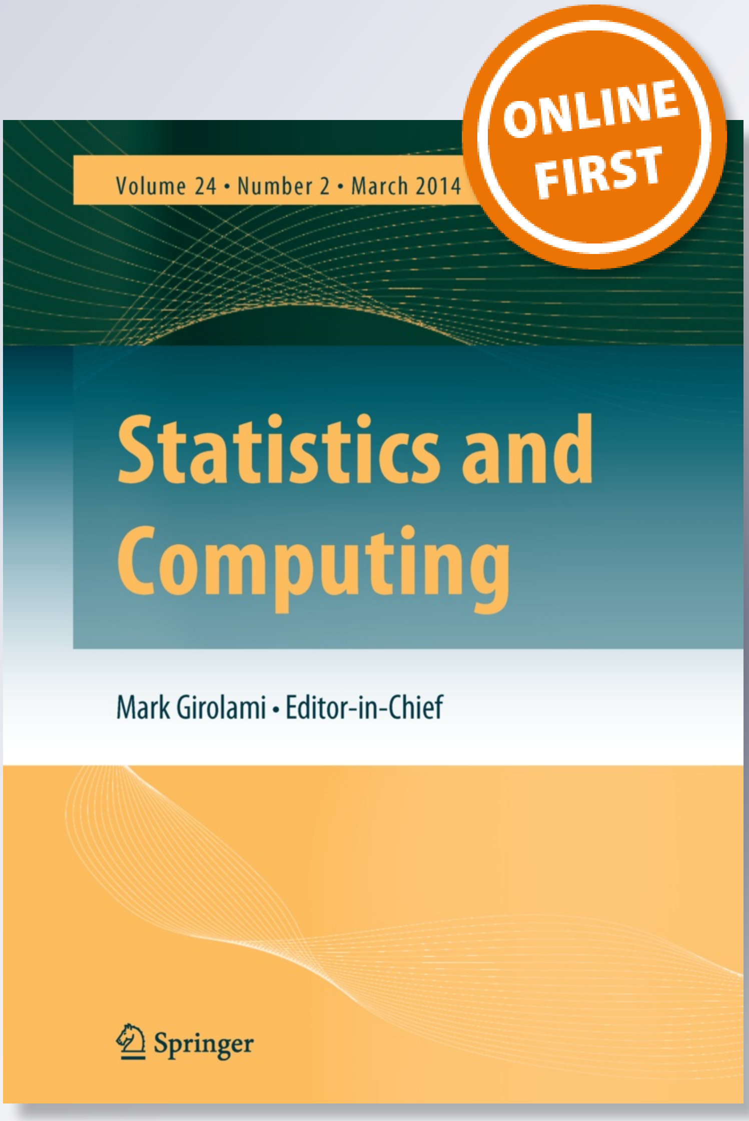
Computational aspects of DNA mixture analysis

Therese Graversen & Steffen Lauritzen

Statistics and Computing

ISSN 0960-3174

Stat Comput
DOI 10.1007/s11222-014-9451-7


ONLINE FIRST
Volume 24 • Number 2 • March 2014
Statistics and Computing
Mark Girolami • Editor-in-Chief
Springer

Stat Comput
DOI 10.1007/s11222-014-9451-7

Computational aspects of DNA mixture analysis

Exact inference using auxiliary variables in a Bayesian network

Therese Graversen · Steffen Lauritzen

Received: 18 July 2013 / Accepted: 22 January 2014

Abstract Statistical analysis of DNA mixtures for forensic identification is known to pose computational challenges due to the enormous state space of possible DNA profiles. 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. We propose a Bayesian network representation for genotypes, allowing computations to be performed locally involving only a few alleles at each step. Exploiting appropriate auxiliary variables in combination with this representation allows efficient computation of the likelihood function and prediction of genotypes of unknown contributors. Importantly, we exploit the computational structure to introduce a novel set of diagnostic tools for assessing the adequacy of the model for describing a particular dataset.

Keywords Bayesian network · Deconvolution · Genotype representation · Junction tree · Model diagnostics · Prequential monitor · Triangulation

1 DNA mixture analysis

The idea behind using DNA for forensic identification is that a person may be identified by a DNA profile, describing specific features of the genome of that person. A sample from a crime scene may contain DNA from more than one person, in which case identifying the single contributors becomes a challenge. DNA mixture analysis addresses the question of identifying the contributors to such a mixed sample of DNA and in this paper we develop methods for exact inference in statistical analysis of DNA mixtures. Using efficient computational techniques we are able to perform the necessary calculations for several unknown contributors without introducing heuristic approximations.

1.1 Identification through DNA profiles

The human nuclear DNA consists of 23 pairs of chromosomes, each with one chromosome inherited from the mother and one inherited from the father. A chromosome can be thought of as a sequence of the letters (bases) A, G, C, and T, and the entire DNA as uniquely identifying the person; an exception being the case of identical twins. For forensic identification only 10–15 specific parts of the chromosome-pairs are considered, the idea being that these *markers* exhibit enough variation to discriminate between people with high probability. The DNA sequence found on one of the two chromosomes at a particular marker is called an *allele*, and the unordered pair of alleles constitutes the *genotype* at that marker. The set of genotypes across markers is the *DNA profile* of a person.

Depending on the context in which the DNA sample is analysed, the focus of an analysis could, for example, be to compare two competing explanations of the composition of the sample or to predict the DNA profiles of the contributors to the sample.

1.2 The measurement process

The markers used for identification are short tandem repeat (STR) markers, which are characterised by variation in terms of the number of times that a certain motif of typically four letters is repeated. The alleles at an STR marker are named by the number of repeats.

T. Graversen (✉) · S. Lauritzen
Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK
e-mail: therese.graversen@stats.ox.ac.uk; graversen@stats.ox.ac.uk

Published online: 20 February 2014

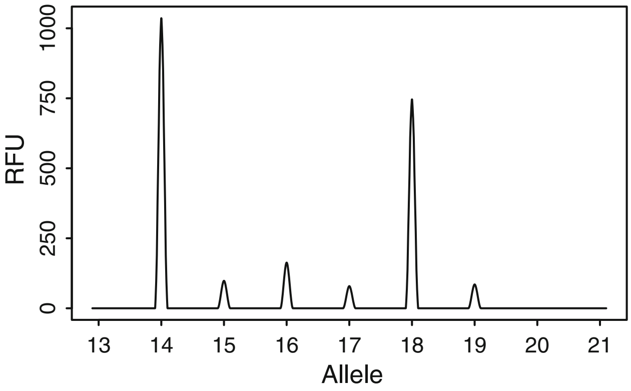


Fig. 1 Stylized electropherogram exhibiting peaks for the alleles at one marker

The alleles present in the DNA sample are located and copied in the polymerase chain reaction (PCR) process, which is in essence a branching process that produces multiple copies of the isolated sequence constituting the allele, all with a dye attached. The different alleles can then be identified through the combination of dye and length; this separation is done by electrophoresis. Alleles are detected as they move through a capillary in batches of alleles of identical length and the fluorescent intensity is recorded in units of RFU (relative fluorescence units). The resulting *electropherogram* (EPG) displays a peak for each detected allele at each marker.

Figure 1 represents a schematic illustration of an EPG corresponding to a single marker.

The height of a peak is roughly proportional to the amount of the allele in the mixture, and so the set of peak heights reflects the allelic composition of the mixture. However, due to imperfections in the PCR process known as *artefacts*, the presence or absence of peaks may not give a true picture of the allelic composition of the mixture. Usually a detection threshold is applied, under which peaks are not registered. If an allele is indeed present in the mixture, though not registered, it is said to have *dropped out*. Another common artefact is *stutter*, which is a result of the PCR process occasionally producing a copy one repeat shorter than the original; the stutter-products from allele a will contribute to the peak for allele $a-1$. Stutter-products more than one repeat shorter or longer are also possible, but are sufficiently infrequent to be ignored.

In summary, the individual contributions of DNA are not observable, and furthermore the amounts of each allele present in the mixture are only observable through the peak heights. In this paper, we consider a joint model for the DNA profiles of contributors to the sample and the observed peak heights across all markers and alleles.

