. Clearly $Q_m > 0$ and $\mathrm{pr}(f_m|\xi_m, N) \geq Q_m \,\mathrm{pr}(f_m|\xi_m, P)$.

For $\xi_m > 0$, define

$$k_m(\xi_m) := \frac{\mathrm{pr}(f_m|\xi_m, P)}{\mathrm{pr}(f_m|\xi_m, N)}, \tag{A.6}$$

the likelihood ratio in favour of paternity, based on the finding for marker m, taking into the possibility of mutation. Then we have

$$0 < k_m(\xi_m) \leq \frac{1}{Q_m}. \tag{A.7}$$

It readily follows that $d_N^m \leq d_P^m$, and then, by l'Hôpital's rule, we have that

$$\kappa_m := \lim_{\xi_m \downarrow 0} k_m(\xi_m) = \frac{\mathrm{pr}^{(d_N^m)}(f_m|\xi_m = 0, P)}{\mathrm{pr}^{(d_N^m)}(f_m|\xi_m = 0, N)} \tag{A.8}$$

exists and is finite; and, further, that, on defining $k_m(0) = \kappa_m$, the function $k_m(\cdot)$ is then differentiable at 0.

From (A.1) we have, for all $\boldsymbol{\xi}$,

$$l(\boldsymbol{\xi}) = (1 - \pi_0) \prod_m \mathrm{pr}(f_m \mid \xi_m, N) \times g(\boldsymbol{\xi}), \tag{A.9}$$

with

$$g(\boldsymbol{\xi}) := 1 + \left(\frac{\pi_0}{1 - \pi_0}\right) \prod_m k_m(\xi_m). \tag{A.10}$$

By assumption, $g(\cdot)$ is finite, positive and differentiable at $\mathbf{0}$, and so, by Lemma A.1, has the approximate factorisation property. From (A.9) the same property then follows for $l(\cdot)$.

A.4 Identification of factors

We now suppose that $l(\boldsymbol{\xi})$, defined by (A.1), has the approximate factorisation property (A.2), and identify the leading terms in it.

From (A.1) and (A.2), for any m we have a likelihood factorisation of the form

$$\begin{aligned}
&\pi_0 \operatorname{pr}(\boldsymbol{f}_{\setminus m} \mid \boldsymbol{\xi}_{\setminus m}, P) \operatorname{pr}(f_m \mid \xi_m, P) + (1-\pi_0) \operatorname{pr}(\boldsymbol{f}_{\setminus m} \mid \boldsymbol{\xi}_{\setminus m}, N) \operatorname{pr}(f_m \mid \xi_m, N) \\
&\quad = a_m(\xi_m)\, a_{\setminus m}(\boldsymbol{\xi}_{\setminus m})\, \{1 + o(\boldsymbol{\xi})\}. \qquad \text{(A.11)}
\end{aligned}$$

Assume first that $a_{\setminus m}(\mathbf{0}) \neq 0$: this will be the case if all the findings on markers other than m are compatible with maternity, so that $\operatorname{pr}(\boldsymbol{f}_{\setminus m} \mid \boldsymbol{\xi}_{\setminus m} = \mathbf{0}, N)$, and thus also $\operatorname{pr}(\boldsymbol{f}_{\setminus m} \mid \boldsymbol{\xi}_{\setminus m} = \mathbf{0})$, is positive. Setting $\boldsymbol{\xi}_{\setminus m} = \mathbf{0}$ in (A.11) we see that we can take

$$a_m(\xi_m) \propto \operatorname{pr}^*_{\setminus m}(f_m \mid \xi_m) := \pi^*_{\setminus m} \operatorname{pr}(f_m \mid \xi_m, P) + (1 - \pi^*_{\setminus m}) \operatorname{pr}(f_m \mid \xi_m, N), \qquad \text{(A.12)}$$

where

$$\begin{aligned}
\pi^*_{\setminus m} &= \frac{\pi_0 \operatorname{pr}(\boldsymbol{f}_{\setminus m} \mid \boldsymbol{\xi}_{\setminus m} = \mathbf{0}, P)}{\pi_0 \operatorname{pr}(\boldsymbol{f}_{\setminus m} \mid \boldsymbol{\xi}_{\setminus m} = \mathbf{0}, P) + (1-\pi_0) \operatorname{pr}(\boldsymbol{f}_{\setminus m} \mid \boldsymbol{\xi}_{\setminus m} = \mathbf{0}, N)} \\
&= \operatorname{pr}(P \mid \boldsymbol{\xi}_{\setminus m} = \mathbf{0}, \boldsymbol{f}_{\setminus m}), \qquad \text{(A.13)}
\end{aligned}$$

the posterior probability of paternity based on the findings on all markers other than m, on the assumption of the mutation rate being zero for these. That is, $\operatorname{pr}^*_{\setminus m}(f_m \mid \xi_m)$ is the marginal probability of obtaining the finding f_m, as a function of the mutation parameter ξ_m, when the "prior" probability of paternity is taken as $\pi^*_{\setminus m}$.

