

Sensitivity of Inference in Forensic Genetics to Assumptions about Founding Genes

Peter Green and Julia Mortera

University of Bristol and Università Roma Tre
P.J.Green@bristol.ac.uk

Annals of Applied Statistics, **3**, 731–763 (2009)

Avon local RSS group, Bristol, 10 May 2011

Forensic Identification

Given data E in the form of DNA profiles, we want to make inferences testing hypotheses ('queries') H of these kinds:

Criminal case Did individual A leave the DNA trace found at the scene of the crime?

Criminal case – mixed trace: Did A and B both contribute to a stain found at the scene of the crime? Who did?

Disputed paternity: Is individual A the father of individual B ?

Disputed inheritance: Is A the daughter of deceased B ?

Immigration: Is A the mother of B ? How is A related to B ?

Disasters: Was A among the individuals involved in a disaster? Who were those involved?

Computation of LR

The **weight of the evidence** is reported as a **likelihood ratio**

$$LR = \frac{P(E|H = \text{true})}{P(E|H = \text{false})}.$$

This can be computed in a Bayesian network from:

$$LR = \frac{\Pr(E | H = \text{true})}{\Pr(E | H = \text{false})} = \frac{\Pr(H = \text{true} | E)}{\Pr(H = \text{false} | E)} \div \frac{\Pr(H = \text{true})}{\Pr(H = \text{false})}.$$

$$\text{Posterior odds} = LR \times \text{Prior odds}$$

This talk is on **methods** for assessing **sensitivity to assumptions** in such LR calculations.

Overview

- Bayesian networks
- Genetic background
- Scenario 1: Criminal Identification
- Uncertainty in Allele Frequency
- *Identity by Descent, Subpopulations*
- *Scenario 2: DNA Mixtures*
- *Scenario 2a: DNA Mixtures using peak areas*
- *Paternity cases*

Focus is on methodology: numerical results are only illustrative.

A more general setting - Bayesian networks

- Directed Acyclic Graph (DAG)
- Nodes V represent discrete (finite) random variables $X_v, v \in V$
- Specify conditional distributions of children given parents:
 $p(x_v | x_{\text{pa}(v)})$
- Joint distribution is then $p(x) = \prod_{v \in V} p(x_v | x_{\text{pa}(v)})$
- Lauritzen & Spiegelhalter algorithm enables efficient computation of $p(x_v | x_A)$ for all $v \in V$ and $A \subseteq V$ by *probability propagation*.
- Mortera, Dawid, Lauritzen, etc., have demonstrated convenience and flexibility of using BNs to compute forensic genetic inferences.

Genetic Background

An identified area (locus) on a chromosome is a *gene* and the DNA composition on that area is an *allele*.

A DNA *marker* is a known locus where the allele can be identified in the laboratory.

Short Tandem Repeats (STR) are markers with alleles given by integers. If an STR allele is 5, a certain word (e.g. **CAGGTG**) is repeated exactly 5 times at that locus:

... **CAGGTGCAGGTGCAGGTGCAGGTGCAGGTG** ...

Standard Assumptions

A **genotype** of an individual at a locus is an unordered pair of genes, e.g.

Marker	Genotype
D13	{9, 14}
FGA	{21, 22}

It is customary to assume that all actors are drawn from a *homogeneous population* in *Hardy-Weinberg equilibrium*, with *known* allele frequencies, e.g. for D13 in Caucasian populations:

allele	8	9	10	11	12	13	14
frequency	.113	.075	.051	.339	.248	.124	.048

Scenario 1: Criminal Identification

A simple case of criminal identification:

we have a DNA profile found at the scene of the crime which matches the DNA profile of a suspect. We denote this evidence by E .

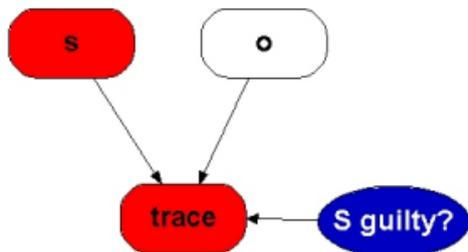
The query or hypothesis H to be investigated: Did the suspect leave the trace at the crime scene? (“suspect is guilty”?)

