

New statistical approaches for estimating mutation parameters

Nuove metodologie per la stima dei parametri di mutazione

Paola Vicard¹²

Dipartimento di Economia

Università Roma Tre

vicard@uniroma3.it

Riassunto: In questo lavoro vengono presentate due metodologie per la stima del tasso di mutazione genetica. La stima viene effettuata sulla base di dati raccolti presso i laboratori forensi, relativi a casi di paternità discussa. Vengono evidenziati quei fattori, quali paternità incerta e mutazioni nascoste, che complicano la stima del tasso di mutazione. Le metodologie proposte tengono conto di questi fattori: la prima usa esclusivamente misure di conteggio mentre la seconda è sviluppata mediante le reti bayesiane.

Keywords: Bayesian networks, defective cases, likelihood function, mutation models, uncertain paternity.

1. Introduction

Short Tandem Repeat (STR) human DNA polymorphisms have become a powerful tool for parentage tests and forensic identification. These markers have a discrete number of possible values (alleles) and are known to be prone to mutation. The overall mutation rate of STR markers used for forensic purposes may range between 5×10^{-4} and 7×10^{-3} per generation (Brinkmann *et al.* (1998), Henke and Henke (1999), Sajantila *et al.* (1999)).

The possibility of mutation in DNA gene transmission is a problem that complicates forensic inference from DNA profiles. It is particularly important in cases of paternity testing, when a putative father may seem not to be the true father because a mutation has led to his passing on a seemingly impossible allele to the child. An *incompatibility* is a genotype configuration of a putative father-mother-child triplet that, in absence of mutation and under certain paternity, could not occur under Mendelian segregation.

In Dawid *et al.* (2001) it is shown how to solve paternity testing problems accounting for mutation; in Dawid *et al.* (2002) this and more complex problems (where one of the parents is not available and one or more relatives are observed instead) have been tackled using Bayesian networks. Here we focus on mutation rate estimation problem taking into account uncertain paternity together with other complicating factors that will be illustrated in what follows.

Data collected at forensic laboratories can be used to assess the overall mutation rate. The combined mutation rate τ , that is the sum of maternal and paternal mutation ($\mu_M + \mu_P$) at some locus is commonly estimated simply by the observed relative frequency of incompatibility at that locus in a sample of families. However, this is a crude estimate, and does not take into account several complicating features such as:

¹Address for correspondence: Via Ostiense 139, 00154 Roma

²This research was supported by MIUR-Cofin2003 and Leverhulme Trust

- *Hidden mutation.* A mutation need not necessarily result in an incompatibility; a genotype configuration may only be seemingly compatible and an undetectable mutation might have occurred.
- *Differential mutation.* Mutation rates not only vary among the different STR markers, but also between male and female germ lines, at the ratio of about 17 : 3 Brinkmann *et al.* (1998).
- *Uncertain paternity.* The putative father may not be the true father since the analysed triplets are collected at forensic laboratories for paternity testing purpose.

In this paper we show how to estimate τ taking into account these sources of bias. The various mutation models for interallelic transitions are described in Section 2, and an introduction to the mutation rate estimation problem is given in Section 3. The methodology based on summary data is presented in Section 4 while the Bayesian network approach is explained in Section 5.

2. Genetic background and mutation models

An STR marker is an area on a chromosome (*locus*) whose DNA composition (*alleles*) is given by integers. The genotype of an individual at some locus is an unordered pair of alleles, one inherited from the mother and one from the father but it is not possible to distinguish which is which. If both alleles have the same value, the genotype is homozygous, otherwise it is heterozygous. Here we assume both linkage and Hardy-Weinberg equilibrium, *i.e.* independence between and across markers.

Consider a marker with a finite repertory of alleles $i = 1, \dots, K$ whose frequencies are p_i , $i = 1, \dots, K$. When a mutation takes place, the allele inherited by the child is different from that of its parents. Data about interallelic transitions are very sparse and not sufficient to estimate $K(K - 1)$ allele-specific mutation rates. Therefore the interallelic transition probabilities $q_{i \rightarrow j}$ ($i, j = 1, \dots, K$) are expressed by means of mutation models with a small number of parameters. Various mutation models have been proposed in the literature (Egeland and Mostad (2002), Dawid *et al.* (2001), Dawid *et al.* (2002), Ohta and Kimura (1973), Valdes *et al.* (1993)).

