Computational aspects of DNA mixture analysis

Exact inference using auxiliary variables in a Bayesian network

Therese Graversen^{*} University of Oxford Steffen Lauritzen University of Oxford

August 30, 2018

Abstract

Statistical analysis of DNA mixtures is known to pose computational challenges due to the enormous state space of possible DNA profiles. We propose a Bayesian network representation for genotypes, allowing computations to be performed locally involving only a few alleles at each step. In addition, we describe a general method for computing the expectation of a product of discrete random variables using auxiliary variables and probability propagation in a Bayesian network, which in combination with the genotype network allows efficient computation of the likelihood function and various other quantities relevant to the inference. Lastly, we introduce a set of diagnostic tools for assessing the adequacy of the model for describing a particular dataset.

Keywords: Bayesian network; genotype representation; junction tree; model diagnostics; prequential monitor; triangulation.

1 Introduction

In this paper we demonstrate methods for exact computation in statistical analysis of DNA mixtures, where the need for summation over the space of possible DNA profiles for unknown contributors traditionally has involved some degree of approximation (Bill et al., 2005; Tvedebrink et al., 2010; Puch-Solis et al., 2012) and has only been made for two or three unknown contributors (Cowell et al., 2011). In contrast, using the methodology presented here and the corresponding implementation by Graversen (2013) in the R-package DNAmixtures, Cowell et al. (2013) were able to perform exact evaluation and subsequent numerical maximisation of the likelihood function for up to six unknown contributors.

The present paper develops a suite of tools for inference in the statistical model described in Cowell et al. (2013) enabling evaluation of the likelihood function, computation of posterior probability of genotypes given a set of observed peak heights, and assessment of model adequacy. We exploit introduction of auxiliary variables combined with an efficient representation of the genotypes as a Bayesian network. The implementation in DNAmixtures interfaces the HUGIN API (Hugin Expert A/S, 2013) via RHugin (Konis, 2013).

^{*}Corresponding author: Therese Graversen, Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, United Kingdom, email: graversen@stats.ox.ac.uk.



Figure 1: Stylized electropherogram exhibiting peaks for the alleles at one particular marker.

The plan of the paper is as follows: Section 2 briefly describes the relevant model for DNA mixture analysis and the computational methods are detailed in Section 3. In Section 4 we show how the methodology can be extended to calculate various quantities of interest; in particular we develop methods for assessing the adequacy of the model.

2 A statistical model for mixed traces of DNA

In statistical analysis of DNA mixtures it is of interest to draw inference about individual DNA profiles in a mixed trace of DNA. The observations consists of a set of peak heights in an electropherogram (EPG) produced after a chemical duplication process known as a polymerase chain reaction (PCR). Figure 1 represents a schematic illustration of part of an EPG.

The DNA sequences at Short Tandem Repeat (STR) *markers* are characterised by a motif of base pairs repeated a number of times, so that a specific repeat number corresponding to an *allele* and a peak in the EPG typically indicates presence of the corresponding allele.

A pair of alleles is called a *genotype*, and the genotypes across a set of markers constitute the *DNA profile*. The markers used for forensic identification are typically located at different chromosome-pairs or, if not, on well separated locations, rendering it reasonable to assume independence of genotypes across markers.

The observed EPG is prone to artefacts known as *stutter* and *dropout*: Stutter refers to the phenomenon that some of the DNA alleles may lose a repeat motif during the PCR process and thus contribute to a peak at a lower repeat number. If there are allelic types $\{1, \ldots, A\}$ we thus assume that any allele *a* receives stutter from the amplification of allele a + 1. Dropout refers to the fact that peak heights occasionally are too small for the peak to be registered.

We distinguish between known and unknown contributors to the sample, depending on whether their DNA profile is considered known or not. The computational complexity of the problem is directly associated with the huge number of possible allocations of genotypes to the unknown contributors.

The genotypes of a DNA profile are assumed independent across markers so we briefly describe the model for one marker only, following Cowell et al. (2013).

2.1 Model for the genotypes of unknown contributors

We assume that the alleles of an unknown person are sampled from a reference population in Hardy–Weinberg equilibrium so that the two alleles can be considered sampled independently. Denote by n_{ia} the number of alleles of type a for contributor i. A genotype (n_{i1}, \ldots, n_{iA}) for an unknown contributor follows a multinomial distribution with allele frequencies (q_1, \ldots, q_A) and $\sum_a n_{ia} = 2$. Unknown contributors are assumed unrelated so their genotypes are independent.

2.2 Peak height distribution for fixed genotypes

Analysing DNA that contains alleles of type a results in a peak at position a, and possibly also a smaller peak at position a - 1 due to stutter during the PCR process; thus the height of the peak $H_a \ge 0$ for allele a depends on the presence of alleles of both type a and type a + 1. Peaks of height H_a below a chosen threshold C are not registered and so the observed peak heights are $Z_a = H_a \mathbb{1}_{\{H_a \ge C\}}$.

For given genotypes of the contributors we assume that the peak height H_a at allelic type *a* is gamma distributed with shape and scale parameters depending on the numbers n_{ia} , $n_{i,a+1}$ of alleles of type *a* and a + 1 that each unknown contributor *i* possesses, as well as a set of model parameters, $\psi = (\rho, \eta, \xi, \phi)$; more precisely we assume that $H_a \sim \Gamma(\lambda_a, \eta)$, where

$$\lambda_a = \rho \sum_{i=1}^k \left\{ (1-\xi)n_{ia} + \xi n_{i,a+1} \right\} \phi_i.$$
(1)

Here, and in the following, we have let $n_{i,A+1} = 0$; the parameter ξ is the mean stutter percentage, ρ is related to the general peak variability, ϕ_i denotes the fraction of DNA from individual i, and η is the scale. If $\lambda_a = 0$ the gamma distribution $\Gamma(0, \eta)$ is considered degenerate at 0.

2.3 Likelihood function

The likelihood function is determined by the distribution of the observed peak heights. The observed peak heights are independent across markers $m = 1, \ldots, M$, and thus the likelihood function factorises accordingly. Using this fact in combination with (1) we find

$$\ell(\psi) = \prod_{m=1}^{M} f_{\psi}(Z_{1}^{m}, \dots, Z_{A_{m}}^{m})$$

$$= \prod_{m=1}^{M} \mathbb{E} \left\{ f_{\psi}(Z_{1}^{m}, \dots, Z_{A_{m}}^{m} \mid \boldsymbol{n}) \right\}$$

$$= \prod_{m=1}^{M} \mathbb{E} \left\{ \prod_{a=1}^{A_{m}} f_{\psi}(z_{a}^{m} \mid \boldsymbol{n}_{a}, \boldsymbol{n}_{a+1}) \right\}, \qquad (2)$$

where the expectation is taken with respect to the distribution of genotypes of the unknown contributors. Here and in the following n denotes the full set of genotypes for all individuals and n_a the vector $n_a = (n_{ia}, i \in I)$ of allele-counts of type a. The expectation in (2) involves summation over all combinations of possible genotypes of potential contributors. There are $\{A_m(A_m+1)/2\}^k$ possible combinations of genotypes

at a marker, and thus there are this many terms in the sum, each being a product of A_m factors. Direct computation is typically infeasible when there are many alleles and many unknown contributors. We attack this computational problem by appropriate use of Bayesian network techniques, as detailed in Section 3 below.

