© Board of the Foundation of the Scandinavian Journal of Statistics 2002. Published by Blackwell Publishers Ltd, 108 Cowley Road, Oxford OX4 1JF, UK and 350 Main Street, Malden, MA 02148, USA Vol 29: 577–595, 2002

Probabilistic Expert Systems for Forensic Inference from Genetic Markers

A. P. DAWID

University College London

J. MORTERA

Università Roma Tre

V. L. PASCALI

Università Cattolica del "Sacro Cuore", UCSC, Rome

D. VAN BOXEL

Carnegie Mellon University

ABSTRACT. We present a number of real and fictitious examples in illustration of a new approach to analysing complex cases of forensic identification inference. This is effected by careful restructuring of the relevant pedigrees as a Probabilistic Expert System. Existing software can then be used to perform the required inferential calculations. Specific complications which are readily handled by this approach include missing data on one or more relevant individuals, and genetic mutation. The method is particularly valuable for disputed paternity cases, but applies also to certain criminal cases.

Key words: Bayesian networks, DNA profile, forensic identification, incomplete evidence, mutation, paternity testing

1. Introduction and background

In a simple problem of forensic DNA identification, we have a "trace" biological sample of unknown provenance, and further "reference" samples obtained from known individuals. All the samples are typed by DNA profiling, and it is desired to use the resulting data to shed light on the origin of the trace sample. More complex problems also arise: thus in a case of disputed paternity, DNA information on the child can be regarded as supplying partial information about its father's DNA (Dawid & Mortera, 1998).

DNA profiles as currently used consist of measurements on a number of genetic markers, typically "short tandem repeat" (STR) markers (Weber & May, 1989), chosen by forensic geneticists for their usefulness. For each marker we can observe its genotype, comprising two genes (or bands), one inherited from the mother and the other from the father (although it is not possible to observe which is which). Each marker has a finite number (up to around 20) of possible values (alleles), generally taking positive integral values, for each of its two constituent bands. The markers used for forensic identification are chosen to be located on different chromosomes, and hence segregate independently. It is often reasonable to assume random mating within an appropriate population, which induces both Hardy–Weinberg and linkage equilibrium, so that different markers, as well as different bands of the same marker, behave entirely independently. We shall assume this throughout.

Databases have been gathered from which the frequency distributions of the different markers, in various populations, can be estimated. In this paper we use estimates based on data collected by the Forensic Genetics Laboratory, Catholic University of the Sacred Heart

(UCSC), Rome. However, as more data are collected, so these estimates may be refined (up-to-date information on the UCSC database may be obtained from: http://www.mclink.it/personal/MD1696/data/freqask.htm). Values used in this paper should not be regarded as definitive.

We are here concerned with the general statistical problem of inferring identification on the basis of the DNA and other evidence at hand (Dawid & Mortera, 1996). This can in principle be solved by determining the relative likelihoods, induced by the full data, for the various competing hypotheses. However, in many cases samples are not available for one or more individuals of interest, and instead we only have indirectly relevant information, perhaps through genetic typing of their relatives. Calculation of the desired likelihoods, on the basis of such partial or incomplete data, can then become both conceptually and computationally demanding, particularly when we allow for the possibility of mutation during gene transmission.

Here we present a new approach to solving such a problem, by reformulating it as a Probabilistic Expert System (PES): a joint graphical and numerical representation that can be implemented and processed in general-purpose computer software. In a PES, complex inter-relationships are broken down into simple modular units, out of which the entire graphical representation is constructed. The resulting representation then forms a framework for the application of fast and efficient computational algorithms.

A complex genetic pedigree fits particularly smoothly into the PES model. The nuclear family relationships constitute natural modular building blocks of the representation, to such an extent that terms such as "parent" and "child" have become part of the general terminology of PES, even for entirely non-genetic applications. The conditional probability tables required are simple and uncontroversial, being given by Mendelian laws of inheritance and logical relationships between genes and genotypes. And the conditional independence relations forming the backbone of a PES representation are likewise natural in the setting of genetic inheritance, since, conditional on its parent's genes, a child's genes are entirely independent of those of all other individuals (more precisely: of those of its non-descendants).

In this paper we exhibit how to pass from an initial pedigree representation of a forensic identification problem to an appropriate graphical PES representation, and how to use information on gene frequencies, Mendelian inheritance and mutation processes to set up the numerical part of the PES specification. We describe how one can apply existing software to calculate the desired likelihoods. The method is illustrated on a number of DNA identification problems of varying degrees of complexity, including real paternity testing cases and an artificial criminal identification example.

2. Disputed paternity

2.1. Basic set-up

In the simplest case of disputed paternity a man is alleged to be the father of a child, but disputes this. DNA profiles are available on the mother `m`, the child `c`, and the putative father `pf`. The disputed pedigree can be represented as in Fig. 1, where a square indicates a male and a circle a female, and `tf` denotes "true father"; grey indicates that a DNA profile is available for that individual. On the basis of these data, we need to assess the likelihood function over possible hypotheses as to the true father. Often these hypotheses are reduced to two: the true father either is the putative father, or else is drawn randomly from the population, being unrelated to the mother or putative father. Throughout this paper we shall make the latter simplifying assumption whenever the true father is not otherwise identified.

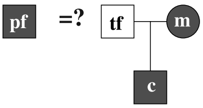

Fig. 1. Simple paternity pedigree.

Because the markers used are on different chromosomes, and we are assuming random mating, we have complete independence across markers. It follows that we can consider the markers one at a time: we simply have to obtain the likelihood ratio on the basis of the data for each marker separately, and finally multiply these values together to obtain the overall likelihood ratio based on all the data on the full collection of markers.

So consider now the measured genotypes, from all three parties, for some fixed marker. To find the associated likelihood function we need to calculate the joint probability of this triplet of observed genotypes, under either hypothesis as to paternity. Making the reasonable assumption that, before we have any data on the child, the identity of the true father is independent of the profiles of the mother and the putative father, we can transfer attention to the conditional probability of the child's genotype, given the other two. Under either hypothesis, this is calculated simply: under paternity, we just apply Mendel's laws of segregation; under non-paternity, we require (estimates of) the frequencies of relevant marker alleles among the population at large. From these likelihoods we readily obtain the desired likelihood ratio. Using Bayes's Theorem, this can then be combined with the prior odds of paternity, based on external evidence, in order to obtain the posterior odds for paternity.

As an illustrative example, suppose that the data, for marker FES, are: child's genotype $=$ $\{12, 12\}$, mother's genotype $= \{10, 12\}$, putative father's genotype $= \{10, 12\}$. The population frequencies of alleles 10 and 12 are, respectively, 0.28425 and 0.25942. In this case, conditioning on the genotypes of mother and putative father, we see that the child's genotype will be as observed if and only if both the mother and the true father contributed allele 12 to the child. This event has probability 0.5×0.5 if the true father is the putative father, and probability 0.5×0.25942 if the true father is, instead, some unrelated individual from the population. Thus the likelihood ratio in favour of paternity (based on these data for marker FES alone) is $0.5/0.25942 = 1.9274$.

