

Auto generation of large Bayesian networks for problems in forensic genetics

Robert G. Cowell¹

Faculty of Actuarial Science and Insurance, City University London.

e-mail: rgc@city.ac.uk

Abstract: In recent years, Bayesian networks have proved useful in analysing problems in forensic genetics. Typically these networks can be quite large, involving hundreds of nodes, or probability tables with perhaps millions of entries, making their construction by hand a time consuming and error-prone process. Although each instance of such problems is unique, the Bayesian networks can be deconstructed into simple building blocks. Alternatively, such building blocks may be put together in an automated fashion to generate in an error free manner the Bayesian networks. In this talk I discuss some software I have written to automate this process, and some of the issues involved.

Keywords: Bayesian networks, DNA mixture analyses, forensic genetics, paternity identification, STR markers.

1. Introduction

In recent years significant progress has been made in applying graphical models, in particular Bayesian networks Cowell *et al.* (1999), to problems of interest to forensic geneticists. Among such problems are included:

1. Given the DNA profiles of a child, and of PF , a man alleged to be the father, is PF the true father of the child?
2. Bodies are recovered from a mass disaster (eg plane crash). DNA profiles are taken of relatives of the known deceased in the disaster. For a particular body, which set of relatives, if any, is the deceased related to?
3. A mixed DNA sample(that is, a sample containing the DNA from two or more people) is recovered from a crime scene. What are the DNA profiles of the contributors to the sample? If a suspect is held, does the DNA profile of the suspect match that of the mixture sample?

Although courts would prefer definitive YES/NO answers to such questions, instead a Bayesian network analysis provides a likelihood ratio for competing hypotheses. For example 1 above, the competing hypotheses would usually be:

H_o The man PF is the true father of the child.

H_a An unspecified man unrelated to PF is the true father of the child.

What makes Bayesian networks particularly suitable for such problems is that a useful analysis can still be carried out in the absence of genetic profiles of all individuals involved. For example, in the paternity case, an analysis is possible even if the genetic

¹Address of correspondence: Faculty of Actuarial Science and Insurance, Cass Business School, City University London, 106 Bunhill Row, London EC1Y 8TZ

profile of the child's mother is not known. It can even be carried out if the DNA profile of the alleged father is not known, but if the DNA profiles of known relatives of the alleged father are available. Another factor that makes the analysis of a variety of such problems possible is that they can usually be decomposed into small building blocks; the key paper that showed this is Dawid *et al.* (2002).

In the next section I give a very brief summary background of relevant genetics. I will present some software I have written that automates the types of analyses given above.

2. Genetic background

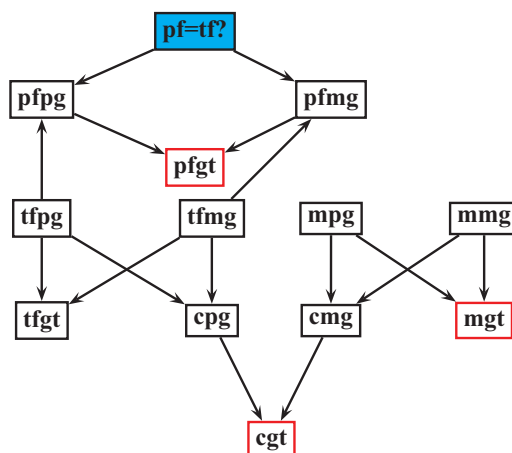
The genetic material that makes up a person is inherited from the parents in equal measure. The genetic profile of an individual that is stored, for example in a DNA database such as the American CODIS system, consists of a set of unordered pair of numbers, with each pair the measured genetic profile of a specific region of a DNA chromosome of the individual called a *short tandem repeat marker*, abbreviated to *STR*. Different markers correspond to different regions, either on the same or different chromosomes. The CODIS system uses 10 markers, plus the special Amelogenin marker that determines sex. The UK national criminal database uses 10 markers plus Amelogenin. Apart from the Amelogenin marker, each number in the pair is either an integer, or a number of the designated form $n.x$ where n is an integer, and $x \in \{1, 2, 3\}$. The two numbers in a pair could be the same. One of the pair is a value inherited from the mother and is called the *maternal gene*, the other from the father and is called the *paternal gene*. It is unordered because typically it is not possible to tell, in the absence of the DNA profiles of the parents, which came from the mother and which came from the father. The combined pair is called the *genotype* on the marker. These genes occur with varying frequency in a population, with some being rare and others more common. We refer to this distribution as the *gene frequency*. This is one of the probabilistic inputs into the Bayesian networks. Another describes the conditional probability tables for the *Mendelian inheritance* of genes passing from a parent to a child. Yet a third set of tables codify *mutation rates* of one gene to another, although for simplicity details this will not be given in this talk.