Whenever there is some marker j on which the finding f_j is incompatible with paternity, then, for any other marker m, $\pi^*_{\setminus m} = 0$, so that we can just take $a_m(\xi_m) = \operatorname{pr}(f_m \mid \xi_m, N)$, *i.e.* we can assume non-paternity (when there are two or more such incompatible markers, we can assume non-paternity in analysing any marker). This is exactly equivalent to just having mother-child data. Even when, as will typically hold, mother and child are compatible, such cases should not be ignored, since they indicate that a possible mutation in the maternal line has *not* occurred, and so constitute evidence in favour of smaller values for maternal mutation rates (this is only an indication, on account of the possibility of "hidden mutation": the correct likelihood contribution can be obtained from a network propagation).

When all markers other than m exhibit compatibility, then for complete triplet data $\pi^*_{\setminus m}$ is readily calculated by the formula of Essen-Möller (1938); for reduced findings, such as simple compatibility/non-compatibility, $\pi^*_{\setminus m}$ could be obtained by using a collection of propagations in suitable networks, one for each other marker, to calculate likelihood ratios in favour of paternity (ignoring mutation) after entering the reduced evidence.

In case $a_{\setminus m}(\mathbf{0}) = 0$ (*i.e.* we have maternal incompatibility on one or more of the other markers), we can proceed as above, except that instead of setting

$\boldsymbol{\xi}_{\setminus m} = \mathbf{0}$ in (A.13) we take a limit as $\boldsymbol{\xi}_{\setminus m} \downarrow \mathbf{0}$. We obtain

$$\pi^*_{\setminus m} = \frac{\pi_0 \, \kappa_{\setminus m}}{\pi_0 \, \kappa_{\setminus m} + (1 - \pi_0)}, \tag{A.14}$$

where $\kappa_{\setminus m} := \prod_{j \neq m} \kappa_j$, with κ_j (assumed to exist) defined as in (A.8). We will typically have $\kappa_j > 0$ if, assuming paternity, the finding f_j can be fully explained by maternal mutation, as for the triple: mother AB, putative father AC, child CD, where A, B, C, D represent distinct alleles. We shall have $\kappa_j = 0$ if, exceptionally, under paternity f_j can only be explained by mutations in both germlines, as for the triple: mother AB, putative father AC, child DE. Whenever this happens for some marker j, (A.14) gives $\pi^*_{\setminus m} = 0$ for all $m \neq j$, so that we can assume non-paternity when analysing the other markers. And again, if it happens for two or more different markers, we can always assume non-paternity.

Once $\pi^*_{\setminus m}$ is found, calculation of the approximate likelihood term $\mathrm{pr}^*_{\setminus m}(f_m \mid \xi_m)$ need only take account of the structure and model for the findings on marker m. As described in the body of the paper, this can typically be accomplished by means of probability propagation in a suitable network constructed for marker m.

We have thus shown that, whenever the joint likelihood $l(\xi)$ has the approximate factorisation property, it can be explicity represented as the approximate product:

$$l(\boldsymbol{\xi}) \propto \prod_{m=1}^{M} l_m(\xi_m) \times \{1 + o(\boldsymbol{\xi})\}, \tag{A.15}$$

with

$$l_m(\xi_m) := \mathrm{pr}^*_{\setminus m}(f_m \mid \xi_m). \tag{A.16}$$

A.5 Additional parameters

We here turn to consider the case that there are additional parameters, θ say, in the mutation model. For example we might have $\theta = (h_m, \rho_m : m = 1, \ldots, M)$, with h_m the mixing parameter in the mixed mutation model, and ρ_m the proportion of total mutation attributable to the male line, for marker m. (We might assume further structure, such as $\rho_m = \rho$, all m.)

We must now replace (A.15) by

$$l(\boldsymbol{\xi}, \theta) = a(\theta) \prod_{m=1}^{M} l_m(\xi_m, \theta) \times \{1 + o(\boldsymbol{\xi})\}. \tag{A.17}$$

We want conditions under which $a(\theta)$ does not depend on θ, so that we can represent the joint likelihood in the form

$$l(\boldsymbol{\xi}, \theta) \propto \prod_{m=1}^{M} l_m(\xi_m, \theta) \times \{1 + o(\boldsymbol{\xi})\}. \tag{A.18}$$

When this holds, and further $\theta = (\theta_1, \ldots, \theta_M)$, with θ_m only affecting the mutation process for marker m, we will have a complete likelihood separation:

$$l(\boldsymbol{\xi}, \theta) \propto \prod_{m=1}^{M} l_m(\xi_m, \theta_m) \times \{1 + o(\boldsymbol{\xi})\}. \tag{A.19}$$

We shall impose the following requirement.