Here we consider the *scalar mutation models* (see Vicard and Dawid (2004)) for which $q_{i \rightarrow j} = \lambda \times s_{i \rightarrow j}$, $j \neq i$, where the nonnegative quantities $\{s_{i \rightarrow j}, j \neq i\}$ are supposed known and λ is the unknown *mutation parameter* to be estimated. The mutation rate is

$$\mu = \kappa \lambda \quad (1)$$

with $\kappa = \sum_{(i,j): i \neq j} p_i s_{i \rightarrow j}$. Specific mutation models can be identified defining the quantities ($s_{i \rightarrow j}$). In this paper we consider first the *proportional model* where $s_{i \rightarrow j}^{prop} = p_j$, $i \neq j$, and

$$\kappa = 1 - \sum p_i^2. \quad (2)$$

This model has the property of stationarity, *i.e.* gene frequencies are not altered by the mutation process. Moreover it is analytically tractable but not totally biologically realistic. Previous analyses have shown that most mutations are towards a neighbouring allele, *i.e.* involve addition or deletion of one single repeat. An alternative model is the

single-step mutation model where $\kappa = 1$ and

$$s_{i \rightarrow j}^{SMM} = \begin{cases} \frac{1}{2} & \text{if } |i - j| = 1, i \neq 1 \text{ or } K \\ 1 & \text{if } |i - j| = 1, i = 1 \text{ or } K \\ 0 & \text{otherwise} \end{cases}$$

This model is too extreme because two or more step mutations are not allowed. Therefore we consider a mixture of the single-step and the proportional models with weight h and $1 - h$ respectively, named *mixed model*. For this model

$$s_{i \rightarrow j}^{mix} = h s_{i \rightarrow j}^{SMM} + (1 - h) s_{i \rightarrow j}^{prop} \quad \text{with} \quad \kappa = (1 - h)(1 - \sum p_i^2) + h.$$

The mixed model is more biologically realistic but non stationary.

In this paper we focus on the estimation of the combined mutation parameter τ . We assume a mutation model common to the paternal and maternal line with parameters λ_P and λ_M . The paternal and maternal mutation rates are expressed in terms of τ and of the paternal to total mutation ratio $\rho = \mu_P / \tau$ as follows: $\mu_P = \rho\tau$ and $\mu_M = (1 - \rho)\tau$. The total mutation parameter is $\xi = \lambda_P + \lambda_M$ with $\tau = \kappa\xi$ and $\rho = \lambda_P / \xi$. Hence the paternal and maternal mutation parameters are $\lambda_P = \rho\xi$ and $\lambda_M = (1 - \rho)\xi$.

3. Estimation of mutation rates from paternity casework

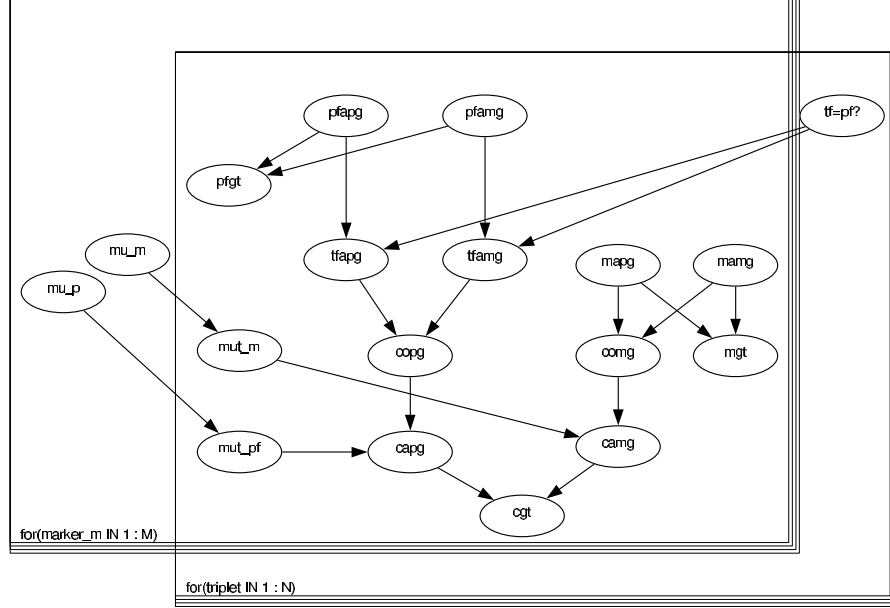
When estimation of mutation rates is carried out on the basis of paternity casework, we face a duality problem between uncertain paternity and mutation rate to be assessed. This problem is represented in Figure 1. The graph shows the general structure of the model and allows us to distinguish between carrying out paternity testing and learning about mutation model parameters. Suppose that M STR markers are measured on N triplets. The graph has two plates, one for the markers and one for the triplets. The intersection of the plates represents the pedigree of each single triplet. When the hypothesis of interest is whether the putative father is the true father, $\mathbf{tf}=\mathbf{pf}?$, we consider one triplet, *i.e.* work inside a triplet plate, and the corresponding likelihood ratio for paternity is derived on the basis of the M markers measured on that triplet.