The LR that is reported to help answer this question compares the probability ($= 1$) of the evidence given that the suspect left the trace, with the probability ($\ll 1$) that a randomly-chosen member of (a suitable) population left the trace.

Criminal Identification

Crime and suspect's DNA profile (excerpt)

Marker	D13	D3	D5	D7	FGA
Evidence E_m	9 14	11 17	9 11	10	21 22
Frequency	.08 .05	.002 .125	.05 .38	.24	.19 .22

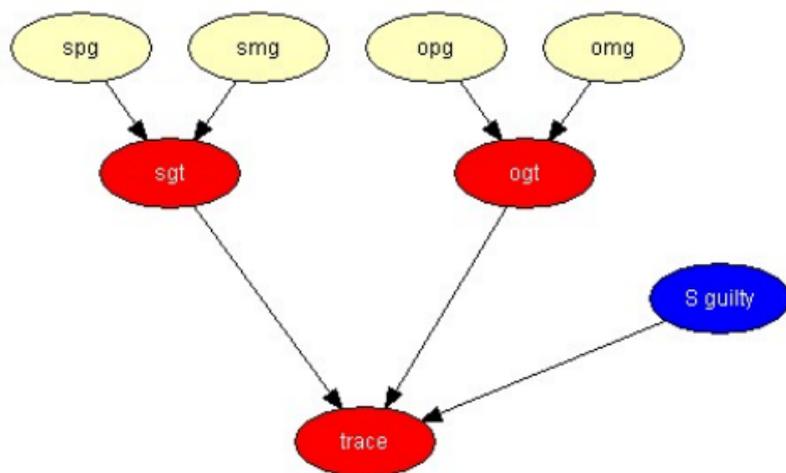


Criminal Identification

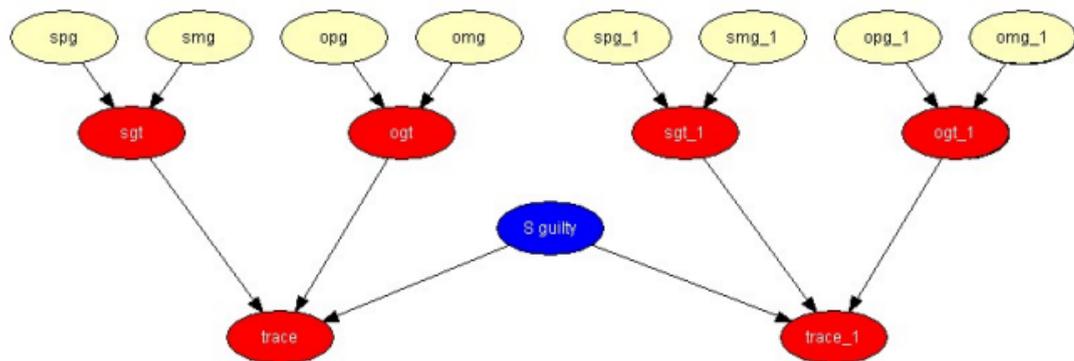
How might the standard assumptions be questioned?

- Allele **frequencies are not known** perfectly
- Individuals (“actors”) in the model are not independent – concept of **identity by descent**
- Population is **heterogeneous** – a mixture of subgroups
- Other failures of Hardy-Weinberg equilibrium

