International Congress Series 1288 (2006) 484–491

www.ics-elsevier.com

Representing and solving complex DNA identification cases using Bayesian networks

A.P. Dawid [a,*], J. Mortera [b], P. Vicard [b]

[a] *Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK*
[b] *Università Roma Tre, Rome, Italy*

Abstract. Object-oriented Bayesian networks (OOBNs) can be used to model and solve a wide variety of complex forensic DNA identification problems, involving such complications as missing individuals, mutation, and null alleles. We provide a brief overview of the approach and illustrate its use. In particular, we investigate the effect on paternity ratios of allowing for silent alleles, and show that this can be substantial even when the probability of silence is very small.

Keywords: Bayesian network; DNA profile; Missed allele; Mutation; Null allele; Object-oriented; Paternity testing; Silent allele

1. Introduction

Forensic DNA identification and parentage testing are currently conducted using DNA profiles comprising several highly polymorphic short tandem repeat (STR) genetic markers [1]. The forensic impact of such DNA evidence is captured by the associated likelihood ratio for comparing rival hypotheses [2,3]. However computing this becomes challenging in the presence of such features as missing individuals, mixed trace evidence, mutation, silent alleles, etc. For example, in a paternity case the true father may appear to be excluded when in fact a mutation has taken place, or an allele has not been recorded.

Here we show how the computational technology of *Bayesian networks* (BNs), and especially *object-oriented Bayesian networks* (OOBNs), can be used to model and solve such problems. We briefly describe a construction set of basic OOBN modules for DNA identification, and its application to some problem cases. In particular we show that, in

* Corresponding author.
E-mail addresses: dawid@stats.ucl.ac.uk (A.P. Dawid), mortera@uniroma3.it (J. Mortera), vicard@uniroma3.it (P. Vicard).
URL: http://www.ucl.ac.uk/~ucak06d/ (AP. Dawid).

0531-5131/
doi:10.1016/j.ics.2005.09.115

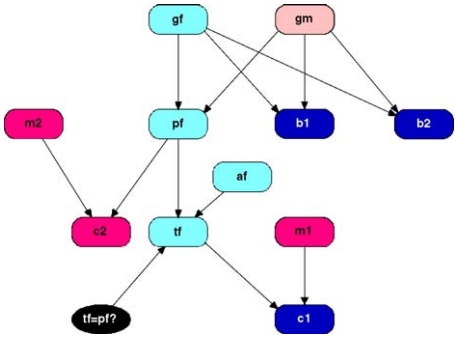

Fig. 1. Pedigree for incomplete paternity case.

paternity testing where we can also observe the putative father's brother, properly allowing for even a very small probability of an inherited silent allele can have a dramatic effect on the strength of the evidence for paternity.

For fuller background and details of the forensic use of BNs the reader is referred to [4–8].

2. Disputed paternity with missing father

Fig. 1 is a "disputed pedigree" representation of a paternity case (originally treated in [4] using a non-object-oriented BN) where we have DNA profiles from a disputed child `c1` and from its mother `m1`, but not from the putative father `pf`. We do however have DNA from `c2`, an undisputed child of `pf` by a different, observed, mother `m2`, as well as from two undisputed full brothers `b1` and `b2` of `pf`. (The sibling relationship is made explicit by the incorporation of the unobserved grandfather `gf` and grandmother `gm`, parents of `pf`, `b1` and `b2`.) The "hypothesis node" `tf=pf`? indicates whether the true father `tf` is `pf`, or is an alternative father `af`, treated as randomly drawn from the population.

The DNA evidence $\mathcal{E}$ consisted of the 6 DNA profiles, each comprising 12 STR markers, from `m1`, `m2`, `c1`, `c2`, `b1` and `b2`. We need to compute the impact of this evidence $\mathcal{E}$ on the case, as measured by the corresponding paternity ratio (likelihood ratio in favour of paternity): $LR = \Pr(\mathcal{E}|H_0) = \Pr(\mathcal{E}|H_1)$: the methods routinely used in simple cases [9] do not apply or readily extend to cases such as this.