2.2. From pedigree to expert system

In the above simple problem the calculations are trivial, and have long been widely implemented much as we have described (Essen-Möller, 1938): we certainly do not need to develop any clever new methods of solution. However, we wish to extend our analysis to more complex problems, particularly those with missing data, for which the calculations are by no means trivial. Purely as a gentle lead-in to this extension, it will be valuable to reformulate the simple paternity problem of section 2.1 above as a "Probabilistic Expert System", or PES (Cowell *et al.*, 1999).

A PES is a representation of a complex probability structure by means of a directed acyclic graph, having a node for each variable, and directed links describing probabilistic causal relationships between variables. The overall probability structure is completely determined by specifying, as desired, the conditional probability tables for each variable given its "parents" in the graph. On the basis of probabilistic conditional independence properties embodied in the graph, the complex global model then decomposes into simpler localized submodels, providing a framework for the application of fast and efficient computational algorithms for exact calculation of marginal and conditional probabilities, and much else beside (Dawid, 1992; Spiegelhalter *et al.*, 1993). The calculations can be described as effected by ''propagation'' of information through the network: this involves efficient organization of simple calculations affecting one local cluster of variables at a time, but spreading throughout the whole network to yield the correct overall answers. These propagation algorithms have now been implemented in widely available software, such as Hugin (http://www.hugin.dk), GeNIe (http://www2.sis.pitt.edu/~genie) or Xbaies (http://www.staff.city.ac.uk/~rgc), thus enabling many otherwise intractable complex problems to be solved.

Because we are at liberty to choose which unobserved variables to include in a PES representation, there can be many such representations, some more manageable than others. Finding an appropriate representation is crucial, as the efficiency, or even the viability, of the computational routines is highly sensitive to the topology of the graphical structure. PES construction is to some extent an art-form, but can be guided by scientific and logical considerations. The main contribution of this paper—beyond advertising the simplicity and benefits of general PES ideas, methods and technology—is to present and analyse what we consider to be good representations for the type of paternity and identification problems we address. However, for other problems, other kinds of representation may be better. The search for good representations for specific problems is an important task for continuing research in this area.

Our graphical PES representation of the simple disputed paternity problem (for a single marker) is displayed in Fig. 2. Purely for presentational purposes we colour-code the nodes to distinguish different types: grey again marks a node which is observed, while black represents a disputed hypothesis. The arrows represent (sometimes degenerate) probabilistic influences. For a detailed description of the semantics of such diagrams, see Cowell *et al.* (1999).

To attain the most efficient and simplified representation, we have aimed to represent the problem at as deep and disaggregated a level as possible. First, we need to identify, and create nodes for, all the interesting variables in the problem. These do not necessarily have to correspond to observables, although all relevant observables must be represented. In our model, in order to maximize the efficiency of the calculations as well as the logical clarity of the representation we choose to disaggregate each individual's genotype into its constituent, unobserved, paternally and maternally inherited genes. We thus have a node `pfmg` representing ''putative father's maternal gene'', etc., as well as nodes for the observed genotypes: `mgt` represents ''mother's genotype'', etc. Table 16 provides a summary of all the notation used in this article.

A related representation (see e.g. Thompson, 2000) has nodes for genotypes, but not for their constituent bands: information as to whether a gene was inherited from the mother or father is represented by means of additional (unobserved) binary meiosis or segregation

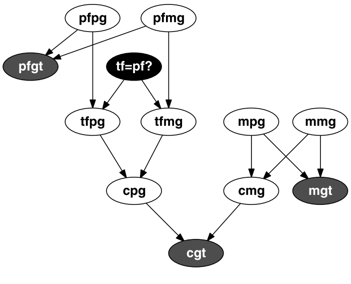

Fig. 2. Simple paternity network.

indicators. Our own representation is perhaps a little more transparent and straightforward, at least for present purposes.

An important feature of Fig. 2 is the explicit introduction into the graph of the "target" (or "hypothesis" or "query") node, `tf = pf?` ("true father = putative father?"), representing the hypotheses of interest. We are thus considering this query on exactly the same footing as observable or latent genetic information. This is a natural step within the general PES approach, although perhaps less natural from the standpoint of genetics, where graphical modelling is usually restricted to representing pedigree structure. There is a large number of powerful specialist pedigree analysis programs available that work with such pedigree representations (e.g. LINKAGE (http://linkage.rockefeller.edu/soft/linkage/), MENDEL (http://www.biomath.medsch.ucla.edu/faculty/klange/software.html), among others (see list at http://linkage.rockefeller.edu/soft/list.html)); and any single one of the problems we consider in this paper could probably be handled better by some one of those programs. However, none of these has the degree of generality and extendibility of the more broadly-based PES technology.

There are alternative ways of building and analysing PES representations, without explicitly representing the hypothesis node in the graph. Indeed, for some problems, as in section 4.1 below, this may be the only efficient way to proceed. We prefer to include an explicit hypothesis node wherever possible, since this is simpler to interpret, and allows one to read off directly, simply by querying the "target" node, the quantity of most interest: the likelihood ratio (for the relevant marker) in favour of paternity, on the basis of the observed evidence.

To complete our PES representation, we need to supply the numerical part of the specification. This requires that we give, for each node in the network, the table of probabilities for its various values, conditional on each configuration of values at its "parent" nodes (if any). We specify these tables as follows. At nodes corresponding to "founder" genes we use population gene frequencies. A child's maternal gene is obtained by drawing, at random, one of it's mother's two (paternal and maternal) genes, `mmg` and `mpg`, and similarly for its paternal gene (drawing from its father's genes). The table for a genotype node is degenerate, encoding the simple deterministic relationship between an unordered pair of values and its two constituents. The true father's paternal (maternal) gene is either identical with the corresponding gene of the putative father, or else generated from the relevant population gene-frequency distribution, depending on the value of the hypothesis node `tf = pf?`.

At the hypothesis node itself, we could set an initial distribution to represent actual prior beliefs about paternity, on the basis of other evidence in the case. However, we shall instead insert a purely formal uniform prior at this node (prior probability of paternity = 0.5, as in Table 2)—for then the formally calculated "posterior odds" on paternity will in fact be numerically identical with the likelihood ratio in favour of paternity based on this marker. Under our assumption of independent markers, these values can then simply be multiplied together to yield the overall likelihood ratio, based on all the markers; and this can then, if desired, be combined with genuine prior beliefs to obtain the correct overall posterior odds on paternity. (It would be possible, but more complicated, to handle the multi-marker data directly in a sequential process: thus we could start with a genuine prior probability distribution at the hypothesis node, and, introducing one new marker at a time, propagate the associated evidence to obtain the induced posterior distribution at the hypothesis node, based on that marker's data—this then becoming the prior to be used with the next marker.)