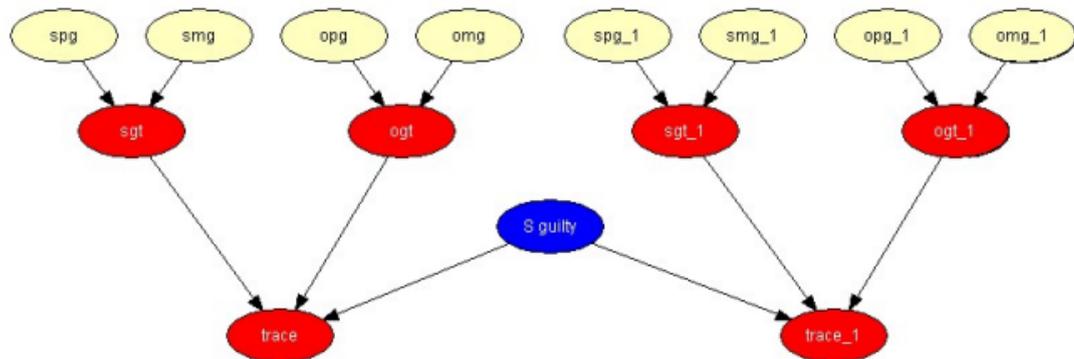
BN for Criminal Identification - 1 marker



BN for Criminal Identification - 2 markers



Joint distribution of all variables



$$p(S \text{ guilty?}) \prod_m [p(\text{spg}_m)p(\text{smg}_m)p(\text{opg}_m)p(\text{omg}_m)] \\ \times \prod_m [p(\text{sgt}_m | \text{spg}_m, \text{smg}_m)p(\text{ogt}_m | \text{opg}_m, \text{omg}_m) \\ \times p(\text{trace}_m | \text{sgt}_m, \text{ogt}_m, S \text{ guilty?})]$$

Sensitivity in Bayesian Networks

Interest is in

$$\log LR = \log \frac{P\{T = 1|E\}}{P\{T = 0|E\}}$$

regarded as a function of the distribution f of the *founding genes*, in this case, $\prod_m [p(\text{spg}_m)p(\text{smg}_m)p(\text{opg}_m)p(\text{omg}_m)]$.

All of our questions about sensitivity can be expressed through alternative settings for f .

Some generate **dependence between founding genes**.

Methods for assessing sensitivity

In our paper, we present:

1. *Algebraic/analytic methods*
 - *Constrained steepest descent*
 - *Linear fractional programming*
2. Structural methods

for different variations on standard assumptions, illustrated on a range of criminal identification and disputed paternity scenarios.

Assessing sensitivity through structure

Here the idea is to model alternatives to the standard assumptions by elaborating the Bayesian Network structure, and using probability propagation on the augmented models.

But probability propagation can no longer be used when we have continuously distributed parents for discrete nodes.

Variation 1: Uncertain Allele Frequencies

Allele frequencies are *not* fixed probabilities, but empirical frequencies in a database.

Assuming the idealisation of a **Dirichlet prior and multinomial sampling** the posterior distribution of a set of probabilities r is Dirichlet $(M\rho(1), M\rho(2), \dots, M\rho(k))$.

The founding genes (spg, smg, opg, omg) are drawn (conditionally) i.i.d. from the distribution r across alleles, which has the above Dirichlet distribution where M is the (posterior) **sample size** and ρ are the **database allele frequencies (posterior means)**.

This corresponds to the standard set-up for a Dirichlet process model and *can be represented in a BN using a Pòlya urn scheme* (thus integrating out r).

UAF: Pólya urn scheme

Founding genes g_1, g_2, \dots are identically distributed, and exchangeable, but not independent.

$$g_1 \sim \rho$$

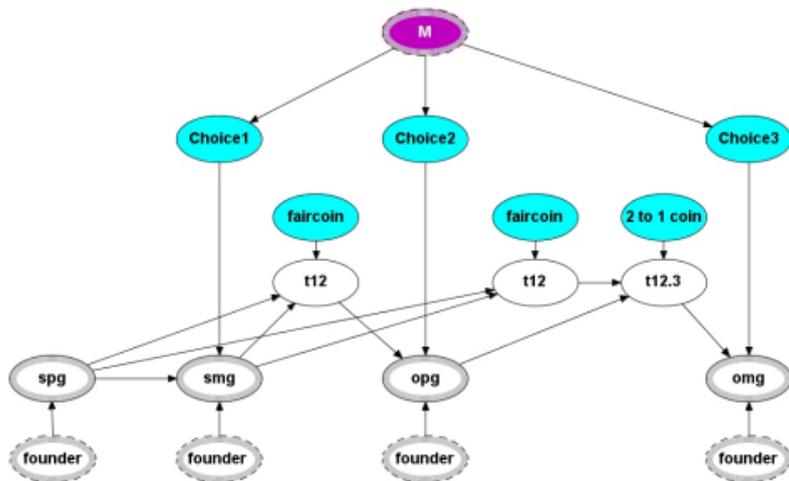
$$g_2|g_1 \sim \frac{1}{1+M}\delta_{g_1} + \frac{M}{1+M}\rho$$