Note that our methodology can be used directly with other choices of distribution for the peak heights, provided that the distribution of the peak height for allele a depends only on the genotypes through the number of alleles of types a and a + 1.

3 Computational methods

As a consequence of (2), and for other purposes, the computational task in DNA mixture analysis involves repeated computation of the expectation $\mathbb{E}\{h(X)\}$ of non-negative functions h of a set of discrete variables $X = \{X_v\}_{v \in V}$. We describe our computational approach in thid general setting before returning to the DNA mixture model in Section 3.2, where we give a network representation of a genotype for an unknown contributor to the trace.

3.1 Computation by auxiliary variables

Let $X = \{X_v\}_{v \in V}$ be a collection of discrete variables with a distribution represented by a Bayesian network. For $B \subseteq V$, we denote by X_B the collection of variables $\{X_v\}_{v \in B}$.

Let h be a non-negative function which can be written on the form

$$h(x) = \prod_{B \in \mathcal{B}} h_B(x_B)$$

for some set \mathcal{B} of subsets of V and real-valued, non-negative functions h_B .

For each $B \in \mathcal{B}$ we introduce binary random variables $Y^B \in \{0, 1\}$ which are conditionally independent given the network and have conditional distributions

$$\mathbb{P}(Y^B = 1 \mid X = x) = \mathbb{P}(Y^B = 1 \mid X_B = x_B) = h_B(x_B)/k^B.$$
(3)

Here, the constant k^B is chosen such that $h_B(x_B)/k^B \in [0, 1]$ over all states x_B and so (3) defines a valid probability distribution. A simple choice would be $k^B = \max_{x_B} h_B(x_B)$, i.e. the largest value that h_B attains over the state space of X_B . We use the state space $\{0, 1\}$ for auxiliary variables, but note that this choice is unimportant for the method itself.

The desired expectation $\mathbb{E}\{\prod_{B\in\mathcal{B}}h_B(X_B)\}\$ can now be expressed as the probability of a specific configuration of the binary variables introduced. As Proposition 1 reveals, this is also the case for the expectation of a product of any subset of the variables $h_B(X_B)$.

Proposition 1. For all $\mathcal{B}' \subseteq \mathcal{B}$ it holds that

$$\mathbb{E}\left\{\prod_{B\in\mathcal{B}'}h_B(X_B)\right\} = \mathbb{P}\left(\bigcap_{B\in\mathcal{B}'}\{Y^B=1\}\right)\prod_{B\in\mathcal{B}'}k_B$$

Proof. Using (3) and the fact that Y^B are conditionally independent given X we get

$$\mathbb{E}\left\{\prod_{B\in\mathcal{B}'}h_B(X_B)\right\} = \mathbb{E}\left\{\prod_{B\in\mathcal{B}'}\left(\mathbb{P}(Y^B=1 \mid X_B)k_B\right)\right\}$$
$$= \mathbb{E}\left\{\prod_{B\in\mathcal{B}'}\mathbb{P}(Y^B=1 \mid X)\right\}\prod_{B\in\mathcal{B}'}k_B$$
$$= \mathbb{E}\left\{\mathbb{P}\left(\bigcap_{B\in\mathcal{B}'}\{Y^B=1\} \mid X\right)\right\}\prod_{B\in\mathcal{B}'}k_B$$
$$= \mathbb{P}\left(\bigcap_{B\in\mathcal{B}'}\{Y^B=1\}\right)\prod_{B\in\mathcal{B}'}k_B$$

as desired.

If the distribution of the variables $\{X_v\}_{v \in V}$ is modelled by a Bayesian network, this network can be extended to include the variables $\{Y^B\}_{B \in \mathcal{B}}$ by for each B adding Y^B as a child of $\{X_v\}_{v \in B}$ with conditional distributions of Y^B in (3). As the auxiliary variables are added as children of existing network nodes, no directed cycles are created and the extended network is a correct representation of the joint distribution of (X, Y)since, given X_B , Y^B is conditionally independent of all other variables in the extended network.

Figure 2 illustrates how the network is extended in case of a function h factorising over two sets of variables (X_2, X_3) and (X_3, X_4, X_5) .



Figure 2: Extending a network with two binary variables for computation of $\mathbb{E}(h_{\{2,3\}}(X_2, X_3)h_{\{3,4,5\}}(X_3, X_4, X_5))$. Here $\mathcal{B} = \{\{2,3\}, \{3,4,5\}\}$

3.1.1 Probability propagation

We now briefly describe probability propagation and explain how to exploit the normalising constants arising as a by-product of the propagation algorithm. We refer for example to Cowell et al. (1999) for further details.

A computational structure is set up in the form of a so-called *junction tree* of subsets of the variables involved: first an undirected graph, the *moralised graph*, is constructed by adding undirected links between nodes that have a common child and removing directions for existing edges. Subsequently links are added to ensure that the resulting graph is chordal. This process is known as *triangulation* and can generally be done in many ways. Finally the cliques in the triangulated graph are arranged in a junction tree.

In the situation described above X_B is the parent set of Y^B in the extended network and the node set X_B will thus be a complete set in the triangulated graph, hence contained in some clique. The efficiency of the method depends crucially on the size of cliques for the chosen triangulation, see further discussion in Section 3.5.1 below.

A distribution p(x) is represented by an unnormalised probability function

$$p(x) \propto g(x) = \frac{\prod_{C \in \mathcal{C}} \zeta_C(x_C)}{\prod_{S \in \mathcal{S}} \zeta_S(x_S)}$$

where S denotes the set of *separators*, i.e. intersections of pairs of neighbouring cliques in the junction tree. The corresponding normalising constant is $N_1 = \sum_x g(x)$. The function g(x) is known as the *charge* and the functions ζ as *potentials*.

A message passing operation referred to as *propagation* brings the charge on a canonical form, where all potentials of the charge are equal to the function g marginalised onto the corresponding clique or separator, i.e.

$$\zeta_D(x_D) = \sum_{y:y_D=x_D} g(y) \text{ for all } D \in \mathcal{C} \cup \mathcal{S}.$$

The normalising constant can then be computed efficiently after propagation as $\sum_{x_D} \zeta_D(x_D)$, for instance choosing D as a separator $S \in \mathcal{S}$ with minimal state space.

The charge g can be modified by entering so-called *likelihood evidence* $\ell_v(x_v)$ on single nodes leading to the charge

$$\tilde{g}(x) = g(x) \prod_{v \in V} \ell_v(x_v)$$

with normalising constant

$$N_2 = \sum_x g(x) \prod_{v \in V} \ell_v(x_v).$$

Taking the ratio of the normalising constants before and after propagating the likelihood evidence yields the expectation of the product of the likelihood evidence with respect to the distribution p(x):

$$\frac{N_2}{N_1} = \frac{\sum_x g(x) \prod_{v \in V} \ell_v(x_v)}{\sum_y g(y)} = \sum_x \frac{g(x)}{\sum_y g(y)} \prod_{v \in V} \ell_v(x_v)$$
$$= \sum_x p(x) \prod_{v \in V} \ell_v(x_v) = \mathbb{E} \bigg\{ \prod_{v \in V} \ell_v(X_v) \bigg\}.$$

As shown in Proposition 2 below, this fact now ensures that the expectation of interest can be calculated by propagating likelihood evidence on the auxiliary variables.