Condition 1 *When the mutation parameter ξ_m is 0, the probability of any findings f_m on marker m does not depend on θ, under both paternity P and non-paternity N.*

This holds for $\theta = (h_m, \rho_m : m = 1, \ldots, M)$ as above, but would not hold if, for example, θ were to include unknown allele frequencies.

A.6 Maternal compatibility

By Condition 1, both $l(\boldsymbol{\xi}, \theta) = \mathrm{pr}(\boldsymbol{f} \mid \boldsymbol{\xi}, \theta)$ and $l_m(\xi_m, \theta) = \mathrm{pr}^*_{\setminus m}(f_m \mid \xi_m, \theta)$ do not depend on θ when $\boldsymbol{\xi} = \mathbf{0}$. So long as the case findings $\boldsymbol{f}$ are compatible with maternity, so that $l(\mathbf{0}, \theta) > 0$, we can set $\boldsymbol{\xi} = \mathbf{0}$ in (A.17) to deduce that $a(\theta)$ is a constant, independent of θ. In this case we thus have a factorisation of the joint likelihood as in (A.18).

A.7 Maternal incompatibility

For a case of maternal incompatibility, as in § A.3.2, an additional condition is needed. We first introduce $K(\theta) := \prod_m \kappa_m(\theta)$, with $\kappa_m(\theta)$ defined as in (A.8).

Lemma A.2 *We shall have a likelihood factorisation of the form (A.18) if and only if $K(\theta)$ does not depend on θ.*

Proof. We have

$$\begin{aligned} l(\boldsymbol{\xi}, \theta) &= (1 - \pi_0) \prod_m \mathrm{pr}(f_m \mid \xi_m, \theta, N) \times \left(1 + \frac{\pi_0}{1 - \pi_0} \prod_m k_m(\xi_m, \theta)\right) \\ l_m(\xi_m, \theta) &= (1 - \pi^*_{\setminus m}) \, \mathrm{pr}(f_m \mid \xi_m, \theta, N) \times \left(1 + \frac{\pi^*_{\setminus m}}{1 - \pi^*_{\setminus m}} k_m(\xi_m, \theta)\right). \end{aligned}$$

Inserting into (A.17), cancelling the positive term $\prod_m \mathrm{pr}(f_m \mid \xi_m, \theta, N)$ and letting $\boldsymbol{\xi} \downarrow \mathbf{0}$ we obtain

$$(1 - \pi_0) \left(1 + \frac{\pi_0}{1 - \pi_0} K(\theta)\right) = a(\theta) \prod_m (1 - \pi^*_{\setminus m}) \prod_m \left(1 + \frac{\pi^*_{\setminus m}}{1 - \pi^*_{\setminus m}} \kappa_m(\theta)\right). \tag{A.20}$$

But for each marker m

$$\frac{\pi^*_{\backslash m}}{1-\pi^*_{\backslash m}}\kappa_m(\theta) = \frac{\pi_0}{1-\pi_0}K(\theta), \tag{A.21}$$

since each side gives the posterior odds on paternity given the findings on all markers when there is no mutation. We deduce

$$a(\theta) = \frac{1-\pi_0}{\prod_m (1-\pi^*_{\backslash m})}\left(1+\frac{\pi_0}{1-\pi_0}K(\theta)\right)^{-(M-1)}. \tag{A.22}$$

Since by Condition 1 $\pi^*_{\backslash m}$ does not depend on θ the result follows. □

A.7.1 Scalar model

Now suppose we are using a scalar model as in §§ 3.1 and 3.4, and that the additional parameter is $\theta = (\rho_m : m = 1, \ldots, M)$.

For a specific marker with parameters (ρ, ξ), suppose we have detailed genotype findings f given by: mother AB, putative father CD, child EF. We allow some of the constituent genes A, B, C, D, E, F to be identical, except we require that there be no overlap between AB and EF — the case of maternal incompatibility.

Using (10) and (11), the quantity $k \equiv k_m$ of (A.6) is

$$k = \frac{(q^M_{A\to E} + q^M_{B\to E})(q^P_{C\to F} + q^P_{D\to F}) + (q^M_{A\to F} + q^M_{B\to F})(q^P_{C\to E} + q^P_{D\to E})}{2\{(q^M_{A\to E} + q^M_{B\to E})\,p_F + (q^M_{A\to F} + q^M_{B\to F})\,p_E\}}. \tag{A.23}$$

