Probabilistic Expert Systems for Handling Artifacts in Complex DNA Mixtures

R. G. Cowell*

Faculty of Actuarial Science and Insurance

Cass Business School

106 Bunhill Row

London EC1Y 8TZ, UK.

S. L. Lauritzen

Department of Statistics

University of Oxford

1 South Parks Road

Oxford OX1 3TG, UK.

J. Mortera

Dipartimento di Economia

Università Roma Tre

Via Silvio D'Amico, 77

00145 Roma, Italy

February 1, 2010

*Corresponding author: email: rgc@city.ac.uk, tel: +44 (0)20 7040 8454, fax: +44 (0)20 7040 8572.

Abstract

This paper presents a coherent probabilistic framework for taking account of allelic dropout, stutter bands and silent alleles when interpreting STR DNA profiles from a mixture sample using peak size information arising from a PCR analysis. This information can be exploited for evaluating the evidential strength for a hypothesis that DNA from a particular person is present in the mixture. It extends an earlier Bayesian network approach that ignored such artifacts. We illustrate the use of the extended network on a published casework example.

Keywords

Allelic dropout, artifacts, Bayesian networks, DNA mixtures, peak area, probabilistic expert systems, silent alleles, STR marker, stutter bands.

1 Introduction

When interpreting the output from a PCR analysis of a DNA mixture, the peak sizes obtained provide useful information regarding the relative amounts of DNA in the mixture originating from the contributors. This information can be exploited to make inferences regarding the genetic profiles of unknown contributors to the mixture, or for evaluating the evidential strength for a hypothesis that DNA from a particular person is present in the mixture. However, a variety of complications may occur during the PCR amplification

process, collectively referred to in the forensic genetics literature as *artifacts*. The presence of such artifacts makes it difficult to carry out these inference tasks. In this paper we describe Bayesian networks for analysing complex DNA mixtures which incorporate possible allelic dropout, stutter bands, and silent alleles in a comprehensive and fully probabilistic analysis of such mixtures.

Gill *et al.* [1] give recommendations on DNA mixture interpretation including some general guidelines for handling artifacts such as dropout and stutter bands. In a recent paper Gill *et al.* [2] illustrate a method to interpret complex DNA profiles where peak height information is used in the preprocessing of PCR output to identify potential stutter bands and other artifacts. In contrast, we present a probability model that handles these artifacts simultaneously.

Our model is also applicable in the investigative phase of DNA mixtures. This could involve a separation analysis to determine the profiles of unknown contributors, or a consideration of various scenarios to determine the most likely contributors to the DNA mixture amongst a group of individuals whose DNA profiles are known.

The plan of the paper is as follows. In the next section we briefly describe the Bayesian network of Cowell *et al.* [3] for modelling peak area values in the absence of artifacts. Then in § 3 we show how to extend this model to handle silent alleles, dropout and stutter bands. In § 4 we analyze the two mixtures taken from casework presented in [2] using our extended model.

Concluding remarks and suggestions for future work are given in § 5.

2 Gamma model

In this section we present an overview of the basic model for peak areas. For a detailed exposition of this model and the implementation as a Probabilistic Expert System (PES) [4] we refer the reader to Cowell *et al.* (2006, 2007a, 2007b).

The gamma model of [3] considers I potential contributors to a DNA mixture. Let there be M markers to be used in the analysis of the mixture with marker m having A_m allelic types, $m = 1, \ldots, M$. Let θ_i denote the proportion of DNA from individual i prior to PCR amplification, with $\theta = (\theta_1, \theta_2, \ldots, \theta_I)$ denoting the vector of proportions from all contributors. Thus $\sum_{i=1}^{I} \theta_i = 1$. It is assumed that this pre-amplification proportion of DNA is constant across markers.

For a specific marker m, the model is describing the *peak weight* W_{+a} at allele a which here is the *peak height* h_a multiplied by a. We have earlier used the peak area instead of peak height, but this makes little practical difference. The model makes the following further assumptions:

- W_{+a} is approximately proportional to the amount of DNA of type a after PCR amplification.
- If W_{ia} denotes the contribution of individual i to peak weight at allele a, then $W_{+a} = \sum_i W_{ia}$.

- Each contribution W_{ia} from individual i to peak weight at allele a has a gamma distribution, $W_{ia} \sim \Gamma(\rho\gamma_i n_{ia}, \eta)$, where:
  - $\gamma_i = \gamma\theta_i$ is the amount of DNA from individual i in the mixture, γ being the total amount of DNA;
  - n_{ia} is the number of alleles of type a carried by individual i;
  - η determines scale and ρ is the amplification factor. Both may be marker dependent.

It follows from properties of the gamma distribution assumption that

$$W_{+a} = \sum_i W_{ia} \sim \Gamma\left(\rho \sum_i \gamma_i n_{ia}, \eta\right)$$

and since $\sum_a \sum_i \theta_i n_{ia} = 2$, that

$$W_{++} = \sum_a W_{+a} = \sum_a \sum_i W_{ia} \sim \Gamma(2\rho\gamma, \eta).$$

By scaling the weight of each allele by the total marker weight we obtain *relative weights* R_a:

$$R_a = W_{+a}/W_{++} \sim Dir(\rho B_a).$$

Here $B_a = \sum_i \gamma_i n_{ia}$ is the weighted allele number, and $B_+ = 2\gamma$ is twice the total amount of DNA γ and is marker independent. It follows that, observed

values r_a of R_a give a contribution to the likelihood of

$$L(\mu|W) \propto \prod_a \frac{r_a^{\mu_a(1/\sigma^2-1)}}{\Gamma\{\mu_a(1/\sigma^2-1)\}} \tag{1}$$

where $\mu_a = B_a/B_+ = \sum_i \theta_i n_{ia}/2$ and $\sigma^2 = 2\rho\gamma$, where we call σ^2 the *variance factor.*

Figure 1 illustrates the structure of the basic Bayesian network in the case of two contributors to a mixture in which three peaks are observed on a single marker.

Insert Figure 1 about here.

Recently, Cowell [7] has carried out an analysis of the validity of the gamma model by comparing with simulations of mixtures using a stochastic procedure presented by Gill *et al.* [8]. He found that the gamma model gives a good description of the distribution of peak area values arising from the PCR process in amplifying moderate to large quantities of (simulated) DNA, but becomes inadequate in the low-copy-template regime where dropout becomes a significant factor.

3 Artifacts

Here we extend the Bayesian network based on the gamma model to deal with the possible artifacts of dropout, silent alleles, and stutter. These artifacts are all handled simultaneously in our PES. An important feature of using the

PES is that it provides posterior probabilities for the presence, absence, or degree, of the occurrence of such features: it does not have to be assumed at the outset of the analysis that such features are definitely present or definitely not present. For ease of exposition, we will explain how these artifacts may be incorporated in the PES one at a time as network fragments.