$$g_3|g_1, g_2 \sim \frac{1}{2+M}\delta_{g_1} + \frac{1}{2+M}\delta_{g_2} + \frac{M}{2+M}\rho$$

In general, suppose that n genes have been drawn at random, of which m are allele a , then the probability that the next gene is also allele a is

$$\frac{m + M\rho(a)}{n + M}$$

UAF: Pólya urn scheme as a BN



There are other ways to represent this model – but in this version, all choices are binary, thus reducing the clique table sizes and hence computational burden.

Variation 2: Identity by descent

Near relatives show **positive dependence** between their genes, through **shared ancestry**. For example, two siblings have the same paternal gene with probability 0.5 by virtue of inheritance from their common father, **on top** of the possibility of equality arising from two independent draws from the gene pool.

The traditional way to quantify this is by means of a scalar quantity variously denoted θ or F_{ST} , that we call the coancestry coefficient. $\theta = F_{ST} = 0$ expresses independence; positive values quantify the amount of relatedness or in-breeding in a population (we think of this as ambient IBD).

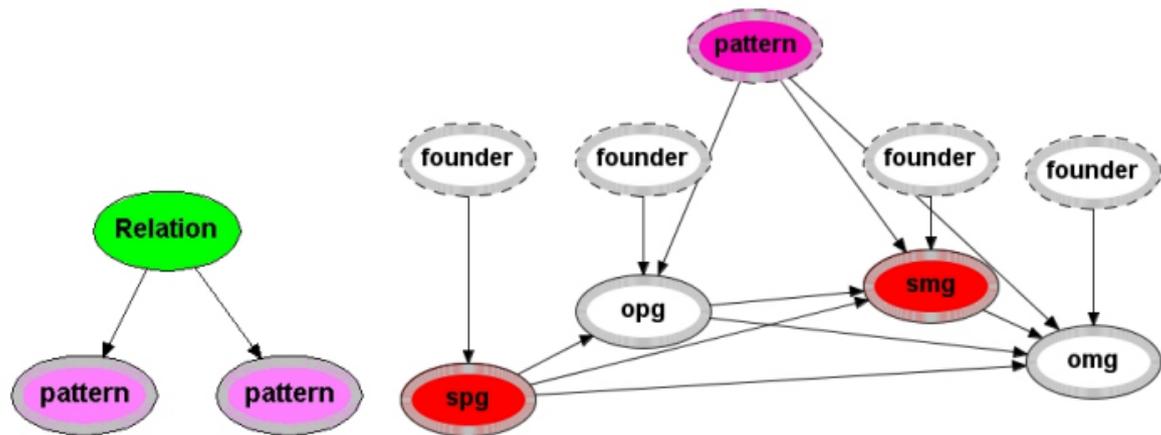
In forensic genetics calculations, likelihood ratios are often adjusted for $\theta = F_{ST} > 0$ using correction formulae due to Balding and Nicholls. These formulae have to be derived from scratch for each new situation: can be quite intricate calculations.