Consider again the illustrative example of section 2.1, with case data: `cgt` $= \{12, 12\}$, `mgt` $= \{10, 12\}$, `pfgt` $= \{10, 12\}$. Tables 1, 2, 3, 4 and 5 show the (conditional) probability tables corresponding to nodes `pfpg`, `tf = pf?`, `tfpg`, `pfgt` and `cpg`, respectively, for

Table 1. *Probability table for* `pfpg`

`pfpg`:	10	12	x
	0.28425	0.25942	0.45634

Table 2. *Probability table for* `tf=pf`*?*

`tf=pf?`:	yes	no
	0.5	0.5

Table 3. *Conditional probability table for* `tfpg` *given* `tf=pf`*? and* `pfpg`

	`tf=pf?`	yes			no		
	`pfpg`	10	12	x	10	12	X
`tfpg`:	10	1	0	0	0.28425	0.28425	0.28425
	12	0	1	0	0.25942	0.25942	0.25942
	x	0	0	1	0.45634	0.45634	0.45634

Table 4. *Conditional probability table for* `pfgt` *given* `pfmg` *and* `pfpg`

	`pfmg`:	10			12			x		
	`pfpg`:	10	12	x	10	12	x	10	12	x
`pfgt`:	10–10	1	0	0	0	0	0	0	0	0
	10–12	0	1	0	1	0	0	0	0	0
	10–x	0	0	1	0	0	0	1	0	0
	12–12	0	0	0	0	1	0	0	0	0
	12–x	0	0	0	0	0	1	0	1	0
	x–x	0	0	0	0	0	0	0	0	1

Table 5. *Conditional probability table for* `cpg` *given* `tfmg` *and* `tfpg`

	`tfmg`:	10			12			x		
	`tfpg`:	10	12	x	10	12	x	10	12	x
`cpg`:	10	1	0.5	0.5	0.5	0	0	0.5	0	0
	12	0	0.5	0	0.5	1	0.5	0	0.5	0
	x	0	0	0.5	0	0	0.5	0.5	0.5	1

marker FES. Only the actually observed alleles 10 and 12, and the aggregation of all unobserved alleles, indicated by x, need to be represented in the probability tables.

After entering the case evidence at the relevant genotype nodes, we "propagate" it throughout the network, using the software. We can then interrogate the hypothesis node to find its updated probability distribution, conditional on the evidence. The required likelihood ratio, based on the data for this marker, is obtained from the marginal formal posterior distribution at node `tf = pf?`, as given in Table 6: $\mathrm{LR} = 0.6584/0.3416 = 1.9274$—in agreement with the analysis of section 2.1.

Our purely technical use of an "artificial" uniform prior probability of 0.5, to derive likelihood ratios, should be clearly distinguished from the common forensic practice of calculating, and quoting in court, a "probability of paternity" based on this uniform prior

Table 6. *Posterior probability table for* `tf=pf`?

`tf=pf`?:	yes	no
	0.65840	0.34160

(Essen-Möller, 1938). Such a prior assumption will often be unreasonable. All the non-DNA evidence in the case should be used to construct a sensible and defensible prior probability which is then combined with the likelihood ratio (summarizing all DNA evidence) to form a posterior probability. However, this is a task that should properly be left to the judge or jury. We advocate the presentation, along with the overall likelihood ratio, of a table, such as Table 8 (see section 2.3.1 below), where a range of posterior probabilities is given corresponding to a range of possible prior probabilities. For a general discussion of the presentation of statistical evidence in court, see Dawid (2002).

We again stress that we are not recommending the use of a PES for the simple paternity case illustrated in this section, which can be solved by a couple of lines of simple algebra. It is presented merely as an illustrative basic model of how a PES can be formulated, which is then suitable for extension to the more complex cases that we treat below—cases that can not be handled by simple algebra. In particular, the PES for these more complex problems can be built out of the same fundamental local modules that we have already described for the simple problem above.

2.3. Missing data

In certain cases, the DNA profiles of one or more of the "principal actors" in the story are not available, but there is indirect evidence, in the form of DNA profiles of various known relatives. Then simple arguments such as those of section 2.1 can no longer be applied, and their appropriate extension to such a case may not be obvious or practicable. However, it is still straightforward to construct and apply a Probabilistic Expert System representation of the problem, this being now expanded to include the other measured individuals and the relationships between all individuals involved. We illustrate this with two real examples from the case-work of the UCSC Forensic Genetics Laboratory.

2.3.1. Paternity case 1

In this case, the only DNA samples available were on the disputed child `c1`, on the putative father `pf`'s undisputed child `c2` by a different mother, and on `pf`'s brother `b`. In particular, no samples were available for the putative father `pf`, nor for his parents `gf` and `gm`, nor for the mother of either child. The case data, and relevant allele frequencies, are given in Table 7. Figure 3 displays the pedigree, and Fig. 4 the corresponding Probabilistic Expert System network. The necessary conditional probability tables are formed exactly as before. The Probabilistic Expert System again readily supports entering of the evidence (at the grey "observation nodes"), rapid propagation of this evidence through the network, and interrogation of the black "target" node to obtain the desired inference. Assuming independence across markers, the overall likelihood ratio in favour of paternity is obtained as the product of all terms in the last column of Table 7, viz. 13.066. Table 8 shows the implied posterior probability of paternity ($\mathtt{tf} = \mathtt{pf?} = \text{yes}$), for various values of the prior probability. Since the jury may hold or wish to consider a range of prior probabilities, such a table can be a useful way of presenting the impact of the DNA evidence to the court.

Table 7. *Observed genotypes, and their frequencies, for incomplete paternity data, Case 1*

Individual:	b	Frequency	c1	Frequency	c2	Frequency	Likelihood ratio
Marker							
TH01	7	0.165	6	0.263	6	0.263	
	9	0.200	6	0.263	9	0.200	0.672
VWA	17	0.298	15	0.107	15	0.107	
	17	0.298	17	0.298	17	0.298	1.788
D3S1358	15	0.278	15	0.278	15	0.278	
	17	0.220	17	0.220	16	0.233	1.450
FGA	18	0.006	22	0.163	23	0.163	
	26	0.028	26	0.028	26	0.028	8.954
TPOX	8	0.508	8	0.508	9	0.107	
	9	0.107	11	0.271	10	0.089	0.459
CSF1PO	10	0.250	10	0.250	10	0.250	
	12	0.380	12	0.380	12	0.380	1.504
D5S818	11	0.392	11	0.392	11	0.392	
	11	0.392	13	0.165	11	0.392	1.207
D7S820	9	0.130	10	0.240	9	0.130	
	12	0.160	13	0.027	10	0.240	0.428
D13S317	11	0.312	11	0.312	11	0.312	
	11	0.312	11	0.312	12	0.286	2.346
Overall							13.066

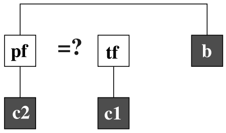

Fig. 3. Pedigree for incomplete paternity data. Case 1.