3.1 Dropout

In amplifying a sample of DNA in the PCR apparatus, one of the first steps is to extract the nuclear DNA from the cells using enzymes, and then to transfer a sample of the extract (aliquot) to the PCR apparatus. A major source of dropout, which is particularly acute in the low-template scenario, is the failure of some alleles to get selected for input into the PCR apparatus. The differential selection of alleles is also a factor in the stochastic variability of peak size values. Before presenting our dropout model, we summarize how dropout arises in the simulation model of [8].

The selection of DNA material for the PCR apparatus proceeds in two stages. First, an enzyme is added to the DNA sample to break up the nuclei of the cells. Then some of the aliquot is taken to be put into the PCR apparatus. Both allele sampling processes are stochastic and may be modelled mathematically by binomial sampling. Thus suppose initially there are n_0 alleles of type a in the nuclei of the cells. Let π_e be the probability that a particular allele is extracted into the solution. Further, let π_a denote the probability that a particular allele in the aliquot is put into the PCR

apparatus. Then, with alleles selected independently, the total number N of alleles of type a in the apparatus is binomially distributed as

$$N \sim \text{Bin}(n_0, \pi_e \pi_a).$$

The allele will not be selected and hence drop out when $N = 0$, which happens with probability $P(N = 0) = (1 - \pi_e \pi_a)^{n_0}$. Gill *et al.* [8] estimated $\pi_e = 0.6$ and $\pi_a = 20/66$ for their laboratory procedures, giving $\pi_e \pi_a = 0.182$, and hence a dropout of $0.818^{n_0} \approx \exp(-0.2 n_0)$.

Our model for pre-PCR dropout assumes for each marker that all the copies of the maternal allele from a person present in the aliquot are either all selected or all not selected for amplification, and similarly for the paternal allele. This *all-or-nothing* selection of the maternal contribution of an individual of a given marker is assumed to be independent of the selection or otherwise of the individual's paternal contribution from the marker, and of the selection or otherwise of any of the individual's other marker alleles; more precisely, this independence is conditional on the total amount of DNA from the individual in the mixture. Our model can thus be viewed as a crude approximation to the Gill *et al.* [8] model, which models the partial selection of alleles in the aliquot. The remaining variation in the number of amplified alleles is modelled by the gamma distribution of peak heights. Note, however, the difference that we are modelling allelic selection at an individual contributor level, whereas in the Gill *et al.* [8] model the selection is from

the combined contributions from all of the mixture contributors.

To model this allelic dropout in our network, we introduce nodes n_{ia}^{amp} with values in $\{0, 1, 2\}$. The various values may occur in the following manner. A value of $n_{ia}^{amp} = 0$ arises in two ways: (i) neither the maternal nor paternal allele is of type a; (ii) one or both of the maternal and paternal allele is of type a but is not selected for amplification. A value of $n_{ia}^{amp} = 1$ indicates that alleles of type a from person i are selected for amplification. This can happen in two ways: (i) either the person is heterozygotic with one allele of type a, in which case all of the a allele contribution is amplified; (ii) the person is homozygote (a, a), in which case only half is amplified. In the latter case, total dropout does not occur, but the peak height associated with the a allele is lower than would be expected given the person's genotype and his/her relative contribution of DNA to the mixture. A value of $n_{ia}^{amp} = 2$ arises if both the maternal and paternal alleles are of type a and they are all selected for amplification.

The network fragment modelling our dropout process is illustrated in Figure 2, which for simplicity of display assumes that only two allelic types, a and b, are seen. The nodes n_{ia} and n_{ib} are the same as occur in Figure 1, and count up the number of (maternal or paternal) alleles of type a and b for person i. Each n_{ia} takes values in $\{0, 1, 2\}$, with the conditional probability table $P(n_{ia} \,|\, p_i gt)$ having entries 0 or 1. Our dropout nodes n_{ia}^{amp} depend on the n_{ia} and the amount of DNA. Now as shown above, from Gill *et al.* [8] the sampling dropout probability has the form $\exp(-0.2n_0)$ in which

n_0 is proportional to the amount of DNA. This exponential dependence on amount is used in our conditional probability table $P(n_{ia}^{amp}|n_{ia}, \theta_i)$ shown in Table 1, which introduces a new parameter λ. Note the binomial distribution in the final column, modelling the independent dropout of the maternal and paternal alleles for profiles homozygote in allele a.

Insert Figure 2 about here.

Insert Table 1 about here.

Since $\sum_a n_{ia} = 2$ for every contributor, this implies that $\sum_i \sum_a \theta_i n_{ia} = 2$. This value of 2 was used as a common fixed normalization to the means of the relative weight in our earlier model, illustrated in Figure 1, but is not appropriate when taking dropout into account. Instead, with the potential for dropout, we have that $n_{tot} = \sum_i \sum_a \theta_i n_{ia}^{amp} \leq 2$. The node n_{tot} in Figure 2 stores this sum, conditional on the value of θ and the n_{ia} nodes, and is used for normalization of nodes μ_a at the bottom of the figure as

$$\mu_a = \frac{\sum_i \theta_i n_{ia}^{amp}}{\sum_i \sum_a \theta_i n_{ia}^{amp}} = \frac{\sum_i \theta_i n_{ia}^{amp}}{n_{tot}}$$

where we define $0/0 \equiv 0$. The likelihood factors (1) are then applied to the modified mean values.

There is a second source of dropout, in which some alleles are sampled but their amplified number is below the threshold detection level to register a distinct peak on the PCR output. Our extended Bayesian network does not model this source of dropout.

3.2 Silent alleles

A *null* or *silent* allele is one that is not recorded by the equipment used. When this can happen, what appears to be a homozygous genotype at some marker may not be so: an alternative explanation is that we are seeing just one band of a heterozygous genotype, the other band being missed. This phenomenon will clearly affect the evidential interpretation of certain patterns of DNA mixture profiles. Several papers in the literature have dealt with genetic aspects of this, see for example [9]. A possible explanation for a silent or null allele is sporadic failure of the apparatus to record the correct allele value or primer binding site mutations. Accounting for the possibility that a silent allele is present can easily be accommodated by including an additional allele in each marker which never gets amplified, corresponding to dropout with a probability of 100%. This in turn potentially affects the normalization $n_{tot} = \sum_i \sum_a \theta_i n_{ia}^{amp} \leq 2$ of the mean nodes as in the case of dropout and can be handled in exactly the same way.

3.3 Stutter

Following [10] and [11], stutter bands are understood to be allelic in origin and arise from slippage of the Taq polymerase enzyme. Only a single stutter band is typically observed and is four bases shorter than the associated "true" peak allele band, i.e., stutters are one repeat unit (allele value) less than the associated peak. In amplifying normal amounts of DNA, stutter tends to be

less than 15% of the size of the associated allelic peak [1], but if the amount of DNA is small, as in Low Template Analysis, this figure can be larger, and in very rare cases a stutter peak can be greater than the peak of the allele from which it arose.