On the other hand, when we wish to estimate the paternal and maternal mutation rates, *i.e.* the target nodes $\mathbf{mu_p}$, $\mathbf{mu_m}$, we work inside a marker plate. For each of the M markers, $\mathbf{mu_p}$, $\mathbf{mu_m}$ are derived from the observed genotypes on all triplets. When an inconsistency is observed on just one out of a series of otherwise consistent markers all pointing to paternity, it is standard practice to assume that a mutation has occurred. However, since non-paternity could also be invoked, inference on mutation rates is “disturbed” by the node $\mathbf{tf}=\mathbf{pf}?$, which lies outside the markers plate.

4. Analysis of summary data

We base our analysis on a sample of N triplets. Most of them are compatible, *i.e.* the genotype of the child can be explained by Mendelian inheritance from those of the parents (assuming that the putative father is the true father). The remaining triplets are incompatible in the sense specified in Section 1. Compatibility and incompatibility are denoted with C and I respectively. We assume here that incompatibilities can be explained by a single mutational event - a full list of incompatible configurations with this property is in

Figure 1 Graph representing the two learning levels of the mutation rate estimation problem.



Dawid *et al.* (2001). Essentially here all the information about mutation is carried by the incompatible triplets which are classified as follows:

- paternal exclusions (I_P), explicable only by a paternal mutation
- maternal exclusions (I_M), explicable only by a maternal mutation
- ambiguous exclusions (I_E), explicable by either paternal or maternal mutation.

This approach to mutation rate estimation is based on the following summary data: N triplets; r incompatibilities out of which r_P paternal exclusions and r_M maternal exclusions. This methodology is mainly aimed at providing factors correcting biases proper of the naïve estimator $\tilde{\tau} = r/N$. This estimator implicitly assumes that every mutation results in an incompatibility, which is not true due to hidden mutations. Hence $\text{pr}(I) < \tau$ ($\text{pr}(I)$ being the probability of observing an incompatibility) and $\tilde{\tau}$, which is an unbiased estimate of $\text{pr}(I)$, underestimates τ . Chakraborty *et al.* (1996) presented an algebraic formula to compute $\text{pr}(I)$ when $\mu_M = 0$ and a specific mutation model is assumed. In Vicard and Dawid (2004) scalar mutation models are considered and it is shown that for $\mu_M = 0$, $\text{pr}(I) = \alpha \lambda_P$, where α is a function of (p_i) and of $(s_{i \rightarrow j})$. For the proportional model

$$\alpha = 1 - 2a_2 - 2a_2^2 + a_3 + 3a_2a_3 + 2a_4 - 3a_5 \quad \text{with} \quad a_m = \sum_i p_i^m. \quad (3)$$

Therefore an unbiased estimate of $\mu_P = \kappa \lambda_P$ is $(\kappa/\alpha) \tilde{\tau}$. Only I_P cases are considered here, since maternal mutation is not allowed.

In Vicard and Dawid (2004) this analysis is extended to the case when $\mu_M \neq 0$ and maternal and ambiguous exclusions are analysed. First we reinterpret the previous analysis, *i.e.* still assuming $\lambda_M = 0$, to get $\text{pr}(I_P) + \text{pr}(I_E) = \alpha \lambda_P$. Then with a variant

of that analysis (under the same assumptions) we can compute $\text{pr}(I_E) = \beta\lambda_P$ and find that $\text{pr}(I_P) = (\alpha - \beta)\lambda_P$. When $\lambda_M \neq 0$, $\text{pr}(I_P)$ is not affected, and by symmetry $\text{pr}(I_M) = (\alpha - \beta)\lambda_M$, $\text{pr}(I_E) = \beta(\lambda_P + \lambda_M)$ and $\text{pr}(C) = 1 - \alpha(\lambda_P + \lambda_M)$. Expressing the parameters λ_P and λ_M in terms of the total mutation parameter ξ and of the paternal to total mutation ratio ρ , we have

$$\text{pr}(I_P) = (\alpha - \beta)\rho\xi \quad \text{pr}(I_M) = (\alpha - \beta)(1 - \rho)\xi \quad \text{pr}(I_E) = \beta\xi \quad \text{pr}(C) = 1 - \alpha\xi.$$

Therefore when paternity is undisputed we have a simple multinomial distribution with

Cell probabilities	$\text{pr}(I_P)$	$\text{pr}(I_M)$	$\text{pr}(I_E)$	$\text{pr}(C)$
Cell counts	r_P	r_M	r_E	$N - r$

and the likelihood function for ξ and ρ is

$$L \propto \xi^r (1 - \alpha\xi)^{N-r} \times \rho^{r_P} (1 - \rho)^{r_M} \quad (4)$$

$$\propto \tau^r ((1 - (\alpha/\kappa)\tau)^{N-r} \times \rho^{r_P} (1 - \rho)^{r_M}) \quad (5)$$

It is easy to see that the likelihood is factorised in: 1) one term, L_τ , for the total mutation rate; 2) one term, L_ρ , for the paternal to total mutation ratio. Notice that L_τ takes into account the total number of incompatibilities without distinguishing them among paternal, maternal or ambiguous. In particular ambiguous exclusions are neither in L_τ nor in L_ρ although they are informative for ρ , if full genotype data are analysed (see next section). Moreover, since r_P and r_M are generally very small, L_ρ will be rather uninformative.