1.3 Modelling DNA mixtures

We now turn to the problem of statistical analysis of a DNA mixture in the context of a crime case. As *evidence* E we consider the peak heights as observed in the EPG as well as the DNA profiles of individuals associated with the case. A model of the mixture naturally involves an explanation of the sample in terms of a set of assumed contributors. In such an explanation, we distinguish between *known* and *unknown* contributors to the sample, depending on whether their DNA profile is considered known or not. One part of the explanation is a specification of the distribution of DNA profiles of the unknown contributors. In addition we need a specification of the distribution of peak heights given the allelic composition of the mixture.

A given explanation is referred to as a *hypothesis*. To assess the *weight of evidence* against a specific person K having contributed to the sample we may formulate two hypotheses: the *prosecution* hypothesis H_p which specifies that the DNA profile of K is present in the sample, i.e. K is among the known contributors; and the *defence* hypothesis H_d, typically replacing K with an unknown contributor U. The DNA profile of an unknown contributor is considered to be randomly selected from a suitable reference population. We then report the weight of evidence against K (Lindley 1977; Balding 2005) as a *likelihood ratio*

$$LR = L(H_p)/L(H_d) = \Pr(E \mid H_p)/\Pr(E \mid H_d). \qquad (1)$$

For a hypothesis involving unknown contributors, it is also of interest to predict their genotypes by finding the posterior distribution of these given the evidence.

1.4 Computational issues

The size of the state space of possible DNA profiles for unknown contributors severely restricts the complexity of models that can be handled and demands development of efficient computational methods.

In Sect. 3 we present a general approach for computing expectations of products of non-negative functions in Bayesian networks using auxiliary variables.

In Sect. 4 we propose a Bayesian network representation of a genotype that exhibits a Markovian structure. This representation in combination with appropriate auxiliary variables constitute a Bayesian network representation of the entire statistical model, and Sect. 5 illustrates how this representation is flexible enough to provide a unified framework for computation of various quantities relevant to a case analysis: likelihood functions, posterior distributions of genotypes given a set of observed peak heights, and predictive distributions of peak heights. In particular, Sect. 5.2 presents the development of novel methods for systematic assessment of the adequacy of the model to the specific case at hand.

The methodology described in this paper has been implemented in the R-package `DNAmixtures` (Graversen 2013), which interfaces the HUGIN API (Hugin 2013) via `RHugin` (Konis 2013).

2 A statistical model for DNA mixtures

The model for the distribution of DNA profiles and peak heights in the EPG is composed of a model for the DNA profiles for the contributors and a model for the conditional distribution of peak heights given the composition of the mixture. These two elements are further described below.

2.1 Statistical model for DNA profiles

We adopt the standard model for DNA profiles: genotypes in DNA profiles are mutually independent and independent across markers; DNA profiles of known contributors are fixed; and the two alleles of an unknown person are considered sampled independently from a reference population, i.e. the population is assumed to be in Hardy–Weinberg equilibrium.

For a given marker, we represent the pair of alleles that constitute a genotype by a vector of allele counts $(n_{i1}, \ldots, n_{iA})$ for alleles $a = 1, \ldots, A$, where n_{ia} denotes the number of alleles a that contributor i possesses. In the following we shall use the terms genotype and allele counts interchangeably. Also, the number A of possible alleles shall be changing depending on the marker considered. We let $\boldsymbol{n}$ denote the full set of genotypes for all individuals and $\boldsymbol{n}_a$ the vector $\boldsymbol{n}_a = (n_{ia}, i \in I)$ of allele counts for allele a across individuals.

A consequence of the standard model is that a genotype $(n_{i1}, \ldots, n_{iA})$ for an unknown contributor follows a multinomial distribution with allele frequencies $(q_1, \ldots, q_A)$ and $\sum_a n_{ia} = 2$. It is customary to assume population allele frequencies q_a to be known and equal to values obtained from a database of individuals, possibly adjusted to correct for potential relatedness in the population (Balding 2013).

2.2 Peak height distribution conditional on genotypes

Analysing DNA that contains alleles of type a may result 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 $H_a \geq 0$ for allele a naturally depends on the presence of alleles a and $a+1$.

Our model for the distribution of peak heights for fixed composition follows Cowell et al. (2013). The key assumption is that the peak heights H_a are mutually independent with a distribution depending only on the numbers n_{ia}, $n_{i,a+1}$ of alleles a and $a+1$ that each unknown contributor i possesses, as well as a set of model parameters. Further we assume that a peak height H_a is gamma distributed with scale parameter η and shape parameter

$$\lambda_a = \rho \sum_{i=1}^{k} \left\{(1-\xi)n_{ia} + \xi n_{i,a+1}\right\} \phi_i, \tag{2}$$

where k denotes the number of individuals considered and we have let $n_{i,A+1} = 0$. Further, ϕ_i denotes the fraction of DNA from individual i, ξ is the *mean stutter proportion*, and ρ is related to the general peak variability. If $\lambda_a = 0$ the gamma distribution is considered degenerate at 0.

Using additivity properties of the gamma distribution, the model may be interpreted in terms of a peak H_a being composed partly by contributions from individuals i, each having a mean proportional to $(1-\xi)\phi_i n_{ia}$, and partly by stutter $\xi\phi_i n_{i,a+1}$ received from the allele $a+1$. Thus the mean peak heights reflect the allelic composition of the mixture as modified by stutter.

Peaks of height H_a below the applied detection threshold C are not registered so the observed peak heights are $Z_a = H_a \mathbb{1}_{\{H_a \geq C\}}$. Thus the distribution of Z_a conditionally on the genotypes depends only on the allele counts for alleles a and $a+1$ and is for $a \leq A$ a mixture with density

$$f_\psi\left(z_a \middle| \boldsymbol{n}_a, \boldsymbol{n}_{a+1}\right) = \begin{cases} g_\psi\left(z_a \middle| \boldsymbol{n}_a, \boldsymbol{n}_{a+1}\right) & z_a \geq C \\ G_\psi\left(C \middle| \boldsymbol{n}_a, \boldsymbol{n}_{a+1}\right) & z_a = 0, \end{cases} \tag{3}$$

where g denotes the density and G the cumulative distribution function for the gamma distribution as in (2), and $\psi = (\phi, \xi, \rho, \eta)$ are the parameters described above. We emphasise that our methodology can be used directly with other choices of distribution for the peak heights, provided that the conditional distribution of the peak height for allele a given genotypes depends only on the genotypes through the number of alleles of types a and $a+1$.

2.3 Likelihood function

The likelihood function is determined by the observed peak heights $\{z_a^m,\ m = 1, \ldots, M;\ a = 1, \ldots, A_m\}$ across all markers m and the possible alleles a under each of these. 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 (3) we find

$$\ell(\psi) = \prod_{m=1}^{M} f_\psi(z_1^m, \ldots, z_{A_m}^m)$$

$$= \prod_{m=1}^{M} \mathbb{E}\left\{f_{\psi}\left(z_1^m, \ldots, z_{A_m}^m \middle| \boldsymbol{n}^m\right)\right\}$$

$$= \prod_{m=1}^{M} \mathbb{E}\left\{\prod_{a=1}^{A_m} f_{\psi}\left(z_a^m \middle| \boldsymbol{n}_a^m, \boldsymbol{n}_{a+1}^m\right)\right\}, \tag{4}$$

where the expectations are taken with respect to the distribution of genotypes $\boldsymbol{n}^m$, $m = 1, \ldots, M$ of the unknown contributors. The expectation in (4) involves summation over all configurations of possible genotypes of potential contributors. At a marker with A_m possible alleles, there are $\{A_m(A_m+1)/2\}^k$ possible combinations of genotypes, 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 Sect. 3 below. We note that it is also of specific interest to find the conditional distribution of the genotypes of unknown individuals—represented by the sets of allele counts $\boldsymbol{n}^m$ across markers—given the observed peak heights $\{z_a^m,\ m = 1, \ldots, M;\ a = 1, \ldots, A_m\}$. Our emphasis in this paper is on exact computation but our methods are equally useful in connection with potential Monte Carlo evaluation of relevant integrals.