Because of maternal incompatibility we have $i \neq j$, and so $q^M_{i\to j} = (1-\rho)\,\xi\, s^M_{i\to j}$, for all the $q^M_{i\to j}$ terms entering into (A.23). Thus

$$k = \frac{(s^M_{A\to E} + s^M_{B\to E})(q^P_{C\to F} + q^P_{D\to F}) + (s^M_{A\to F} + s^M_{B\to F})(q^P_{C\to E} + q^P_{D\to E})}{2\{(s^M_{A\to E} + s^M_{B\to E})\,p_F + (s^M_{A\to F} + s^M_{B\to F})\,p_E\}}. \tag{A.24}$$

Since each $q^P_{i\to j}$ is a linear function of $\lambda_P = \rho\,\xi$, with its constant and slope coefficients depending only on the $(s_{i\to j})$, so is k. We readily deduce that $\kappa = \lim_{\xi\downarrow 0} k(\xi)$ is independent of ρ (though it may depend on other parameters in the mutation model, such as the parameter h in the mixed model of § 3.3). It thus follows from Lemma A.2 that we again have a full likelihood factorisation of the form (A.18).

This result extends, by suitable mixing, to coarser findings that imply maternal incompatibility.

A.8 The overall likelihood

The overall joint likelihood function $L(\boldsymbol{\xi})$, based on all cases in the data-set, is the product of all the various case-specific joint likelihood terms, $l(\boldsymbol{\xi})$. Assuming we have extensive data that are reasonably consistent with existing knowledge of the range of plausible values for $\boldsymbol{\xi}$, $L(\boldsymbol{\xi})$ will be negligible unless each $\xi_m < 0.01$ (roughly). Within this range the approximation which ignores the $o(1)$ term in (A.2) is likely to be excellent, so that the overall likelihood can likewise be regarded as factorising in this region; outside this range both the correct and the approximate overall likelihood are in any case negligible. Even with non-extensive or untypical data, any realistic prior distribution would serve to discount values for ξ outside this region. Consequently it should be safe to proceed as if we had a factorisation

$$L(\boldsymbol{\xi}) \propto \prod_m L_m(\boldsymbol{\xi}), \tag{A.25}$$

with each marker-specific factor $L_m(\boldsymbol{\xi})$ obtained as a product, over the cases in the data-set, of the terms $l_m(\xi_m)$, given by (A.16). For each case the factor $l_m(\xi_m)$ can be found by treating the finding f_m on marker m as if it were the full data available, after first updating the paternity probability from π_0 to $\pi^*_{\backslash m}$, to take account of the findings on the other markers.

If we assume that there are no *a priori* relations between the mutation rates across markers, inference for the different rates can thus proceed completely separately, each based on its own likelihood factor of the form L_m.

References

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J., and Rolf, B. (1998). Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat. *American Journal of Human Genetics*, **62**, 1408–15.

Chakraborty, R., Stivers, D. N., and Zhong, Y. (1996). Estimation of mutation rates from parentage exclusion data: Applications to STR and VNTR loci. *Mutation Research*, **354**, 41–8.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer, New York.

Dawid, A. P. (1979). Conditional independence in statistical theory (with Discussion). *Journal of the Royal Statistical Society, Series B*, **41**, 1–31.

Dawid, A. P. (2003). An object-oriented Bayesian network for estimating mutation rates. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Jan 3–6 2003, Key West, Florida,* (ed. C. M. Bishop and B. J. Frey). `http://tinyurl.com/39bmh`.

Dawid, A. P. and Evett, I. W. (1997). Using a graphical method to assist the evaluation of complicated patterns of evidence. *Journal of Forensic Sciences*, **42**, 226–31.

Dawid, A. P., Mortera, J., and 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., and van Boxel, D. W. (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics*, **29**, 577–95.

Di Rienzo, A., Donnelly, P., Toomajiana, C., Siska, B., Hillc, A., Petzl-Erlerd, M. L., Hainese, G. K., and Barchf, D. H. (1998). Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. *Genetics*, **148**, 1269–84.

Durret, R. and Kruglyak, S. (1999). A new stochastic model of microsatellite evolution. *Journal of Applied Probability*, **36**, 621–31.

Egeland, T. and Mostad, P. F. (2002). Statistical genetics and genetical statistics: A forensic perspective. *Scandinavian Journal of Statistics*, **29**, 297–308.

Essen-Möller, E. (1938). Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis. Theoretische Grundlagen. *Mitteilungen der Anthropologischen Gesellschaft*, **68**, 9–53.

Fimmers, R., Henke, L., Henke, J., and Baur, M. (1992). How to deal with mutations in DNA testing. In *Advances in Forensic Haemogenetics 4*, (ed. C. Rittner and P. M. Schneider), pp. 285–7. Springer-Verlag, Berlin.

Geada, H., Viriato, L., Vieira-Silva, C., Cruz, C., Lucas, I., Ribeiro, T., and Espinheira, R. (2003). STR mutations in paternity investigations: A study of 1-year consecutive cases. In *Progress in Forensic Genetics 9*, International Congress Series, Vol. 1239, (ed. B. Brinkmann and A. Carracedo), pp. 657–60.