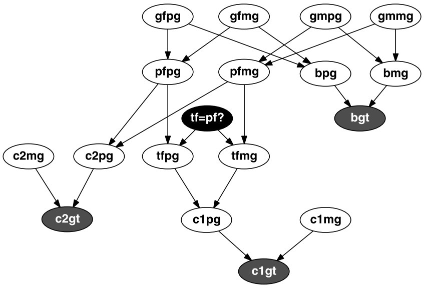

Fig. 4. Network for incomplete paternity data. Case 1.

Table 8. *Posterior probability of paternity for Case 1*

Prior probability:	0.001	0.01	0.1	0.3	0.5	0.7	0.9
Posterior probability:	0.013	0.117	0.592	0.848	0.929	0.968	0.992

2.3.2. Paternity case 2

In this problem, DNA profiles were available on the disputed child `c1` and its mother `m1`, on the putative father's two full brothers `b1` and `b2`, on the undisputed child `c2` of the putative father by another mother `m2`, and on `m2`. No samples were available for the putative father pf, nor for his parents `gf` and `gm`. The pedigree is given in Fig. 5. The network for the corresponding Probabilistic Expert System is now given by Fig. 6. The detailed case data and allele frequencies (for 10 markers) are not given here, but can be supplied on request. Notwithstanding the greatly increased complexity, it is once again completely straightforward, using PES software, to enter the observed evidence, propagate, and interrogate the target node to obtain the required likelihood ratio. This was done using GeNIe: the overall likelihood ratio obtained in favour of paternity was 1303.

Networks such as the above can also be used to analyse the possibility that the true father is some individual in the pedigree other than the putative father. Further relatives, measured or unmeasured, could be investigated by suitable extension of the pedigree and corresponding elaboration of the network.

3. Mutation

A problem that can complicate forensic inference from DNA profiles is the possibility of mutation of the DNA between generations. Indeed, the microsatellite markers typically used for forensic purposes are known to be particularly prone to mutation, with overall mutation rates of between 5×10^{-4} and 7×10^{-3} per generation (Brinkmann *et al.*, 1998). It is thus possible, for example, that a putative father may seem to be excluded by the evidence, whereas in fact he is the true father, but mutation has led to his passing on a *prima facie* impossible

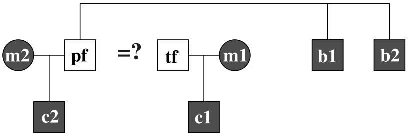

Fig. 5. Pedigree for incomplete paternity data. Case 2.

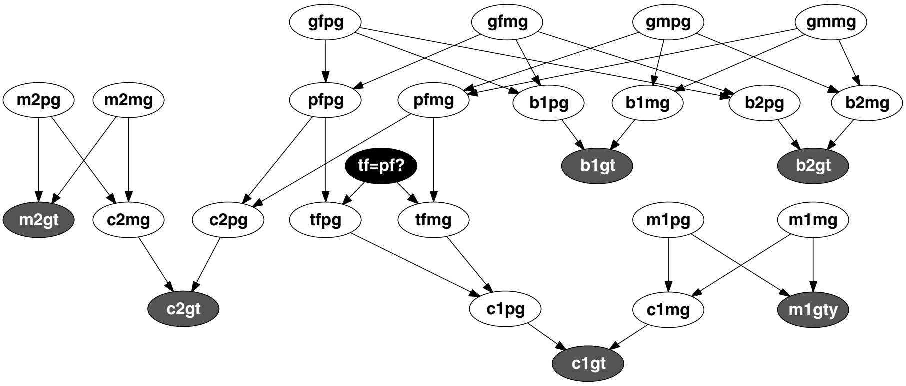

Fig. 6. Network for incomplete paternity data. Case 2.

allele to the child. Particularly in cases where such "exclusion" is based on a single marker, all other markers yielding data consistent with paternity, the possibility of a mutation may need to be taken seriously.

Within a Probabilistic Expert System representation, mutation may be incorporated into the process whereby a child's (say) maternal gene is determined by his mother's two genes. First one of these is chosen at random, as before, to form the "original" (transmitted) gene; then this spawns a new "actual" (inherited) gene, according to some specified mutation process. (In fact our ordering of these two stages is the opposite of Nature's, but the end result will be equivalent so long as the same mutation process operates on the male and female germlines. Our construction involves fewer mutation events, thus yielding a simpler and more efficient graphical representation.) The "actual" genes are thus those which can be observed as constituents of genotypes, while the "original" genes are unobservable.

With this extension, and using e.g. `pfopg` and `pfapg` to represent "putative father's original paternal gene" and "putative father's actual paternal gene", the network for the "simple" paternity problem of section 2.2 is transformed into that shown in Fig. 7. In fact in this new problem it is still possible to perform the calculations directly, without using a PES, although the expressions now become considerably more complex (Dawid *et al*., 2001). Once again, we develop the expert system formulation of this problem mainly as a model for handling more complex cases.

In addition to the previous probability specifications, we need to specify the transition matrix of mutation rates, whereby an original gene mutates into an actual gene. There are various sources of data that can be used to supply estimates of overall mutation rates, but data on rates of transition between specific alleles are sparse, so that we need to make some tentative assumptions on the structure of the transition matrix. For real applications it will be important to investigate the sensitivity of the conclusions to varying assumptions about overall and transition-specific mutation rates (Dawid *et al*., 2001).

Table 9 gives estimated overall mutation rates obtained from data from UCSC, and from Brinkmann *et al*. (1998), who investigated 11 000 meioses for nine STR markers. We have used the estimate $\mu = (s + 0.5)/(n + 1)$, where s is the number of observed mutations and n the number of observed meioses, which can be regarded as a Bayesian estimate under the Jeffreys Beta$\left(\frac{1}{2}, \frac{1}{2}\right)$ prior distribution for a binomial proportion. In particular, this simple Bayesian

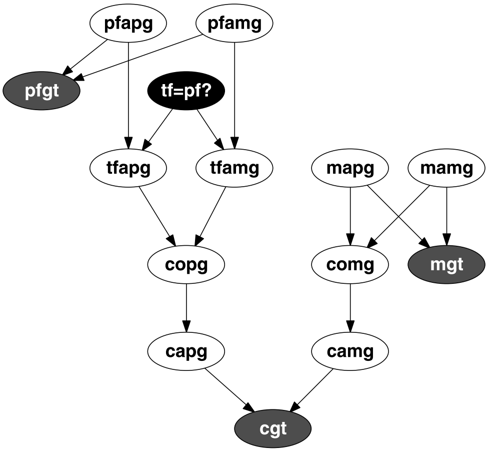

Fig. 7. Simple network with mutation.