Proposition 2. Let likelihood evidence for each node Y^B , $B \in \mathcal{B}' \subseteq \mathcal{B}$ be given as:

$$\ell_B(Y^B) = \begin{cases} k_B, & Y^B = 1\\ 0, & Y^B = 0 \end{cases}$$

and let N_1 and N_2 be the normalising constants before and after propagation of the likelihood evidence. Then we have

$$\mathbb{E}\left\{\prod_{B\in\mathcal{B}'}h_B(X_B)\right\} = \frac{N_2}{N_1}$$

Proof.

$$\frac{N_2}{N_1} = \mathbb{E} \left\{ \prod_{B \in \mathcal{B}} \ell_B(Y^B) \right\}$$
$$= \mathbb{E} \left(\prod_{B \in \mathcal{B}'} k_B \mathbb{1}_{\{Y^B = 1\}} \right)$$
$$= \mathbb{P} \left(\bigcap_{B \in \mathcal{B}'} \{Y^B = 1\} \right) \prod_{B \in \mathcal{B}'} k_B$$

which by Proposition 1 equals the desired expectation.

3.2 A Bayesian network representation of genotypes

The multinomial distribution of allele-counts (n_{i1}, \ldots, n_{iA}) representing the genotype of individual *i* does not in itself have Markovian properties. However, if we define the partial sums $S_{ia} = \sum_{b=1}^{a} n_{ia}$ counting the number of alleles of type up to and including *a* that person *i* possesses, we can represent the genotype in a Bayesian network as displayed in Figure 3.



Figure 3: Network representation of a genotype at a marker with A = 6 allelic types.

If we imagine the two alleles in the genotype being allocated sequentially, then the number of alleles that a person has of type a + 1 only depends on how many alleles of the total two are left to allocate, and the allocation happens according to a binomial distribution. In Proposition 3 we establish the formal correctness of the network specification.

Proposition 3. The distributions of genotypes and partial sums satisfy the following relations

$$S_{i1} = n_{i1},$$

 $n_{i1} \sim \operatorname{bin}(2, q_1),$

and for $a \in \{2, \ldots, A\}$

$$S_{ia} = S_{i,a-1} + n_{ia},$$

$$n_{ia} | S_{i,a-1} \sim \operatorname{bin} \left(2 - S_{i,a-1}, q_a / \sum_{b=a}^{A} q_b \right).$$
(4)

Finally, we have the conditional independence relations

$$n_{ia} \perp (n_{i1}, \dots, n_{i,a-1}, S_{i1}, \dots, S_{i,a-2}) \mid S_{i,a-1}$$

$$S_{ia} \perp (n_{i1}, \dots, n_{i,a-1}, S_{i1}, \dots, S_{i,a-2}) \mid (S_{i,a-1}, n_{ia}).$$
(5)

Proof. The unnumbered relations follow directly from the definition of the quantities involved. We further have

$$p(n_{ia} | n_{i1}, \dots, n_{i,a-1}) = \frac{p(n_{i1}, \dots, n_{i,a-1}, n_{ia})}{p(n_{i1}, \dots, n_{i,a-1})}$$

$$= \frac{\frac{2!}{(2-S_{i,a-1}-n_{ia})!\prod_{b=1}^{a}n_{ib}!} \left(\sum_{b=a+1}^{A} q_b\right)^{2-S_{i,a-1}-n_{ia}} \prod_{b=1}^{a} q_b^{n_{ib}}}{\frac{2!}{(2-S_{i,a-1})!\prod_{b=1}^{a-1}n_{ib}!} \left(\sum_{b=a}^{A} q_b\right)^{2-S_{i,a-1}} \prod_{b=1}^{a-1} q_b^{n_{ib}}}$$

$$= \frac{(2-S_{i,a-1})!}{n_{ia}!(2-S_{i,a-1}-n_{ia})!}$$

$$\times \left(1 - \frac{q_a}{\sum_{b=a}^{A} q_b}\right)^{2-S_{i,a-1}-n_{ia}} \left(\frac{q_a}{\sum_{b=a}^{A} q_b}\right)^{n_{ia}}.$$

The conditional independence (5) follows from the fact that the conditional distribution of n_{ia} given $n_{i1}, \ldots, n_{i,a-1}$ only depends on the condition through $S_{i,a-1}$; inspection of the expression for the conditional distribution yields (4).

3.3 Auxiliary variables for computing the likelihood function

In order to compute the inner expectation in (2), we note that this is an expectation of a product over alleles, where each factor is a function of the variables n_a and n_{a+1} , and so we can compute this expectation using auxiliary variables as described in Section 3.1: For each allele a, we add an auxiliary variable O_a with parents n_{ia} and $n_{i,a+1}$ for all unknown contributors i, except for O_A that is given only one parent n_{iA} per contributor. Figure 4 shows the network for modelling one marker of a mixture with two contributors and six alleles. Note that O_a and its parents n_{ia} , $n_{i,a+1}$, $i \in \{1, \ldots, k\}$



Figure 4: Bayesian network modelling the genotypes of 2 unknown contributors i and j for a marker with 6 possible allelic types.

are necessarily contained in the same clique, implying that any valid junction tree will

contain cliques with an associated state space that is exponential in the number k of unknown contributors. Unfortunately, as the moralised graph is not chordal – for instance $(S_{i1}, n_{i1}, n_{j2}, n_{i3}, S_{i2}, S_{i1})$ is a cycle – further edges need to be added, resulting in an additional increase in the size of the cliques. We shall return to this issue in Section 3.5.1.

The distribution of a peak height Z_a conditionally on the allele-counts is for a < A

$$f_{\psi}(z_{a} | \boldsymbol{n}_{a}, \boldsymbol{n}_{a+1}) = \begin{cases} g_{\psi}(z_{a} | \boldsymbol{n}_{a}, \boldsymbol{n}_{a+1}), & z_{a} \ge C \\ G_{\psi}(C | \boldsymbol{n}_{a}, \boldsymbol{n}_{a+1}), & z_{a} < C \end{cases}$$
(6)

where g and G denotes the density respectively the cumulative distribution function for the gamma distribution with parameters as in (1).

Define the distribution of O_a for an observed allele, where $z_a \ge C$, as

$$\mathbb{P}(O_a = 1 \mid \boldsymbol{n}_a, \boldsymbol{n}_{a+1}) = g_{\psi}(z_a \mid \boldsymbol{n}_a, \boldsymbol{n}_{a+1}) / k_a^{\psi}, \tag{7}$$

noting the dependence of the scaling factor k_a^{ψ} on ψ . For an unobserved allele, where $z_a^m = 0$, let the distribution of O_a be defined as

$$\mathbb{P}(O_a = 0 \mid \boldsymbol{n}_a, \boldsymbol{n}_{a+1}) = G_{\psi}(C \mid \boldsymbol{n}_a, \boldsymbol{n}_{a+1}).$$
(8)

For convenience we have defined the auxiliary variables so that for all alleles the event $O_a = 1$ corresponds to the event that the peak at allele *a* is above the threshold *C*.