3. Bayesian networks

Fig. 1 is in fact the user interface of an object-oriented Bayesian network, constructed using the OOBN software Hugin version 6.[1] We build a separate such network for each STR marker. On entering the available DNA data, we can compute the associated paternity ratio. Finally we multiply these together across all markers to obtain the overall paternity ratio.

Each node in Fig. 1 is itself an "instance" of another generic ("class") network, with further internal structure. We describe only selected features here. A complete description of our networks can be found in [8], while the underlying computational theory is described in [10].

Nodes `gf`, `gm`, `m1`, `m2` and `af` are all instances of a class **founder**; and `pf`, `b1`, `b2`, `c1` and `c2` are instances of a class *child*; `tf` is an instance of class **query**.

[1] Obtainable from www.hugin.com.

Within **founder** (not shown) we have two instances (maternal and paternal genes) of a class **gene** which embodies the relevant repertory of alleles and their associated frequencies.

The internal structure of *child* is displayed in Fig. 2. On the paternal (left-hand) side of *child*, the input nodes `fpg` and `fmg` represent the child's father's paternal and maternal genes (an arrow such as that from `pf` to `c2` in Fig. 1 serves to copy the relevant values from `pf` into `c2`). These are then copied into nodes `pg` and `mg` of an instance `fmeiosis` of a class network **mendel**, whose output node `cg` is obtained by flipping a fair coin (node `cg=pg`?) to choose between `pg` and `mg`; this is then copied to `pg` (child's paternal gene) in network **child**. A similar structure holds for the maternal (right-hand) side of **child**. Finally `pg` and `mg` are copied into an instance `gt` of a network class *genotype*, which forgets the information on parental origin (this is also a feature of **founder**). Any DNA evidence on the individual is entered here.

The hypothesis node `tf=pf`? embodies H_0 (`tf=pf`) when it takes the value *true* and H_1 (`tf=af`) when *false*; it feeds into the instance `tf` of class **query** to implement this selection. We initially, and purely nominally, set both hypotheses as equally probable, so that, after propagation of evidence, the ratio of their posterior probabilities yields the paternity ratio based on this marker. By entering the data for each marker into the appropriate Bayesian network, we can thus easily calculate the associated paternity ratio. The overall paternity ratio, given by their product, was around 1300 for this particular case.

It should be clear that, once supplied with the basic building blocks **founder**, **child** and **query**, we can connect them together in different ways, much like a child's construction set, to represent a wide range of similar problems, including the other cases treated in [4].

4. Mutation

It is easy to modify our networks to account for possible mutation of genes in transmission from parent to child. We distinguish between a child's *original gene* `cog`, identical with one of the parent's own genes, and the *actual gene* `cag` available to the child, which may differ from `cog` because of mutation. We elaborate the class network **mendel** of Fig. 2 as shown in Fig. 3, by passing its original output `cog` ("child's original gene") through an instance `cag` ("child's actual gene") of a new network **mut**, constructed to implement whatever model is used to describe how the value of `cog` is stochastically altered by mutation. The output of mut is then copied to `cg`. Thus **mendel** now represents the result of mutation acting on top of Mendelian segregation.

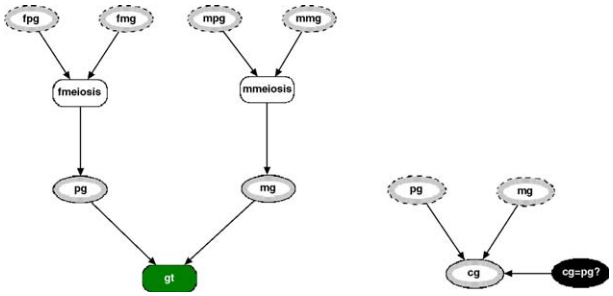

Fig. 2. Networks *child* and *mendel*.

