Current challenges in statistical DNA evidence evaluation
Cereda, G.

Citation

Cover Page

The handle http://hdl.handle.net/1887/45172 holds various files of this Leiden University dissertation.

Author: Cereda, G.
Title: Current challenges in statistical DNA evidence evaluation
Issue Date: 2017-01-12

Chapter 8

A solution for the rare type match problem when using the DIP-STR marker system

This chapter is based on: Cereda, G., Gill, R. D., and Taroni, F. "A solution for the rare type match problem when using the DIP-STR marker system". Submitted to *Forensic Science International: Genetics*.

Abstract

The rare type match problem is an evaluative challenging situation in which the analysis of a DNA profile reveals the presence of (at least) one allele which is not contained in the reference database. This situation is challenging because an estimate for the frequency of occurrence of the profile in a given population needs sophisticated evaluative procedures.

The rare type match problem is very common when the DIP-STR marker system, which has proven itself very useful for dealing with unbalanced DNA mixtures, is used, essentially due to the limited size of the available database. The object-oriented Bayesian network proposed in Cereda et al. (2014b) to assess the value of the evidence for general scenarios, was not designed to deal with this peculiar situation. In this paper, the model is extended and partially modified to be able to calculate the full Bayesian likelihood ratio in presence of any (observed and not yet observed) allele of a given profile. The method is based on the approach developed in Cereda (2016a) for Y-STR data. Alternative solutions, such as the plug-in approximation and an empirical Bayesian methodology are also proposed and compared with the results obtained with the full Bayesian approach.

8.1 Introduction

The most common task of a forensic scientist or statistician is to quantify the probative value of the observation of some scientific findings (e.g. DNA profiles, fragment of paint, fibres), under the hypotheses of interest for the court of justice. This is done through the quantification of the likelihood ratio. In case the hypotheses of interest deal with whether the recovered material has the same origin as some control material, it is important to be able to quantify the rarity of the corresponding characteristics. For instance, the evidence can be the correspondence between the DNA profile of a crime stain and of a suspect: the rarer the profile, the more probative is the scientific finding regarding propositions about the source. The rarity of the profile of interest is often used to assign the probability of the random occurrence of the given stain, and some available (and relevant) database is used to support the scientist's assignment.

The 'rare type match problem', also called 'the fundamental problem of forensic mathematics' (Brenner, 2010) is the situation in which the corresponding characteristic has not been observed in the relevant reference database for the case. One example is the DIP-STR marker system, a rather novel genotyping technique, proposed in Castella et al. (2013), which turned out to be very useful to analyse DNA mixtures if the proportion of the DNA quantities of the, say, two contributors is more extreme than 1:10. Due to limited size of available databases, rare DIP-STR profiles are often encountered.

A Bayesian framework for evaluating DIP-STR results was developed in Cereda et al. (2014b), using object-oriented Bayesian networks, with the aim of calculating the likelihood ratio for mixtures of two contributors, when the major contributor's genotype is known and the two competing hypotheses are 'the minor contributor is the suspect' (h_p) and 'the minor contributor is an unknown person, unrelated to the suspect' (h_d), also extended to cases where the suspect is missing.

This paper proposes a Bayesian solution for assigning the likelihood ratio for mixture results in presence of a rare type match, that is when at least one of the DIP-STR alleles of the contributors is not present in the reference database. This situation was not covered by Cereda et al. (2014b).

The Bayesian model adopted is based on a similar one proposed in Cereda (2016a). Several issues concerning Bayesian methodology, and notation, have been improved.

The paper is structured as follows. Section 8.2 discusses the use of the DIP-STR marker system for extremely unbalanced mixtures, while Section 8.3 describes the object-oriented Bayesian network that was built to evaluate DIP-STR profiling results in Cereda et al. (2014b). The chosen notation and the definition of what a full Bayesian approach to likelihood ratio assessment is, can be found in Sections 8.4 and 8.5, respectively. The model developed to evaluate results from mixtures of two contributors in presence of the rare type match problem (described in Section 8.6) is detailed in Sections 8.7 and 8.8. More detailed descriptions of the development of the full Bayesian likelihood ratio, which takes advantage of the Lemma introduced in Section 8.9, are confined to the Appendix. A discussion about the choice of the prior distribution for the parameters is also provided in Section 8.10, while conclusions can be found in Section 8.11.