3. FINEX and FSS-ibd

FINEX was a prototype system for forensic identification, and is described in Cowell (2003). Figure 3. shows the Bayesian network structure for a simple paternity calculation, in which the genotypes of the mother m , child c and putative father pf are known, and it is assumed no mutation takes place. The special shaded *query node* $pf=cf?$ will yield the desired likelihood ratio. We see for this case that apart from the query node, there are twelve other nodes, three for each person (rising to five nodes per person if mutation is also modelled). Bearing in mind that this network is for a single marker, the Bayesian network for 10 markers will have 10 times as many nodes, that is 121 in total (and a probability table for each node). This makes for a relatively large network, quite a task to make by hand. However the very regular structure allows such a network to be constructed automatically given suitable input, which is what FINEX achieved.

Despite FINEX being a useful program, it had some limitations. It has been superseded in a commercial product called **FSS-ibd** that I developed together with the UK Forensic

Figure 1: *Single marker Bayesian network for paternity identification*



Science Service, and hopefully this will be available by the time of the conference. One extra feature is the ability to carry out a database search for relatives, so called *familial searching*. Suppose that a DNA sample is recovered from a crime scene, but it is found not to match any previous offenders whose profiles are stored in a DNA database. Might the offender be a close relative of someone in the database? **FSS-ibd** can carry out such a search looking for possible parent-child or sibling relationships, and returns a list, ordered by likelihood ratio, of possible matching relatives, taking into account mutation. It can carry out a search on several thousand profiles in a database in a matter of minutes, and was instrumental in identifying the so-called *Ilkley Moor rapist*²

4. MAIES: mixture analysis in expert systems

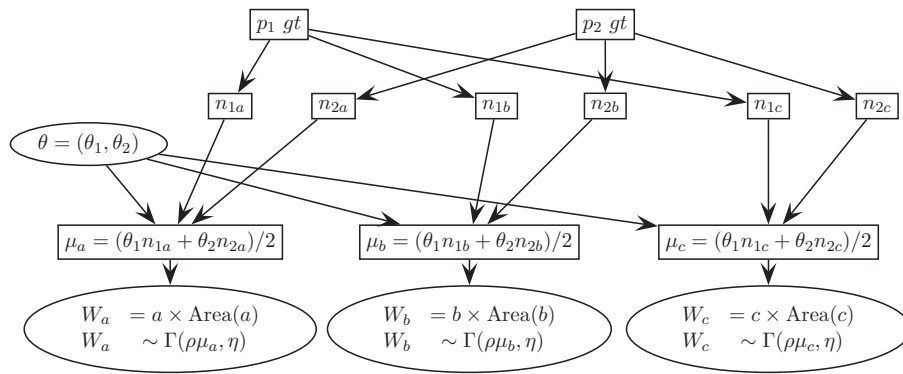
To automate the analysis of DNA mixtures using Bayesian networks, I have been developing a prototype software called *MAIES*. A parallel system built using the object oriented framework of *Hugin*³ is also being developed: see Cowell et. al (2006, 2007a, 2007b); The analysis of DNA mixtures proceeds using so called *peak area values* that are output from the DNA measuring apparatus, and indicate the relative amounts of STR genes (alleles) in the DNA sample: this can be used to infer the genetic identity of contributors to the DNA sample. For this Bayesian networks are more complicated than for the paternity identification networks, they require the presence of continuous nodes to represent the measured peak area values. Other complicating effects concerning the DNA measuring process are also modelled. Despite the networks being more complicated, they can still be built up from small components. Figure 2 shows a simplified network for a single marker. The modularity in the figure indicates how the network would grow with more markers, however it does not convey the numerical complexity. For illustration, carrying out the simple paternity analysis using the networks of Section 3. requires the construction of a Bayesian network having a total number of states in the probability tables numbering in

²<http://www.westyorkshire.police.uk/section-item.asp?sid=12&iid=6363>

³<http://www.hugin.dk>

the hundreds or low thousands. The network can be constructed and analysed in a second. In contrast, in analysing one case of a single DNA mixture with contribution from three people, the probability tables grew to a total in excess of 70 million values; in another case of a pair of three-person mixtures, where the same three people contributed to two distinct DNA samples, the total state space was in excess of 2 billion numbers.

Figure 2: High level representation of a Bayesian network for modelling DNA mixtures, for a mixture of two peoples' DNA on single marker



References

- Cowell R.G. (2003) FINEX: A probabilistic expert system for forensic identification, *Forensic Science International*, 134, 196–206.
- Cowell R.G., Dawid A.P., Lauritzen S.L. and Spiegelhalter D.J. (1999) *Probabilistic networks and expert systems*, Springer–Verlag.
- Cowell R.G., Lauritzen S. and Mortera J. (2006) MAIES: A tool for DNA mixture analysis, in: *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, Dechter R. and Richardson T., eds., AUAI Press, 90–7.
- Cowell R.G., Lauritzen S.L. and Mortera J. (2007a) A Gamma Bayesian network for DNA mixture analyses, *Bayesian Analysis*, 2, 333–48.
- Cowell R.G., Lauritzen S.L. and Mortera J. (2007b) Identification and separation of DNA mixtures using peak area information using peak area information, *Forensic Science International*, 166, 28–34.
- Dawid A.P., Mortera J., Pascali V.L. and van Boxel D. (2002) Probabilistic expert systems for forensic inference from genetic markers, *Scandinavian Journal of Statistics*, 29, 577–595.