The maximum likelihood estimates (MLEs) obtained from (5) are

$$\hat{\tau} = \frac{\kappa}{\alpha} \tilde{\tau} \quad \hat{\rho} = \frac{r_P}{r_P + r_M} \quad (6)$$

It is easy to see that μ_P and μ_M can be estimated with $\hat{\mu}_P = \rho\hat{\tau}$ and $\hat{\mu}_M = (1 - \rho)\hat{\tau}$ respectively, having fixed a value for ρ or using $\hat{\rho}$.

The analysis becomes more complex when the hypothesis of certain paternity is removed. We should compute, for each single case, the probability of paternity on the basis of the other markers excluding the one whose mutation rate we are estimating. In this approach, in order to simplify the analysis, we assume a probability of paternity, π , common to all cases.

For certain non-paternity, $\text{pr}(I_P)$, $\text{pr}(I_M)$, $\text{pr}(I_E)$ and $\text{pr}(C)$ can be simply obtained from the analysis conducted under certain paternity observing that paternal mutation can be described with the proportional model with $\lambda_P = 1$.

Finally for uncertain paternity with probability π , the probabilities of the different kinds of exclusions are computed combining the corresponding results under paternity and under non-paternity with weight π and $1 - \pi$ respectively.

In general the overall likelihood will not have a simple form. However under the proportional model it simplifies remarkably. In particular we find that (for more details see Vicard and Dawid (2004)):

$$\begin{aligned} \text{pr}(I_P) &= (\alpha - \beta)(1 - \pi + \pi\lambda_P) & \text{pr}(I_M) &= (\alpha - \beta)\lambda_M \\ \text{pr}(I_E) &= \beta(1 - \pi + \pi\lambda_P + \lambda_M) & \text{pr}(C) &= 1 - \alpha(1 - \pi + \pi\lambda_P + \lambda_M) \end{aligned}$$

Defining $\lambda_P^* = 1 - \pi + \pi\lambda_P$, we notice that the resulting expressions of $\text{pr}(I_P)$, $\text{pr}(I_M)$, $\text{pr}(I_E)$ and $\text{pr}(C)$ are identical to those computed assuming paternity, with λ_P^* instead of λ_P .

Therefore it is sufficient to change some parameter definition in order to account for the possibility of non paternity, when the proportional model is used.

The new likelihood L^* is easily obtained from (4) replacing ξ and ρ by $\xi^* = \lambda_P^* + \lambda_M$ and $\rho^* = \lambda_P^*/\xi^*$ respectively:

$$L^* \propto \xi^{*r} (1 - \alpha\xi^*)^{N-r} \times \rho^{*r_P} (1 - \rho^*)^{r_M}. \quad (7)$$

The associated likelihood factorisation is now less useful on account of the constraint $\lambda_P^* \geq 1 - \pi$. However we can see the effect of uncertain paternity in general terms: in particular, it can not be ignored unless the probability of non paternity is much less than the paternal mutation rate. Since $\lambda_P = (\lambda_P^* - (1 - \pi))/\pi$, the estimate of λ_P is computed as follows

$$\hat{\lambda}_P = \frac{\hat{\lambda}_P^* - (1 - \pi)}{\pi}$$

and the total mutation rate is estimated by

$$\hat{\tau} = \frac{\hat{\mu}_P - \kappa(1 - \pi)}{\pi} + \hat{\mu}_M. \quad (8)$$

4.1. An application

We apply this methodology to real data provided to us by Bern Brinkmann of the Institut für Rechtsmedizin of Münster. The data set consists of $N = 966$ mother-putative father-child triplets from German/Austrian population, with $r = 4$ incompatibilities, out of which two paternal exclusions, $r_P = 2$, and two ambiguous exclusions, $r_E = 2$.

In this analysis we consider a specific marker, VWA, whose allele frequency distribution is given in Table 1 of Vicard and Dawid (2004). Here we consider the proportional model since the analysis accounting for uncertain paternity has been developed under this model. Using expressions (2) and (3) we obtain $\kappa = 0.807$ and $\alpha = 0.620$. The MLEs of τ are computed using (6) and (8) for certain and uncertain paternity respectively. We consider specific values of π : 1, 0.9997 and 0.997 to investigate the effect of uncertain paternity on the mutation rate. The estimates are given in Table 1

Table 1 Maximum likelihood estimates of τ

	$\pi = 1$	$\pi = 0.9997$	$\pi = 0.997$
<i>MLE</i>	0.005412	0.005172	0.003

Notice that the naïve estimate is $\tilde{\tau} = 4/962 = 0.004141$. Comparing $\tilde{\tau}$ with the MLE for $\pi = 1$ we see that accounting for hidden mutations pushes the mutation rate towards higher values. On the other hand considering uncertain paternity, $\pi < 1$, reduces the estimates.