Table 9. *Estimated overall mutation rates per thousand generations*

Marker:	F13	VWA	D21S11	D1S80	TH01	FES	MBP	APO-B	COL2A1
Mutation rate:	2.55	2.23	2.69	2.69	2.49	1.76	2.69	2.69	2.69

computation avoids zero estimates. For markers where mutation data were not available the highest observed rate, 2.69×10^{-3}, was taken.

For a given marker, let the mutation transition matrix be denoted by $Q = (q_{ij})$, where q_{ij} denotes the probability of a mutation to "actual" allele j, from "original" allele i. It is reasonable to assume equilibrium, i.e. that the vector of gene frequencies $\pi = (\pi_1, \ldots, \pi_k)^{\mathrm{T}}$ is constant over time. Then π must be a stationary distribution for the associated transition matrix Q, i.e. $\pi^{\mathrm{T}}Q = \pi^{\mathrm{T}}$.

Given π and the overall mutation rate μ, we can construct a transition matrix Q having reasonable properties. Stationarity will be assured if Q is chosen to satisfy the detailed balance condition:

$$\pi_i q_{ij} = \pi_j q_{ji}. \tag{1}$$

We can restrict attention to states having $\pi_i > 0$. Let $S = (s_{ij})$ be a symmetric matrix having $s_{ij} \geqslant 0$ for $i \neq j$ and $\sum_j s_{ij} = 0$ for all i, and let λ be an adjustable positive parameter. We define

$$q_{ij} = \lambda s_{ij}/\pi_i \quad (i \neq j), \tag{2}$$

and

$$q_{ii} = 1 - \sum_{j \neq i} q_{ij} = 1 + \lambda s_{ii}/\pi_i. \tag{3}$$

The detailed balance equation (1) will then be satisfied. The overall mutation rate is

$$\mu = 1 - \sum_i \pi_i q_{ii} = \lambda \times \left(- \sum_i s_{ii} \right), \tag{4}$$

so that we must take

$$\lambda = \mu \bigg/ \left(- \sum_i s_{ii} \right). \tag{5}$$

In order to ensure $q_{ii} \geqslant 0$, all i, we require $\lambda \leqslant \min_i\{-\pi_i/s_{ii}\}$. Consequently, for given S there is an upper limit on the overall mutation rate that can be obtained from this model.

It is biologically reasonable to assume that an allele is more likely to mutate to a neighbouring allele than to one further away, and we choose the matrix S accordingly. Specifically, we take $s_{ij} = \alpha^{|i-j|}$, for $i \neq j$, where α is a fixed constant: this is similar to the step-wise mutational model of Valdes *et al.* (1993). The smaller is α, the greater is the probability that a mutation will be to a closely neighbouring allele.

Using the overall mutation rates in Table 9 and the procedure described above, we obtain, for each marker, the mutation transition matrix to be used at all nodes representing the "actual" transmitted gene, such as `capg` of Fig. 7. Table 10 shows a real example from UCSC casework, analysed using the PES in Fig. 7. Tables 11 and 12 give the allele frequency distribution and the corresponding mutation transition matrix, calculated as described above with $\alpha = 0.5$, for marker VWA. Again, only the observed alleles in the case at hand, and x, representing the aggregation of all other alleles, are required. (In principle such aggregation could now lead to violation of the conditional independence properties represented in the

Table 10. *Paternity case data*

					Putative		Likelihood
Marker	Mother		Child		father		ratio
F13	7	16	5	7	7	7	0.000749
VWA	14	17	17	19	16	17	0.00277
D21S11	28	34.2	32.2	34.2	32	32.2	5.013
D1S80	24	24	18	24	18	24	2.391
TH01	6	7	6	7	6	6	2.360
FES	11	11	11	11	10	11	1.358
MBP	1	5	1	5	1	1	1.491
APO-B	45	47	37	47	37	41	1.341
COL2A1	10	14	10	14	8	14	1.629
Overall							0.00026
Excluding F13							0.347

Table 11. *Selected allele frequencies for marker VWA*

Allele:	14	16	17	19	*x*
Frequency:	0.089	0.197	0.298	0.067	0.349

Table 12. *Marker VWA: Conditional probability table for* capg *[resp.* camg*] given* copg *[resp.* comg*]*

	j				
i	14	16	17	19	*x*
14	0.997272	0.000391	0.000196	0.0000489	0.00209
16	0.000177	0.998652	0.000354	0.0000884	0.000729
17	0.000058	0.000234	0.999109	0.000117	0.000482
19	0.000065	0.000260	0.000520	0.996377	0.00278
x	0.000533	0.000412	0.000412	0.000533	0.998110

Expert System. Full correctness could be ensured by avoiding the aggregation step; however, because of the low mutation rates, in practice the effect of such aggregation will be entirely negligible.)

Note that, for markers F13 and VWA, if mutation was not a possibility the putative father would be excluded by the evidence. For the "non-excluding" markers, i.e. those other than F13 and VWA, introducing the possibility of mutation does not affect the likelihood ratio, to 3 significant figures. It does, of course, have a profound effect for F13 and VWA, changing a zero to a non-zero value.

Using the above mutation model, the overall likelihood ratio, obtained as the product of the entries in the last column of Table 10, is 0.00026, a very small value which effectively excludes this putative father. However, if the data on F13 had been absent the likelihood ratio would have been 0.347, which is by no means negligible if there is other incriminating evidence in the case. We thus see that a single seeming exclusion may not in reality exclude, when the possibility of mutation is taken into account.

When there are no "seeming exclusions" the effect of accounting for mutation will usually be negligible—certainly so for simple cases with no missing individuals—and it can then safely be ignored.

4. Inference about identity

We now turn to consider a problem of a somewhat different nature.

Suppose a DNA trace t belonging to an unknown criminal or from an unidentified body is found. One wishes to know if the trace belongs to any individual in a given pedigree Π, on some of whom we have DNA data; or to some other individual in the overall population $\mathscr{I}$.