Now Proposition 2 can readily be used to evaluate the contribution to the likelihood from marker m for a given value of ψ by propagating likelihood evidence

$$\ell_a(O_a) = \begin{cases} k_a^{\psi} \mathbb{1}_{\{O_a=1\}}, & \text{if } a \text{ is seen} \\ \mathbb{1}_{\{O_a=0\}}, & \text{if } a \text{ is unseen.} \end{cases}$$
(9)

3.4 Posterior distribution of genotypes

When entering and propagating likelihood evidence as in (9) for a in a set of alleles B, we obtain a representation of the conditional distribution of the full network given the relevant state of the auxiliary variables $O_a, a \in B$. Furthermore, this distribution is identical to the conditional distribution of the nodes in the network given the peak height information $\{z_a\}_{a \in B}$:

$$p(x \mid \{z_a\}_{a \in B}) = p\left(x \mid \bigcap_{\substack{a \in B, \\ z_a > C}} \{O_a = 1\} \bigcap_{\substack{a \in B, \\ z_a = 0}} \{O_a = 0\}\right)$$
(10)

This follows from the following argument:

$$\begin{split} p(x) \prod_{a \in B} \ell_a(O_a) \\ \propto & p\left(x \mid \bigcap_{\substack{a \in B, \\ z_a > C}} \{O_a = 1\} \bigcap_{\substack{a \in B, \\ z_a = 0}} \{O_a = 0\}\right) \\ \propto & p(x) \mathbb{P}\left(\bigcap_{\substack{a \in B, \\ z_a > C}} \{O_a = 1\} \bigcap_{\substack{a \in B, \\ z_a = 0}} \{O_a = 0\} \mid x\right) \\ &= & p(x) \prod_{\substack{a \in B, \\ z_a > C}} \mathbb{P}(O_a = 1 \mid x) \prod_{\substack{a \in B, \\ z_a = 0}} \mathbb{P}(O_a = 0 \mid x) \\ &= & p(x) \prod_{\substack{a \in B, \\ z_a > C}} \mathbb{P}(O_a = 1 \mid n_a, n_{a+1}) \prod_{\substack{a \in B, \\ z_a = 0}} \mathbb{P}(O_a = 0 \mid n_a, n_{a+1}) \\ &= & p(x) \prod_{\substack{a \in B, \\ z_a > C}} \{g_{\psi}(z_a \mid n_a, n_{a+1}) / k_a^{\psi}\} \prod_{\substack{a \in B, \\ z_a = 0}} G_{\psi}(C \mid n_a, n_{a+1}) \\ &\propto & p(x) \prod_{\substack{a \in B, \\ z_a > C}} f_{\psi}(\{z_a\}_{a \in B} \mid x) \\ &\propto & p(x \mid \{z_a\}_{a \in B}). \end{split}$$

As a consequence, we can easily sample from the conditional distribution of genotypes given peak height information, which we shall exploit in Sections 4.2 and 4.3 below.

3.5 Network complexity considerations

The main concerns when applying computation by auxiliary variables to a specific problem are that the junction tree representation of the network may not fit in the physical memory, and propagation and other network operations may take prohibitively long. Both of these issues are directly related to the *total size* of the network junction tree. An additional concern lies in finding a good triangulation, as this can be both timeand memory-consuming; we eliminate this additional cost by specifying triangulations directly.

The total size of the junction tree is the sum of the sizes of state spaces for all cliques and separators and determines how many numbers are needed to store the clique and separator tables.

Once a junction tree has been created for a network, computation by auxiliary variables involves setting the conditional probability tables for each auxiliary variable and propagating evidence. The number of elementary arithmetic operations for propagation is linear in the total size. Also, the number of cells that need updating when the conditional probability tables for the auxiliary variables change is, in the worst case, determined by the total size.

In the following we study the relation of the total sizes of junction tree representations used for DNA mixture analysis to the number A of possible alleles at a marker and the number k of unknown contributors.

3.5.1 Junction tree sizes for DNA mixtures

We shall consider three different triangulations of networks of the type discussed in Section 3.3 and investigate the behaviour of the total sizes of the corresponding junction trees. We restrict attention to mixture networks where any allele a — apart from the last allele A — can receive stutter from a + 1.

Any triangulation must necessarily have cliques that contain auxiliary variables with their parent sets as these are complete sets in the moralised graph. For all our junction trees we avoid adding additional variables to all such sets and simply combine any auxiliary variable with its parent set to form a clique. We can thus focus the discussion on triangulating the part of the moralised graph that does not involve auxiliary variables.

If we have N binary auxiliary variables per allele, their cliques and corresponding separators contribute to the total size of the junction tree by

$$TS_{\text{aux}} = 3N\left\{ (A-1)3^{2k} + 3^k \right\},\$$

since there are N(A-1) cliques containing an auxiliary variable along with its 2k parents, and each is separated from the remaining junction tree by a separator containing the 2kparents. The N auxiliary variables for the last allele have only k parents.

Bearing Figure 3 in mind, the structure of the genotype networks requires upper triangle sets $\{S_{i,a-1}, S_{ia}, n_{ia}\}$ to be in a clique as they are complete sets. If allele a - 1receives stutter from a, then the lower triangle set $\{n_{i,a-1}, n_{ia}, S_{ia}\}$ is also complete in the moralised graph and must be contained in some clique.

The first triangulation method we shall consider, uses the simple idea of slicing the network into cliques

$$\{S_{ia}, S_{i,a+1}, n_{ia}, n_{i,a+1}\}_{i=1}^{k}$$

for a = 1, ..., A. The corresponding junction tree, which we shall refer to as the *slice* tree, is displayed in Figure 5. In addition to the cliques and separators arising from the auxiliary variables, the slice tree has A - 1 cliques each consisting of 4k nodes, and A - 2 separators between them, each consisting of 2k nodes. Thus the total size of the slice tree becomes

$$TS_{\text{slice}} = (A-1)3^{4k} + (A-2)3^{2k} + TS_{aux}.$$



Figure 5: Slice junction tree for k = 3 contributors, A = 4 alleles, and N = 1 auxiliary variable per allele.

However, we can improve on this triangulation by splitting each slice into two cliques as Figure 6 illustrates. The resulting *triangle tree* has 2(A-1) cliques of each 3k nodes

$S_{1,a} S_{1,a+1}$		$S_{1,a}$	$S_{1,a} S_{1,a+1}$
$n_{1,a} n_{1,a+1}$		$n_{1,a} n_{1,a+1}$	$n_{1,a+1}$
$S_{2,a} S_{2,a+1}$	\longmapsto	$S_{2,a}$	$S_{2,a} S_{2,a+1}$
$n_{2,a} n_{2,a+1}$		$n_{2,a} n_{2,a+1}$	$n_{2,a+1}$
$S_{3,a} S_{3,a+1}$		$S_{3,a}$	$S_{3,a} S_{3,a+1}$
$n_{3,a} n_{3,a+1}$		$n_{3,a} n_{3,a+1}$	$n_{3,a+1}$

Figure 6: Splitting each slice into two cliques consisting of lower and upper for a reduction in total size.

and 2(A-1) separators of each 2k nodes, and thus the total size

$$TS_{\text{triangle}} = 2(A-1)3^{3k} + \{2(A-1)-1\}3^{2k} + TS_{\text{aux}}$$

grows significantly slower with the number of unknown contributors than the slice tree; see Figure 10.