3 Computation by auxiliary variables

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 $X = \{X_v\}_{v \in V}$ of discrete random variables. We describe our computational approach in this general setting before returning to the specific DNA mixture model in Sect. 4.

3.1 Auxiliary variables

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

Let h be a non-negative function which can be factorised as

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

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

For each $B \in \mathscr{B}$ we introduce a binary random variable $Y^B \in \{0, 1\}$. These *auxiliary* variables are assumed to be mutually conditionally independent given the network and distributed as

$$\mathbb{P}\left(Y^B = 1 \middle| X = x\right) = \mathbb{P}\left(Y^B = 1 \middle| X_B = x_B\right) = h_B(x_B)/k_B. \tag{5}$$

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 (5) 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 $\mathbb{E}\{\prod_{B \in \mathscr{B}} h_B(X_B)\}$ can now be expressed in terms of the probability of a specific configuration of the auxiliary variables introduced. As Lemma 1 reveals, this is also the case for the expectation of a product of any subset of the variables $h_B(X_B)$.

Lemma 1 *For all $\mathscr{B}' \subseteq \mathscr{B}$ it holds that*

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

Proof Using (5) and the fact that Y^B, $B \in \mathscr{B}$ are mutually conditionally independent given X we get

$$\begin{aligned}
\mathbb{E}\left\{\prod_{B \in \mathscr{B}'} h_B(X_B)\right\} &= \mathbb{E}\left\{\prod_{B \in \mathscr{B}'} \left(\mathbb{P}\left(Y^B = 1 \middle| X_B\right) k_B\right)\right\} \\
&= \mathbb{E}\left\{\prod_{B \in \mathscr{B}'} \mathbb{P}\left(Y^B = 1 \middle| X\right)\right\} \prod_{B \in \mathscr{B}'} k_B \\
&= \mathbb{E}\left\{\mathbb{P}\left(\bigcap_{B \in \mathscr{B}'} \left\{Y^B = 1\right\} \middle| X\right)\right\} \prod_{B \in \mathscr{B}'} k_B \\
&= \mathbb{P}\left(\bigcap_{B \in \mathscr{B}'} \left\{Y^B = 1\right\}\right) \prod_{B \in \mathscr{B}'} k_B
\end{aligned}$$

as desired. □

Lemma 1 allows the interpretation of the expectation of interest as a scaled probability, which can be computed in various ways; Proposition 1 below provides one such way.

The Bayesian network representing the distribution of $\{X_v\}_{v \in V}$ can be extended to include the variables $\{Y^B\}_{B \in \mathscr{B}}$ by for each B adding Y^B as a child of $\{X_v\}_{v \in B}$ with conditional distributions of Y^B as given in (5). 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).

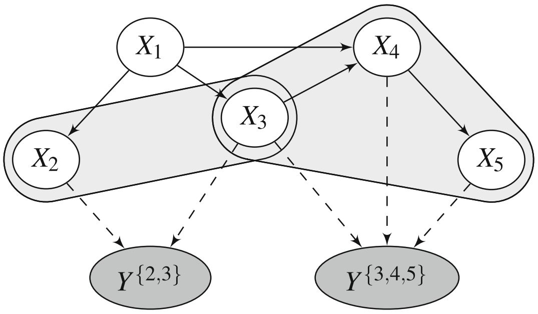


Fig. 2 Extending a network with two binary variables for computation of $\mathbb{E}\left(h_{\{2,3\}}(X_2, X_3)h_{\{3,4,5\}}(X_3, X_4, X_5)\right)$. Here $\mathscr{B} = \{\{2,3\},\{3,4,5\}\}$

3.2 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 are largely following the exposition of Dawid (1992) as described in Cowell et al. (1999), Sect. 6.3 where also further details of Bayesian networks and probability propagation can be found.

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 edges between nodes that have a common child and removing directions for existing edges. Subsequently edges are added to ensure that the resulting graph is *chordal*, i.e. that in any cycle of four or more distinct nodes there exist two non-consecutive nodes which are joined by an edge. This process is known as *triangulation* and can generally be done in many ways; finding an optimal triangulation for a given optimality criterion is an NP-hard problem (Yannakakis 1981). Finally the (maximal) *cliques*, i.e. maximal complete sets, 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 Sect. 4.4.1 below.

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

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

where $\mathscr{C}$ denotes the set of cliques and $\mathscr{S}$ 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 to a canonical form, where all potentials of the charge are equal to the function g marginalised onto the corresponding clique or separator, i.e.

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

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 $S \in \mathscr{S}$ with minimal state space.

The charge g can be modified by *propagating likelihood evidence* $\ell_v(x_v)$ on nodes $v \in V$, denoting the process of multiplying the charge by non-negative functions $\ell_v(x_v)$ followed by a propagation to bring the modified charge to its canonical form. The modified charge is

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

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 likelihood evidence yields the expectation of the product of the likelihood evidence with respect to the distribution $p(x)$:

$$\begin{aligned}\frac{N_2}{N_1} &= \frac{\sum_x g(x) \prod_{v\in V} \ell_v(x_v)}{\sum_y g(y)} \\ &= \sum_x \frac{g(x)}{\sum_y g(y)} \prod_{v\in V} \ell_v(x_v) \\ &= \sum_x p(x) \prod_{v\in V} \ell_v(x_v) \\ &= \mathbb{E}\left\{ \prod_{v\in V} \ell_v(X_v) \right\}. \end{aligned} \tag{6}$$

3.3 Calculation of expectations by propagation

As shown in Proposition 1 below, the property (6) ensures that the expectation of interest can be calculated by propagating likelihood evidence on the auxiliary variables.

Proposition 1 *Let likelihood evidence for each node* Y^B, $B \in \mathscr{B}' \subseteq \mathscr{B}$ *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 likelihood evidence. Then we have*

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

Proof

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

which by Lemma 1 equals the desired expectation. □

We note that in the special case where the original charge is normalised so that $N_1 = 1$ and the evidence functions ℓ_v are indicator functions for nodes being in particular states, it is well established that N_2 yields the joint probability for the specific configuration of states (Cowell et al. 1999, p. 36).

Proposition 1 leads to a practical way of computing the desired expectation: for each auxiliary variable Y^B we compute the conditional probabilities $\mathbb{P}\left(Y^B \middle| X_B\right)$ for all configurations of (Y^B, X_B) after which the expectation can be obtained in a single propagation of the corresponding likelihood evidence.

4 Bayesian networks for DNA mixtures

We describe here how our genotype representation in combination with appropriate auxiliary variables yields a powerful and flexible Bayesian network representation of the joint model for peak heights and genotypes. This representation allows efficient evaluation of a wide range of quantities of interest in a case analysis.

4.1 A Bayesian network representation of genotypes

The multinomial distribution of allele counts $(n_{i1}, \ldots, n_{iA})$ representing the genotype of an unknown contributor i for a given marker does not in itself have Markovian properties. However, if we define the partial sums

$$S_{ia} = \sum_{b=1}^{a} n_{ib}$$

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 Fig. 3.

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 2 we establish the formal correctness of the network specification.