In DNA mixtures a stutter peak could be indistinguishable from the minor contributor's allele peak or could be masked by a "true" allele peak thus leading to a higher allele peak size.

The proportion of stutter band increases with the length of the allele so that the longer allele in heterozygous sample has a higher percentage of stutter than the shorter allele [12]. For simplicity, we consider the prior distribution of stutter percentage to be constant across alleles and markers.

Insert Figure 3 about here.

Figure 3 shows a fragment of the Bayesian network that represents our stutter model. The layer of mean nodes μ_a at the top are augmented by stutter nodes s_a. The stutter node s_a gives the proportion that is lost by stuttering of allele a and contributes to the peak for the allele one repeat value lower, increasing the mean μ_{a-1}.

Without stutter, the mean value at allele a is μ_a. With stutter, then two things can happen: (i) part of the DNA of allele type a is amplified as stutter, so this decreases the effective mean of the measured peak weight associated with a and increases that associated with $a-1$; (ii) the allele $a+1$ can also stutter and increase the mean associated with a. Thus the effective mean μ_a^*

depends on the stutter losses s_a and s_{a+1} and the values of μ_a, μ_{a+1} as

$$\mu_a^* = (1 - s_a)\mu_a + s_{a+1}\mu_{a+1}.$$

Having found the μ_a^* values, the likelihood factors in (1) should now refer to μ^* rather than μ.

4 Application to a case

We shall apply our model to the challenging example presented by Gill *et al.* [2], who describe the case background as follows.

"An incident had occurred in a public house where the deceased had spent the evening with some friends. There was an altercation in the car park between the deceased (K_1) and several others resulting in the death of the victim. The alleged offenders then left the scene and went to another public house where they were seen to go into the lavatory to clean themselves."

Two known individuals, K_2 and K_3, alleged to be present at the time of the offence, were typed and their profiles together with that of the victim K_1 are shown in Table 2: all were males.

Insert Table 2 about here.

Two blood stains, called $MC18$ and $MC15$ were found at the public house lavatory and were typed using the SGM plus system. The results of

the typing are show in Table 3 and Table 4. Both blood stains indicated that they were DNA mixtures of at least three individuals.

In the following we shall compare various scenarios to the specific baseline hypothesis H_b that the contributors to the mixture are exactly the individuals K_1, K_2 and K_3. One alternative scenario is that the stains originate from three unknown individuals U_1, U_2, and U_3, but there are several other possibilities. We also discuss the question of whether or not the individual K_3 has contributed to the traces, but prior distributions for various scenarios must be specified for this to be answered.

Under the base hypothesis H_b, ten of the allele peaks in Table 3 would need to be interpreted as stutter alleles, for example allele 22 in marker D2, as none of the three individuals possess this allele. Also, if there are at most three contributors there must be stutter peaks in markers D8 and FGA, as seven alleles are observed. Similarly, for mixture $MC15$ in Table 4 seven of the peaks would have to be stutter peaks. In addition the following alleles, assumed to belong to K_2, would have to have dropped out: 16.2(D19), 22(FGA) and 9(TH0). There are also other alleles in $MC18$ that are not present in $MC15$.

Insert Table 3 about here.

We shall analyse these stains in turn, both individually beginning with $MC18$, and in combination. In all the analyses we use the following parameters. The vector of contributor fractions $\theta = (\theta_1, \theta_2, \theta_3)$ is uniform on the

set of positive coordinates which add to 1, discretized with intervals of size 0.1. To get parameters for the dropout probabilities we argue as follows. For $n_0 = 20$ molecules the estimates in [8] give a probability that none are selected of $(1 - 0.182)^{20} = 0.0179$. We then let $\lambda\gamma \approx -\log_e(0.0179) \approx 4$. The frequency of a silent allele has been set to 0.005 in all markers. The stutter nodes are crudely modelled with three states: (No stutter, 5%, 10%) having prior probability distribution $(0.98, 0.01, 0.01)$. We used $\sigma^2 = 0.03$ for the variance factor, and the allele frequencies are based on the Caucasian population in Appendix II of [12]. For each single mixture and the combined mixture analyses, we considered the eight different possible scenarios involving exactly three contributors. For each of these, the ratio of the likelihood of the base hypothesis H_b to that of the scenario was evaluated.

4.1 Analysis of $MC18$

This would appear to be the simplest of the two blood samples to analyse since, although stutter peaks must be present under H_b, there is no overt dropout. However, in our analyses we assumed the potential presence of all three artifacts in our PES model: stutter, dropout and silent alleles. The second column of Table 5 shows the likelihood ratio of the base hypothesis H_b to the other seven alternative scenarios involving exactly three people. From this table we see that the most likely scenarios, by a wide margin, are those that assume both the victim K_1 and K_3 contributed to the mixture whereas it is less definite whether K_2 has contributed to the mixture. The

most likely of the scenarios not involving K_3 is that the mixture consists of contributions from K_1 and two unknown persons. The likelihood ratio in favour of the base hypothesis H_b when compared to this scenario is equal to 3.74×10^8. Alternatively, if we assume a uniform prior over the scenarios so that it is assumed every contributor is independently either in or out of the mixture with equal probability, the posterior probability that K_3 is *not* in the mixture is found to be 1.89×10^{-9}, rendering it extremely unlikely that K_3 did not contribute to the trace.

4.2 Analysis of $MC15$

This would appear to be the more challenging of the two stains as both dropout and stutter must be involved under H_b. We analysed this mixture for the same scenarios as the $MC18$ mixture. The likelihood ratio in favour of the base hypothesis against each of the seven alternative scenarios are shown in column three of Table 5. The most likely scenarios by a wide margin are those that assume both the victim K_1 and K_3 contributed to the mixture, the same conclusion as reached for $MC18$. The overall most likely scenario is that the mixture consists of contributions from the victim K_1, K_3, and an unknown person. We note that this scenario is about twice as likely as the original base hypothesis H_b. The likelihood ratio of the base hypothesis to the scenario $K_1U_1U_2$ being the contributors is 1.14×10^5. Using again a uniform prior over the hypotheses as in § 4.1, the posterior probability that K_3 is *not* in the mixture is 1.24×10^{-8}, yielding strong evidence that K_3 is

in the mixture.