S_{11}	$S_{11} S_{12}$	S_{12}	$S_{12} S_{13}$	S_{13}	$S_{13} S_{14}$
$n_{11} n_{12}$	n_{12}	$n_{12} n_{13}$	n_{13}	$n_{13} n_{14}$	n_{14}
S_{21}	$S_{21} S_{22}$	S_{22}	$S_{22} S_{23}$	S_{23}	$S_{23} S_{24}$
$n_{21} n_{22}$	n_{22}	$n_{22}n_{23}$	n_{23}	$n_{23} n_{24}$	n_{24}
S_{31}	$S_{31} S_{32}$	S_{32}	$S_{32} S_{33}$	S_{33}	$S_{33} S_{34}$
$n_{31} n_{32}$	n_{32}	$n_{32} n_{33}$	n_{33}	$n_{33}n_{34}$	n_{34}
$n_{11} n_{12}$		$n_{12} n_{13}$		$n_{13} n_{14}$	$\left[n_{14} \right]$
$n_{21} n_{22}$		$n_{22} n_{23}$		$n_{23} n_{24}$	n_{24}
$n_{31} n_{32}$		$n_{32}n_{33}$		$n_{33}n_{34}$	n_{34}
O_1		O_2		O_3	O_4

Figure 7: Triangle junction tree for k = 3 contributors, A = 4 alleles, and N = 1 auxiliary variable per allele.

In the case of only one unknown contributor, the total size of the triangle tree cannot be reduced. However, with more than one unknown contributor, each clique containing k upper triangles can be further split into k cliques as in Figure 8. Note that the cliques

$S_{1,a} S_{1,a+1}$		$S_{1,a} S_{1,a+1}$	$S_{1,a+1}$		$S_{1,a+1}$
$n_{1,a+1}$		$n_{1,a+1}$	$n_{1,a+1}$		$n_{1,a+1}$
$S_{2,a} S_{2,a+1}$	\longmapsto	$S_{2,a}$	$S_{2,a} S_{2,a+1}$		$S_{2,a+1}$
$n_{2,a+1}$	1 '	$n_{2,a+1}$	$n_{2,a+1}$		$n_{2,a+1}$
$S_{3,a} S_{3,a+1}$		$S_{3,a}$	$S_{3,a}$		$S_{3,a} S_{3,a+1}$
$n_{3,a+1}$		$n_{3,a+1}$	$n_{3,a+1}$	ļ	$n_{3,a+1}$

Figure 8: Splitting upper triangle cliques for a further reduction in total size.

containing k lower triangle sets cannot be split in a similar fashion. The resulting junction tree then has A - 1 cliques of each 3k nodes, a further k(A - 1) of each 2k + 1 nodes, and (k + 1)(A - 1) - 1 separators of 2k nodes between them. The total size of the tree is thus

$$TS_{\text{opt}} = (A-1)3^{3k} + \{(4k+1)(A-1) - 1\}3^{2k} + TS_{\text{aux}}.$$

A further slight reduction of the total size can be obtained by a small alteration in the cliques that cover nodes from the first two and last three alleles; the resulting tree is seen in Figure 9. We shall refer to this tree as the *optimal* tree, as this is the best junction tree we have been able to construct. We have also investigated junction trees found by using triangulation algorithms implemented in HUGIN but none have smaller total size than our optimal tree.



Figure 9: Optimal junction tree for a DNA mixture network with k = 3, A = 6, and N = 1.

The optimal junction tree can be generated by an elimination sequence which first eliminates all the auxiliary variables and then proceeds through the network nodes as

$$m{S}_{A}, m{S}_{A-1}, m{S}_{1}, m{n}_{1}, \{m{n}_{a}, m{S}_{a}\}_{a=2}^{A-2}, m{n}_{A-1}, m{n}_{A}\}_{a=2}^{A-2}$$

where S_a denotes $\{S_{ia}\}_{i=1}^k$ etc.

The exponential growth of the total size of the three types of junction tree is illustrated in Figure 10. Our numerical examples all include N = 3 auxiliary variables for each allele to reflect the size of the networks used in the R-package DNAmixtures. The choice of N makes little difference to the total size as this in all cases grows linearly with N.

The network representations constructed for the genotypes have a large number of state combinations that are impossible, for example due to the constraint that $\sum_a n_{ia} = 2$ for all *i*. In HUGIN there is a facility to *compress* the domain, such that only configurations of clique and separator states with non-zero probability are stored, thus reducing the effective size of the junction tree. There is a slight cost in terms of book-keeping, but for our purposes this cost is negligible.

As is apparent from Figure 10, the exponential growth pattern prevails for the compressed domains. Note that after compression all three junction trees are approximately of the same size. Also, the reduction of total size obtained by compression is itself growing exponentially; ignoring any slight reduction in total size from compressing states with probability zero in the cliques with auxiliary variables, the total size for the compressed slice tree is

$$TS_{\text{compr.slice}} = (A-3)10^k + \{3N(A-1) + A\}6^k + 3N3^k.$$

In general, to make a compression, one single propagation has to be performed and therefore the uncompressed networks set the limit for computational feasibility. When numbers are represented in single precision of each four bytes, the horisontal band in Figure 10 represents a range of capacities from 2GB to 512 GB of memory.

Figure 10 indicates that using the optimal junction tree should enable computation for up to k = 6 unknown contributors, whereas using the slice tree restricts computation to around k = 4.

There is a simple way of compressing the slice tree in that there are at most 10 possible configurations of the states in each of $\{S_{ia}, S_{i,a+1}, n_{ia}, n_{i,a+1}\}$. So if the state



Figure 10: Total sizes of junction trees as a function of the number k of unknown contributors, in the case of A = 25 allelic types and N = 3 auxiliary variables per allele. Solid lines are uncompressed sizes and dashed lines compressed sizes. The horisontal band indicates total sizes ranging from 2GB to 512GB assuming numbers are represented in single precision.

space is defined by these from the outset, it would in principle be possible to handle up to k = 9 unknown contributors, as it the compressed network would determine the maximal capacity; however, the general flexibility of the representation would be reduced.

3.5.2 Other representations of genotypes

Clearly, the network that represents the genotype of an unknown contributor could be replaced by a different representation than the one suggested here and connected to the auxiliary variables in an appropriate way. We shall briefly consider two alternative representations of a genotype.

Allele-pair representation More commonly, a genotype has been represented directly as an unordered pair of alleles; this representation has for example been used in Cowell et al. (2011). Including A alleles in the model there are A(A + 1)/2 possible unordered pairs. If an allele-pair is represented by a single node for each contributor, the parent set for each auxiliary variable in this network is the collection of the k unknown genotype-nodes, resulting in a junction tree where each clique and each separator contains all of the k genotype-nodes. Adding N auxiliary variables for each of A alleles yields the total size

$$TS_{\text{allele-pair}} = (3NA - 1) \{A(A + 1)/2\}^k$$

We note that this junction tree exhibits polynomial rather than linear growth in A, rendering the representation less efficient for markers with a large number of possible allelic types. For a fixed number of alleles, the growth in the number k of unknown

contributors is still exponential; see Figure 10. For junction trees based on the Markov representation of genotypes, the number of alleles makes a neglible impact on the total size. However, for the allele-pair representation the rate of growth depends heavily on the number of allelic types: For 25 alleles as in Figure 10 it is feasible to handle up to about 3 unknown contributors, whereas if only 10 allelic types are needed, then 4-5 unknown contributors can be handled. For 7 or more allelic types, the Markov representation in combination with optimal triangulation is superior to the allele-pair representation regardless of the number of unknown contributors. As the allele-pair representation is compressed by construction, there is no possibility of further compression of the junction tree.