Identity by descent

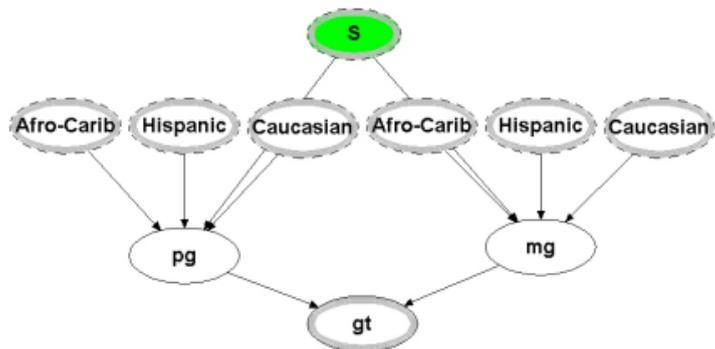
The Balding/Nicholls approach can only be approximate given realistic patterns of relatedness, is only suitable for low levels of relatedness, and ignores the fact that **IBD introduces dependence between markers when relationships are uncertain.**

We consider instead some explicit patterns of close relatedness (parent/child, siblings, half-siblings, . . .) with various probabilities and compute LRs exactly, even in the multiple marker case, using an elaborated BN.

Partial networks representing relation R and IBD



Variation 3: Heterogeneity – uncertainty in subpopulation



This induces dependence between markers, m . S is same for all m so mixing across subpopulations is not the same as using mixture of allele frequencies.

Marker data may not be (conditionally) independent

Usually, the likelihood ratio LR for $E = \{E_m\}$ on $m = 1, 2, \dots, M$ markers is given by the *product rule*:

$$LR = \frac{P\{E|T = 1\}}{P\{E|T = 0\}} = \prod_{m=1}^M \left\{ \frac{P\{E_m|T = 1\}}{P\{E_m|T = 0\}} \right\}.$$

For **IBD** and **HET** the **product rule (PR)** fails to apply (they have latent variables common to all markers).

In such cases, we have to either build a huge network including all markers at the same time, or loop over markers, running separate BNs for each, averaging resulting joint probabilities appropriately before forming LRs.

Some numerical results for criminal identification

LR	Standard	UAF	IBD	HET
D13	138.9	106.6	88.7	126.7
D3	1162.8	194.6	111.9	3488.4
D5	27.7	23.6	20.5	35.6
D7	16.9	14.6	13.7	11.8
...

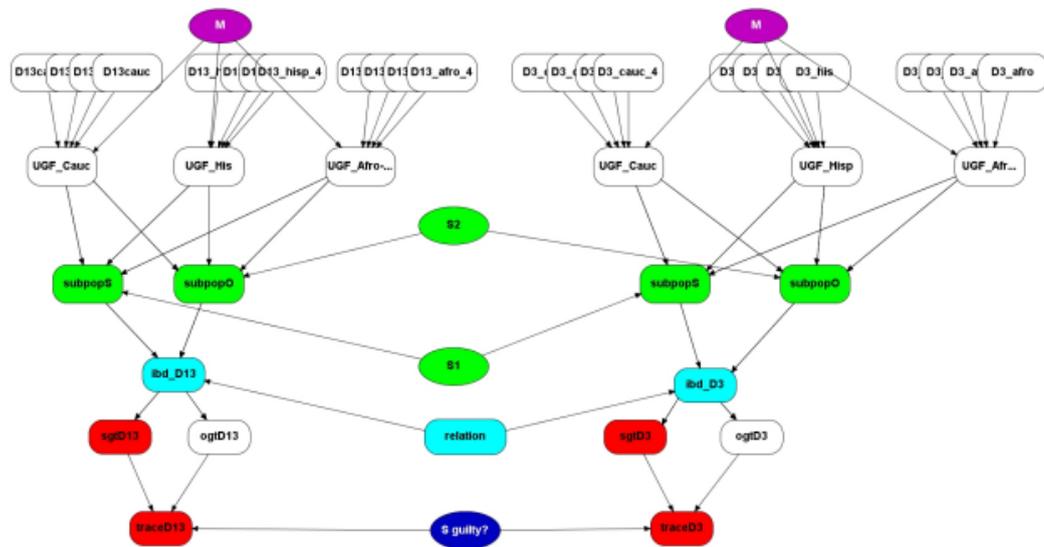
Overall $\text{Log}_{10}LR$ for 8 markers

exact	13.38	12.10	7.71	13.85
product rule	13.38	12.10	11.54	13.57

[Basis: UAF: $M = 100$; IBD: parent/child and half-sibs w.p. 5% each.]