8.2 DIP-STR marker system for extremely unbalanced mixtures

A DIP-STR marker is a compound marker made of a DIP (Deletion/Insertion polymorphism, Weber et al. (e.g., 2002)), and of a standard STR polymorphism. These two polymorphisms are chosen less than 500 bp apart, in order to be dependent on one another.

Standard methods for the analysis of DNA mixtures, such as STR markers (Butler, 2011), fail to detect the DNA of contributors whose DNA constitutes less than the 10% of the total DNA material (Clayton and Buckleton, 2005). On the other hand, as long as the minor contributor has at a specific locus at least one DIP allele different from the DIP alleles of the suspect, the DIP-STR marker system allows the selected amplification of its DIP-STR genotype, up to mixture proportions as extreme as 1:1000,

At each DIP-STR locus, the possible configurations are the following (summarized in Table 8.1).

- In the case where the major and minor contributors are DIP homozygous with different alleles (i.e., one is L-L and the other is S-S) both DIP-STR alleles of the minor can be detected. This is the best scenario the scientist can be faced with.

- If the major is DIP homozygous (for instance L-L) and the minor is DIP heterozygous, only one of the two DIP-STR alleles of the minor can be detected: the one with the other DIP allele (in the example the allele S).

- The worst situation is the one in which the major is DIP heterozygous, or both contributors are homozygous for the same DIP alleles: in these cases the DNA profile of the minor cannot be obtained.

DIP genotype of major/minor contributor	DIP-STR alleles observed in the trace	Information gained for the second contributor
Hom/Hom (different kind)	2 (if STR het)	Yes completely
	1 (if STR hom)	Yes
Hom/Het	1 (regardless STR)	Yes
Hom/Hom (same kind)	0 (regardless STR)	No
Het/Hom	0 (regardless STR)	No
Het/Het	0 (regardless STR)	No

Table 8.1: Informativeness of genotypic configurations. 'Hom' denotes homozygous for the DIP allele, and 'Het' heterozygous.

A first panel of 10 DIP-STR markers was presented in Castella et al. (2013). A second panel with 9 additional DIP-STR markers has recently been provided in Oldoni et al. (2015). When one analyses a mixed stain, each of the 19 available markers may present one of the three situations described above.

8.3 Bayesian network for evaluating DIP-STR profiling results from unbalanced DNA mixtures.
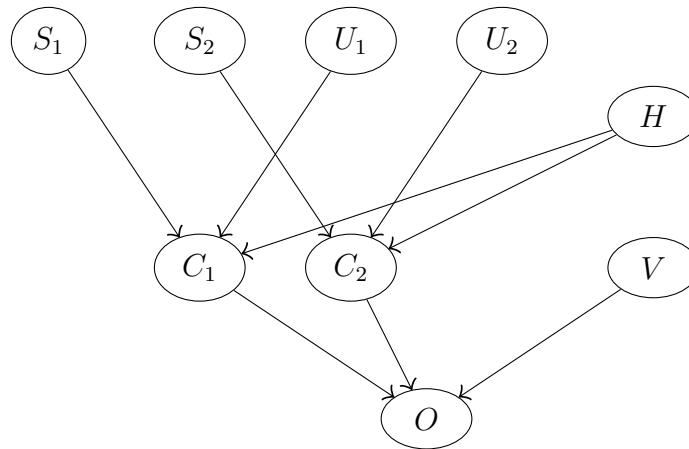

Figure 8.1: Bayesian network corresponding to the object-oriented Bayesian network of Cereda et al. (2014b). The meaning of the nodes is described in Section 8.3.

In Cereda et al. (2014b) a locus specific object-oriented Bayesian network (OOBN), designed to assist the evaluation of DIP-STR results obtained from mixtures with two contributors, is proposed. The network, reproducing the mechanism described in Section 8.2 and in Table 8.1, is proposed here in the form of a Bayesian network (see Figure 8.1). It is suitable for a situation in which the DIP-STR profile of a suspect (potential contributor to the mixture) is available. The two hypotheses of interest are 'the minor contributor is the suspect' (h_p) and 'the minor contributor is an unknown person, unrelated to the suspect' (h_d). The major contributor is often referred to as "the victim", taken as a known contributor, and his/her DIP-STR profile is generally available.

It is important to notice that the only information needed from the known major contributor regards his DIP alleles. Hence, the only node in the network which concerns the victim, V, has three possible states: *HomoL*, *HomoS* and *Hetero*. The node H represents the two hypotheses of interest defined above.