Henke, L. and Henke, J. (1999). Mutation rate in human microsatellites. *American Journal of Human Genetics*, **64**, 1473. With reply by B. Rolf and B. Brinkmann, 1473–4.

Mortera, J. (2003). Analysis of DNA mixtures using Bayesian networks. In *Highly Structured Stochastic Systems*, (ed. P. J. Green, N. L. Hjort, and S. Richardson), chapter 1B, pp. 39–44. Oxford University Press.

Mortera, J., Dawid, A. P., and Lauritzen, S. L. (2003). Probabilistic expert systems for DNA mixture profiling. *Theoretical Population Biology*, **63**, 191–205.

Ohta, T. and Kimura, M. (1973). A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. *Genetical Research*, **22**, 201–4.

Sajantila, A., Lukka, N., and Syvanen, A. C. (1999). Experimentally observed germline mutations at human micro- and minisatellite loci. *European Journal of Human Genetics*, **7**, 263–6.

Valdes, A. M., Slatkin, M., and Freimer, N. B. (1993). Allele frequencies at microsatellite loci: The stepwise mutation model revisited. *Genetics*, **133**, 737–49.

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

Weber, J. L. and May, P. E. (1989). Abundant classes of human DNA polymorphism which can be typed by the polymerase chain reaction. *American Journal of Human Genetics*, **44**, 388–97.

Whittaker, J. C., Harbord, R. M., Boxall, N., Mackay, I., Dawson, G., and Sibly, R. M. (2003). Likelihood-based estimation of microsatellite mutation rates. *Genetics*, **164**, 781–7.

Wilson, I. J., Weale, M. E., and Balding, D. J. (2003). Inference from DNA data: Population histories, evolutionary processes and forensic match probabilities (with Discussion). *Journal of the Royal Statistical Society, Series A*, **166**, 155–201.

Notation	Definition
`pf`	putative father
`tf`	true father
`m`	mother
`c`	child
`mg`	maternal gene
`pg`	paternal gene
`amg` [`apg`]	actual maternal [paternal] gene
`omg` [`opg`]	original maternal [paternal] gene
original maternal gene	
`mming`, *etc.*	mother's minimum gene, *etc.*
`mmaxg`, *etc.*	mother's maximum gene, *etc.*
`tf=pf?`	query: "true father = putative father?"
`newmg`, `newpg`, `otherpg`	new founder genes with population gene frequency distribution
`segr`	segregation indicator, a fair coin toss representing Mendel's law
`fcoin`	fair coin toss
`bcoin`	biased coin toss
`1step`	1-step mutated gene
`mutmg` [`mutpg`]	mutated maternal [paternal] gene
`mut_m` [`mut_pf`]	maternal [putative paternal] mutation indicator
`h`	mixing parameter for mutation model
`rho`	ratio of paternal to total mutation rate
`xi`	total mutation parameter ξ (proportional to total mutation rate τ)
`comp`	compatibility of triplet genotypes
`comp_m` [`comp_pf`]	mother [putative father] compatibility with child
`cmg_m` [`cpg_m`], *etc.*	child compatibility with mother's [putative father's] maternal gene, *etc.*
`switch`	used to block mutation in unobserved mother

Table 1: Glossary

allele	frequency
11.22	0.0003
13	0.0018
14	0.1009
15	0.1004
16	0.1949
17	0.2834
18	0.2162
19	0.0866
20	0.0137
21	0.0015
22	0.0003

Table 2: Probability table for founder gene nodes, based on allele frequencies for marker VWA in a German-Austrian reference population (supplied by B. Brinkmann)

	h			
	1	0.9	0.5	0
α'	0.790	0.788	0.780	0.768
β'	0.620	0.616	0.607	0.590
γ'	0.561	0.561	0.557	0.548

Table 3: Coefficients α', β' and γ' for different values of h

	Child	Mother	Putative father
Case 1	18 20	18 19	14 19
Case 2	16 18	16 17	17 19
Case 3	18 19	17 18	16 18
Case 4	15 18	18 18	17 18

Table 4: Findings on incompatible cases for marker VWA

	Case 1	Case 2	Case 3	Case 4
π^*	0.99984	0.99959	0.99920	0.99612

Table 5: Prior probability π^* of paternity in the four incompatible cases

	h			
	1	0.9	0.5	0
π	(SMM)			(Prop. model)
0.9	0.01165	0.01263	0.02002	0.18207
0.95	0.00565	0.00612	0.00969	0.08556
0.995	0.00055	0.00060	0.00094	0.00812
0.9995	0.00005	0.00006	0.00009	0.00081
0.9997	0.00003	0.00004	0.00006	0.00048
0.99984	0.00002	0.00002	0.00003	0.00026
1	0	0	0	0