5. Analysis with Bayesian networks

In the previous section we illustrated the main biases affecting the standard estimate $\tilde{\tau}$ and provided correcting factors that are easy to be used at forensic laboratories. Several simplifying assumptions and approximations have been made:

- analysis of summary data only;
- use of the proportional mutation model to deal with uncertain paternity;
- probability of paternity common to all cases;
- at most one mutational event occurring in a triplet.

Moreover, the methodology is developed to deal with complete cases only. In general some of the cases are *defective*, that is either the mother or the putative father are not observed. Incomplete cases contain information about mutation and a methodology to analyse them is needed.

In order to remove or relax the assumptions and approximations introduced in Section 4, we propose the use of *Bayesian networks* (BN) or *Probabilistic Expert Systems* (see Cowell *et al.* (1999)). A BN is a graphical and numerical representation of the joint distribution of a set of variables. BNs are particularly important and useful to learn and represent the dependency structure proper of these variables. In general a BN is given by:

- a) a direct acyclic graph incorporating the set of dependencies among variables,
- b) an inferential engine to make inference on the model parameters.

BNs are a natural tool to represent a pedigree structure; the arrow direction is suggested by family relationships and the conditional probabilities are given by simple genetic laws, such as Mendel's law. One of the main features of BNs is their *modularity*, that is the possibility to decompose a complex relation structure in a number of modules embodying simple local dependencies. BNs have been applied to parentage tests and forensic identification problems; for a detailed description see Dawid *et al.* (2002), while for specific applications see Mortera *et al.* (2003), Mortera (2003).

Here we use Object-Oriented Bayesian networks (OOBN, see Koller and Pfeffer (1997)) which are supported by the software Hugin (<http://www.hugin.it>). In these networks the *objects* are instances of Bayesian networks. By using OOBNs, solving complex (forensic) problems is a straightforward task since the modularity property of BNs can be exploited. In Dawid (2003) a OOBN for mutation rate estimation has been presented. Here we start our analysis from this network and modify it to account for non stationarity of the more biologically realistic mixed mutation model and to perform a correct and general analysis of the defective cases (putative father-child and mother-child pairs). The top-level network describing the problem for one marker is shown in Figure 2; for a detailed description of the whole OOBN structure see Dawid (2003).

In the top left-hand side of the graph there are the unobservable maternal genes (mapg and mamg) determining the observed genotype mgt. mapg and mamg (and all *founder* gene nodes pfapg, pfamg and otherpg) are given the population allele frequency distribution. By Mendelian segregation the maternal genes give rise to the *original* child maternal gene comg. The instance camg represents the gene *actually* inherited by the child from the mother. This gene coincides with the original gene if no mutation occurs, otherwise the child inherits a mutated gene generated from the original according to the adopted mutation model. In this BN the mixed mutation model is used and the mixing parameter h is defined with three possible states (0 for the proportional model, 0.9 for

Figure 2 Object Oriented Bayesian network to asses mutation rate: top level network.

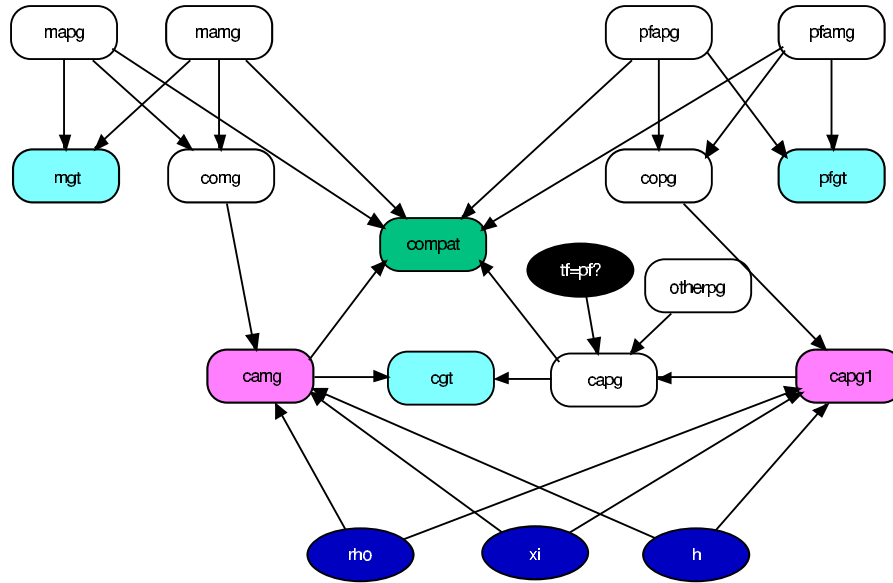
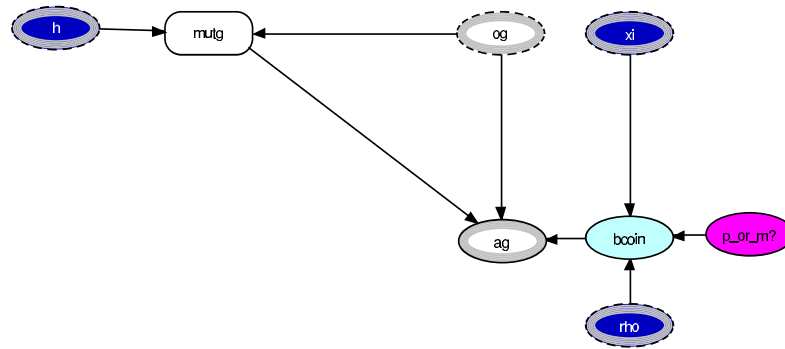