Single gene representation Another possibility, used for example in Dawid et al. (2002) and Mortera et al. (2003), is to model the genotype at the single gene level. A single gene can be represented by the same Markovian network structure as that in Figure 3 used for a genotype, just that each node n_{ia} or S_{ia} has state space $\{0, 1\}$ rather than $\{0, 1, 2\}$. However, there is a cost in that two such networks are needed per unknown contributor, resulting in a total size with growth-rate $O(A \times 2^{3(2k)})$ compared to $O(A \times 3^{3k})$ when using the genotype representation. Thus, the single gene network will always be inferior to the genotype network.

The total size of the optimal single gene tree renders computations feasible for up to about 5 unknown contributors. Compression of the single gene slice tree yields a growth rate of $O(A \times 16^k)$, which still is considerably higher than $O(A \times 10^k)$ for the corresponding compressed genotype slice tree. It would stay feasible if $k \leq 7$.

For $A \ge 11$ allelic types, the single gene representation compares favourably to the allele-pair representation.

Although inefficient, the single gene network representation may be preferable for other reasons; for example in cases where the two genes might be selected from different populations, if sensitivity to uncertainty or population structure should be investigated as in Green and Mortera (2009), or if there is additional complexity involving family relations etc. as in Mortera et al. (2003).

4 DNA mixture analysis

The analysis of a mixed trace can have different objectives depending on the context. The objective can be a quantification of the strength of *evidence* for a given hypothesis over another, or the objective may be a *deconvolution* of the trace, i.e. that one wishes to predict genotypes of unknown contributors.

As a generic example we consider a trace MC15 from Gill et al. (2008), also analysed in Cowell et al. (2013). The trace is believed to contain DNA from at least three contributors, and the victim, who we shall denote K_1 , is assumed present along with another contributor K_2 . We shall here deal with the question of the identity of the third contributor. The peak heights from one marker are given in Table 1 along with the allele-counts for each of three genotyped individuals.

The available *evidence* E consists of the peak heights as observed in the EPG as well as the genotypes of individuals associated with the case. It is customary to assume relevant population gene frequencies to be known.

Table 1: Peak heights for marker D2S1338 above threshold in trace MC15, and genotypes of associated individuals.

Allele	Peak height	Allele-count		
a	Z_a	K_1	K_2	K_3
16	64	0	0	1
17	96	0	0	1
23	507	1	0	0
24	524	1	2	0

Strength of evidence. We now consider two competing explanations to the trace.

The prosecution hypothesis $H_p: K_1\&K_2\&K_3$ claims that the trace has exactly three contributors who are identical to the three known individuals K_1, K_2 , and K_3 .

An alternative explanation of the trace is the *defence hypothesis* H_d : $K_1\&K_2\&U$ that the trace contains the DNA of K_1 , K_2 , as well as that of an unknown and unrelated individual U, whereas K_3 has not contributed.

The strength of the evidence is reported as a *likelihood ratio*:

$$LR = L(\hat{H}_p)/L(\hat{H}_d) = \Pr(E \mid \hat{H}_p)/\Pr(E \mid \hat{H}_d)$$

where \hat{H}_i indicates that we use the maximum likelihood estimates of the parameters under the hypothesis H_i , see Table 2 below.

Deconvolution. Under the defence hypothesis we are interested in determining the identity of the unknown contributor U. This could for example be done by finding the most probable genotypes for U given the evidence, i.e. those with the highest values of $\Pr(U | \hat{H}_d, E)$. We shall return to this issue in Section 4.3 below.

Estimation. In order to calculate the relevant quantities for any of the above questions, we need to estimate the unknown parameters of the model. Being able to evaluate the likelihood function, this can be done by numerical maximisation. The maximum likelihood estimates and standard errors obtained under the defence hypothesis H_d and prosecution hypothesis H_p are given in Table 2. The resulting likelihood ratio is $\log_{10}(LR) = 12.12$.

Defence hy	γ pothesis	Prosecution hypothesis			
Parameter	Estimate	Parameter	Estimate		
ρ	26.95	ρ	33.86		
η	33.86	η	26.94		
ξ	0.086	ξ	0.076		
ϕ_{K_1}	0.823	ϕ_{K_1}	0.825		
ϕ_{K_2}	0.055	ϕ_{K_2}	0.049		
ϕ_U	0.122	ϕ_{K_3}	0.126		
$\log_{10} L(\hat{H})$	-130.21	$\log_{10} L(\hat{H})$	-118.09		

Table 2: Maximum likelihood estimates based on MC15.

4.1 Model Diagnostics

In the assessment of forensic evidence, little attention has been devoted to demonstrate the adequacy of a proposed model used to analyse a specific case or, of equal importance, to assert that data have been correctly recorded for the analysis. This may partly be due to the unavailability of useful methods for the purpose. However, we believe this aspect to be of utmost importance; in particular we find it reasonable that one should not only compare the prosecution and defence hypothesis, but there should also be an effort to demonstrate that neither hypothesis represents an implausible explanation of the trace under analysis.

Previously we have introduced auxiliary variables O_a , to enable simple computation of the likelihood function (2) and representation of evidence from observed peak heights (10). We shall in the following introduce further auxiliary variables such as binary variables D_a which indicate whether or not a peak was observed for allele a, and variables Q_a which indicate whether a peak observed at allele a was less than a specified value. Both of these types of auxiliary variables shall prove to be useful for model validation; in addition, the variables D_a can be used in an analysis which refrains from exploiting the peak heights but is based only on peak presence; see Section 4.4 below.

4.1.1 Assessing peak height distributions

First, we wish to investigate whether our model appropriately predicts the observed peak heights. Given $Z_a \ge C$, the peak height follows a continuous distribution and thus the probability transform $\mathbb{P}(Z_a \le z_a | Z_a \ge C)$ follows a uniform distribution.

To express the probability in a way suitable for computation with auxiliary variables we first note that for $z \ge C$ we have

$$\mathbb{P}(Z_a \le z \mid Z_a \ge C) = \frac{\mathbb{P}(Z_a \le z) - \mathbb{P}(Z_a < C)}{\mathbb{P}(Z_a \ge C)}.$$

Thus all we need to evaluate is the distribution function in the observed value z_a and at the threshold C. The distribution function

$$\mathbb{P}(Z_a \le z) = \mathbb{E}\left\{ \mathbb{P}(Z_a \le z \mid \boldsymbol{n}_a, \boldsymbol{n}_{a+1}) \right\}$$
(11)

is the expectation of a trivial product of one factor, and to compute this we add an auxiliary variable Q_a with the same parents as for O_a and with conditional probability

$$\mathbb{P}(Q_a = 1 \mid \boldsymbol{n}_a, \boldsymbol{n}_{a+1}) = \mathbb{P}(Z_a \leq z \mid \boldsymbol{n}_a, \boldsymbol{n}_{a+1}).$$

Similarly, we add a binary variable D_a allowing the evaluation of both $\mathbb{P}(Z_a \geq C)$ and $\mathbb{P}(Z_a < C)$.