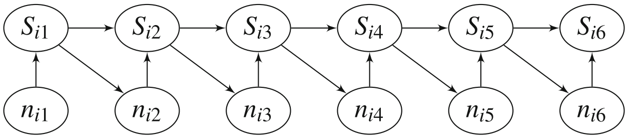


Fig. 3 Network representation of a genotype at a marker with $A = 6$ allelic types

Proposition 2 *The distributions of genotypes and partial sums satisfy the following relations*

$$\begin{aligned}&S_{i1} = n_{i1},\\ &n_{i1} \sim \text{bin}\left(2, q_1\right),\\ &\textit{and for}\;\; a \in \{2, \ldots, A\}\\ &S_{ia} = S_{i,a-1} + n_{ia},\\ &n_{ia} \mid S_{i,a-1} \sim \text{bin}\left(2 - S_{i,a-1}, q_a \Big/ \sum_{b=a}^{A} q_b\right).\end{aligned} \tag{7}$$

Finally, we have the conditional independence relations

$$\begin{aligned}&n_{ia} \perp\!\!\!\perp (n_{i1}, \ldots, n_{i,a-1}, S_{i1}, \ldots, S_{i,a-2}) \mid S_{i,a-1}\\ &S_{ia} \perp\!\!\!\perp (n_{i1}, \ldots, n_{i,a-1}, S_{i1}, \ldots, S_{i,a-2}) \mid (S_{i,a-1}, n_{ia}).\end{aligned} \tag{8}$$

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

$$\begin{aligned}&p(n_{ia} \mid n_{i1}, \ldots, n_{i,a-1})\\ &\quad= \frac{p(n_{i1}, \ldots, n_{i,a-1}, n_{ia})}{p(n_{i1}, \ldots, n_{i,a-1})}\\ &\quad= \frac{\frac{2!}{(2-S_{i,a-1}-n_{ia})!\prod_{b=1}^{a} n_{ib}!}\left(\sum_{b=a+1}^{A} q_b\right)^{2-S_{i,a-1}-n_{ia}} \prod_{b=1}^{a} q_b^{n_{ib}}}{\frac{2!}{(2-S_{i,a-1})!\prod_{b=1}^{a-1} n_{ib}!}\left(\sum_{b=a}^{A} q_b\right)^{2-S_{i,a-1}} \prod_{b=1}^{a-1} q_b^{n_{ib}}}\\ &\quad= \frac{(2-S_{i,a-1})!}{n_{ia}!(2-S_{i,a-1}-n_{ia})!}\\ &\qquad\times\left(1-\frac{q_a}{\sum_{b=a}^{A} q_b}\right)^{2-S_{i,a-1}-n_{ia}}\left(\frac{q_a}{\sum_{b=a}^{A} q_b}\right)^{n_{ia}}.\end{aligned}$$

The conditional independence (8) 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 (7). □

4.2 Auxiliary variables for computing the likelihood function

To compute the inner expectation in the expression (4) for the likelihood function, we note that this is an expectation of a product over alleles, where each factor is a function of the variables $\boldsymbol{n}_a$ and $\boldsymbol{n}_{a+1}$, and so we can compute this expectation using auxiliary variables as described in Sect. 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

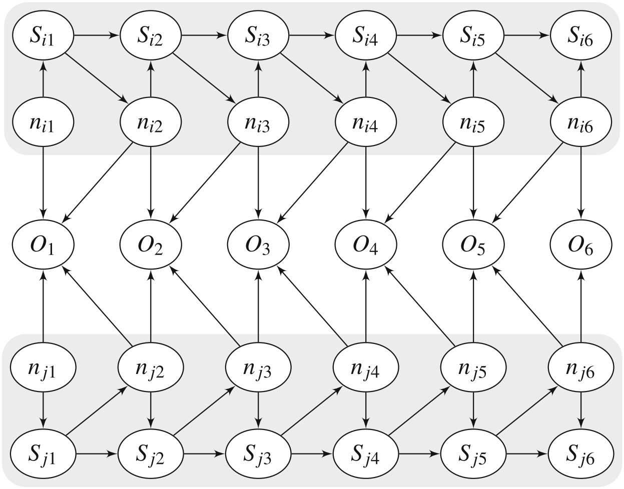


Fig. 4 Bayesian network modelling the genotypes of two unknown contributors i and j for a marker with six possible allelic types

shows the network for modelling one marker of a mixture with two contributors and six alleles.

The structure displayed in Fig. 4 can be seen as representing our model as parallel and coupled hidden Markov models; indeed the probability propagation algorithm we use is one of a variety of more general variants of the forward–backward or BCJR algorithm used for inference in hidden Markov models (Bahl et al. 1974). Cowell et al. (1999), Sect. 6.7 contains a more detailed discussion of the relation between algorithms of this type.

Note that O_a and its parents $n_{ia}, n_{i,a+1}, i \in \{1, \ldots, k\}$ 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 Sect. 4.4.1.

The distribution of an auxiliary variable O_a conditionally on the allele counts is defined using the distribution (3) of the peak height Z_a conditionally on the allele counts as follows:

If a peak for allele a has been observed above the detection threshold C, i.e. $z_a \geq C$, the distribution of O_a is defined as

$$\mathbb{P}\left(O_a = 1 \middle| \boldsymbol{n}_a, \boldsymbol{n}_{a+1}\right) = g_\psi\left(z_a \middle| \boldsymbol{n}_a, \boldsymbol{n}_{a+1}\right) / k_a^\psi, \tag{9}$$

noting the dependence of the scaling factor k_a^ψ on ψ. For an undetected peak, i.e. $z_a^m = 0$, the distribution of O_a is defined as

$$\mathbb{P}\left(O_a = 0 \middle| \boldsymbol{n}_a, \boldsymbol{n}_{a+1}\right) = G_\psi\left(C \middle| \boldsymbol{n}_a, \boldsymbol{n}_{a+1}\right). \tag{10}$$

Now Proposition 1 can readily be used to evaluate the contribution to the likelihood from marker m for a given value of ψ by propagating likelihood evidence

$$\ell_a(O_a) = \begin{cases} k_a^\psi \mathbb{1}_{\{O_a=1\}}, & \text{if } z_a \geq C \\ \mathbb{1}_{\{O_a=0\}}, & \text{if } z_a < C. \end{cases} \tag{11}$$

4.3 Posterior distribution of genotypes

The inclusion of auxiliary variables may serve other purposes than merely as a device for calculating an expectation. In DNA mixture analysis, we are interested in the conditional distribution of the genotypes $\boldsymbol{n}$ for contributors given—possibly only a subset of—the observed peak heights.

By propagating likelihood evidence (11) for a set of alleles $a \in \mathscr{A} \subseteq \{1, \ldots, A\}$, we obtain a representation of the conditional distribution of the full network given the relevant state of the auxiliary variables $O_a, a \in \mathscr{A}$. We have defined the auxiliary variables so that for all alleles the event $O_a = 1$ corresponds to the event $Z_a \geq C$ that the peak at allele a is above the threshold C. Therefore conditioning on auxiliary variables $O_a, a \in \mathscr{A}$ yields the conditional distribution of the nodes in the network given the peak height information $\{z_a\}_{a\in\mathscr{A}}$ as formalised in the following:

Proposition 3 *For an arbitrary subset $\mathscr{A}$ of alleles we have*

$$p\left(x \middle| \{z_a\}_{a\in\mathscr{A}}\right) = p\left(x \middle| \bigcap_{\substack{a\in\mathscr{A},\\ z_a\geq C}} \{O_a = 1\} \bigcap_{\substack{a\in\mathscr{A},\\ z_a<C}} \{O_a = 0\}\right). \tag{12}$$

Proof This follows from the following argument:

$$\begin{aligned}
& p(x) \prod_{a\in\mathscr{A}} \ell_a(O_a) \\
& \propto p\left(x \middle| \bigcap_{\substack{a\in\mathscr{A},\\ z_a\geq C}} \{O_a = 1\} \bigcap_{\substack{a\in\mathscr{A},\\ z_a<C}} \{O_a = 0\}\right) \\
& \propto p(x)\,\mathbb{P}\left(\bigcap_{\substack{a\in\mathscr{A},\\ z_a\geq C}} \{O_a = 1\} \bigcap_{\substack{a\in\mathscr{A},\\ z_a<C}} \{O_a = 0\} \middle| x\right) \\
& = p(x) \prod_{\substack{a\in\mathscr{A}\\ z_a\geq C}} \mathbb{P}\left(O_a = 1 \middle| x\right) \prod_{\substack{a\in\mathscr{A}\\ z_a<C}} \mathbb{P}\left(O_a = 0 \middle| x\right) \\
& = p(x) \prod_{\substack{a\in\mathscr{A}\\ z_a\geq C}} \mathbb{P}\left(O_a = 1 \middle| \boldsymbol{n}_a, \boldsymbol{n}_{a+1}\right) \prod_{\substack{a\in\mathscr{A}\\ z_a<C}} \mathbb{P}\left(O_a = 0 \middle| \boldsymbol{n}_a, \boldsymbol{n}_{a+1}\right) \\
& = p(x) \prod_{\substack{a\in\mathscr{A}\\ z_a\geq C}} \left\{g_\psi\left(z_a \middle| \boldsymbol{n}_a, \boldsymbol{n}_{a+1}\right) / k_a^\psi\right\} \prod_{\substack{a\in\mathscr{A}\\ z_a<C}} G_\psi\left(C \middle| \boldsymbol{n}_a, \boldsymbol{n}_{a+1}\right)
\end{aligned}$$

$$
\begin{aligned}
&\propto p(x) \prod_{a \in \mathscr{A}} f_{\psi}\left(\{z_a\}_{a \in \mathscr{A}} \middle| x\right) \\
&\propto p\left(x \middle| \{z_a\}_{a \in \mathscr{A}}\right)
\end{aligned}
$$

as desired. □

As a consequence, we can for example easily simulate from the conditional distribution of genotypes given observed peak heights, which we shall exploit in Sects. 5.3 and 5.4 below.