Figure 3 Network describing the process whereby the actual gene is determined from the original gene.



the mixed model and 1 for the single-step mutation model). We will fix h although this network can also be used to make inference on it. The process whereby the original gene is passed down to the child generating the actual gene is represented in the instance in Figure 3.

The node `bcoin` is a Bernoulli variable representing whether mutation occurs. The parameter of this variable is specified according to the node `p_or_m?` which has three possible states:

- state 2 has to be instantiated when defective cases are analysed. Suppose a putative father-child pair is observed. The actual maternal gene is not observed, then it has to be taken at random from the reference population. As it is shown in Figure 2, `camg` results from a process where mutation is allowed. When the mutation model is not stationary, as is the mixed model, the actual gene allele frequencies may be

altered by the mutation process. Consequently they may be different from those of the reference population. In order to avoid this possibility and properly model the unobserved maternal gene, we have inserted a “switch” (state 2 of `p_or_m?`). When `p_or_m?`=2, any mutation process on the maternal side is forbidden. Notice that, due to the generality of the instance, the same holds when mother-child pairs are analysed and mutation has to be blocked on the paternal side.

- if `p_or_m?` is not set to 2, the parameter of the Bernoulli variable is given by the expression $\xi \cdot (\rho \cdot \text{p_or_m} + (1 - \rho) \cdot (1 - \text{p_or_m}))$. `p_or_m?` is set to 0 [1] when the maternal [paternal] line is specified. In this way we find that the parameter of the `bcoin` variable is $(1 - \rho)\xi = \lambda_M$ [$\rho\xi = \lambda_P$] when maternal [paternal] mutation side of the graph is considered.

Going back to Figure 2, the total mutation parameter (node `xi`) has been given 43 possible states (the minimum is 0, the maximum is 0.0198), more concentrated around the mutation rate values reported in the literature. This node is given a uniform prior distribution, allowing us to obtain the likelihood for ξ once evidence is inserted and propagated through the network. Finally `rho` is given a uniform prior over the states 0, 0.5, 0.7, 0.9, 1. As for h , ρ will be generally fixed but inference can be performed on it using the same network.

The paternal line structure (right hand side of the network) is similar to the maternal. In the top right-hand side of the graph we have the putative father genes and genotype nodes. An original gene `copg` is generated by Mendelian segregation and it then gives rise (by the process described for the maternal line) to the actual gene `capg1` deriving from the putative father. The provenance of the child actual paternal gene `capg` is then obtained according to the *query* node `tf=pf?`. `capg` is `capg1` if the true father is the putative father, `tf=pf? = true`, otherwise the child actual paternal gene is taken at random from the reference population, `otherpg`. Notice that having placed the query instance after that of mutation, guarantees that if `tf=pf? = false`, the paternal gene is taken at random from the population gene frequencies (*i.e.* unaltered by a non stationary mutation process). In the network in Figure 2 the node `compat` represents the compatibility status of the child’s genotype with: 1) those of the mother and of the putative father when a triplet is analysed; 2) the putative father genotype when a putative father-child pair is given; 3) the maternal genotype when a mother-child pair is analysed.

5.1. Data analysis

The network described above can be used to analyse the full genotype data relative to triplets and pairs. For an analytic approach to mutation rate estimation see Vicard *et al.* (2004). The data set analysed in Section 4.1 has 4 incompatible and 962 fully compatible triplets. The data given to us by Bernd Brinkmann also comprised $N^p = 81$ fully compatible pairs, out of which $N_M^p = 41$ mother-child pairs. As in Section 4.1 the methodology is applied to assess the mutation rate of a specific marker, VWA.

