Bayesian Networks for Forensic Identification Problems

Aarhus University, Fall 2003, Lecture 7

Steffen L. Lauritzen, Aalborg University

Overview

- Forensic identification

- DNA profiles

- Basic paternity cases

- Indirect information

- Mutation

- Body identification

- Mixtures

- Other issues

Forensic identification

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

Immigration cases: Is A the mother of B? Are A and B related at all? If so, how?

Criminal cases: Did person A contribute to a given stain, found at the scene of the crime? Who contributed to the stain?

Disasters: Was A among the individuals found in a grave? How many of a named subset of individuals were in the grave? Who were found in a grave?

STR markers

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

A locus thus corresponds to a (random) variable and an allele to its realised state.

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...

Mitochondrial DNA

The human cell also contains DNA molecules outside the nucleus, known as *mitochondrial* DNA (mtDNA).

mtDNA is maternally inherited, i.e. it is passed in identical form from mother to child, ignoring mutation.

This makes mtDNA important for evolutionary genetics. But it is also significant for forensic identification:

Two persons which are related through a maternal line will have (almost) identical mtDNA.

Inheritance of DNA

As mentioned, *mtDNA is maternally inherited* and passed unchanged from mother to child.

Similarly, *the Y-chromosome is paternally inherited,* i.e. passed from father to son in identical form.

So *two male individuals related through a paternal line will have identical Y-chromosomes.*

The homologous chromosome pairs are inherited in a more complex fashion, where *recombination* can occur during the process of forming gametes, known as *meiosis.*

Meiosis

A	10	8	11

Paternal
Maternal

B	8	6	14

A	8	6	11

Gamete 1
Gamete 2

B	10	8	14
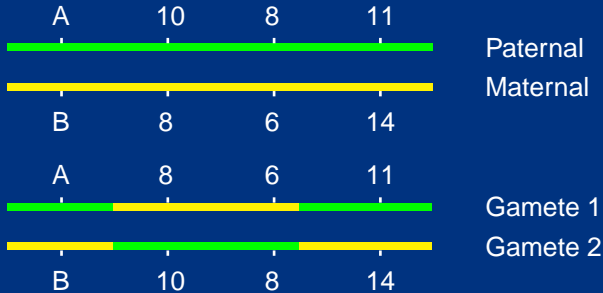
During human reproduction cells form *gametes*, where maternal and paternal DNA is mixed. A child receives one randomly chosen gamete from mother and one from father, to form a new homologous pair.

DNA profile and genotypes

The *genotype* of an individual at a given locus is the unordered pair of alleles at that locus. One cannot measure which allele originated from the mother and which from the father.

The genotype is typically reported as $(12, 14)$ or (A, B), so that the smallest is mentioned first.

A *DNA profile* consists of measurements of the genotype at a number of marker loci. Standard kits use 9 or 10 markers, but occasionally more markers are measured.

Markers are generally chosen on different chromosomes, to avoid problems of *linkage*, i.e. dependence created in the process of meiosis.

Classical paternity case

- DNA profiles of *mother*, a *child*, and a *male* individual, known as the *putative father*. Denote this *evidence* by E.

- Query Q to be investigated :

  > *Is the putative father equal to the true father?*

- Weight of evidence reported as a *likelihood ratio:*

$$L = \frac{P(E \mid Q = \texttt{true})}{P(E \mid Q = \texttt{false})}.$$

Using Bayesian networks

- Make BN using genetic laws with Q representing *query node* assuming random genes of putative father.

- Let $P(Q = \texttt{true}) = P(Q = \texttt{false})$ so we have

$$L = \frac{P(E \mid Q = \texttt{true})}{P(E \mid Q = \texttt{false})} = \frac{P(Q = \texttt{true} \mid E)}{P(Q = \texttt{false} \mid E)}$$

and compute the latter by probability propagation.

We can make a network for each independent marker and multiply likelihood ratios, or we can make a network incorporating all markers at once.

Object-oriented specification of BN
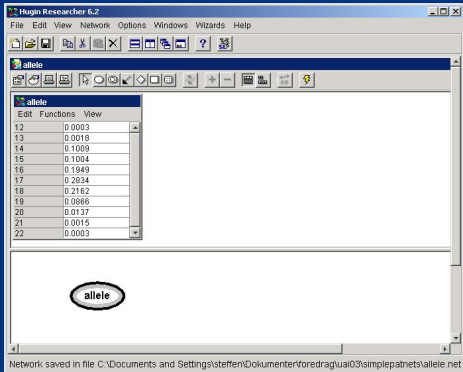
- Objects are *instances* of BNs of certain class

- Objects have *input nodes* and *output nodes*, and also ordinary BN nodes

- Instances of a given class have *identical conditional probability tables* for non-input nodes

- Objects are connected by directed links from output nodes to input nodes. The links represent *identification* of nodes, so nodes must be of same type and have the same states.