4.4 Network complexity considerations

The main concerns when applying the methodology of Sect. 3 to a specific problem are that the junction tree representation of the network may not fit in the physical memory, and that 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, defined as the sum of the sizes of state spaces for all cliques and separators. 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 total size determines how many numbers are needed to store the clique and separator tables and the number of elementary arithmetic operations for propagation is linear in the total size. In the worst case, total size also determines the number of cells that need updating when changing the conditional probability tables for the auxiliary variables.

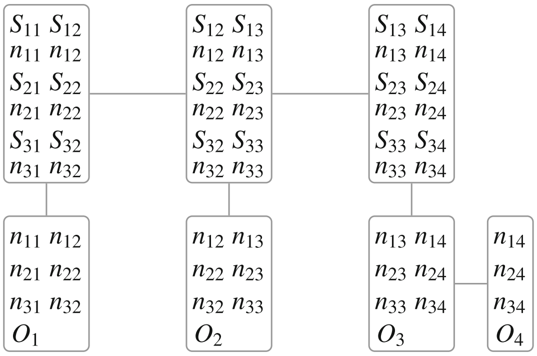
An additional concern lies in finding a good triangulation, as this can be both time- and memory-consuming; we eliminate this additional cost by specifying triangulations directly.

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.

4.4.1 Junction tree sizes for DNA mixtures

We shall consider three different triangulations of networks of the type discussed in Sect. 4.2 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 an auxiliary variable together with its parent set, as these are complete 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.



S_{11} S_{12} n_{11} n_{12} S_{21} S_{22} n_{21} n_{22} S_{31} S_{32} n_{31} n_{32}
S_{12} S_{13} n_{12} n_{13} S_{22} S_{23} n_{22} n_{23} S_{32} S_{33} n_{32} n_{33}
S_{13} S_{14} n_{13} n_{14} S_{23} S_{24} n_{23} n_{24} S_{33} S_{34} n_{33} n_{34}
n_{11} n_{12} n_{21} n_{22} n_{31} n_{32} O_1
n_{12} n_{13} n_{22} n_{23} n_{32} n_{33} O_2
n_{13} n_{14} n_{23} n_{24} n_{33} n_{34} O_3
n_{14} n_{24} n_{34} O_4

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

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 $2k$ parents. The N auxiliary variables for the last allele have only k parents.

Bearing Fig. 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 $a-1$ receives 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 $a = 1, \ldots, A-1$. The corresponding junction tree, which we shall refer to as the *slice tree*, is displayed in Fig. 5. We note that using the propagation algorithm on the slice tree is effectively equivalent to using the forward–backward algorithm on the hidden Markov chain with these cliques representing the hidden states. 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

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

However, we can improve on this triangulation by splitting each slice into two cliques as Fig. 6 illustrates. The resulting *triangle tree* in Fig. 7 has $2(A-1)$ cliques of each $3k$ nodes and $2(A-1)$ separators of each $2k$ nodes, and thus the total size

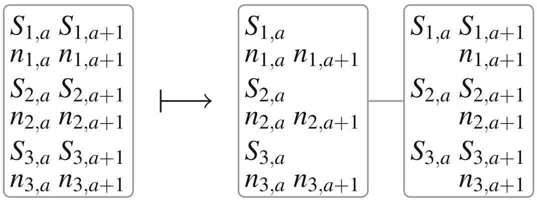
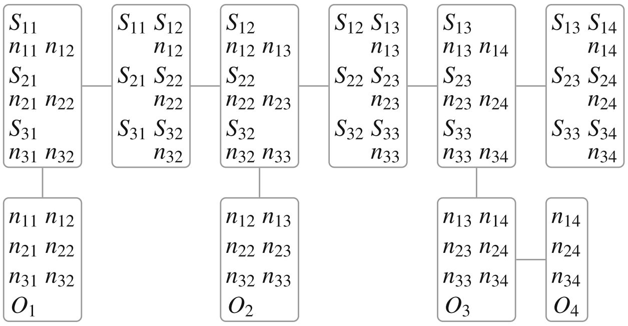


Fig. 6 Splitting each slice into two cliques consisting of lower and upper for a reduction in total size



S_{11} n_{11} n_{12} S_{21} n_{21} n_{22} S_{31} n_{31} n_{32}
S_{11} S_{12} n_{12} S_{21} S_{22} n_{22} S_{31} S_{32} n_{32}
S_{12} n_{12} n_{13} S_{22} n_{22} n_{23} S_{32} n_{32} n_{33}
S_{12} S_{13} n_{13} S_{22} S_{23} n_{23} S_{32} S_{33} n_{33}
S_{13} n_{13} n_{14} S_{23} n_{23} n_{24} S_{33} n_{33} n_{34}
S_{13} S_{14} n_{14} S_{23} S_{24} n_{24} S_{33} S_{34} n_{34}
n_{11} n_{12} n_{21} n_{22} n_{31} n_{32} O_1
n_{12} n_{13} n_{22} n_{23} n_{32} n_{33} O_2
n_{13} n_{14} n_{23} n_{24} n_{33} n_{34} O_3
n_{14} n_{24} n_{34} O_4

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

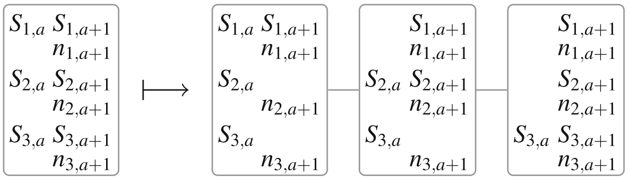


Fig. 8 Splitting *upper triangle* cliques for a further reduction in total size

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

grows less quickly with the number of unknown contributors than the slice tree; see Fig. 10.

In the case of only one unknown contributor, the total size of the triangle tree cannot be reduced (7). However, with more than one unknown contributor, each clique containing k upper triangles can be further split into k cliques as in Fig. 8.

Note that the cliques containing k lower triangle sets cannot be split in a similar fashion. The resulting junction tree—the *split 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

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

A further slight reduction of the total size can be obtained by a small alteration in the cliques that cover nodes for the first two and last three alleles; the resulting tree is seen in Fig. 9. This is the best junction tree we have been able to construct. We have investigated junction trees found by the algorithm for minimizing total clique size (excluding separators) as implemented in HUGIN, but none have smaller total size than our split tree.

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

$$\boldsymbol{S}_A, \boldsymbol{S}_{A-1}, \boldsymbol{S}_1, \boldsymbol{n}_1, \{\boldsymbol{n}_a, \boldsymbol{S}_a\}_{a=2}^{A-2}, \boldsymbol{n}_{A-1}, \boldsymbol{n}_A,$$

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

The exponential growth of the total size of the three types of junction tree is illustrated in Fig. 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 configurations 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 Fig. 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

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

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 horizontal band in Fig. 10 represents a range of capacities from 2 to 512 GB of memory.