It can be of interest to consider the distribution of the peak height in the light of other observed peaks, and not just the marginal distribution of the peak itself. For instance, we can condition on the peak heights of all other alleles to get $\mathbb{P}(Z_a \leq z \mid Z_b = z_b, b \neq a, Z_a \geq C)$, or we could include this information for only the preceding alleles in the ordering to get $\mathbb{P}(Z_a \leq z \mid Z_b = z_b, b \leq a, Z_a \geq C)$. These distributions can all be obtained simply through conditioning on relevant variables O_a as described in Section 3.4.

In Figure 11, quantile-quantile plots for the conditional distribution of a peak height given observed peak heights for all other alleles are shown for H_p and H_d using trace MC15 and the associated maximum likelihood estimates in Table 2.

We note that in both diagrams the points are close to the identity line and there is no indication that the peak height distributions are inadequately modelled under either of the hypotheses.



Figure 11: Quantile-quantile plots for the prosecution and defence hypotheses for MC15.

We can also take a closer look at the distribution of the peak height at any single allele, for example to identify outlying observations. This is illustrated in Figure 12. Boxes indicate quartiles and whiskers indicate 0.5% and 99.5% prediction limits for the conditional distributions of peak heights $\mathbb{P}(Z_a \leq z | Z_b = z_b, b \neq a, Z_a \geq C)$. The quantiles are found by numerical inversion of the distribution function (11).



Figure 12: Comparison of observed peak heights to their predictive distribution conditionally on all other observed peak heights for marker D2S1338. The bar below each peak indicates the probabilities of observing (grey) and not observing (black) a peak at this allele.

We note that although the observed peak heights at alleles 23 and 24 are somewhat lower than expected, there are no observations that are clear outliers, conforming with the quantile-quantile plots in Figure 11. Note that the prosecution hypothesis predicts complete absence of peaks at alleles 18–21 and 25-27, whereas this is not the case for the defence hypothesis involving alleles from unknown contributors; hence under this hypothesis peaks are *a priori* possible at any allele.

4.1.2 Prequential monitoring of peak presence

Next, we wish to investigate whether our model correctly predicts absence and presence of peaks in the EPG. We use the prequential theory of Dawid (1984) with so-called prequential monitors (Seillier-Moiseiwitsch and Dawid, 1993).

Using some arbitrary ordering, we consider the set of alleles across all markers and the probability that a peak has been seen for allele a given the peak heights observed on all preceding alleles,

$$p_a = \mathbb{P}(Z_a \ge C \mid z_i, i < a) = \mathbb{P}(D_a = 1 \mid z_i, i < a)$$

which can be obtained by propagation as described in Section 3.4. For each allele a, we then consider the logarithmic score

$$Y_a = \begin{cases} -\log p_a, & \text{if } z_a \ge C\\ -\log(1-p_a), & \text{if } z_a < C \end{cases}$$

so that Y_a is always non-negative and higher values of Y_a represent a large penalty for assigning a small probability $(p_a \text{ or } 1 - p_a)$ to the event that actually happens.

The cumulative logarithmic score, adjusted for incremental expectations,

$$M_a = \sum_{i=1}^{a} \{ Y_i - \mathbb{E}(Y_i \,|\, Z_b, b < i) \}$$

is a martingale with respect to the sequence of peak heights.

As $\mathbb{V}(M_a - M_{a-1} | Z_b, b < a) = \mathbb{V}(Y_a | Z_b, b < a)$, the distribution of the normalised cumulative score

$$\frac{\sum_{i=1}^{a} Y_i - \sum_{i=1}^{a} \mathbb{E}(Y_i \mid Z_b, b < i)}{\sqrt{\sum_{i=1}^{a} \mathbb{V}(Y_i \mid Z_b, b < i)}}$$

approaches a standard normal distribution as the denominator becomes infinitely large (Seillier-Moiseiwitsch and Dawid, 1993). Thus for $q_{1-\alpha}$ being the $1-\alpha$ quantile of the standard normal distribution,

$$q_{1-\alpha} \sqrt{\sum_{i=1}^{a} \mathbb{V}(Y_i \,|\, Z_b, b < i)}$$

is an approximate pointwise $1 - \alpha$ upper predictive limit for the cumulative score at allele a.

The cumulative score can easily be calculated using that if $p_a \in \{0, 1\}$ we have $Y_a = 0$ and otherwise

$$\mathbb{E}(Y_a \mid Z_b, b < a) = -p_a \log p_a - (1 - p_a) \log(1 - p_a),$$

$$\mathbb{V}(Y_a \mid Z_b, b < a) = p_a(1 - p_a) \{\log p_a - \log(1 - p_a)\}^2$$

Prequential monitor plots of the prosecution and defence hypothesis for MC15 are displayed in Figure 13.



Figure 13: Prequential monitor plots of the prosecution and defence hypotheses for MC15. The dashed horisontal lines indicate upper 95% and 99% pointwise predictive limits based on the approximating normal distribution.

A negative jump in the score means that we have observed what the model predicts as most likely, whereas a positive jump means that we have observed the opposite of what is most likely according to the model. If it is equally likely for a peak to fall above and below the threshold, or there is only one possible outcome — i.e. if $p_a \in \{0, 1/2, 1\}$ — there is no jump. The size of an upward jump indicates the level of disagreement between model and observations. Note that for the defence hypothesis, the monitors cross the upper limits towards the end of the plot, indicating that this hypothesis may not adequately describe the pattern of observed peaks. Further investigation may reveal whether upward jumps are due to observation of rare alleles or, for example, due to recording errors in the data.

4.2 Simulation

As noted in Section 3.4, introducing evidence on the auxiliary variables O_a yields a representation of the posterior distribution of the genotypes of the unknown contributors. This in turn enables simulation of a full DNA trace including peak heights, either marginally or conditionally on relevant subsets of the observed peak heights. More generally, we have for any event *B* that

$$f_{\psi}(\{z_a\}_{a \in A}, \boldsymbol{n} \mid B) = f_{\psi}(\{z_a\}_{a \in A} \mid \boldsymbol{n}, B)p(\boldsymbol{n} \mid B).$$

If conditioning with B can be represented by propagation in our Bayesian network, for example if $B = \{Z_b = z_b, b \neq a\}$, we can easily simulate from $p(\boldsymbol{n} | B)$ by standard methods (Cowell et al., 1999, Section 6.4.3). Thus to sample a full DNA trace, we just further need a method for sampling from $f_{\psi}(\{z_a\}_{a \in A} | \boldsymbol{n}, B)$.

This method of simulation can for example be used in a bootstrap analysis of the estimation uncertainty as in Graversen and Lauritzen (2013). Simulation could also be

relevant for assessing the discriminatory ability of the calculated likelihood ratio, for illustration of peak height variability, and other forms of model validation. Below we are exploiting simulation in the prediction of profiles of unknown contributors.