Consider first the incompatible cases. With the BN approach, we can take into account the different *posterior* probabilities of paternity based on all markers other than VWA. In doing so we assume that no mutation occurs in these markers so that the probability of paternity is simply computed using paternity testing formulas (Essen-Möller (1938)). Suppose that these probabilities are 0.9999, 0.9995, 0.9992, 0.997. To obtain the contribution to the likelihood from each single incompatible case, we simply insert in

Table 2 Maximum likelihood estimates of τ computed using the Bayesian network in Figure 2 assuming the proportional model ($\hat{\tau}_{prop}$) and mixed model ($\hat{\tau}_{mix}$).

	$\pi = 1$	$\pi = 0.9997$	$\pi = 0.997$	π based on the other markers
$\hat{\tau}_{prop}$	0.005238	0.004858	0.001619	0.004244
$\hat{\tau}_{mix}$	0.005106	0.004390	0.002783	0.003706

the network, for each case, evidence about: the genotypes, π , mutation model choice (*i.e* value of h), value of ρ (here we set it to 0.5, assuming $\mu_P = \mu_M$), value of `p_or_m` (0 and 1 in `camg` and `capg1` respectively). Once evidence is inserted, we propagate it and read the likelihood for ξ , $\ell_i^I(\tau)$, $i = 1, \dots, r$, that of τ is simply obtained recalling that $\tau = \kappa\xi$.

Due to the large number of compatible cases, we only analyse information about the compatibility status, represented in the network by the node `compat`. In principle these cases can be analysed by the same procedure used for the incompatible triplets. The contribution to the likelihood from a compatible triplet, $\ell^{CT}(\tau)$, is obtained by entering and propagating in the network evidence about: triplet genotypes compatibility, π , h , $\rho = 0.5$, `p_or_m` as above. When a mother-child [putative father-child] compatible pair is analysed, the following evidence is entered: mother-child [putative father-child] compatibility, π (for paternal pairs only), h , $\rho = 0.5$, `p_or_m` = 2 in the paternal [maternal] line to cut off the unobserved parent. Propagating this evidence we get $\ell^{CP_M}(\tau)$ [$\ell^{CP_P}(\tau)$]. Notice that since summary data are analysed, we assume a π common to all compatible cases. The full likelihood for τ is $\ell(\tau) = \prod_{i=1}^r \ell_i^I(\tau) \left(\ell^{CT}(\tau)\right)^{N-r} \left(\ell^{CP_M}(\tau)\right)^{N_M^P} \left(\ell^{CP_P}(\tau)\right)^{N_P^P}$.

In Table 2 we report the maximum likelihood estimates of τ computed under the proportional and mixed model and considering various probabilities of paternity. The value π affects the estimates and its effect is stronger under the mixed model. This is due to the fact that in the observed data, three out of the four incompatible cases can be explained by two or more-step mutations. Under the mixed model the probability of this kind of mutation is very small. Therefore even a little departure from paternity, causes a relevant reduction in the mutation rate estimate. The relative likelihoods are shown in Figure 4. This graph shows that for smaller π , the more to the left the curve is shifted, implying that smaller values of τ are more likely. The likelihoods in Figure 4 are obtained assuming $\pi = 0.9997$ in the compatible cases. If this probability is changed to $\pi = 1$ or $\pi = 0.997$, leaving the probability of paternity of the incompatible cases unaltered, the likelihood is nearly unaffected and the estimates change only in the fourth significative figure.

In Figure 5 the conditional likelihood for τ given $\rho = 0.5$ and $\pi = 0.9997$ is compared with the likelihood computed using the BN approach. We see that even when the same π and mutation model are chosen, full genotype and defective case analyses carry some information on τ . The comparison of curve (a) with (c) shows that it is very important to consider the probabilities of paternity of each single incompatible case. Finally when a biologically realistic mutation model is adopted all the information contained in the incompatible cases can be conveyed to produce the estimate of τ , since this model and its BN representation are able to account for the genotypic configuration of the triplet.

Figure 4 Relative likelihood for τ for different values of π under the mixed model