4.3 Combined analysis of $MC15$ and $MC18$

Cowell *et al.* [13] showed that combining peak information from two mixtures having some contributors in common can strongly enhance the mixture analysis. Here we apply this idea to a combined analysis of the two mixtures $MC15$ and $MC18$. Figure 4 illustrates the combined analysis of a pair of three-person mixtures. Each mixture has its own vector, θ and ϕ, of pre-amplification proportion of DNA from the three contributors. We have initially based the analysis on the assumption that the two traces have the same three contributors. The corresponding likelihood ratios in favour of the base hypothesis versus the other seven scenarios are displayed in the final column of Table 5. Assuming once more a uniform prior over the hypotheses, the posterior probability that K_3 has not contributed to both mixtures is 1.88×10^{-10}, a smaller posterior probability that obtained from each mixture analysed separately.

Notice however that the likelihood ratio for H_b compared to the scenario of contributors $K_1U_1U_2$ is now 1.40×10^3 which is a lot smaller than when each mixture was analysed separately. This weakening of the evidence in favour of H_b when combining the traces reflects that it is highly unlikely that K_2 has contributed to both traces. Indeed the likelihood ratio in favour of the scenario that $K_1K_3U_1$ were the contributors against H_b, is $1/2.62 \times 10^{-7} = 3.82 \times 10^6$, yielding the base hypothesis H_b quite unlikely. From the values

in the final column of Table 5 (which give the likelihood ratio *in favour* of the base hypothesis H_b) the scenario $K_1K_3U_1$ is seen to be the most likely out of the eight because it has the *smallest* value.

Insert Figure 4 about here.

It is worth emphasizing that our simultaneous analysis of a pair of mixtures does not require that an allele present in one mixture is also present in the other, even though we assume that the contributors to the two traces are the same. This is in contrast to the interpretation of duplicated STR analyses of low copy number (LCN) STR samples using the conventional consensus approach [9].

Insert Table 5 about here.

4.4 Other scenarios

In the combined analyses of the previous section, we assumed that each mixture had the same three contributors. This is not an essential requirement of our PES, we merely imposed this to limit the number of scenarios under consideration to a manageable number. Table 6 shows some further combined analyses, in which the contributors to the mixtures were not all assumed to be identical.

Insert Table 6 about here.

The likelihoods for some of these combined mixture analyses may be found from the likelihoods of individual mixture analyses by exploiting conditional independence properties of the networks. This applies for example to the first combination of Table 6, because the peak areas in the two mixtures are independent given K_1 and K_3, and both K_1 and K_3 are known. Similarly, the likelihood for the second combined analysis can be found from the individual mixture analyses, because the mixtures are independent given the profiles of the three known persons. The final combination is really two independent mixture analyses, and may be obtained from multiplying the values in columns 2 and 3 of the final row of Table 5. (Compare this to non-independent analysis value in the last row, column 4, of Table 5.) Only for the third combination in Table 6, having a partial overlap of two common unknown contributors to each mixture, is it not possible to find the likelihood from the likelihoods of the single mixture analyses.

We also note that the only likely scenarios are those involving the individual K_3 as a contributor to both mixtures. The two first combinations in Table 6 are both more likely than H_b whereas the most likely scenario, as seen from Table 5, still is that K_1, K_3, and an unknown individual has contributed to both traces, the likelihood ratio in favour of the scenario being 1.09×10^4 compared to the most likely scenario in the second row of in Table 6 with the two unknown contributors being different.

4.5 Identification of artifacts

Recall that our PES does not impose at the outset that a particular peak is or is not a stutter peak, or that a particular allele has dropped out, or that a silent allele is present. Instead, it is possible to query the PES, after entering the peak area information and profile information on known persons, by examining the posterior marginal distributions of the variables making up the network. It thus becomes possible to gauge the occurrence or otherwise of the various artifacts. We illustrate this by some examples.

If we assume that mixture $MC18$ is made of DNA from the three known persons, then on comparing the profiles in Table 2 to the peaks of mixture $MC18$ in Table 3, we see that ten of the peaks have to be stutter peaks. These are 22 (D2); 14,16(D3); 12,15 (D8); 13,15 (D18); 12 (D19); 25 (FGA); and 17 (VWA). These peaks would arise from the allele one repeat number higher stuttering. Table 7 shows the posterior distribution over a selection of these the associated stutter nodes. Note the zero probability on 0% stutter for the alleles that must stutter under the hypothesis H_b. In contrast allele 29 of marker D21 does not have to be explained as a stutter peak under H_b, because K_2 has genotype $(29, 30)$. However the peak height of allele 30 is about 8.5 times that of allele 29, so it could be suspected of giving rise to 29 as a stutter peak. Table 7 also shows the posterior probability for allele 23 (D2) to have stuttered in either mixture from a simultaneous analysis of both traces. In the mixture $MC15$, the posterior probability of allele 23 not stuttering is 1, which is consistent because there is no allele 22 (D2) peak in

$MC15$.

Insert Table 7 about here.

In the single mixture analysis of $MC18$, the posterior probability on the proportion θ showed that K_2 contributed the least amount of DNA, around 15% of the trace. This could make K_2 more susceptible as a source of dropout than the other contributors, although there is no overt allelic dropout in the mixture if the three known persons contributed to the mixture. Examining the posterior probabilities for marker D16 shown in Table 8 indicates that covert dropout most probably occurred. It shows that for individual K_1 both maternal and paternal alleles were amplified with probability 0.844, but there is approximately a 16% chance that one of two alleles dropped out. Similarly for K_3 it is most likely that both alleles were amplified. However, for K_2 there is a posterior probability of 0.342 that his allele 12 was not amplified: indeed from the table the probability is only 0.316 that both of K_2's alleles were amplified, and thus a probability of 0.684 that one or both of the alleles dropped out. Most of the other markers show similar covert dropout for K_2, and to a lesser extent for K_1 and K_3. For example for the marker TH0, the posterior probability that one of K_2's 9 alleles dropped out is 0.91, with a probability of 0.09 that neither dropped out. In contrast, the posterior probability that one of K_1's alleles dropped out is less than 10^{-7}.

Similar results regarding stutter and dropout were obtained for the mixture $MC15$, on the assumption that K_1, K_2 and K_3 were the contributors,

with the additional result that overt allelic dropouts are picked out with posterior probabilities of unity. In a joint mixture analysis, a probability of 0.486 was found that one of K_2's 24 (D2) alleles was not amplified from mixture $MC15$, and 0.380 that both were not amplified. For mixture $MC18$ the corresponding probabilities are 0.516 and 0.313.

Insert Table 8 about here.

These few examples are illustrative of the information about artifacts that may be obtained from examining the posterior probabilities on nodes in the network.