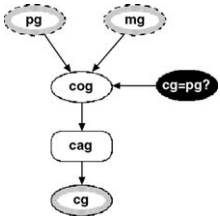

Fig. 3. Revised network *mendel*, incorporating mutation.

Once an appropriate network **mut** has been built, and **mendel** modified as described above, pedigree networks constructed as in §2 will now automatically incorporate the additional possibility of mutation. No other changes are required. We can now use them, for example, to compute the paternity ratio in a seemingly incompatible simple case (all three parties observed) where the incompatibility might be due to mutation. Of course we can deal just as readily with more complex cases such as that of Fig. 1. We have experimented with both a simple "proportional mutation" model and the more realistic "mixed mutation model" [11,12,7]. By including additional adjustable parameter nodes we can also explore the sensitivity of the paternity ratio to the assumed overall mutation rate, or the ratio of the mutation rates for the paternal and maternal lines.

Mutation rates for forensic markers are often estimated from routine case data collected for paternity testing. Then the possibility of non-paternity can itself perturb estimates [12]. We can use essentially the same network to develop estimation methods that properly correct for this [7].

5. Null alleles

A "null" (or "drop-out") allele is one that is not recorded by the equipment used. Then what appears to be a homozygous genotype may in fact be heterozygous, one band being null. This phenomenon will affect the evidential interpretation of DNA profiles. One possible cause is a mutation in the primer binding site leading to failure of the amplification process [13]. Such a null allele will be inherited exactly like any other allele; we then term it *silent*. Another possibility is sporadic failure of the recording apparatus. We refer to such a non-inherited null allele as *missed*.

Again very simple modifications to the lower level networks in our system–see [8] for full details–allow us to incorporate these possibilities, both singly and in combination with each other and/or mutation. Having made such internal modifications, we can continue to use top-level pedigree networks such as Fig. 1 unchanged. In [8] we give several examples: in particular we explore the sensitivity of conclusions to assumed rates of mutation, silence and missingness.

6. Examples

We have built and used OOBNs to analyse a wide variety of complex cases. Here we illustrate the effects of accounting for inherited silent alleles. We use marker VWA, with Austrian–German gene frequencies as given in [8].

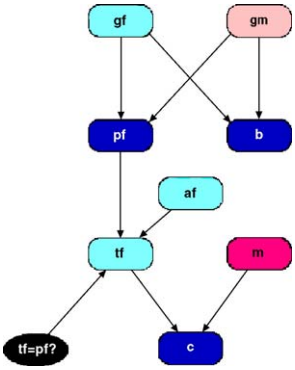

Fig. 4. Pedigree for paternity testing with additional individual.

We consider a simple disputed paternity case where, in addition to the genotypes of the basic triplet m, pf and c, the genotype bgt of the putative father's full brother b may also be observed. The relevant top-level network is shown in Fig. 4. The likelihood ratio in favour of paternity P based on just the triplet data D:=(mgt, pfgt, cgt) is L_D:=$\Pr(D|P)/\Pr(D|\bar{P})$; the impact of the additional information carried by the brother's data B:=(bgt) is measured by L_B:=$\Pr(B|D,P)/\Pr(B|D,\bar{P})$; and the overall paternity ratio, taking account of both D and B, is LR:=$L_D \times L_B$. Under simple Mendelian segregation with fully observed genotypes, the additional information B would be entirely uninformative as to paternity. However this needs no longer be so once we allow for mutation or silent or missed alleles.

Example 1. To illustrate the possible effect of the additional measurement B on the paternity ratio, we consider an example where the triplet evidence D is as follows: mgt={12, 15}, pfgt={14, 14}; cgt={12, 12}. The putative father and child are both apparently homozygous, in a way that would be incompatible with paternity under Mendelian segregation. However pf could still be the true father if he had a silent allele that he passed to the child. Observation of his brother's genotype can help to shed light on this possibility.