Figure 10 indicates that using the split 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 ten possible configurations of the states in each of $\{S_{ia}, S_{i,a+1}, n_{ia}, n_{i,a+1}\}$. So if the state 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.

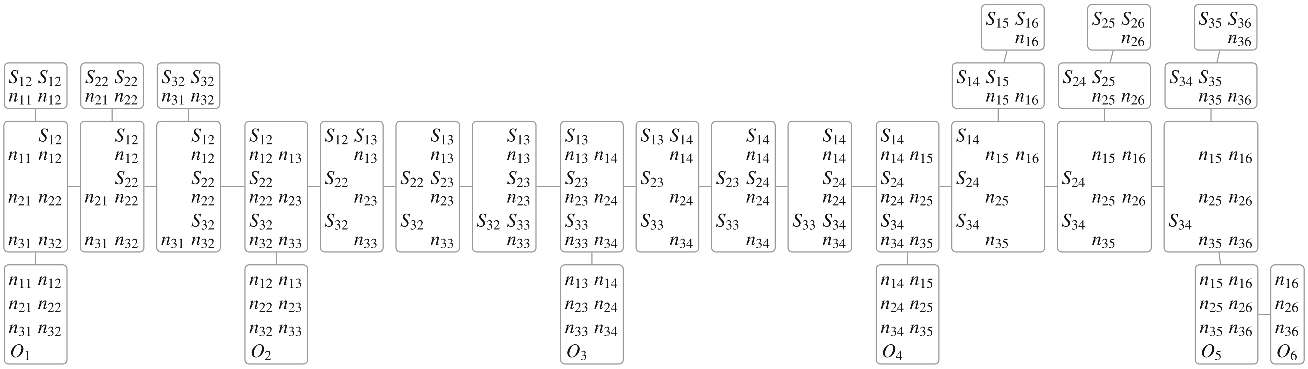

S_{15} S_{16} n_{16}
S_{25} S_{26} n_{26}
S_{35} S_{36} n_{36}
S_{12} S_{12} n_{11} n_{12}
S_{22} S_{22} n_{21} n_{22}
S_{32} S_{32} n_{31} n_{32}
S_{14} S_{15} n_{15} n_{16}
S_{24} S_{25} n_{25} n_{26}
S_{34} S_{35} n_{35} n_{36}
S_{12} n_{11} n_{12} S_{22} n_{21} n_{22} S_{32} n_{31} n_{32}
S_{12} n_{12} S_{22} n_{21} n_{22} n_{31} n_{32}
S_{12} n_{12} S_{22} n_{22} S_{32} n_{31} n_{32}
S_{12} n_{12} n_{13} S_{22} n_{22} n_{23} S_{32} n_{32} n_{33}
S_{12} S_{13} n_{13} S_{22} n_{23} S_{32} n_{33}
S_{13} n_{13} S_{22} S_{23} n_{23} S_{32} n_{33}
S_{13} n_{13} S_{23} n_{23} S_{32} S_{33} n_{33}
S_{13} n_{13} n_{14} S_{23} n_{23} n_{24} S_{33} n_{33} n_{34}
S_{13} S_{14} n_{14} S_{23} n_{24} S_{33} n_{34}
S_{14} n_{14} S_{23} S_{24} n_{24} S_{33} n_{34}
S_{14} n_{14} S_{24} n_{24} S_{33} S_{34} n_{34}
S_{14} n_{14} n_{15} S_{24} n_{24} n_{25} S_{34} n_{34} n_{35}
S_{14} n_{15} n_{16} S_{24} n_{25} S_{34} n_{35}
n_{15} n_{16} S_{24} n_{25} n_{26} S_{34} n_{35}
n_{15} n_{16} n_{25} n_{26} S_{34} n_{35} n_{36}
n_{11} n_{12} n_{21} n_{22} n_{31} n_{32} O_1
n_{12} n_{13} n_{22} n_{23} n_{32} n_{33} O_2
n_{13} n_{14} n_{23} n_{24} n_{33} n_{34} O_3
n_{14} n_{15} n_{24} n_{25} n_{34} n_{35} O_4
n_{15} n_{16} n_{25} n_{26} n_{35} n_{36} O_5
n_{16} n_{26} n_{36} O_6

Fig. 9 Our best junction tree for a DNA mixture network with $k = 3$, $A = 6$, and $N = 1$

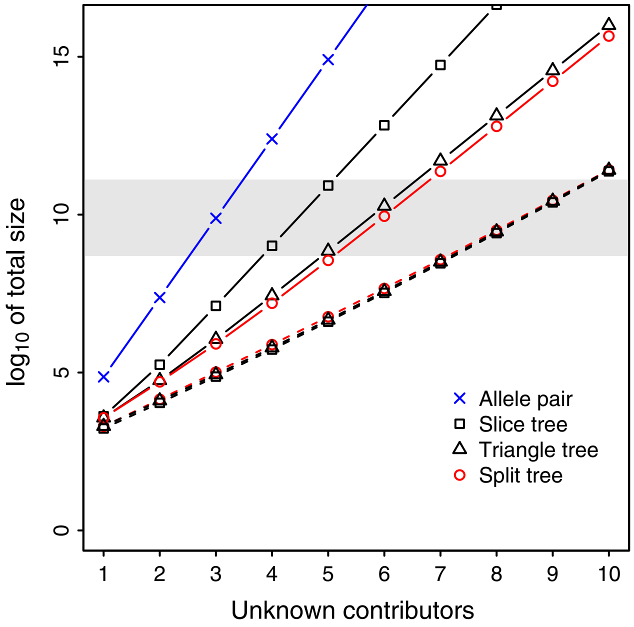


Fig. 10 Total sizes of junction trees as a function of the number k of unknown contributors, in the case of $A = 25$ allelic types and $N = 3$ auxiliary variables per allele. *Solid lines* are uncompressed sizes and dashed lines compressed sizes. The *horizontal band* indicates total sizes ranging from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

4.4.2 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. An alternative possibility is to use a more algebraic representation of the formulae; however, again this would typically reduce flexibility and we shall not discuss these possibilities further here.

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). For A alleles there are $A(A+1)/2$ possible unordered pairs. If an allele pair is represented by a single node for each of the k unknown contributors, the parent set for each auxiliary variable in this network is the collection of these k nodes, resulting in a junction tree where each clique and each separator contains all k 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 alleles. For a fixed number of alleles, the growth in the number k of unknown contributors is still exponential; see Fig. 10. For junction trees based on the Markov representation of genotypes, the number of alleles makes a negligible impact on the total size. However, for the allele pair representation the rate of growth depends heavily on the number of alleles: For 25 alleles as in Fig. 10 it is feasible to handle up to about three unknown contributors, whereas if only 10 allelic types are needed, then 4–5 unknown contributors can be handled. For $A \geq 7$, 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 allele representation Another possibility, used for example in Dawid et al. (2002) and Mortera et al. (2003), is to model the genotype at the single allele level. A single allele can be represented by the same Markovian network structure as that in Fig. 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

Table 1 Peak heights for marker D2S1338 above threshold in sample 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

Table 2 Maximum likelihood estimates based on MC15

Defence hypothesis		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

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 allele network will always be inferior to the genotype network.

The total size of the split tree using the single allele representation renders computation feasible for up to about five unknown contributors. Compression of the slice tree with single alleles 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 slice tree using the genotype representation introduced in Sect. 4.1. It would stay feasible if $k \leq 7$. For $A \geq 11$, the single allele representation compares favourably to the allele pair representation.

Although inefficient, the single allele network representation may be preferable for other reasons; for example in cases where the two alleles 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).

5 DNA mixture analysis

As a generic example we consider the DNA sample *MC15* from Gill et al. (2008), also analysed in Cowell et al. (2013). The sample is believed to contain DNA from at least three contributors. 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.

5.1 Estimation

The model parameters $\psi = (\rho, \eta, \xi, \phi)$ are typically unknown and need to be estimated. In particular, the proportions ϕ_i of DNA from each contributor i are specific to the case at hand.