OOBN for paternity case: single marker
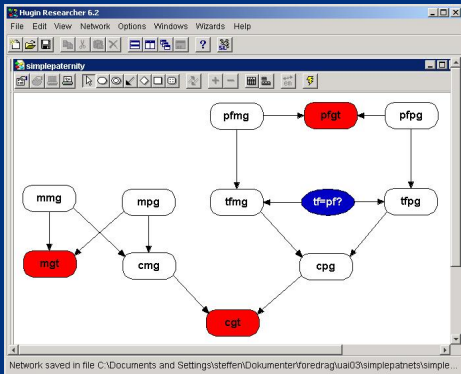
Each node represents itself a Bayesian network.

Allele

This class represents a randomly chosen allele

Faircoin
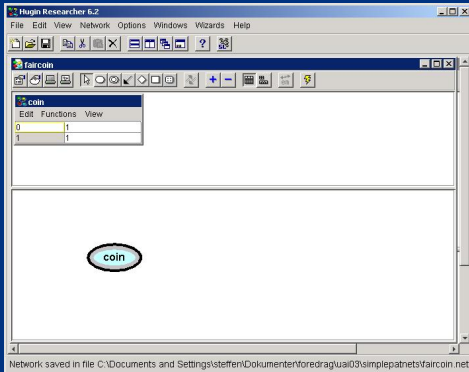
Represents a coin, used to choose allele under meiosis

Meisis
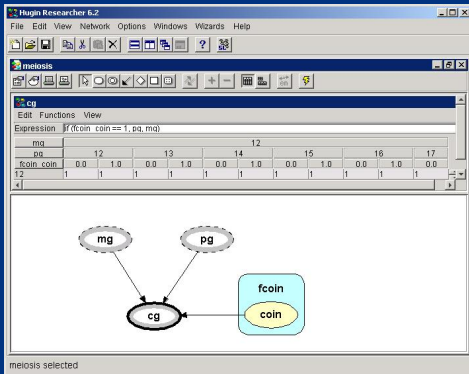
Represents the transmission of allele through meiosis

Who is the father?

Is the allele from the putative father or random?

Genotype

Observation of the smallest and largest allele

Expanded OOBN
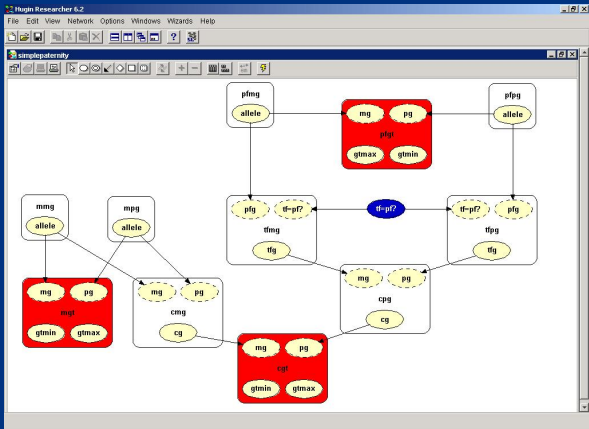
Results
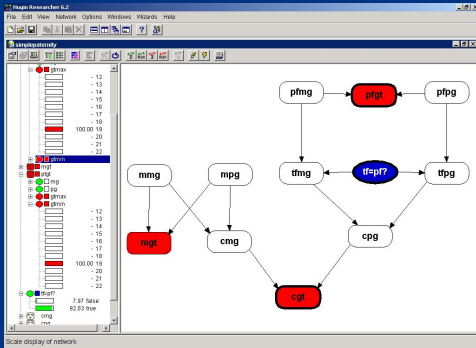
Mother: $(15, 16)$, child: $(15, 19)$, male: $(19, 19)$; $L = 92.03/7.97 = 11.55$.

Indirect evidence: only brother available

Brother of pf: $(19, 19)$; $L = 86.25/13.75 = 6.27$.

Mutation

Possible mutation in transmission of alleles

Mutation in male germline
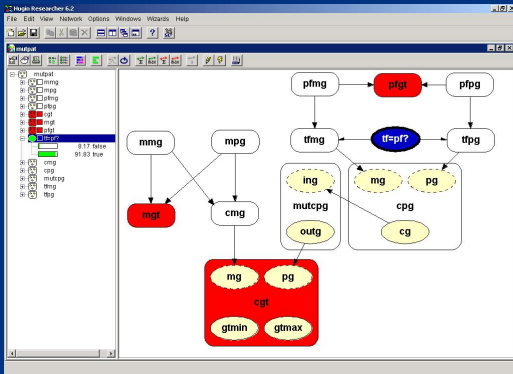

$$L = 91.83/8.17 = 11.24.$$

Body identification

Identification of a *single* dead body is not very different for paternity cases.