Table 6: Intercept a of the likelihood contribution $\ell(\tau) \approx a+\tau$ from incompatible Case 1

	h		
	0.9	0.5	0
π			(Prop. model)
0.9	4.98779	0.30287	0.12480
0.95	1.03600	0.13419	0.05778
0.995	0.06788	0.01217	0.00542
0.99612	0.05228	0.00943	0.00420
0.9995	0.00656	0.00121	0.00054
0.9997	0.00393	0.00072	0.00032
1	0	0	0

Table 7: Intercept a of the likelihood contribution $\ell(\tau) \approx a+\tau$ from incompatible Case 4

π	Proportional model	Mixed model
1	0.0053	0.0052
0.9997	0.0049	0.0044
π^*	0.0042	0.0038
0.995	0	0.0022

Table 8: Maximum likelihood estimates of τ for the proportional and mixed ($h = 0.9$) mutation models

π	All triplets and one pair Prop. model	Mixed model	All pairs Prop. model	Mixed model
1	0.0052	0.0051	0.0073	0.0072
0.9997	0.0049	0.0044	0.0069	0.0064
π^*	0.0042	0.0037	0.0062	0.0053

Table 9: Maximum likelihood estimates of τ when the fully compatible cases are constituted by (a) all triplets and one pair, (b) all pairs

π	Mixed model Case 1	Case 2	Case 3	Case 4	Proportional model Case 1	Case 2	Case 3	Case 4
0.9997	0.00047	0.00047	0.00023	0.00023	0.00048	0.00048	0.00024	0.00024
π^*	0.00025	0.00065	0.00062	0.00302	0.00026	0.00066	0.00065	0.00315

Table 10: Intercept a of the likelihood contribution $\ell(\tau) \approx a + \tau$ from incompatible cases when only the specific type of incompatibility is considered