4.6 A separation analysis

Our analysis of § 4.4 suggests it is much more likely that both mixtures consist of contributions from K_1, K_3 and an unknown individual U. Under this scenario, in contrast to H_b, it is no longer definite that the $MC18$ peak of allele 22 (D2) arises as a stutter peak from allele 23; it could be that the unknown person has one or both alleles. Using methods described in [6], it is possible to use the same network for the *separation* of DNA profiles, and so give an estimate of the most likely DNA profile of the third contributor conditional on the peak height values and the DNA profiles of K_1 and K_3. Such a simultaneous analysis of both mixtures yields the following posterior probability distribution that allele 23 (D2) in $MC18$ stutters: 0.565 (at None), 0.176 (at 5%), and 0.259 (at 10%). This should be contrasted with the values

in Table 7 for H_b. The interpretation is that it is more probable that allele 22 is in the genotype of the unknown individual, rather than arising as a stutter peak.

Tables 9 and 10 show the predictions for U's genotype across all markers from separation analyses of the two mixtures individually, and Table 11 shows the predictions of a simultaneous separation analysis of both mixtures. In all analyses the unknown individual is predicted, with a posterior probability of more than 0.97, to be the minor contributor, contributing 10% of the total DNA to the each mixture. For K_1 the predicted contribution is 70%, and for K_3 the predicted contribution is 20%, in both mixtures.

We consider first the individual mixture analyses. In column 2 of Tables 9 and 10 the most probable selections of alleles (gt^{amp}) from the genotype (gt) of the unknown person that are amplified from each mixture, and in column 3 its marginal posterior probability. An allelic dropout is denoted by D. Note that the gt^{amp} value of person p_i is logically determined by the set of values of his/her n_{ia}^{amp} nodes in Figure 2; indeed for each person a node gt_i^{amp} could be added to Figure 2, with gt_i^{amp} a common child of the n_{ia}^{amp} nodes of person p_i.

The mixture $MC15$ appears to have significant dropout from the unknown person, with complete dropout on markers D2, D16 and TH0, and partial dropout on markers D3, D19, D21 and FGA. Less dropout is indicated for $MC18$, with only the marker D16 suggesting as most probable the total dropout of both maternal and paternal alleles. Such a high level of

dropout in each mixture is consistent with the low amount of DNA that the unknown contributor to each mixture is predicted to have contributed. There is agreement on some markers on the most probable genotype of the unknown person (column 4 of both tables), but also some disagreements. Only for marker VWA do both analyses predict the same genotype $(16, 17)$ with high probability.

Insert Table 9 about here.

Insert Table 10 about here.

The prediction of a $(17, x)$ genotype on marker D2 for $MC15$ may appear puzzling, given that it is most probable that both alleles dropped out. Note though that the associated probability of 0.191 is quite low. (Here the x designation in the predicted genotype for marker D2 indicates an allele of unspecified type *other* than of any peaks appearing in the mixture.) The explanation is found by looking beyond the most probable values of gt^{amp} and gt; Table 12 shows the six most probable values of these quantities for marker D2. We see that the second most probable genotype is (x, x) with a probability only slightly less than for $(17, x)$. It is also worth remarking that allele 17 has the highest population frequency of 0.182. A similar story applies to marker D16 in both mixtures, in which it is most probable that neither of the alleles is amplified in either mixture. Nevertheless, the $(11, 12)$ genotype is predicted by both mixtures separately. The allele frequencies

of 11 and 12 are 0.321 and 0.326 respectively, the two highest amongst the alleles of D16.

Insert Table 11 about here.

Insert Table 12 about here.

We now turn to the simultaneous mixture analysis. For both mixtures the predicted proportion of DNA contributed by the unknown person is 10%, with probabilities 0.99 ($MC15$) and 0.98 ($MC18$). We see that the pattern of most probable allelic selection or dropout exhibited in Table 11 is the same on each marker as for the individual mixture analyses in Table 9 and Table 10. However the posterior probabilities are markedly higher across all markers when compared against the individual mixture analyses. In addition, for five of the markers D3, D18, D19, FGA and VWA, the probabilities for the genotype prediction are greater than 0.9. The genotypes on the remaining markers, for which dropout is quite likely, have lower probabilities. However, some useful partial information is still available from the marginal distribution over the genotype (not tabulated here). For D2, the posterior probability that one of the alleles is a 22 is 0.586. For D16, allele 11 has a probability of 0.754 of being in the genotype; for D21, allele 29 has a probability of 0.998; and for TH0, allele 9 has a probability of almost 1 of being in the genotype. For marker D8, for which the most probable genotype $(12, 14)$ has a probability of 0.630, the second most probable genotype is $(14, 15)$ with a probability of 0.355; taken together these two genotypes account for 98.5%

of the posterior probability on the genotypic possibilities for this marker. The usefulness of such partial information for investigative purposes in a database search is clearly apparent.

5 Discussion

We have presented an extension of the model in [3] which incorporates the possible presence of silent alleles, stutter peaks, and dropout. We have applied this model to the analysis of two complex mixtures found at a crime scene and calculated likelihood ratios necessary for an evidential analysis of the traces, both individually as well as in combination. The networks also give the posterior probabilities for the specific artifacts to have occurred at particular alleles.

For illustrative purposes we have implemented a naïve stutter model with a crude discretization and a simple probability distribution. The analysis can be very sensitive to the values used in the stutter model. If the stutter probability is increased from 0.01 to 0.05, for example, the scenario $K_1K_2K_3$ becomes more likely than K_1K_3U in each single trace analysis as well as in the combined analysis. This reflects that the additional peaks, representing a very small amount of DNA, then most likely are the result stutter rather than additional alleles. The fact that K_3 has contributed to both traces seems robust varying the stutter probability within reasonable values.

For the model to be fully reliable in real casework it should be more

realistic in terms of both the level of discretization and the probability distribution over it. Any analysis should also be supplemented with a study of sensitivity and robustness by varying the parameters involved, for example along the lines of [15] who analyse sensitivity to distributional assumptions about founder genes.

We have ignored that the amount of stuttering tends to increase as the total amount of DNA decreases, and is also marker dependent. Similarly we have assumed the probability of silent alleles to be common for all loci, but there is a known dependence on marker and allele [12, 14]. These and other dependences could be incorporated into the model, for example can the prevalence of silent alleles be made marker dependent; in our exposition we have for simplicity assumed a common prevalence across markers.

An important concern is the computational complexity of our Bayesian networks. For analysing two person mixtures the computations are quite fast, a matter of a few seconds. Increasing the number of contributors to three people in a single mixture significantly increases the computation and memory requirements: typically between 2-3Gb of memory and around 1-2 hours of computation time was required to find each likelihood in Table 5. Unfortunately, we were unable to complete any analyses involving four people with the dropout model due to hardware constraints. A simple measure of the computational complexity involved is to look at the total size of the state space of tables in the junction tree for the Bayesian network [4]. For two-person two-trace networks, state spaces are typically of the order of 10^6

in size. For the three-person two-mixture networks examined in Table 5, the state space sizes were of the order of 10^9. For a four person two-trace network, the state space grows to 10^{11} or more. We believe that some of these difficulties will be addressed in future work.