For example, if a missing person is known to be a specific member of a family (e.g. the father of two children) and DNA profiles can be found for the body, the mother, and the two children, a minor modification of the paternity network yields the solution.

Problems of identification involving *more than one* body, such as in mass graves and in disasters are more difficult because of their complexity.

Unidentified body
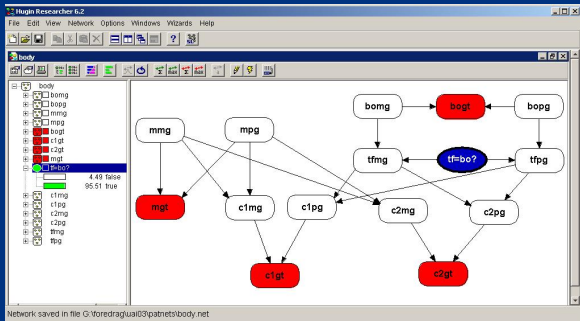
Is the body father of the two children? Same data as for paternity. Second child $(16, 19)$; $L = 95.51/4.49 = 21.27$.

Mixtures

In *criminal cases* it is not uncommon to find traces where the DNA is a mixture of contributions from several individuals.

This happens for example in *rape* cases, where a vaginal swab typically will contain DNA from the victim as well as the perpetrator, and possibly also from a consensual partner.

But it is also common e.g. in *robberies*, where a balaclava is found on the scene of the crime; these have often been used by several persons.
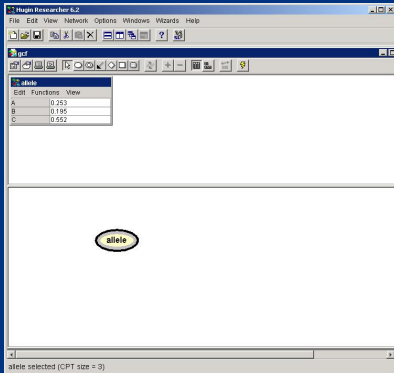
Weir's example

Profile			Marker		
	LDLR	GYPA	HBGG	D7S8	Gc
trace:	B	AB	AB	AB	ABC
victim:	B	AB	AB	AB	AC
suspect:	B	A	A	A	B
p_A	0.433	0.538	0.566	0.543	0.253
p_B	0.567	0.462	0.429	0.457	0.195
p_C	0	0	0.005	0	0.552

This example of a rape case has been used by Weir *et al*. (1997) and Mortera *et al*. (2003).

Mixture net for all markers
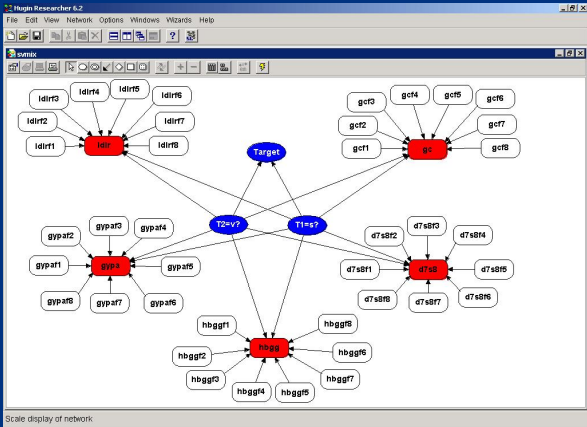
One founder for every marker

Different allele probabilities for the 5 markers. Here Gc.

Who contributed to the mixture?

Either a specified individual or a random allele

Mixing the DNA

This network mixes DNA from 4 alleles, i.e. two persons.

Network for markers
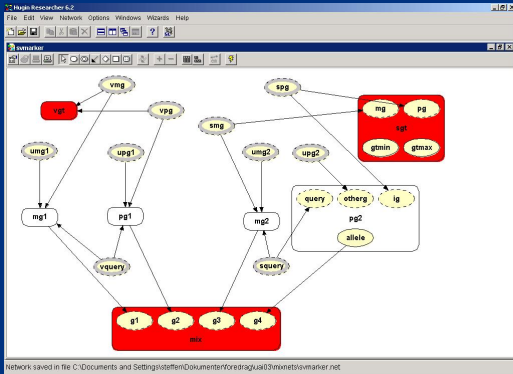
An instance of this network tells the story.