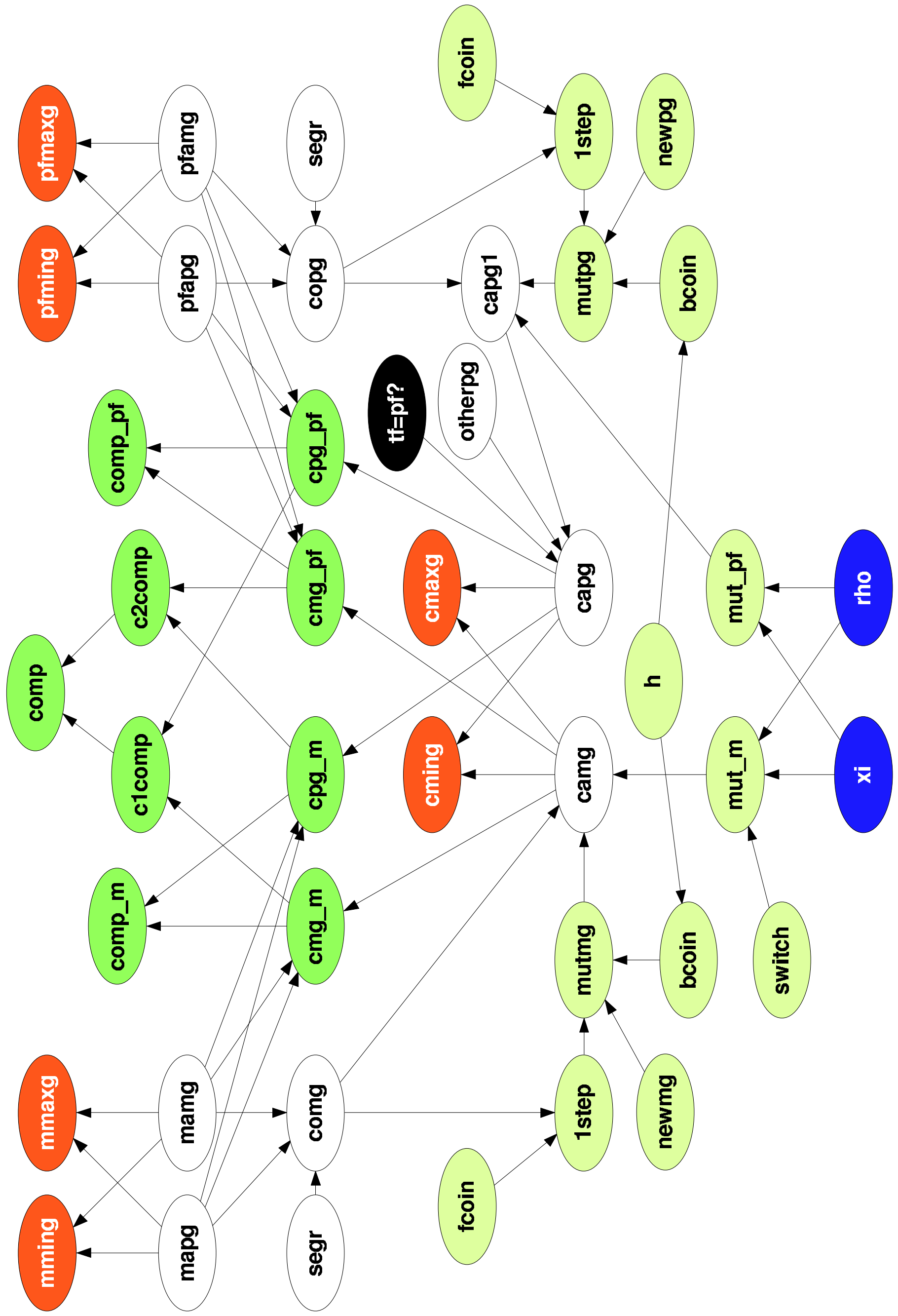

Figure 1: Bayesian network for mutation

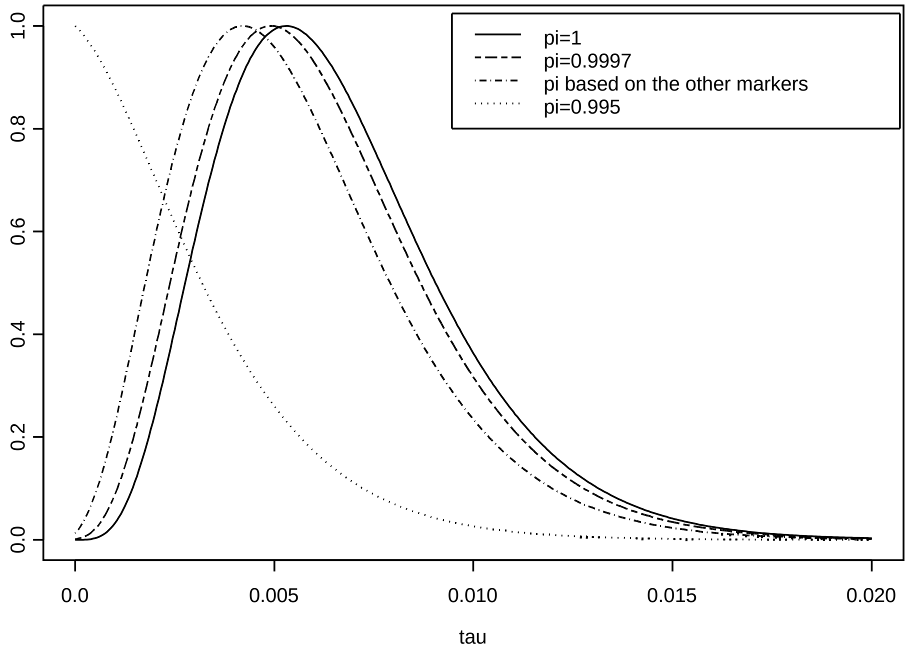

Figure 2: Likelihood for τ, proportional model

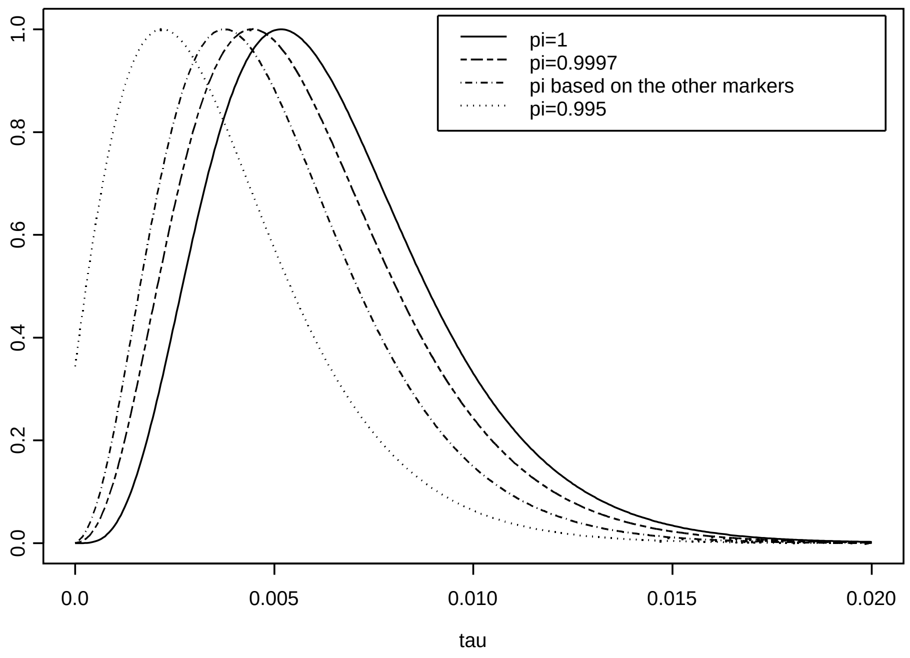

Figure 3: Likelihood for τ, mixed ($h = 0.9$) model

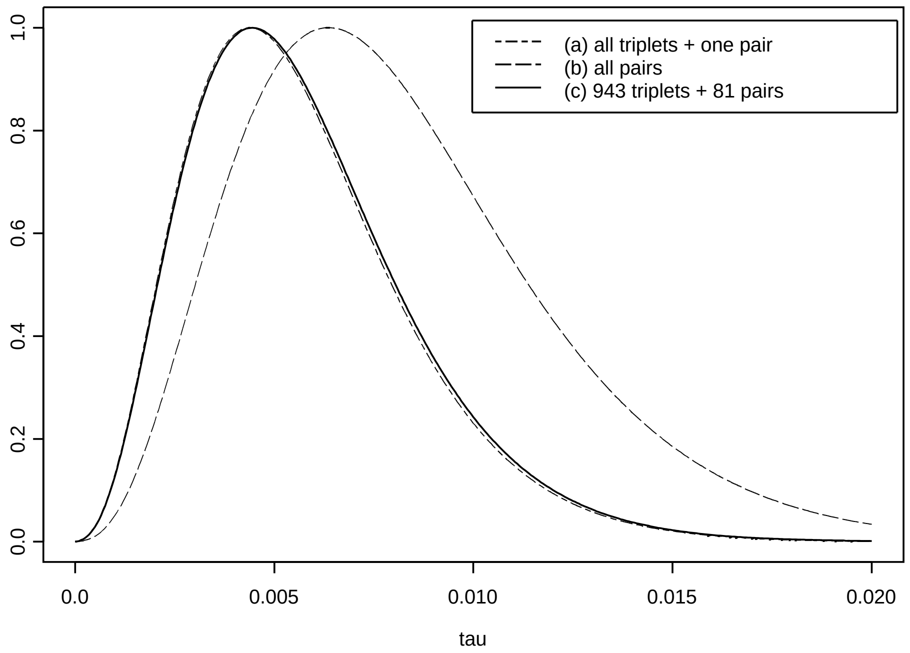