Finally we emphasize that threshold generated dropout should eventually be incorporated as should issues concerning trace contamination.

References

[1] P. Gill, C.H. Brenner, J.S. Buckleton, A. Carracedo, M. Krawczak, W.R. Mayr, N. Morling, M. Prinz, P.M. Schneider, and B.S. Weir. DNA commission of the international society of forensic genetics: Recommendations on the interpretation of mixtures. *Forensic Science International*, 60:90–101, 2006.

[2] Peter Gill, James Curran, Cedric Neumann, Amanda Kirkham, Tim Clayton, Jonathan Whitaker, and Jim Lambert. Interpretation of complex DNA profiles using empirical models and a method to measure their robustness. *Forensic Science International:Genetics*, 2:91–103, 2008.

[3] Robert G. Cowell, Steffen Lilholt Lauritzen, and Julia Mortera. A gamma Bayesian network for DNA mixture analyses. *Bayesian Analysis*, 2:333–348, 2007.

[4] Robert G. Cowell, A. Philip Dawid, Steffen Lilholt Lauritzen, and David J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer, New York, 1999.

[5] Robert G. Cowell, Steffen L. Lauritzen, and Julia Mortera. MAIES: A tool for DNA mixture analysis. In R. Dechter and T. Richardson, editors, *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 90–97, San Francisco, 2006. Morgan Kaufmann Publishers.

[6] Robert G. Cowell, Steffen Lilholt Lauritzen, and Julia Mortera. Identification and separation of DNA mixtures using peak area information. *Forensic Science International*, 166:28–34, 2007.

[7] Robert G Cowell. Validation of an STR peak area model. *Forensic Science International: Genetics*, 3(3):193–199, 2009.

[8] Peter Gill, James Curran, and Keith Elliot. A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci. *Nucleic Acids Research*, 33(2):632–643, 2005.

[9] P. Gill, J. Whitaker, Christine Flaxman, Nick Brown, and J. Buckleton. An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Science International*, 112:17–40, 2000.

[10] P. Gill, R. Sparkes, R. Pinchin, T. Clayton, J. Whitaker, and J. Buckleton. Interpreting simple STR mixtures using allele peak areas. *Forensic Science International*, 91:41–53, 1998.

[11] P. S. Walsh, N.J. Fildes, and R. Reynolds. Sequence analysis and characterization of stutter products. *Nucleic Acid Research*, 24:2807–2812, 1996.

[12] J. M. Butler. *Forensic DNA typing.* Elsevier, USA, 2005.

[13] Robert G. Cowell, Steffen L. Lauritzen, and Julia Mortera. Probabilistic modelling for DNA mixture anlaysis. *Forensic Science International: Genetics Supplement Series*, 1:640–642, 2008.

[14] T. M. Clayton, S. M. Hill, L. A. Denton, S. K. Watson, and A. J. Urquhart. Primer binding site mutations affecting the typing of STR loci contained within the AMPF*l*STR® SGM Plus^{TM} kit. *Forensic Science International*, 139:255–259, 2004.

[15] Peter J. Green and Julia Mortera. Sensitivity of inferences in forensic genetics to assumptions about founding genes. *Annals of Applied Statistics*, 3(2):731–763, 2009.

Figure captions

Figure 1 A single marker network for the generation of peak areas from a mixture of DNA from two people having genotypes p_1gt and p_2gt, with three observed alleles a, b and c.

Figure 2 Illustration of how nodes relate in the dropout network fragment for a marker with two observed alleles, a and b, in a two person mixture.

Figure 3 Network fragment for modelling stutter. The mean μ_a is affected by stuttering in two ways: (i) a reduction due to part of type a alleles amplifying to type $a-1$; (ii) an increase due to part of type $a+1$ alleles amplifying to type a.

Figure 4 Schematic modelling of a pair of three-person mixture samples that simultaneously share DNA from the same three contributors, but in possibly different proportions θ and ϕ.

Table 1: The conditional probability table $P(n_{ia}^{amp}|n_{ia}, \theta_i)$ quantifying dropout with $\delta_i = \exp(-\lambda\gamma\theta_i)$.

	n_{ia}		
n_{ia}^{amp}	0	1	2
0	1	δ_i	δ_i^2
1	0	$1-\delta_i$	$2(1-\delta_i)\delta_i$
2	0	0	$(1-\delta_i)^2$

Table 2: Profiles of the victim K_1 and individuals K_2, and K_3 for the ten markers.

	K_1		K_2		K_3	
D2	23	24	24	24	16	17
D3	15	18	17	17	17	19
D8	13	16	13	14	10	11
D16	12	12	11	12	11	13
D18	14	16	14	14	12	16
D19	13	14	15	16.2	14	15
D21	30	31	29	30	28	30
FGA	24	26	21	22	20	23
TH0	7	8	9	9	9.3	9.3
VWA	14	18	16	16	15	19

Table 3: Blood stain $MC18$, showing alleles a, and peak heights h_a for ten markers.

a	h_a	a	h_a	a	h_a	a	h_a	a	h_a
D2		D8		D18		D21		TH0	
16	189	10	241	12	187	28	304	7	670
17	171	11	192	13	87	29	134	8	636
22	55	12	127	14	997	30	1146	9	99
23	638	13	1092	15	80	31	734	9.3	348
24	673	14	127	16	744	FGA			
D3		15	58			20	99	VWA	
14	50	16	808	D19		21	49	14	876
15	715			12	57	22	76	15	249
16	67	D16		13	775	23	145	16	274
17	479	11	534	14	818	24	412	17	97
18	638	12	1786	15	159	25	39	18	967
19	136	13	265	16.2	76	26	349	19	251

Table 4: Blood stain $MC15$, showing alleles a, and peak heights h_a for ten markers.

a	h_a	a	h_a	a	h_a	a	h_a	a	h_a
D2		D8		D18		D21		TH0	
16	64	10	152	12	99	28	120	7	727
17	96	11	140	13	61	29	89	8	625
23	507	12	76	14	707	30	1010	9.3	165
24	534	13	929	15	107	31	783		
		14	58	16	930				
		15	84					VWA	
D3		16	901			FGA		14	1036
14	79			D19		20	90	15	98
15	993	D16		12	53	21	52	16	163
17	286	11	256	13	546	23	103	17	79
18	689	12	1724	14	655	24	556	18	746
19	135	13	109	15	98	26	392	19	85

Table 5: Likelihood ratios in favour of the base hypothesis H_b that K_1, K_2 and K_3 are the mixture contributors, against alternative scenarios H_a involving contributions from exactly three persons. The ratios are given both for the two single-mixture analyses, and for the simultaneous analysis. The U_i $(i = 1, 2, 3)$ refer to contributors whose profiles are not known.