Table 1 displays the paternity ratio, allowing for silent alleles, for a range of values for p_s=pr(silent), the probability of silence. The second column gives the paternity ratio L_D based on the triplet data only. The later columns show the additional factor L_B for various possible observations on the brother's genotype bgt. The behaviour of this term is

Table 1
Prima facie incompatible case: mgt={12, 15}, pfgt={14, 14}, cgt={12, 12}

pr(silent)	L_D	L_B with bgt=						
		{16, 20}	{12, 17}	{12, 14}	{14, 17}	{14, 14}	{16, 16}	{12, 12}
0	0	1	1	0.546	0.546	1	6.13	3334
0.000015	0.472	1	1	0.546	0.546	1.0000	6.12	1595
0.0001	2.473	1	1	0.546	0.546	0.9999	6.07	403.7
0.001	7.485	1	1	0.551	0.551	0.9992	5.54	46.07
0.01	8.100	1	1	0.590	0.590	0.9932	3.19	5.45

Likelihood ratio in favour of paternity allowing for silent alleles: L_D, without brother's genotype. L_B, additional effect of brother's genotype.

determined by its relationship to the putative father's observed genotype pfgt. In columns 3 and 4 we consider bgt = {16, 20} and bgt = {12, 17}: b is heterozygous, and does not share any allele (in particular, not a silent allele) with pf. We see that the additional observation B makes no difference whatsoever in this case: $L_B = 1$ for all values of pr(silent).

However, when b is heterozygous but shares an allele with pf, the paternity ratio is reduced by this additional knowledge: intuitively this is because it becomes more likely that pf is a true homozygote, and hence excluded from paternity. This effect is seen in columns 5 and 6 of Table 1 for the cases bgt = {12, 14} and bgt = {14, 17}, so that b and pf share allele 14. The additional paternity ratio factor is the same in both cases, and close to 0.5.

Column 7 refers to the case bgt = pfgt (= {14, 14}). Since b could now have a silent allele the additional data do little to distinguish whether or not pf is a true homozygote. Indeed we see that the extra factor L_B is very close to 1, and so essentially uninformative.

Finally we consider the case that b is apparently homozygous, but with bgt different from pfgt. With such a configuration pf and b might still share a silent allele, and the additional observation B therefore renders it more probable that pf is a false homozygote, who could have passed a silent allele down to the child. As a consequence the paternity ratio is increased. In column 8 the brother exhibits a relatively common allele, bgt = {16, 16}, where $p_{16} \approx 20\%$. Even though this renders him likely to be a true homozygote, the effect on the paternity ratio of the uncertainty introduced by this extra information is to introduce a factor of around 6 for small p_s, reducing somewhat as p_s increases. In column 9 we take a very rare allele, bgt = {12, 12}, where $p_{12} = 0.03\%$. The increase in the paternity ratio is now dramatic. The limiting value of the additional factor L_B as the probability p_s of silence approaches 0 is 3334.33, while the overall paternity ratio $LR = L_D \times L_B$ for this prima facie incompatible case achieves a maximum value of 1027.3, at $p_s = 0.0000642$ (!).

A similar but smaller effect can also be seen in compatible cases when pfgt is apparently homozygous. This is most marked when bgt is also apparently homozygous but different from pfgt: in our experiments this modified the paternity ratio by a factor of around one half for $p_s \sim 10^{-4}$.

7. Mixed trace analysis

Bayesian networks have also been constructed to address other challenging problems of forensic DNA identification, for example the interpretation of mixed trace evidence. Fig. 5 shows a (non-object-oriented) Bayesian network that, taking the observed alleles (repeat numbers) as data, can be used to infer which of a suspect, victim and up to 6 possible unknown individuals might have contributed DNA to a mixed crime trace [5].