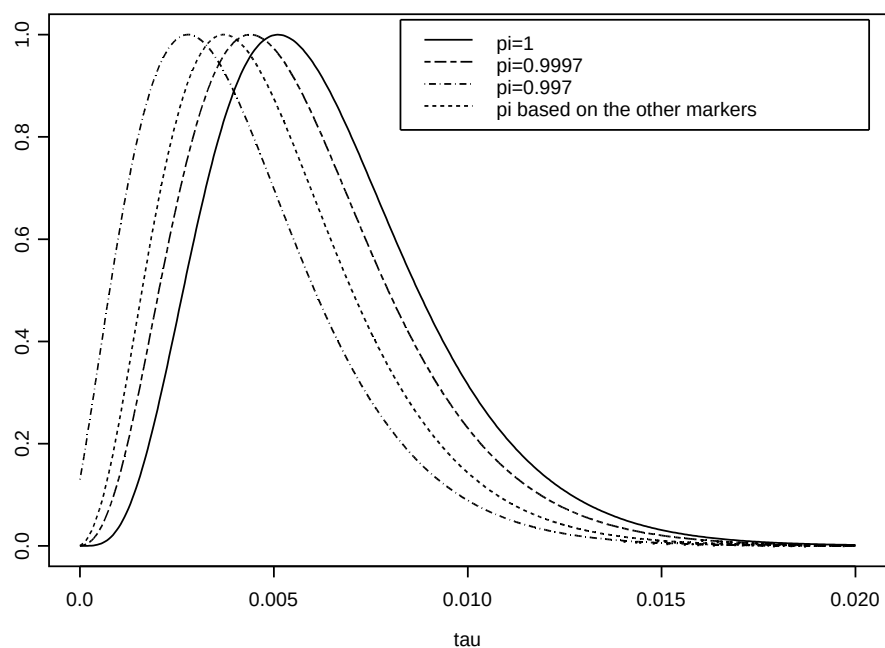
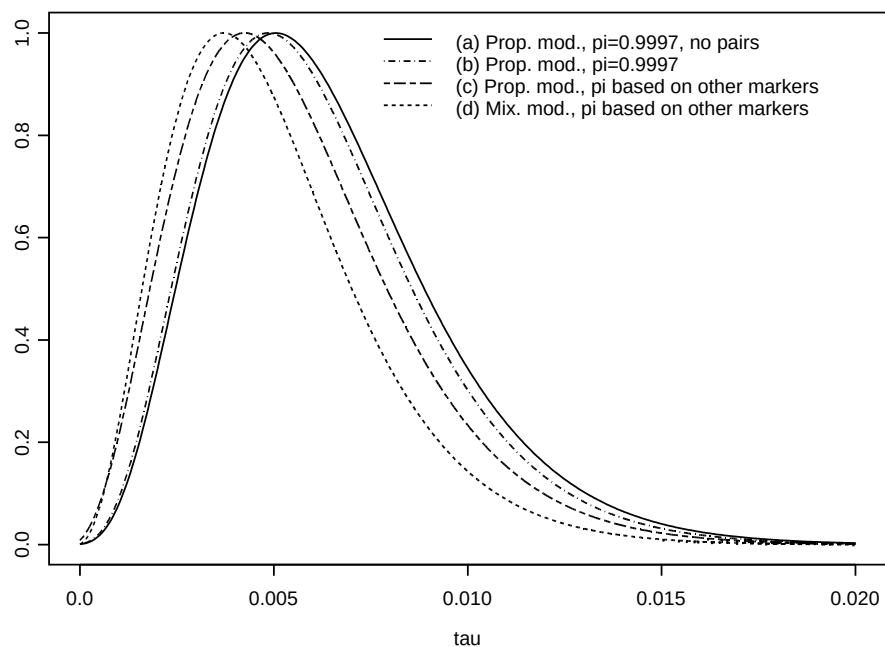


Figure 5 Relative likelihood for τ using: (a) method in Section 4 with $\pi = 0.9997$ and $\rho = 0.5$; BN approach for (b) proportional model, $\pi = 0.9997$, (c) proportional model, π based on the other markers, (d) mixed model, π based on the other markers.



6. Acknowledgements

I am grateful to A. Philip Dawid and Julia Mortera for helpful discussions.

References

- Brinkmann B., Klitsch 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, 1408–1415.
- 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, 41–48.
- 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. (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*, Bishop C.M. and Frey B.J., eds.
- 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, 577–595.
- 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.
- Henke L. and Henke J. (1999) Mutation rate in human microsatellites, *American Journal of Human Genetics*, 64, 1473.
- Koller D. and Pfeffer A. (1997) Object-oriented Bayesian networks, in: *Proceedings of the 13th Annual Conference on Uncertainty in AI (UAI)*, 302–313.
- Mortera J. (2003) Analysis of DNA mixtures using Bayesian networks, in: *Highly Structured Stochastic Systems*, Green P.J., Hjort N.L. and Richardson S., eds., Oxford University Press, chapter 1B, 39–44.
- 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, 201–204.
- 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, 263–266.
- Valdes A.M., Slatkin M. and Freimer N.B. (1993) Allele frequencies at microsatellite loci: The stepwise mutation model revisited, *Genetics*, 133, 737–749.
- Vicard P. and Dawid A.P. (2004) A statistical treatment of biases affecting the estimation of mutation rates, *Mutation Research*, 546, 19–33.
- Vicard P., Dawid A.P., Lauritzen S.L. and Mortera J. (2004) Estimation of mutation rates from paternity cases, manuscript.