H_a	$MC18$	$MC15$	$MC15$ and $MC18$
$K_1K_2K_3$	1	1	1
$K_1K_2U_1$	6.07×10^{9}	4.88×10^{9}	7.29×10^{8}
$K_1K_3U_1$	2.00	1.41×10^{-3}	2.62×10^{-7}
$K_2K_3U_1$	6.59×10^{13}	8.79×10^{13}	9.08×10^{13}
$K_1U_1U_2$	3.74×10^{8}	1.14×10^{5}	1.40×10^{3}
$K_2U_1U_2$	1.43×10^{23}	1.94×10^{23}	3.28×10^{22}
$K_3U_1U_2$	6.84×10^{13}	6.22×10^{10}	1.19×10^{7}
$U_1U_2U_3$	4.90×10^{21}	3.06×10^{18}	4.19×10^{16}

Table 6: Likelihood ratios in favour of the base hypothesis H_b that K_1, K_2 and K_3 are the mixture contributors, against selected alternative scenarios H_a that do not have the same contributors to the two traces. The ratios are based on simultaneous analyses of both traces. The U_i $(i = 1, \ldots, 6)$ refer to contributors whose profiles are not known.

$MC18$	$MC15$	$H_b : H_a$
$K_1K_3U_2$	$K_1K_3U_1$	2.86×10^{-3}
$K_1K_2K_3$	$K_1K_3U_1$	1.42×10^{-3}
$U_1U_2U_4$	$U_1U_2U_3$	3.93×10^{21}
$U_1U_2U_3$	$U_4U_5U_6$	1.51×10^{40}

Table 7: Posterior probabilities for alleles to have stuttered, for a small selection of alleles for the single trace analysis of $MC18$, and for the simultaneous analysis of both traces, assuming in each case that the DNA in the mixtures originates from the three known persons K_1,K_2 and K_3.

	Only $MC18$		
	Stutter amount		
Allele (Marker)	0%	5%	10%
23(D2)	0	0.400	0.600
15(D3)	0	0.431	0.569
17(D3)	0	0.356	0.645
13(D19)	0	0.428	0.572
26(FGA)	0	0.367	0.633
18(VWA)	0	0.382	0.618
30(D21)	0.969	0.014	0.017
	Simultaneous analysis		
$MC18$ 23(D2)	0	0.400	0.600
$MC15$ 23(D2)	1	0	0

Table 8: Posterior probabilities of genotypic contribution to the amplification of mixture $MC18$ for marker D16, assuming K_1, K_2 and K_3 are the contributors. D indicates an allele (paternal or maternal) that has dropped out.

K_1		K_2		K_3	
$12, 12$	0.844	$11, D$	0.342	$11, 13$	0.832
$12, D$	0.156	$11, 12$	0.316	$13, D$	0.168
D, D	3.6×10^{-7}	D, D	0.207		
		$12, D$	0.136		

Table 9: Separation analysis of mixture $MC15$, to predict the most probable profile of the unknown contributor U, assuming that the mixture is made up of DNA from K_1, K_3 and U. For each marker, the second column (gt^{amp}) shows the most likely combination of maternal and paternal alleles that were amplified, with an allelic dropout denoted by D. The posterior probability for the amplified genotype is shown in the third column. The fourth and fifth columns show the most probable genotype (gt) and its posterior probability.

	Alleles amplified from U		Predicted profile of U	
Marker	gt^{amp}	Posterior	gt	Posterior
D2	D, D	0.580	$17, x$	0.191
D3	$14, D$	0.600	$14, 15$	0.248
D8	$12, 14$	0.412	$12, 14$	0.426
D16	D, D	0.457	$11, 12$	0.236
D18	$13, 15$	0.636	$13, 15$	0.673
D19	$12, D$	0.551	$12, 14$	0.322
D21	$29, D$	0.681	$29, 30$	0.301
FGA	$21, D$	0.782	$21, 24$	0.174
TH0	D, D	0.592	$7, 9.3$	0.178
VWA	$16, 17$	0.879	$16, 17$	0.902

Table 10: Separation analysis of mixture $MC18$, to predict the most probable profile of the unknown contributor U, assuming that the mixture is made up of DNA from K_1, K_3 and U. The format of the table is the same as Table 9.

	Alleles amplified from U		Predicted profile of U	
Marker	gt^{amp}	Posterior	gt	Posterior
D2	$22, D$	0.515	$22, x$	0.268
D3	$14, 16$	0.550	$14, 16$	0.581
D8	$12, 14$	0.621	$12, 14$	0.636
D16	D, D	0.421	$11, 12$	0.252
D18	$13, 15$	0.651	$13, 15$	0.689
D19	$12, 16.2$	0.601	$12, 16.2$	0.624
D21	$29, D$	0.704	$29, 30$	0.309
FGA	$21, 22$	0.873	$21, 22$	0.880
TH0	$9, D$	0.701	$9, 9.3$	0.462
VWA	$16, 17$	0.882	$16, 17$	0.906

Table 11: Separation profile analyses for the unknown contributor, from a simultaneous analysis of the two traces. The table shows the most probable combination of alleles (gt^{amp}) from the genotype amplified, together with the associated probability. An allele not amplified is denoted by D. Also shown is the predicted genotype (gt) and its posterior probability. See main text for more details.

Marker	$MC15$		$MC18$		Predicted profile	
	gt^{amp}	Posterior	gt^{amp}	Posterior	gt	Posterior
D2	D, D	0.634	$22, D$	0.432	$22, x$	0.170
D3	$14, D$	0.919	$14, 16$	0.854	$14, 16$	0.903
D8	$12, 14$	0.610	$12, 14$	0.615	$12, 14$	0.630
D16	D, D	0.432	D, D	0.397	$11, 12$	0.269
D18	$13, 15$	0.928	$13, 15$	0.926	$13, 15$	0.981
D19	$12, D$	0.941	$12, 16.2$	0.937	$12, 16.2$	0.973
D21	$29, D$	0.719	$29, D$	0.729	$29, 30$	0.318
FGA	$21, D$	0.992	$21, 22$	0.922	$21, 22$	0.929
TH0	D, D	0.748	$9, D$	0.683	$9, 9.3$	0.503
VWA	$16, 17$	0.970	$16, 17$	0.969	$16, 17$	0.995

Table 12: The six most probable combinations of alleles amplified (gt^{amp}), and the six most probable genotypes (gt), for the unknown contributor on marker D2 obtained from an analysis of $MC15$.

gt^{amp}	Posterior	gt	Posterior
D, D	0.580	$17, x$	0.191
$17, D$	0.143	x, x	0.169
$24, D$	0.097	$24, x$	0.129
$23, D$	0.079	$23, x$	0.116
$16, D$	0.049	$17, 24$	0.072
$17, 24$	0.011	$17, 23$	0.065