Following Dawid & Mortera (1996), we denote by C the random variable indicating the unknown individual in $\mathscr{I}$ leaving the trace. The scene-of-crime or trace DNA evidence is denoted by $\mathscr{E} : \chi_C = t$. Furthermore we let $\mathscr{E}_\Pi$: $\chi_\alpha = \xi$ denote the measured DNA evidence for a set α of identified individuals in pedigree Π. We suppose that, in the absence of the trace evidence χ_C, information on χ, the DNA profiles for all members of the population $\mathscr{I}$, would be irrelevant to the identity of C. Using the notation for conditional independence of Dawid (1979), this property is expressed as:

$$C \perp\!\!\!\perp \chi. \tag{6}$$

Given all the evidence $(\mathscr{E}, \mathscr{E}_\Pi)$, the likelihood L_i of the hypothesis that the trace belongs to individual $i \in \mathscr{I}$ is given by:

$$L_i \propto \mathrm{pr}(\chi_\alpha = \xi, \chi_C = t | C = i) = \mathrm{pr}(\chi_\alpha = \xi, \chi_i = t | C = i) \propto \mathrm{pr}(\chi_i = t | \chi_\alpha = \xi), \tag{7}$$

by (6). For any individual $i \in \mathscr{I}$ unrelated to Π, (7) is just the match probability $\mathrm{pr}(\chi_i = t)$. For each marker, a single propagation in the probabilistic expert system representing the pedigree Π will calculate (7) for all $i \in \Pi$ simultaneously. These likelihoods can then be multiplied across markers (assuming independence), and combined with prior probabilities, using Bayes's theorem, to yield the posterior probability that $C = i$ for various $i \in \mathscr{I}$.

4.1. A murder example

To indicate the complexity which can be handled by this approach, we construct a network for a fictional criminal case, taken from Egeland *et al.* (1997b). A mutilated murdered body has been found, of unknown identity and of male sex. There are a number of individuals who it could be, labelled I.1, II.1, II.2, III.1 and IV.1. There is also a possibility that it is none of these (nor related to any of them), represented by $i \in U$. DNA profiles are available from the body, and from living individuals $\alpha = \{\text{IV.2, IV.3, V.1}\}$. All these individuals are known to be related according to the pedigree Π given in Fig. 8. We construct the associated Probabilistic

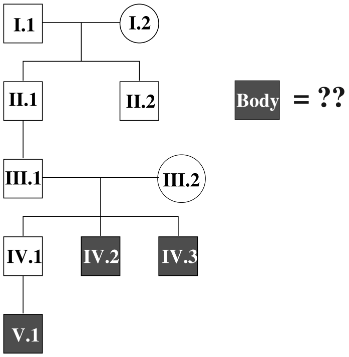

Fig. 8. Murder case: Complex pedigree.

Expert System, incorporating mutation, as in Fig. 9. The (fictional) genotype data, for a single marker, are given in Table 13, and the associated allele frequencies in Table 14.

In the network we merely have to enter and propagate the DNA profile evidence, $\mathscr{E}_\Pi$, observed at the "data nodes" (genotypes for individuals IV.2, IV.3, V.1 in the pedigree), so obtaining the likelihood L_i, as given by (7), for each "suspect" individual $i \in \Pi \cup U$. In particular for any $i \in U$ the likelihood L_i is simply the prior probability $\text{pr}(\chi_i = t)$, i.e. the (estimated) frequency in the data-base of type t genotype. Note that the network also provides likelihoods for the female individuals I.2 and III.2, as well as for the living individuals IV.2,

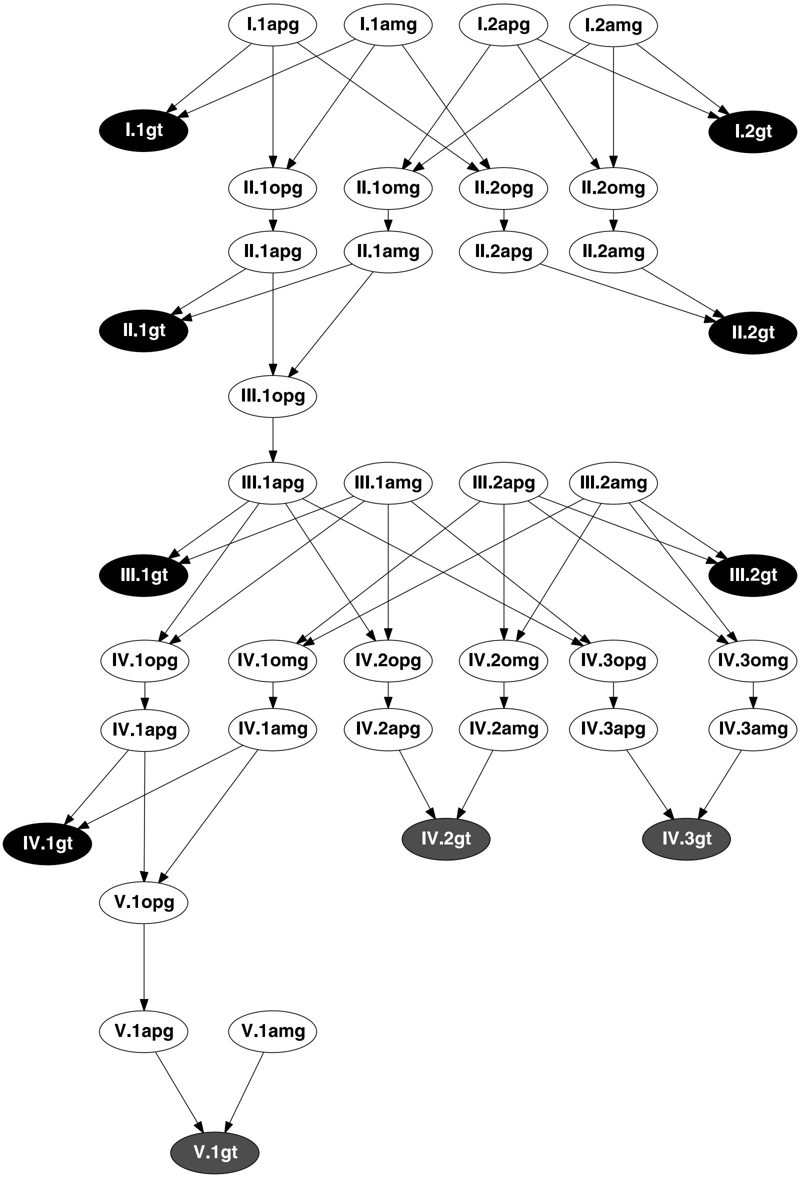

Fig. 9. Network for murder case.

Table 13. *Genotype data, murder case*

Individual:	IV.2	IV.3	V.1	body
Genotype:	*ad*	*ac*	*ab*	*ac*

Table 14. *Allele frequencies, murder case*

Allele:	*a*	*b*	*c*	*d*	*x*
Frequency:	0.01	0.15	0.05	0.35	0.44

IV.3 and V.1, although in all these cases the corresponding prior probabilities $\mathrm{pr}(C = i)$ must be zero. This is not an error, since the likelihood is a measure of the conditional probability of the DNA data observed, under the hypothesis that a specific individual supplied the body; under the assumptions incorporated in the model this conditional probability can be positive, even when the posited hypothesis is known to be false. Of course, in such a case the value of the likelihood is irrelevant: the posterior probability will be zero because the prior is.