Figure 4: Likelihood for mixed ($h = 0.9$) model with π=0.9997 when the fully compatible cases consist of (a) all triplets and one pair, (b) all pairs, (c) 943 triplets and 81 pairs

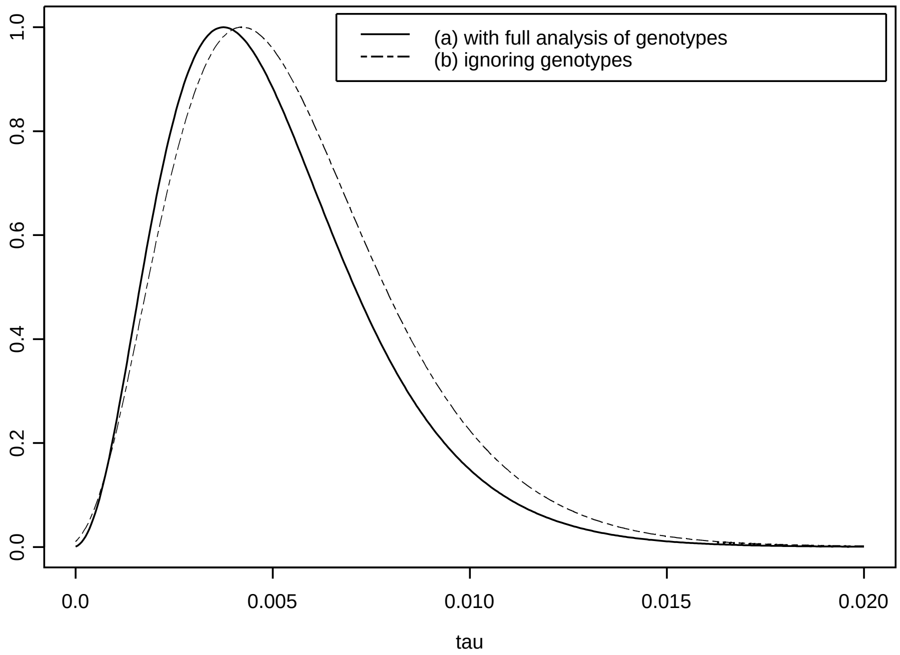

Figure 5: Likelihood for τ, mixed ($h = 0.9$) model with $\pi = \pi^*$, with (a) full analysis of incompatible cases, (b) summary analysis using only the incompatibility type