4.3 Prediction of unknown profiles

In a model involving unknown contributors it can be relevant to investigate the distribution of genotypes for each of these conditionally on the evidence. Focusing on a single or few alleles, we can explore this distribution directly. For any combination of genotypes we can compute its probability exactly by probability propagation. We can identify those of highest probability by sampling genotypes until a proportion p of the probability mass has been visited as then each of the remaining combinations of genotypes must have probability at most 1 - p. Thus the r combinations with probability strictly greater than 1 - p must be among those sampled. They can then be ranked according to their probability and constitute the list of the r most probable combinations. Here the number r depends on the probability p chosen.

Considering the defence hypothesis of trace MC15, we would like to identify the genotype of the unknown contributor U. If we consider the full genotype, at all markers, we often get a very diffuse distribution as for example reported in Cowell et al. (2013).

One reason for this is that, due to dropout, there are generally many unseen alleles that could be present in the mixture without giving rise to a peak. However, if we focus on explaining the peaks actually seen in the EPG we get a more concentrated distribution, as displayed in Table 3, where the total probability of the six combinations add up to one. As the table shows, the probability that the unknown contributor has at

Table 3: Probabilities of genotype at marker D2S1338 for the unknown contributor U under the defence hypothesis. The defendant K_3 has genotype (16,17).

16	17	23	24	D	Prob
1	1	0	0	0	0.5276
0	1	0	0	1	0.1861
0	2	0	0	0	0.1697
0	1	0	1	0	0.0640
0	1	1	0	0	0.0509
1	0	0	0	1	0.0017
Total probability					1.0000

least one allele 17 is .9983, close to certainty. There is some uncertainty concerning the second allele which can be virtually anything although it is by far most probable that the genotype is (16, 17); this genotype is that of the defendant K_3 . The second most probable explanation of the trace is that the other allele has dropped out.

4.4 Strength of evidence when ignoring peak heights

Another potential application of the auxiliary variables is to calculate a likelihood ratio which only uses information about peak presence or absence. This can be done by specifying evidence for the nodes D_a introduced in Section 4.1 rather than for nodes O_a .

It is still necessary to specify a set of model parameters, which for example could be estimated using peak heights. Using the estimates in Table 2 we obtain a likelihood ratio of $\log_{10} LR = 9.85$ which is weaker than the evidence obtained with full peak height information but it is still incriminating for the defendant. Such an analysis is analogous to the one used in **likeLTD** as suggested by Balding (2013), where peak heights are used only to classify peaks as present, absent, or uncertain.

We have used peak heights to estimate the parameters of the model. In principle parameters could also be estimated solely on the peak presence information, possibly in combination with prior information on some of these, although such estimates would be ill-determined and therefore not useful.

4.5 Multiple mixed traces

By adding more auxiliary variables to the model, we can easily extend the model to handle multiple traces, either with independent unknown contributors or where some or all unknown contributors coincide.

We assume that the peak heights across mixed traces are conditionally independent given the genotypes of common contributors. Peak height distributions are allowed to vary across traces through the model parameters.

The network now models the set of all unknown contributors to the mixed traces. Denote by ϕ_i^j the proportion of DNA that contributor *i* has made to trace *j*. Then $\phi_i^j = 0$ corresponds to contributor *i* not being present in trace *j*. Therefore, the case where some or all contributors are distinct to a particular mixed trace is a sub-model corresponding to $\phi_i^j = 0$ for some (i, j).

An advantage of this specification of the joint model is that we do not need to make assumptions about possible common unknown contributors to the traces, but we can let the maximisation of the likelihood point to the relevant scenario. This has been used in Cowell et al. (2013) for a combined analysis of MC15 with another trace pertaining to the same case.

In the case where the traces have completely independent unknown contributors, it is recommendable to represent each trace as a separate network to limit the number of unknown contributors in each network.

5 Discussion

We note that our computational methods are exact throughout under the model adopted, and that the only approximations relate to the model representing an inevitable approximation to reality, and possible imprecision of numerical methods. Nevertheless, using the efficient junction tree representations and exact compression methods as described in Section 3.5.1, we are able to handle more contributors than what has previously been possible.

We have far from exhausted the flexibility and the potential of the Bayesian network model and point out that simple modifications or elaborations of the basic network can readily be used to, say, incorporate the presence of silent alleles simply by including an extra allele in the genotype representation, or to enable the direct computation of the probability that a specific peak is due to stutter or an absent peak is due to random dropout or allele absence; see Cowell et al. (2013) for this and further examples.

References

- Balding, D. (2013). Evaluation of mixed-source, low-template DNA profiles in forensic science. Proceedings of the National Academy of Sciences of the United States of America. Published online doi:10.1073/pnas.1219739110.
- Bill, M., Gill, P., Curran, J., Clayton, T., Pinchin, R., Healy, M., and Buckleton, J. (2005). PENDULUM – a guideline-based approach to the interpretation of STR mixtures. *Forensic Science International*, 148:181–189.
- Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). Probabilistic Networks and Expert Systems. Springer-Verlag, New York.
- Cowell, R. G., Graversen, T., Lauritzen, S., and Mortera, J. (2013). Analysis of DNA mixtures with artefacts. arXiv:1302:4404.
- Cowell, R. G., Lauritzen, S. L., and Mortera, J. (2011). Probabilistic expert systems for handling artifacts in complex DNA mixtures. *Forensic Science International: Genetics*, 5:202–209.
- Dawid, A. P. (1984). Statistical theory. The prequential approach. Journal of the Royal Statistical Society, Series A, 147:277–305.
- 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*, 29:577–595.
- Gill, P., Curran, J., Neumann, C., Kirkham, A., Clayton, T., Whitaker, J., and Lambert, J. (2008). Interpretation of complex DNA profiles using empirical models and a method to measure their robustness. *Forensic Science International: Genetics*, 2:91– 103.
- Graversen, T. (2013). DNAmixtures: Statistical Inference for Mixed Traces of DNA. R package version 0.1-0, dnamixtures.r-forge.r-project.org/.
- Graversen, T. and Lauritzen, S. (2013). Estimation of parameters in DNA mixture analysis. Journal of Applied Statistics. Published online doi:10.1080/02664763.2013.817549.
- Green, P. J. and Mortera, J. (2009). Sensitivity of inferences in forensic genetics to assumptions about founder genes. *Annals of Applied Statistics*, 3:731–763.
- Hugin Expert A/S (2013). Hugin API Reference Manual, Version 7.7. Hugin Expert A/S, Aalborg, Denmark.
- Konis, K. (2013). RHugin. R package version 7.7-5, rhugin.r-forge.r-project.org.
- Mortera, J., Dawid, A. P., and Lauritzen, S. L. (2003). Probabilistic expert systems for DNA mixture profiling. *Theoretical Population Biology*, 63:191–205.
- Puch-Solis, R., Rodgers, L., Mazumder, A., Pope, S., Evett, I., Curran, J., and Balding, D. (2012). Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. Technical report, LGC Research Report LGC/P/2012/138.

- Seillier-Moiseiwitsch, F. and Dawid, A. P. (1993). On testing the validity of sequential probability forecasts. *Journal of the American Statistical Association*, 88:355–359.
- Tvedebrink, T., Eriksen, P. S., Mogensen, H. S., and Morling, N. (2010). Evaluating the weight of evidence by using quantitative short tandem repeat data in DNA mixtures. *Applied Statistics*, 59:855 – 874.