Being able to evaluate the likelihood function as in Sect. 4.2, estimation can be done by numerical maximisation. In `DNAmixtures`, likelihood functions are maximised numerically using `Rsolnp` (Ghalanos and Theussl 2012; Ye 1987). Approximate standard errors of estimates are based on the inverse Hessian of the likelihood function found by numerical derivation using `numDeriv` (Gilbert and Varadhan 2012).

We wish to calculate the likelihood ratio for evidence against K_3 as in (1). We consider the prosecution hypothesis H_p that the contributors to the trace are individuals K_1, K_2, and the defendant K_3, whereas the defence hypothesis H_d replaces the defendant with an unknown contributor U. The maximum likelihood estimates and standard errors obtained under the two hypotheses are given in Table 2. From the last line in the table we see that the likelihood ratio against the suspect K_3 having contributed to the mixture is of the order 10^{12}, giving overwhelming support for the prosecution hypothesis H_p. As the estimates for the parameters are quite similar under the two hypotheses, the large likelihood ratio is reflecting that it is very improbable that a random individual would have this particular genotype by chance.

5.2 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 sample under analysis.

Previously we have introduced auxiliary variables O_a, to enable simple computation of the likelihood function (4) and representation of evidence from observed peak heights (12). We shall in the following introduce further auxiliary variables: binary variables D_a which indicate whether a peak is below threshold at allele a, and variables Q_a which indicate whether a peak observed at allele a is less than a specified value. Both of these sets of auxiliary variables are useful for model validation; in addition, the variables D_a can be used in an analysis which is based only on peak presence; see Sect. 5.5 below.

5.2.1 Assessing peak height distributions

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

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

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

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

$$\mathbb{P}(Z_a \leq z) = \mathbb{E}\left\{\mathbb{P}(Z_a \leq z \mid \boldsymbol{n}_a, \boldsymbol{n}_{a+1})\right\} \tag{13}$$

is the expectation of a product with only one factor, so 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 $\mathbb{P}(Z_a < C)$ and $\mathbb{P}(Z_a \geq C)$.

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 at all other alleles to get $\mathbb{P}\left(Z_a \leq z \middle| Z_b = z_b, b \neq a, Z_a \geq C\right)$, or we could include this information for only the preceding alleles in the ordering to get $\mathbb{P}\left(Z_a \leq z \middle| Z_b = z_b, b \leq a, Z_a \geq C\right)$. Proposition 3 ensures that these distributions can all be obtained simply through conditioning on relevant subsets of variables O_a.

In Fig. 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 sample 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 hypothesis.

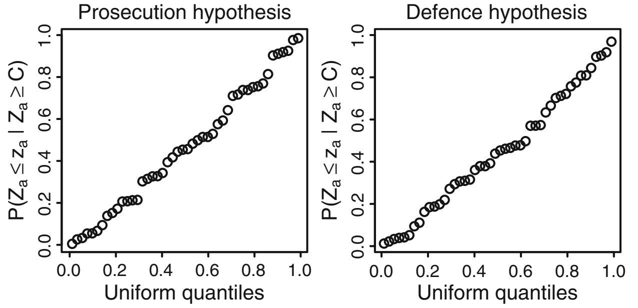


Fig. 11 Quantile–quantile plots for the prosecution and defence hypotheses for MC15

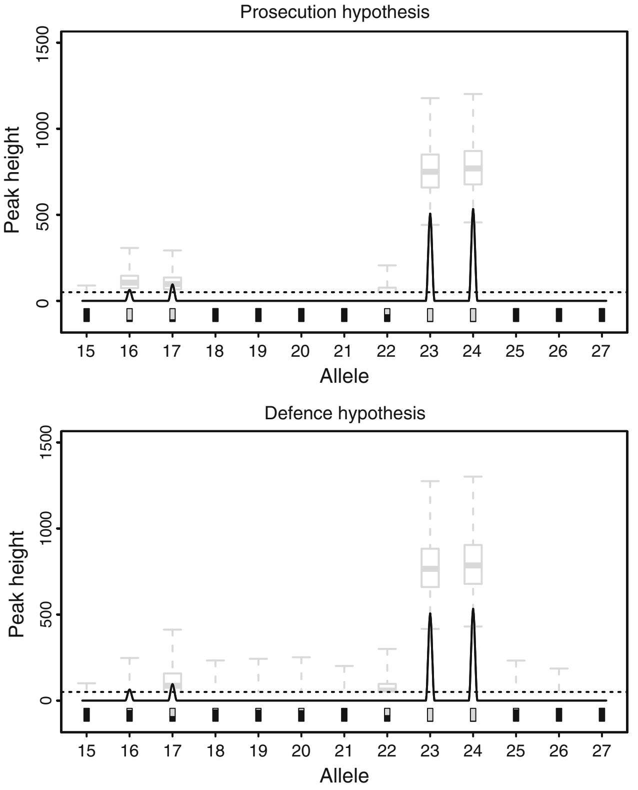


Fig. 12 Comparison of observed peak heights to their predictive distribution conditionally on all other observed peak heights for marker D2S1338. The *bar* below each peak indicates the probabilities of observing (*grey*) and not observing (*black*) a peak at this allele. (Colour figure online)

We can also take a closer look at the distribution of the peak height at any single allele, for example to identify unusual observations. This is illustrated in Fig. 12. Boxes indicate quartiles and whiskers indicate 0.5 and 99.5 % prediction limits for the conditional distributions of peak heights $\mathbb{P}\left(Z_a \leq z \middle| Z_b = z_b, b \neq a, Z_a \geq C\right)$. The quantiles are found by numerical inversion of the distribution function (13).

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 Fig. 11. Expectedly, the prosecution hypothesis predicts complete absence of peaks at alleles 18–21 and 25–27, since these alleles are not present in either genotypes nor are they in a possible stutter position.

In contrast, under the defence hypothesis peaks are *a priori* possible at any allele.

5.2.2 *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 any 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,

$$p_a = \mathbb{P}\left(Z_a \geq C \middle| z_i, i < a\right) = \mathbb{P}\left(D_a = 0 \middle| z_i, i < a\right)$$

which can be obtained by propagation as described in Sect. 4.3. For each allele a, we then consider the logarithmic score

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

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 or $1 - p_a$) to the event that actually happens.

The cumulative logarithmic score, adjusted for incremental expectations,

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

is a martingale with respect to the sequence of peak heights. As $\mathbb{V}\left(M_a - M_{a-1} \middle| Z_b, b < a\right) = \mathbb{V}\left(Y_a \middle| Z_b, b < a\right)$, the distribution of the normalised cumulative score

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

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,

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

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

$$\mathbb{E}\left(Y_a \middle| Z_b, b < a\right) = -p_a \log p_a - (1 - p_a)\log(1 - p_a),$$
$$\mathbb{V}\left(Y_a \middle| Z_b, b < a\right) = p_a(1 - p_a)\left\{\log p_a - \log(1 - p_a)\right\}^2.$$

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

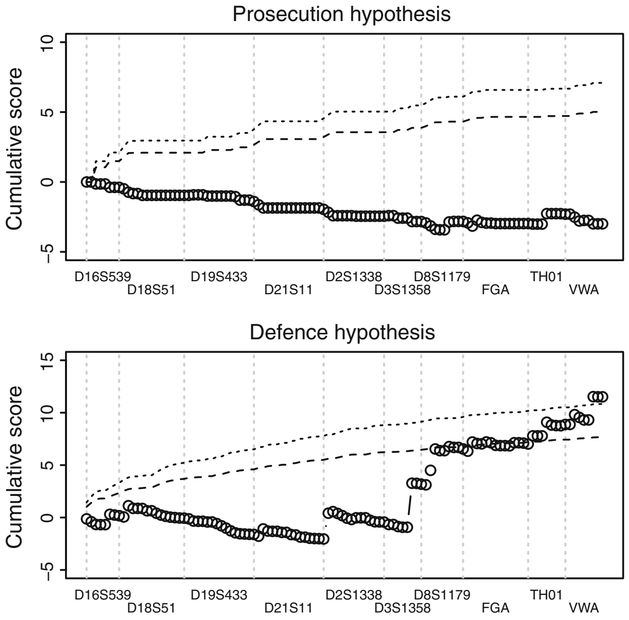


Fig. 13 Prequential monitor plots of the prosecution and defence hypotheses for MC15. The dashed horizontal lines indicate upper 95 and 99 % pointwise predictive limits based on the approximating normal distribution

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 $p_a \in \{0, 0.5, 1\}$—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 monitor crosses 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.