With the exception of node O, the remaining part of the network deals with the unknown minor contributor. Nodes S_1 and S_2 represent the two DIP-STR alleles of the suspect. Nodes U_1 and U_2 represent the two DIP-STR alleles of the alternative (unknown) contributor in a two-person mixture. Nodes C_1 and C_2 represent the DIP-STR alleles of the actual second contributor (for example, the suspect). Depending on the state of node H, the second contributor's allele can be a copy of S_1 and S_2 (under state h_p), or of U_1 and U_2 (under state h_d). The state of node O, which contains results obtained from the mixture, depends on the combination of V, C_1 and C_2 (according to Table 8.1).

The probability tables for the nodes are of different types. In the scenario considered, node V is observed, since the major contributor is known. As such, its probability table is not relevant for the final result because its state is fixed, thus it is filled with equal prior probabilities

for its three states. The same holds for node H, which is in turn instantiated to obtain the numerator and the denominator of the likelihood ratio.

Nodes C_1 and C_2 are deterministic given H, S_1, S_2, U_1, and U_2: if H is in state h_p, then C_1 and C_2 are copies of, respectively, S_1 and S_2, otherwise they are copies of U_1 and U_2. Also node O is deterministic, given nodes V, H, C_1 and C_2: its probability table is filled out with 0's and 1's (according to the conditions defined in Table 8.1). The states of nodes S_1, S_2, U_1, U_2, C_1, and C_2 are La, Lb, Lx, Sa, Sb, Sx. Notice that at each DIP-STR locus there may be more than six possible alleles: La, Lb, Sa, Sb are used to represent the two alleles that at most could be observed, while Sx and Lx represent all the other (not observed) alleles different from a and b. In Cereda et al. (2014b), this solution was preferred to having the entire list of DIP-STR alleles, in order to make the model simpler, and usable for different loci. The disadvantage is that, at each new case, the meaning of these symbols changes, and the probability tables have to be adapted accordingly. In this paper, we will develop a methodology to overcome this constraint.

The probability tables for nodes S_1, S_2, U_1, and U_2 should be filled with the allelic proportions corresponding to the alleles represented by names La, Lb, etc., in the population of interest. These allelic proportions are unknown, but we a have a database of DIP-STR alleles, which we can consider as a random sample from the population of interest. In Cereda et al. (2014b), a Dirichlet distribution with all parameters equal to one was used as prior for the DIP-STR allelic proportions, and the probability tables for nodes S_1, S_2, U_1, U_2 were filled out with the posterior means (conditional to the observation of the database). However, as discussed in Section 8.5, this approach suffers from some limitations and it can be improved and made more consistent with the Bayesian theory. Moreover, the number of possible distinct DIP-STR alleles was chosen by looking at those in the database. Thus, the model was not suitable to be used when new alleles (not previously detected) were observed. This paper aims at solving these problems.

8.4 Notation

Throughout the paper the following notation is chosen: random variables and their values are denoted, respectively, with uppercase and lowercase characters: x is a realization of X. Random vectors are denoted with bold characters: $\mathbf{x}$ is a realization of the random vector $\mathbf{X}$. Probability is denoted with $\Pr(\cdot)$, while density of a continuous random variable X is denoted alternatively by $p_X(x)$ or by $p(x)$ when the subscript is clear from the context. For a discrete random variable Y, the density notation $p_Y(y)$ and the discrete one $\Pr(Y = y)$ will be alternately used. Moreover, we will use shorthand notation like $p(y \mid x)$ to stand for the probability density of Y with respect to the conditional distribution of Y given $X = x$.

Given $k \geq 2$, and $\alpha = (\alpha_1, ..., \alpha_k)$ such that $\alpha_i > 0$,

$$\mathbf{X} \sim \text{Dir}^k(\alpha_1, ..., \alpha_k)$$

means that vector $\mathbf{x}$ follows a k-dimensional Dirichlet distribution (Press, 2009), whose den-

sity is

$$p(\mathbf{x}) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}.$$

In the appendix, we will denote with $\mathbf{z} = (\mathbf{x}, y)$ the vector $\mathbf{z}$ obtained by adding element y at the end of vector $\mathbf{x}$.