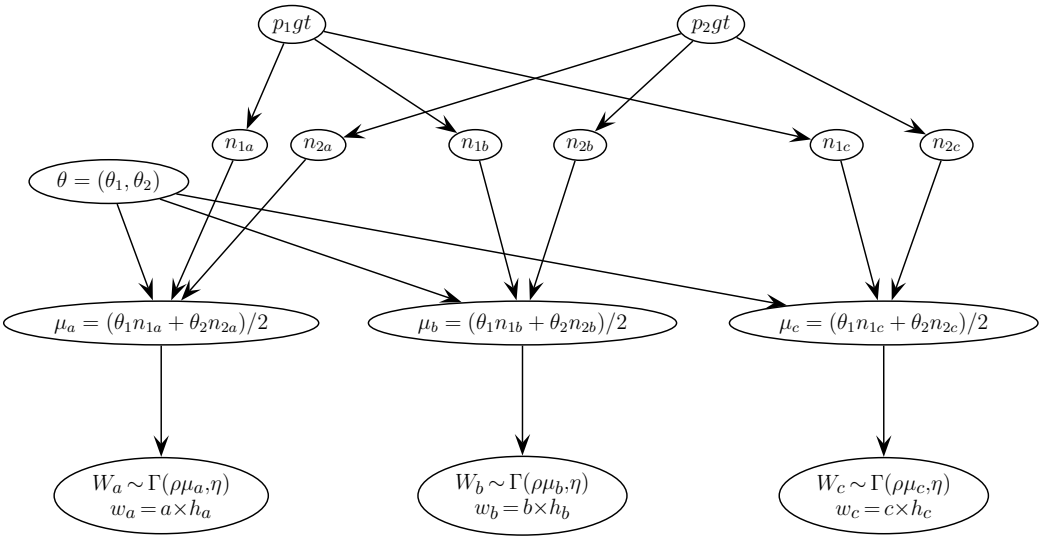

Figure 1: A single marker network for the generation of peak areas from a mixture of DNA from two people having genotypes p_1gt and p_2gt, with three observed alleles a, b and c.

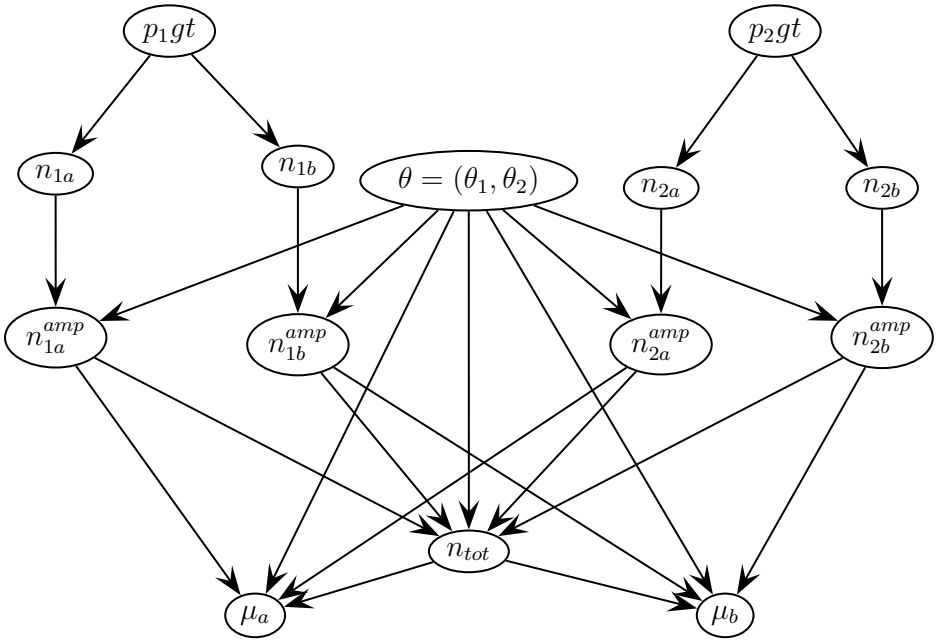

Figure 2: Illustration of how nodes relate in the dropout network fragment for a marker with two observed alleles, a and b, in a two person mixture.

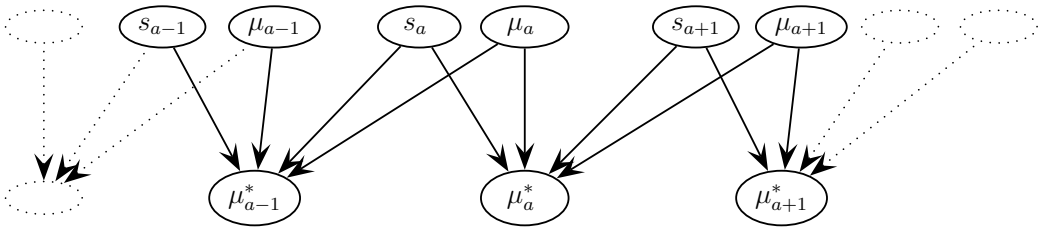

Figure 3: Network fragment for modelling stutter. The mean μ_a is affected by stuttering in two ways: (i) a reduction due to part of type a alleles amplifying to type $a-1$; (ii) an increase due to part of type $a+1$ alleles amplifying to type a.

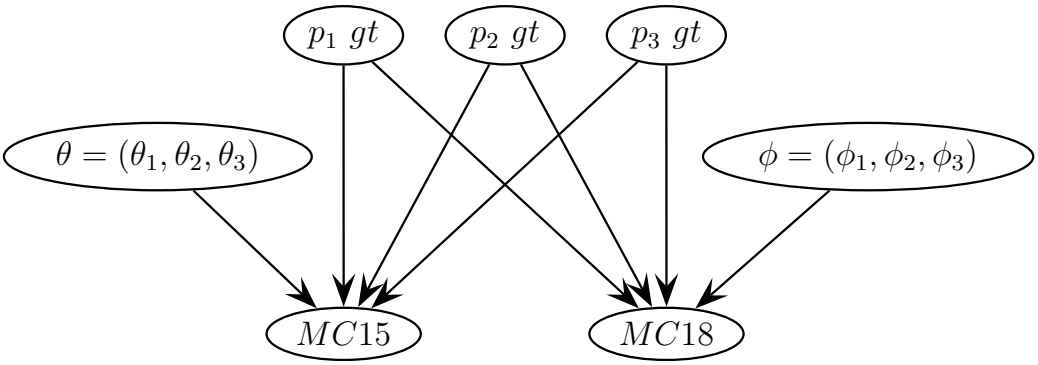

Figure 4: Schematic modelling of a pair of three-person mixture samples that simultaneously share DNA from the same three contributors, but in possibly different proportions θ and ϕ.