Overall LR for UAF is about 20 times smaller than baseline, whereas for IBD it is roughly 460×10^3 smaller than baseline and 7×10^3 smaller than product rule.

BN for several markers, with uncertain allele frequencies, IBD and population heterogeneity



Scenario 2: Identification from a DNA mixture

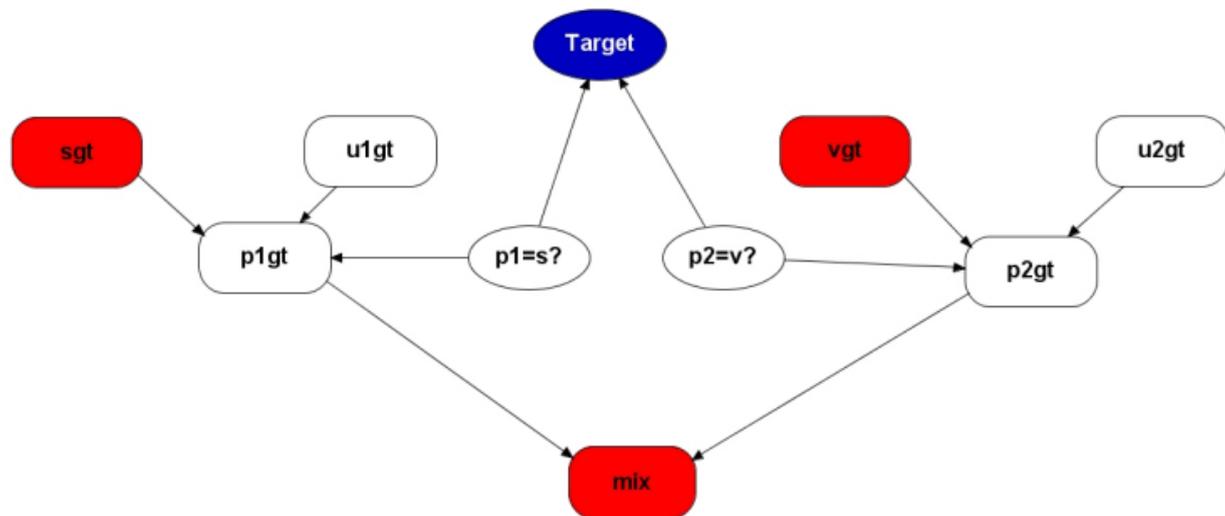
Suppose that analysis of a trace (for example a blood stain) at a crime scene reveals presence of more than two alleles at one or more markers, showing that more than one individual contributed to the trace. The stain (**mix**) can be explained as a combination of the genotypes of a victim and a suspect (**vgt** and **sgt**). How incriminating is this evidence?

We have to assess the probabilities that other members of the population could have contributed to the stain, perhaps in combination with the victim or suspect (and possibly that more than two individuals were involved).

Identification from a DNA mixture

Marker	mix	sgt	vgt
D13	8 11	8 8	8 11
D3	16 18	18 18	16 16
D5	12 13	12 13	12 12
D7	8 10 11	8 10	8 11
FGA	22 24 25 26	22 26	24 25
THO1	6 7	6 7	6 7
TPOX	8 11	8 8	8 11
VWA	17 18	17 17	17 18

BN for DNA mixture



Note: $4 \times 2 = 8$ founding genes in this case.