Table 15 gives the likelihoods L_i, both in the absence of mutation and allowing for the possibility of mutation (using a simplistic 9-allele mutation transition matrix Q having $q_{ii} = 0.96$, $q_{ij} = 0.005$, $i \neq j$).

The representation and analysis adopted for this problem are somewhat different from those used in section 2 and section 3 above. The same result could in principle have been obtained here by introducing, as there, a "query node", which now would have to be a "child" of all individuals to whom the body could belong. However, that creates a large clique of connected nodes, making such a representation inefficient for computation. Conversely, we could have addressed the problems of section 2 and section 3 by the same logic as used here; however, although this would have provided a straightforward and efficient approach to those problems, we consider that the clarity and simplicity obtained by explicitly representing the hypothesis node in the graph is a strong reason to do so whenever feasible.

The above example appeared in the unpublished preliminary version (Egeland *et al.*, 1997b) of the paper Egeland *et al.* (1997a). These papers describe and illustrate the program PATER (Mostad & Egeland, 1998), based on a routine for organization of pedigree calculations—without, however, our more detailed structuring of the network. In Egeland *et al.* (1997b) the authors state that PATER had not been able to solve the above problem incorporating mutation, after running for 12 hours on a Sparc-Solaris workstation (their calculations involve consideration of about 10^{14} different configurations). In contrast, when

Table 15. *Likelihood L_i for pedigree of Fig. 8*

Individual *i*	No Mutation	Mutation
I.1	0.0084	0.0078
I.2	0.0084	0.0078
II.1	0.0158	0.0199
II.2	0.0084	0.0086
III.1	0.0330	0.0450
III.2	0.0330	0.0362
IV.1	0.4906	0.3859
IV.2	1	1
IV.3	1	1
V.1	0	0
Unrelated	0.001	0.001

formulated as we have done, it can be solved, using the HUGIN software, instantaneously with just one propagation. This underlines the vital importance of constructing an appropriate initial representation of the problem, and using good computational algorithms. Mostad and Egeland have since produced a new program, FAMILIAS (http://www.nr.no/familias), that solves the above problem more efficiently, using a genetic "peeling algorithm" (Elston & Stewart, 1971; Cannings *et al.*, 1978) closely related to the general purpose algorithm used by HUGIN.

5. Conclusions and related problems

This article has attempted to demonstrate the value of applying existing general-purpose probabilistic expert system methodology and software to address problems of forensic genetics. Its principal contribution should be seen as comprised in our suggestions for translating such a problem into a suitable PES representation. Although we have largely proceeded by example, certain general principles emerge: in particular, wherever possible one should aim for a fine-grained representation (e.g. incorporating individual genes, not just genotypes), and build the overall structure out of simple repeatable modules, simply connected together. However, the best representation of any specific problem can vary from one problem type to another.

In future work we plan to develop intelligent software to provide a suitable mix of standardized and problem-tailored routines. We also propose to address departures from some of the simplifying assumptions we have made here. Thus we have assumed Hardy–Weinberg and linkage equilibrium, i.e., independence within and across markers; and, further, that all founders in a pedigree, including unrepresented individuals, can be regarded as drawn at random from the same homogeneous population. We aim to relax these requirements, e.g. by considering populations composed of partially distinct subpopulations (Roeder *et al.*, 1998; Foreman *et al.*, 1997; Dawid & Pueschel 1999). We have also assumed here that we have fully accurate and relevant figures for gene frequencies and mutation rates. These assumptions will be relaxed by extending the framework to allow Bayesian and other forms of statistical learning of relevant parameters from data.

There is a wide range of further identification problems that can be tackled using Probabilistic Expert System representations and computations, either along the lines set out above or through extensions, elaborations or variations of them. We plan to study these with the overall aim of attaching them to a suitable general representational framework.

One such problem, readily handled by the methods presented here, arises when a crime suspect's DNA is not available—perhaps he escaped the country—but DNA evidence is available on some of his relatives. One then needs to compute the likelihood that the crime trace belonged to the fugitive, given the DNA traces of some of his close relatives. This problem can be handled in a similar fashion to paternity cases with missing data, thus avoiding complex algebraic case-by-case computations. Analogous forensic problems arise when someone has been kidnapped, a body part has been sent by the kidnappers, and, given DNA profiles from the kidnapped person's parents, sibling's or other kin, one wishes to compute the likelihood that the part belongs to the kidnapped person. The idea of transforming a complex pedigree into a PES also has less gory applications in cases where a prospective immigrant claims relationship with one or more citizens.

Another important and complex problem that can easily be formulated and solved using a PES (Mortera, 2002) is the analysis of DNA profiles containing a mixture of genetic material from two or more persons (Evett & Weir, 1998, ch. 7). This is common in rape cases, where a sample may contain biological material from the victim, multiple perpetrators, and one or more consensual partners; it can also occur in criminal cases where there has been a scuffle or brawl, for example. Here again the most appropriate PES representation (generally close in spirit to those considered here, albeit a little different in detail) may depend on the specific problem and the identification question that is to be resolved.

A somewhat different kind of problem, unfortunately of increasing importance, is that of identification of multiple remains following disasters such as wars, fires, or earthquakes. A celebrated instance was the discovery in Yekaterinburg of human remains thought to be those of the executed last Tsar Nicholas of Russia, his family and servants (Gill *et al.*, 1994). There was a collection of skeletons, of ascertainable sex and (in some cases) approximate age, from which DNA profiles were obtained. A number of possible pedigrees relating the skeletons to each other and to other known individuals (including Prince Philip, Duke of Edinburgh) can be entertained, and a probabilistic network developed to describe each of these, allowing determination of the most likely pedigree given the evidence. The information on sex and age reduces the very large number of possible pedigrees and establishes plausible terminal nodes, i.e. children who cannot have had offspring. The program FAMILIAS can be used to automate the handling of such multiple pedigree problems (Egeland *et al.*, 2000). Further development is envisaged to integrate its features with the approach presented here.

The exact computational approach embodied in a PES analysis, like that of its forerunners in algorithms for peeling genetic pedigrees, is widely but not universally applicable. In particular, for even quite small "loopy" pedigrees exact analysis rapidly becomes computationally intractable even with the most sophisticated algorithms. In such cases it is usual to resort to approximate methods, such as Gibbs sampling or similar Monte Carlo Markov chain techniques (Gilks *et al.*, 1996). Methods such as "blocking Gibbs" (Jensen *et al.*, 1995) combine features of both approaches, and hold out promise of extending the range of problems that can be handled by the type of graphical representation we have considered here. Here too, identification of an appropriate graphical representation will be essential for computational feasibility.