5.3 Simulation

As stated in Proposition 3, 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 set of DNA profiles and corresponding peak heights, either marginally or conditionally on relevant subsets of the observed peak heights. More generally, we have for any event B that

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

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 $p\left(\boldsymbol{n} \middle| B\right)$ by standard methods

(Cowell et al. 1999, Sect. 6.4.3). Thus to sample the the peak heights, we just further need a method for sampling from $f_\psi\left(\{z_a\}_{a\in A}\middle|\boldsymbol{n}, B\right)$.

This method of simulation can for example be used in a bootstrap analysis of the estimation uncertainty or in a Monte-Carlo based fully Bayesian analysis 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.

5.4 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 sample MC15, we would like to identify the genotype of the unknown contributor U. If we consider the full DNA profile 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 for a single marker, we get a more concentrated distribution, as displayed in Table 3, where the total probability of the six combinations add up to one.

Table 3 Probabilities of genotype at marker D2S1338 for the unknown contributor U under the defence hypothesis

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

The defendant K_3 has genotype (16,17)

As the table shows, the probability that the unknown contributor has at least one allele 17 is 0.9983, 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 sample is that the other allele has dropped out.

5.5 Weight 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 Sect. 5.2 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 parameter 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 somewhat analogous to the one used in `likeLTD` as suggested by Balding (2013), where peak heights are used only to classify alleles as definitely absent, definitely present, or possibly present in the mixture.

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.

5.6 Multiple DNA mixtures

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

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

The network now models the set of all unknown contributors to the mixtures. Denote by ϕ_i^j the proportion of DNA that contributor i has made to sample j. Then $\phi_i^j = 0$ corresponds to contributor i not being present in mixture j. Therefore, the case where some or all contributors are distinct to a particular mixture 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 mixtures, 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 DNA mixture pertaining to the same case.

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

6 Discussion

Other authors have addressed statistical analysis of DNA mixtures using various heuristics for limiting the number of terms involved in the computation (Bill et al. 2005; Tvedebrink et al. 2010; Puch-Solis et al. 2013) and have only considered two or three unknown contributors (Cowell et al. 2011).

We note that our computational methods are exact 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 Sect. 4.4.1, we are able to handle more contributors than what has previously been possible. Indeed, using the methodology presented here and the corresponding implementation by Graversen (2013) in `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.

Methods used to calculate approximations to likelihood ratios as in (1) are potentially unreliable, as they are ratios of very small numbers; in particular, small changes in the denominator of the ratio can have drastic effects on the weight of evidence. For other types of calculation, approximate methods could be very useful if they represent major reductions in computational effort. However, it is hard to see how the exponential growth in effort with the number of unknown contributors can be avoided while maintaining sufficient computational accuracy. The basic junction tree representations of the model are equally useful for obtaining Monte Carlo approximations to relevant quantities, for example in the slightly more complex situation where parameters are treated by fully Bayesian methods and likelihood functions should be integrated rather than maximised, see also Graversen and Lauritzen (2013).

We conclude with noting that we have far from exhausted the flexibility and the potential of the Bayesian network representation of the model. 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 direct computation of the probability that a specific peak is due to stutter or an absent peak is due to dropout or allele absence; see Cowell et al. (2013) for this and further examples.

References

Bahl, L., Cocke, J., Jelinek, F., Raviv, J.: Optimal decoding of linear codes for minimizing symbol error rate. IEEE Trans. Inf. Theory **20**, 284–287 (1974)

Balding, D.: Evaluation of mixed-source, low-template DNA profiles in forensic science. Proc. Natl. Acad. Sci. USA. **110**(30), 12,241–12,246 (2013)

Balding, D.J.: Weight-of-Evidence for Forensic DNA Profiles. Statistics in Practice. Wiley, Chichester (2005)

Bill, M., Gill, P., Curran, J., Clayton, T., Pinchin, R., Healy, M., Buckleton, J.: PENDULUM: a guideline-based approach to the interpretation of STR mixtures. Forensic Sci. Int. **148**, 181–189 (2005)

Cowell, R.G., Dawid, A.P., Lauritzen, S.L., Spiegelhalter, D.J.: Probabilistic Networks and Expert Systems. Springer, New York (1999)

Cowell, R.G., Lauritzen, S.L., Mortera, J.: Probabilistic expert systems for handling artifacts in complex DNA mixtures. Forensic Sci. Int. **5**, 202–209 (2011)

Cowell, R.G., Graversen, T., Lauritzen, S., Mortera, J.: Analysis of forensic DNA mixtures with artefacts. arXiv **1302**, 4404 (2013)

Dawid, A.P.: Statistical theory. The prequential approach. J. R. Stat. Soc. Ser. A **147**, 277–305 (1984)

Dawid, A.P.: Applications of a general propagation algorithm for probabilistic expert systems. Stat. Comput. **2**, 25–36 (1992)

Dawid, A.P., Mortera, J., Pascali, V.L., van Boxel, D.W.: Probabilistic expert systems for forensic inference from genetic markers. Scand. J. Stat. **29**, 577–595 (2002)

Ghalanos, A., Theussl, S.: Rsolnp: General Non-linear Optimization Using Augmented Lagrange Multiplier Method. R package version 1.12 (2012)

Gilbert P, Varadhan R (2012) numDeriv: Accurate Numerical Derivatives. `R`-package version 2012.9-1.

Gill, P., Curran, J., Neumann, C., Kirkham, A., Clayton, T., Whitaker, J., Lambert, J.: Interpretation of complex DNA profiles using empirical models and a method to measure their robustness. Forensic Sci. Int. **2**, 91–103 (2008)

Graversen, T.: DNAmixtures: Statistical Inference for Mixed Traces of DNA. `R`-package version 0.1-0, `dnamixtures.r-forge.r-project.org/` (2013)

Graversen, T., Lauritzen, S.: Estimation of parameters in DNA mixture analysis. J. Appl. Stat. **40**(11), 2423–2436 (2013)

Green, P.J., Mortera, J.: Sensitivity of inferences in forensic genetics to assumptions about founder genes. Ann. Appl. Stat. **3**, 731–763 (2009)

Hugin Expert A/S (2013) Hugin API Reference Manual, Version 7.7. Hugin Expert A/S, Aalborg, Denmark.

Konis, K.: RHugin. `R`-package version 7.7-5 (2013)

Lindley, D.: A problem in forensic science. Biometrika **64**(2), 207–213 (1977)

Mortera, J., Dawid, A.P., Lauritzen, S.L.: Probabilistic expert systems for DNA mixture profiling. Theor. Popul. Biol. **63**, 191–205 (2003)

Puch-Solis, R., Rodgers, L., Mazumder, A., Pope, S., Evett, I., Curran, J., Balding, D.: Evaluating forensic DNA profiles using peak heights, allowing for multiple donors, allelic dropout and stutters. Forensic Sci. Int. **7**(5), 555–563 (2013)

Seillier-Moiseiwitsch, F., Dawid, A.P.: On testing the validity of sequential probability forecasts. J. Am. Stat. Assoc. **88**, 355–359 (1993)

Tvedebrink, T., Eriksen, P.S., Mogensen, H.S., Morling, N.: Evaluating the weight of evidence by using quantitative short tandem repeat data in DNA mixtures. Appl. Stat. **59**, 855–874 (2010)

Yannakakis, M.: Computing the minimum fill-in is NP-complete. SIAM J. Algebr. Discrete Methods **2**, 77–79 (1981)

Ye, Y.: Interior algorithms for linear, quadratic, and linearly constrained non-linear programming. PhD thesis, Department of Electrical Engineering, Stanford University, Stanford (1987).