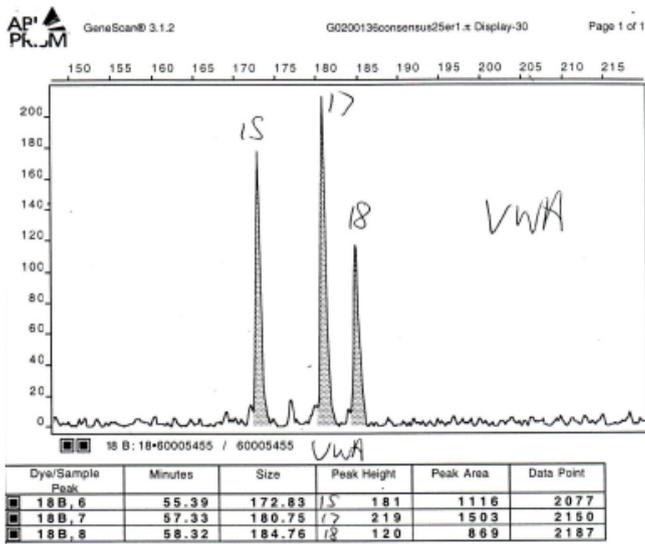
Results for DNA Mixture

Marker	Baseline	UAF	IBD	HET
D13	5.22	4.85	4.83	7.17
D3	7.10	6.38	6.22	6.72
D5	3.63	3.36	3.40	3.53
Overall \log_{10} LR for 8 markers				
exact	6.59	6.33	4.85	6.52
product rule	6.59	6.33	6.22	6.46

The baseline is roughly 1.8, 55 and 1.2 times bigger than those for UAF, IBD and HET. The PR is about 23 times bigger than the exact for IBD; for HET it is about 1.2 times smaller.

Postscript: Using more information - peak areas

The data are not really discrete; we can try to get more out of the analogue profile, especially in the mixed trace identification problem.



Cowell/Lauritzen/Mortera peak area model

In a mixed trace, for each particular marker, let W_{ia} be *peak weight* for individual i and allele a – defined as peak area multiplied by allele number, to correct for preferential amplification. CLM derive an independent Gamma model for the W_{ia} , with means proportional to $\theta_i n_{ia}$ where θ_i is the proportion of DNA in profile that is contributed by individual i (constant across markers), and $n_{ia} = 0, 1, 2$ is the number of alleles a in i 's genotype.

It follows that the relative total weights

$$R_a = \frac{\sum_i W_{ia}}{\sum_a \sum_i W_{ia}},$$

which are observable, have a Dirichlet distribution with parameters $(\sigma^{-2} - 1)\mu_a$ where $\mu_a = (1/2) \sum_i \theta_i n_{ia}$. They fix σ^2 to match reported major/minor peak area ratios.

CLM peak area model as a BN

In place of the discrete mixed trace at each marker, we now have a data node for the relative total weight vector (R_a), whose parents in the DAG are the individuals' genotypes, as before, and an additional node holding the unknown contribution fractions (θ_i).

The latter can be discretised without much loss, but how can we represent the (R_a) in a (discrete) BN?

Fortunately, the (R_a) are observed, so their possible other values are immaterial – all that they contribute to the (posterior) joint probability of all other variables is the likelihood. So we can include in the BN a dummy node for each marker, with parents θ and marker-specific genotypes, providing the appropriate factor for the probability propagation algorithm.

Sensitivity to prior assumptions with peak areas

This can be studied exactly as before.

Markers will always be dependent, because of uncertainty in the shared latent variable θ .

Evelt data

Marker	Relative weights on alleles			
	1	2	3	4
D8	0.435	0.029	0.537	—
D18	0.887	0.054	0.059	—
D21	0.053	0.068	0.428	0.452
FGA	0.570	0.391	0.039	—
THO1	0.402	0.598	—	—
VWA	0.417	0.088	0.475	0.020

Suspect's genotype highlighted in blue.

Is the crime-scene trace a mixture of the suspect and an unknown individual, or two unknowns?