Table 16. *Summary of the notation used in this article*

Notation	Definition
`pf`	putative father
`tf`	true father
`m`	mother
`m1`, `m2`	mother 1, mother 2
`c`	child
`c1`, `c2`	child 1, child 2
`b`	brother
`b1`, `b2`	brother 1, brother 2
`gf`	grandfather
`gm`	grandmother
`mg`	maternal gene
`pg`	paternal gene
`gt`	genotype
`apg`	actual paternal gene
`amg`	actual maternal gene
`opg`	original paternal gene
`omg`	original maternal gene
`pfpg`, `pfmg`, `pfgt`, etc	putative father's paternal gene, maternal gene, genotype, etc.
`mapg`, `mopg`, etc	mother's actual, original paternal gene, etc.
I.x, II.x, etc	individual x, $x = 1, \dots, n$, of generation I, II, etc.
`tf=pf`?	query node: "true father = putative father?"

Acknowledgements

We wish to thank Hugin Expert A/S for allowing free use of their software. Also, since the graphical interfaces to current PES software are not optimized for the kinds of problem we treat here, we are indebted to Robert Cowell for access to a preliminary version of his program FINEX (Cowell, 2001), which eases the specification task, constructs HUGIN networks, and conducts likelihood analyses. We are grateful to Michael Boehnke, Marina Dobosz, Thore Egeland, and Steffen Lauritzen for valuable discussions and assistance. The comments of an anonymous referee helped to improve the paper. The work was supported by grants from MURST, CNR and ESF (HSSS).

References

Brinkmann, B., Klintschar, M., Neuhuber, F., Hühne, J. & Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Amer. J. Human Genet.* **62**, 1408–1415.

Cannings, C., Thompson, E. A. & Skolnick, M. H. (1978). Probability functions on complex pedigrees. *Adv. Appl. Probab.* **10**, 26–61.

Cowell, R. G. (2001). FINEX: Forensic identification by network expert systems. Research Report 22, Department of Actuarial Science and Statistics, The City University, London.

Cowell, R. G., Dawid, A. P., Lauritzen, S. & Spiegelhalter, D. J. (1999). *Probabilistic networks and expert systems*. Springer Verlag, New York.

Dawid, A. P. (1979). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B* **41**, 1–31.

Dawid, A. P. (1992). Applications of a general propagation algorithm for probabilistic expert systems. *Statist. Comput.* **2**, 25–36.

Dawid, A. P. (2002). Bayes's theorem and weighing evidence by juries. In *Bayes's theorem* (ed. R. Swinburne). *Proc. British Acad.* **113**, in press.

Dawid, A. P. & Mortera, J. (1996). Coherent analysis of forensic identification evidence. *J. Roy. Statist Soc. Ser. B* **58**, 425–443.

Dawid, A. P. & Mortera, J. (1998). Forensic identification with imperfect evidence. *Biometrika* **85**, 835–849.

Dawid, A. P. & Pueschel, J. (1999). Hierarchical models for DNA profiling using heterogeneous databases. In *Bayesian statistics 6* (eds J. M. Bernardo, J. O. Berger, A. P. Dawid & A. F. M. Smith), pp. 187–212. Oxford University Press, Oxford.

Dawid, A. P., Mortera, J. & Pascali, V. L. (2001). Non-fatherhood or mutation? A probabilistic approach to parental exclusion in paternity testing. *Forensic Sci. Internat.* **124**, 55–61.

Egeland, T., Mostad, P. F. & Olaisen, B. (1997a). A computerised method for calculating the probability of pedigrees from genetic data. *Sci. Justice* **37**, 269–274.

Egeland, T., Mostad, P. F. & Olaisen, B. (1997b). Probability assessments of family relations. Technical Report, Norwegian Computing Center, Oslo, Norway.

Egeland, T., Mostad, P. F., Mevåg, B. & Stenersen, M. (2000). Beyond traditional paternity and identification cases: selecting the most probable pedigree. *Forensic Sci. Internat.* **110**, 47–59.

Elston, R. C. & Stewart, J. (1971). A general model for the genetic analysis of pedigree data. *Human Heredity* **21**, 523–542.

Essen-Möller, E. (1938). Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis. Theoretische Grundlagen. *Mitt. Anthropol. Ges.* **68**, 9–53.

Evett, I. W. & Weir, B. S. (1998). *Interpreting DNA evidence*. Sinauer, Sunderland, MA.

Foreman, L. A., Smith, A. F. M. & Evett, I. W. (1997). Bayesian analysis of deoxyribonucleic acid profiling data in forensic identification applications (with discussion). *J. Roy. Statist. Soc. Ser. A* **160**, 429–469.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall, London.

Gill, P. E., Ivanov, P. L., Kimpton, C., Piercy, R., Benson, N., Tully, G., Evett, I., Hagelberg, E. & Sullivan, K. (1994). Identification of the remains of the Romanov family by DNA analysis. *Nature Genetics.* **12**, 417–420.

Jensen, C. S., Kong, A. & Kjærulff, U. (1995). Blocking Gibbs sampling in very large probabilistic expert systems. *Internat. J. Human–Comput. Stud.* **42**, 647–666.

Mortera, J. (2002). Analysis of DNA mixtures using probabilistic expert systems. In *Highly structured stochastic systems* (eds P. J. Green, N. L. Hjort & S. Richardson). Oxford University Press, Oxford, in press.

Mostad, P. F. & Egeland, T. (1998). Probability assessments of family relations using the program "pater". Technical Report, Norwegian Computing Center, Oslo, Norway.

Roeder, K., Escobar, M., Kadane, J. B. & Balazs, I. (1998). Measuring heterogeneity in forensic databases using hierarchical Bayes models. *Biometrika* **85**, 269–287.

Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. & Cowell, R. G. (1993). Bayesian analysis in expert systems (with discussion). *Statist. Sci.* **8**, 219–283.

Thompson, E. A. (2000). MCMC estimation of multi-locus genome sharing and multipoint gene location scores. *Internat. Statist. Rev.* **68**, 53–73.

Valdes, A. M., Slatkin, M. & Freimer, N. B. (1993). Allele frequencies at microsatellite loci: the stepwise mutation model revisited. *Genetics* **133**, 737–749.

Weber, J. L. & May, P. E. (1989). Abundant classes of human DNA polymorphism which can be typed by the polymerase chain reaction. *Amer. J. Human Genet.* **44**, 388–397.

Received November 2000, in final form October 2001

A. Philip Dawid, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK.
E-mail: dawid@stats.ucl.ac.uk