A more sensitive analysis uses information on the peak areas measured by an electropherogram in addition to the observed repeat numbers. This requires much more detailed modelling, but again this can be made into a Bayesian network [6]. Fig. 6 shows the top level of a OOBN for two contributors, involving six markers, each represented as an instance of a lower level network **marker**. Because the mixture proportion frac of DNA contributed by one party is a common quantity across markers, we must now handle them all simultaneously within one "super-network".

Cowell et al. [6] analyse the data shown in Table 2 (taken from [14]), involving a 6-marker mixed profile with between 2 and 4 distinct observed bands per marker, and a

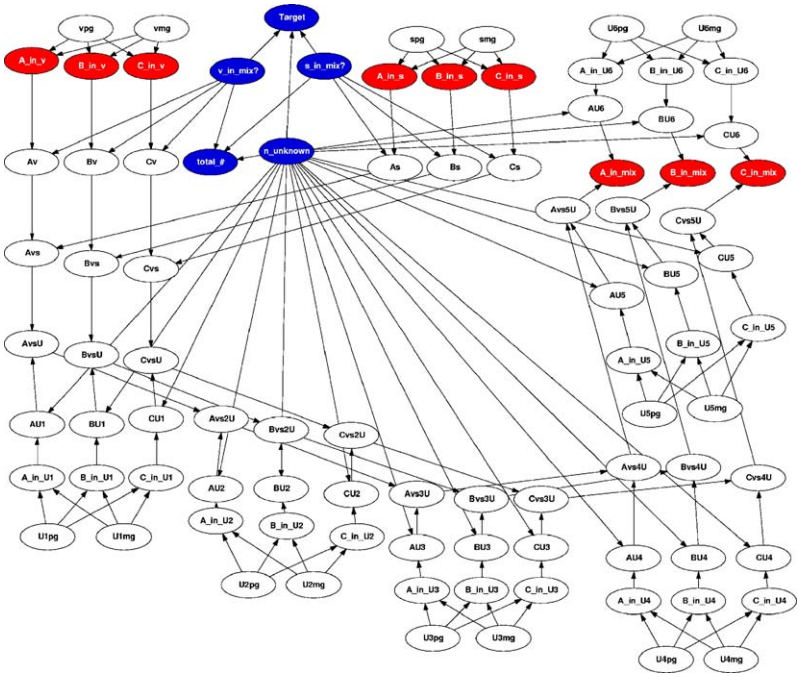

Fig. 5. Single-marker network for mixture, up to 6 unknown contributors.

suspect whose profile is contained in these. It is assumed that this profile is a mixture either of the suspect and one other unobserved contributor, or of two unknowns. Using only the repeat numbers as data, the likelihood ratio for the suspect being a contributor to

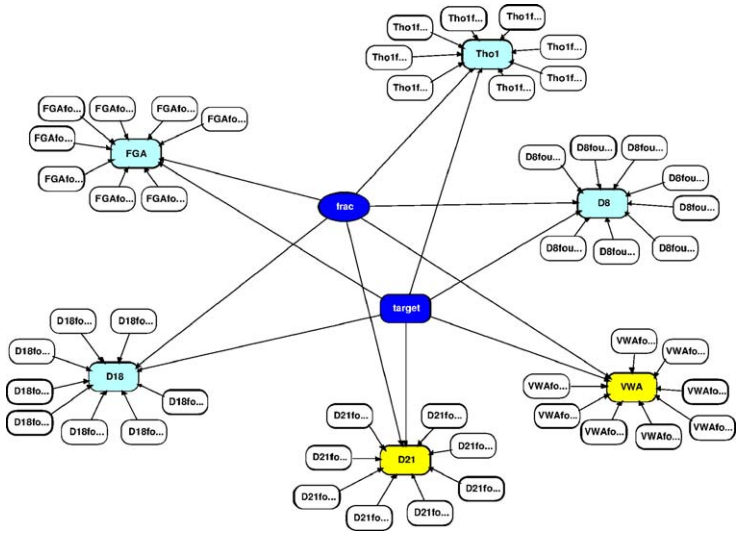

Fig. 6. 6-marker OOBN for mixture using peak areas, 2 contributors (reproduced from [6]).