Results from all markers
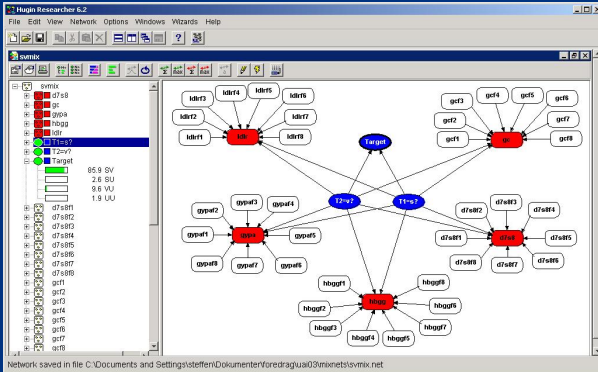
Individual likelihoods for Weir's example

			Marker			
Hyp.	LDLR	GYPA	HBGG	D7S8	Gc	Full
SV	0.573	0.279	0.285	0.280	0.511	0.859
SU	0.184	0.198	0.191	0.197	0.143	0.026
VU	0.184	0.279	0.283	0.280	0.180	0.096
UU	0.059	0.243	0.241	0.243	0.167	0.019

The full likelihood is equal to the posterior probability for the full evidence. Can also be calculated by multiplying individual likelihoods and normalising.

Algebraic alternative

Weir et al. (1997) gives algebraic formulae e.g. for the likelihood for suspect, victim, and 2 unknown contributors

$$12\, p_A\, p_B\, p_C\, (p_A + p_B + p_C + 2\, p_D),$$

while that for the victim and 3 unknown contributors is

$$(p_A + p_B + p_C + p_D)^6 - (p_B + p_C + p_D)^6$$
$$- (p_A + p_C + p_D)^6 - (p_A + p_B + p_D)^6$$
$$+ (p_C + p_D)^6 + (p_B + p_D)^6 + (p_A + p_D)^6 - p_D.$$

Extensions

Modularity and flexibility of Bayesian networks enables easy extensions to cases such as

- More potential contributors (e.g. consensual partner)

- Indirect information on individuals (missing suspect, but relative of suspect available)

- Silent alleles (e.g. behaving as 0 in the ABO-system)

- Incorporating other types of measurement error

FINEX

Alternative to OOBN is to use purpose built software for specifying Bayesian networks for forensic problems.

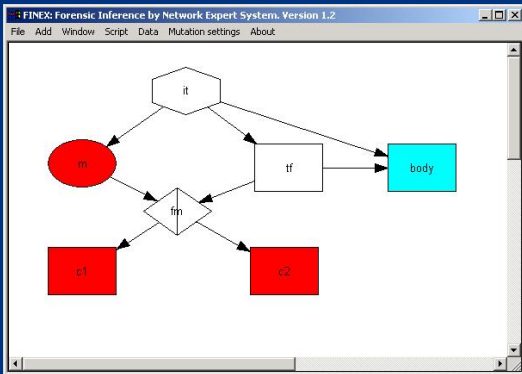*FINEX* is an example of such software, under development by Cowell (2001).

FINEX uses nodes for *individuals* and *genepools*.

Arrows into *query-individuals* denote exclusive or.

Arrows from genepools to individuals identify how genes are drawn.
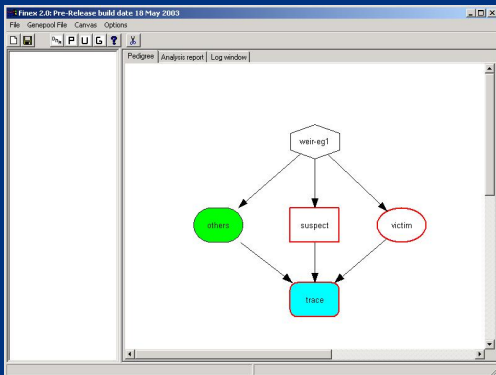
The next overheads show prints from the FINEX canvas of some of the networks previously discussed.

Unidentified body

The body identification problem in FINEX.

Mixture problem

The mixture problem in FINEX.

Problems under current research

- Estimation of mutation rates and influence of mutation rates

- Partial DNA profiles

- Varying population frequencies

- Incorporating information on amount of DNA for separating mixed profiles

- Deconvolution of mixed traces: initialise database search

- Identifying unknown pedigrees, for example in connection with disasters and immigration cases.

References

Cowell, R. G. (2003). FINEX: A probabilistic expert system for forensic identification. *Forensic Science International*, **134**, 196–206.

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

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

Weir, B. S., Triggs, C. M., Starling, L., Stowell, L. I., Walsh, K. A. J., and Buckleton, J. S. (1997). Interpreting DNA mixtures. *Journal of Forensic Sciences*, **42**, 213–22.