8.5 Full Bayesian approach

In the case of interest, for each analysed locus the forensic scientist or statistician is given the following input data: the victim's and the suspect's DIP-STR profile (denoted as E_v and E_s respectively), along with the DIP-STR alleles obtained from the mixture (E_m). This data has to be evaluated in the light of the hypotheses of interest (h_p and h_d) as defined in Section 8.1. The evaluation of such evidence heavily depends on the allelic proportions of the DIP-STR alleles of the trace and of the suspect, which are unknown. The vector $\boldsymbol{\theta}$, containing the population proportions of all the possible DIP-STR alleles at the considered locus, is the nuisance parameter of the model. A database (denoted here as D), consisting of a list of DIP-STR alleles from the population of interest is given to the statistician, in order to support him in the assessment of the uncertainty about $\boldsymbol{\theta}$. The data to evaluate are thus made of $E=(E_v,\ E_s,\ E_m)$ and D. This notation reflects the distinction described in Cereda (2016a) between 'evidence', data directly related to the crime, and 'background', data related only to the nuisance parameter of the model.

The full Bayesian approach consists of modelling all these variables, including $\boldsymbol{\theta}$, as random variables whose joint distribution Pr reflects prior belief of the expert.

The largely accepted method to evaluate the data in order to discriminate between the two hypotheses of interest, is the calculation of the *Bayes factor* (BF), in forensic context regularly called *likelihood ratio* (LR). It is defined as the ratio of the probabilities of observing the data under the two competing hypotheses:

$$\text{LR} = \frac{\Pr(E = e, D = d \mid H = h_p)}{\Pr(E = e, D = d \mid H = h_d)} = \frac{\Pr(E = e \mid D = d, H = h_p)}{\Pr(E = e \mid D = d, H = h_d)}, \tag{8.1}$$

where the last equality holds in virtue of the independence of database and hypotheses.

The nuisance parameter $\boldsymbol{\theta}$ has been integrated out according to its prior distribution. Notice indeed that $\boldsymbol{\theta}$ does not appear in (8.1).

In Cereda et al. (2014b), we used a different approach: Bayesian estimates of the allelic proportions were plugged into the probability tables for nodes S_1, S_2, U_1, and U_2. This is equivalent to using a likelihood ratio for a given $\boldsymbol{\theta}$, such as

$$\text{LR} = \frac{\Pr(E = e \mid \boldsymbol{\Theta} = \boldsymbol{\theta}, D = d, H = h_p)}{\Pr(E = e \mid \boldsymbol{\Theta} = \boldsymbol{\theta}, D = d, H = h_d)},$$

and to plug inside the estimates for $\boldsymbol{\theta}$.

The plug-in method can be seen as an approximation to the full Bayesian method (Cereda, 2016a). To obtain it, a Bayesian network is built which allows one to use an integrated full Bayesian approach, by introducing, among others, a node which represents the database D, and a node that represents the nuisance parameter $\boldsymbol{\theta}$. The full Bayesian approach is then compared to the plug-in method, to check the impact of the approximations.

8.6   Rare type match problem

When the findings to evaluate include a correspondence between the DNA profile of a particular piece of evidence (i.e., a trace of unknown origin) and a suspect's DNA profile, but at least one of the alleles of this profile are not present in the available database, it is difficult to assess the uncertainty over the population proportion of that allele. It is likely to be a rare allele (from which the term *rare type match problem*) but it is challenging to quantify how rare. This assessment is important for the quantification of the likelihood ratio: the rarer the matching profile, the larger is the likelihood ratio.

Using DIP-STR data, it is very likely to encounter the rare type match problem, because the available database size is still limited (Oldoni et al., 2015). The same happens when Y-chromosome (or mitochondrial) DNA profiles are used, since the set of possible Y-STR profiles is extremely large. As a consequence, most of the Y-STR haplotypes are not represented in the database. In Cereda (2016a,b,c) several (Bayesian and frequentist) solutions are proposed for the rare type match problem for Y-STR data. The object-oriented Bayesian network of Cereda et al. (2014b), here presented in Figure 8.1, cannot be used in the case of a rare type match problem: there, the number of different alleles at a given locus was considered as fixed, and equal to that observed in the database. This makes that model useless in cases where new DIP-STR alleles are observed.

As a solution, we will consider the number of different DIP-STR alleles present in the population as random, by introducing additional variables in the model, explained in detail in Section 8.7. This is based on one of the Bayesian methods proposed in Cereda (2016a).