Table 2
Data for mixed trace with two contributors

Marker	D8		D18		D21				FGA				THO1		VWA			
Alleles	10*	11	14*	13*	16	17	59	65	67*	70*	21*	22*	23	8*	9.3*	16*	17	18*	19
Peak area	6416	3835	6593	8,985	19141	19911	22614	34881	68894	16,099	10,538	10141	7,441	22,368	46699	31472	4188		

The starred values are the suspect's alleles.

the mixture is calculated to be around 25,000. On taking account of the peak areas also, this rises to about 170,000,000.

8. Conclusions

We have illustrated the use of object-oriented Bayesian networks to model and solve complex problems of forensic DNA identification and paternity testing involving missing individuals, mutation, silent alleles and mixed samples. The technology could also be applied to model such further artifacts as stutter, drop-in, contamination, laboratory error, etc., and we hope to address these in future work.

Acknowledgements

This work was supported by Leverhulme Trust Research Interchange Grant F/07 134/K. We are grateful to Steffen Lauritzen and Robert Cowell for their valuable inputs.

References

[1] J.S. Buckleton, C.M. Triggs, S.J. Walsh (Eds.), Forensic DNA Evidence Interpretation, CRC Press, 2004.
[2] I.W. Evett, B.S. Weir, Interpreting DNA Evidence, Sinauer, Sunderland, MA, 1998.
[3] N. Morling, et al., Paternity Testing Commission of the International Society of Forensic Genetics: recommendations on genetic investigations in paternity cases, Forensic Science International 129 (2002) 148–157.
[4] A.P. Dawid, et al., Probabilistic expert systems for forensic inference from genetic markers, Scandinavian Journal of Statistics 29 (2002) 577–595.
[5] J. Mortera, A.P. Dawid, S.L. Lauritzen, Probabilistic expert systems for DNA mixture profiling, Theoretical Population Biology 63 (2003) 191–205.
[6] R.G. Cowell, S.L. Lauritzen, J. Mortera, Identification and separation of DNA mixtures using peak area information using a probabilistic expert system, Research Report No. 25, Cass Business School, City University, 2004.
[7] P. Vicard, et al., Estimation of mutation rates from paternity cases using a probabilistic expert system, Research Report No. 249, Department of Statistical Science, University College London, 2004.
[8] A.P. Dawid, J. Mortera, P. Vicard, Object-oriented Bayesian networks for complex forensic DNA profiling problems, Research Report No. 256, Department of Statistical Science, University College London, 2005.
[9] E. Essen-Möller, Die Beweiskraft der Ähnlichkeit im Vaterschaftsnachweis, Theoretische Grundlagen, Mitteilungen der Anthropologischen Gesellschaft 68 (1938) 9–53.
[10] R.G. Cowell, et al., Probabilistic Networks and Expert Systems, Springer, New York, 1999.
[11] A.P. Dawid, J. Mortera, V.L. Pascali, Non-fatherhood or mutation? A probabilistic approach to parental exclusion in paternity testing, Forensic Science International 124 (2001) 55–61.
[12] P. Vicard, A.P. Dawid, A statistical treatment of biases affecting the estimation of mutation rates, Mutation Research 547 (2004) 19–33.
[13] T.M. Clayton, et al., Primer binding site mutations affecting the typing of STR loci contained within the AMPF/STR® SGM Plus™ kit, Forensic Science International 139 (2004) 255–259.
[14] I.W. Evett, P.D. Gill, J.A. Lambert, Taking account of peak areas when interpreting mixed DNA profiles, Journal of Forensic Sciences 43 (1998) 62–69.