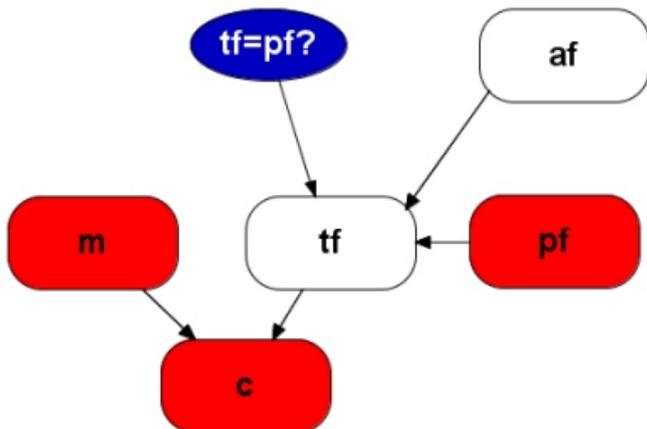
Evet data – provisional results

Marker	Baseline	UAF	IBD
D8	12.73	11.43	11.39
D18	32.00	24.29	26.12
D21	40.26	34.56	34.55
FGA	8.11	7.54	7.49
THO1	7.94	7.27	7.22
VWA	5.28	5.27	5.23
Overall \log_{10} LR for 6 markers			
exact	8.23	7.93	6.20
product rule	6.75	6.44	6.46

Discrete uniform prior on θ ; σ^2 fixed at 0.01.

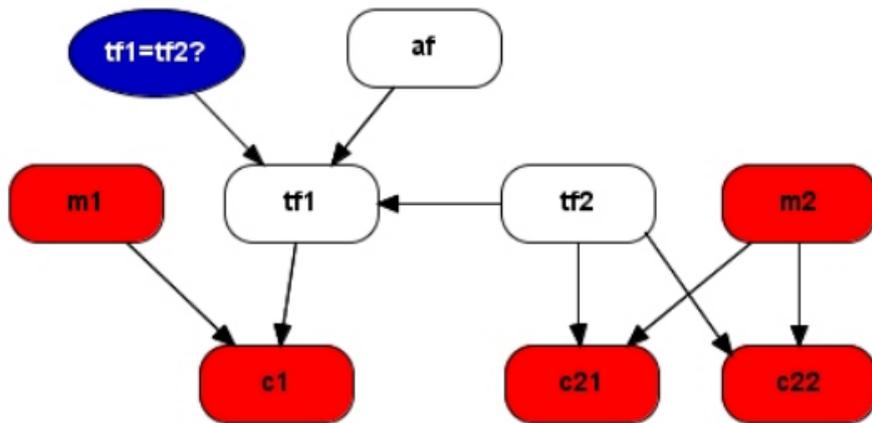
Inference much more certain using peak weights – discrete profiles give $\log_{10}(\text{LR})$ of 4.40 in baseline case.

Other scenarios: paternity cases



A simple disputed paternity case. Some likelihood ratios:
standard: 1318; IBD: 202 (798 by product rule).

Other scenarios: paternity cases



A more complex scenario - disputed sibship.

Discussion

- We have a **range of different methods**. Some of these could be applicable to other domains of application of BNs.
- Results show that effects of IBD, UAF and HET can be quite **dramatic**.
- IBD **more subtle** than the standard θ (F_{ST}) approach.
- The Bayes net approach extends to deal with a number of important variations on standard assumptions.
- IBD and HET induce **dependence among markers** which can be handled it in **one big net** or using **smaller nets and looping over latent variables**.
- **Can infer** the posterior probability of **a specific relationship R** among actors. Useful in immigration cases?
- **Free software GRAPPA in R**
(<http://www.stats.bris.ac.uk/~peter/Grappa>) for construction of and inference in discrete BNs.

To follow up

- “Sensitivity of inferences in forensic genetics to assumptions about founding genes”, by Green and Mortera, *Annals of Applied Statistics*, **3**, 731–763 (2009).
doi: 10.1214/09-AOAS235.
ArXiv: <http://arxiv.org/abs/0908.2862>.
- Webpage: www.stats.bris.ac.uk/~peter/Sensitivity
- Email: P.J.Green@bristol.ac.uk, mortera@uniroma3.it