8.7   A prior for θ

Let us denote with L-STR (or S-STR) the DIP-STR alleles which have the DIP part equal to L (or S). Assume that, at a specific locus, there are at most m theoretically possible L-STR alleles and m theoretically possible S-STR alleles. The random vector $\boldsymbol{\Theta} = (\Theta_1^L, ..., \Theta_m^L, \Theta_1^S, ..., \Theta_m^S)$ contains the population proportions of all the potential $2m$ DIP-STR alleles at that locus (for instance, alphabetically ordered).

Only k^L (k^S) of the m possible L-STR (S-STR) alleles are actually present in nature (or more specifically in the population of interest), but k^L and k^S are unknown. Which of the m L-STR alleles are those k^L and k^S is not known either.

The vector $\mathbf{t}^L$ contains the ordered positions (from 1 to m) of the k^L L-STR alleles present

in the population of interest. $\mathbf{t}^L$ is modelled through a random variable $\mathbf{T}^L$: each possible configuration $\mathbf{t}^L$ is assumed as equiprobable, hence it is chosen uniformly at random from the possible $\binom{m}{k^L}$ configurations. Random vector $\mathbf{T}^S$ is defined similarly. Notice that $\theta_i^L = 0, \forall i \notin \mathbf{t}^L$, and $\theta_i^S = 0, \forall i \notin \mathbf{t}^S$.

Specifying $\mathbf{\Theta}$ is equivalent to specifying three random variables $\mathbf{\Phi}^L$, $\mathbf{\Phi}^S$, Ψ. Ψ is the sum of the occurrence probabilities of the L-STR alleles

$$\psi = \sum_{i=1}^{m} \theta_i^L,$$

while ϕ^L is the normalized vector of the occurrence probabilities of the L-STR alleles. Stated otherwise,

$$\phi^L = (\frac{\theta_1^L}{\psi}, ..., \frac{\theta_m^L}{\psi}).$$

Similarly, ϕ^S is the normalized vector of the frequencies of the S-STR alleles:

$$\phi^S = (\frac{\theta_1^S}{1-\psi}, ..., \frac{\theta_m^S}{1-\psi}).$$

The prior distribution for $\boldsymbol{\theta}$ can be described in terms of the prior over $\mathbf{\Phi}^L$, $\mathbf{\Phi}^S$, and Ψ, which will be taken to be independent. The latter is distributed according to a Beta(1,1), while the positive entries of ϕ^L, i.e., $(\phi_i^L \mid i \in \mathbf{t}^L)$ are Dirichlet distributed, given $\mathbf{t}^L$, with all hyperparameters equal to α. The same holds for ϕ^S given $\mathbf{t}^S$. Hence, the distribution of node $\boldsymbol{\theta}$ can be described in terms of the distribution of seven additional random variables, whose conditional dependencies can be described by the Bayesian network of Figure 8.2. Bayesian networks using beta and Dirichlet distributions in forensic contexts are presented in Biedermann et al. (2011a). Other examples can also be found in Taroni et al. (2014).
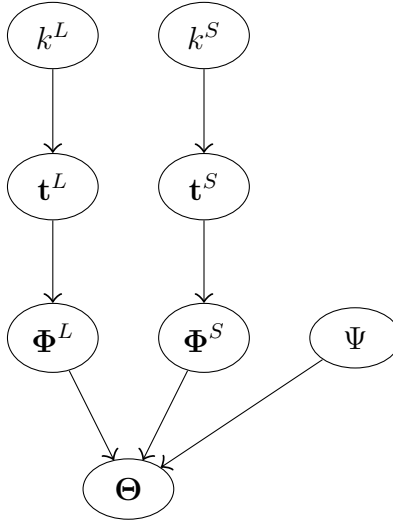

Figure 8.2: The conditional dependency relationships of the random variables used to build the distribution of $\boldsymbol{\theta}$. The definition of the nodes can be found in Section 8.7.
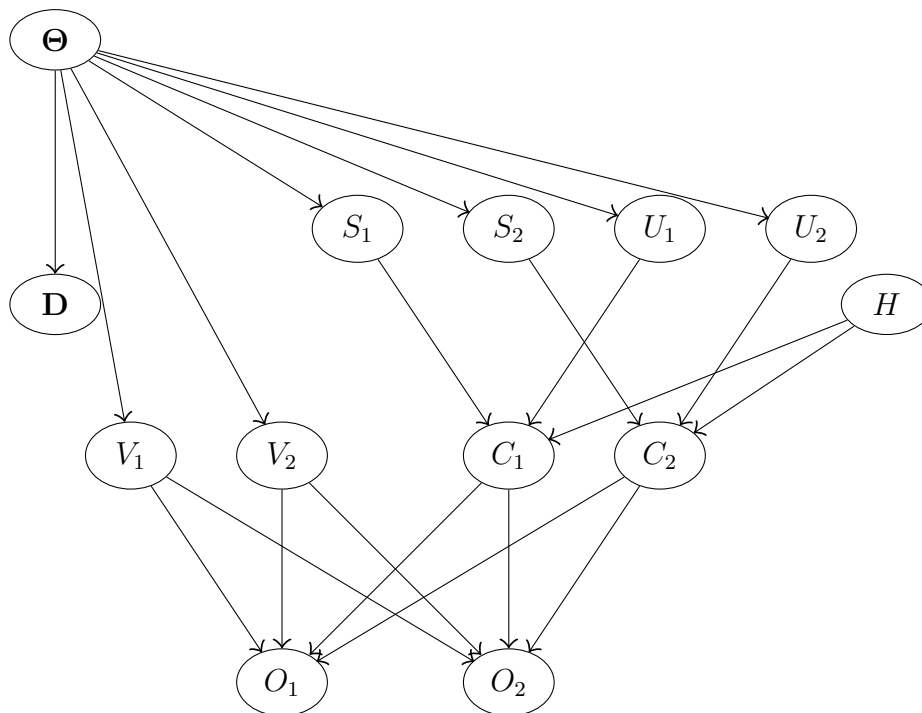
8.8   Full model

Figure 8.3: Bayesian network for the Dirichlet-multinomial model with a random number of types, to be used for DIP-STR data. The definition of the nodes can be found in Section 8.8.

This model is represented by the Bayesian network of Figure 8.3. Notice that there are differences from the model depicted in Figure 8.1, among which is the presence of node $\boldsymbol{\theta}$ distributed as described in Section 8.7. The first difference lies in the definition of nodes S_1, S_2, U_1, U_2, C_1, and C_2. Their values are couples (L,i) or (S,i) where $i \in \{1, ..., m\}$, describing the position in $\boldsymbol{\theta}$ of the corresponding DIP-STR allele. The same holds for nodes V_1 and V_2, which replace node V of Figure 8.1, and represent the two DIP-STR alleles of the victim. All these nodes are now linked to node $\boldsymbol{\Theta}$ because, given $\boldsymbol{\Theta} = \boldsymbol{\theta}$, the random variables S_1, S_2, U_1, U_2, V_1, V_2 have the following density (with parameter $\boldsymbol{\theta}$):

$$p((j, i) \mid \boldsymbol{\theta}) = \theta_i^j, \quad \forall j \in \{L, S\}, \forall i \in \{1, ..., m\}. \tag{8.2}$$

The second difference is the presence of two nodes O_1 and O_2, instead of a single node O as in Figure 8.1. O_1 represents one of the DIP-STR alleles observed from the mixture (if any, 0 otherwise). O_2 is always 0 unless we are in the situation described by the first row of Table 8.1, where two DIP-STR alleles are observed. In this case, the convention is for O_1 and O_2 to be ordered alphabetically.

The random vector $\mathbf{D}$ represents the available database of size n, through the list of labels $((L, i)$ or $(S, i))$ of the DIP-STR alleles contained in the database. The order does not matter, so we can choose the order in which the alleles appear in the database. A particular configuration of $\mathbf{D}$ is denoted as $\mathbf{d} = (d_1, ..., d_n)$, where, given $\boldsymbol{\theta}$, each component is i.i.d. with the same density as in (8.2).

According to this notation, the likelihood ratio for the scenario of interest can be written as

$$\text{LR} = \frac{p(o_1, o_2, s_1, s_2, v_1, v_2, \mathbf{d} \mid h_p)}{p(o_1, o_2, s_1, s_2, v_1, v_2, \mathbf{d} \mid h_d)} = \frac{p(o_1, o_2 \mid s_1, s_2, v_1, v_2, \mathbf{d}, h_p)}{p(o_1, o_2 \mid s_1, s_2, v_1, v_2, \mathbf{d}, h_d)}. \tag{8.3}$$

Due to the complexity of the chosen distribution, the Bayesian network cannot be treated with available software, such as Hugin, or OpenBUGS. However, the likelihood ratio can be obtained analytically using the Lemma presented in Section 8.9.
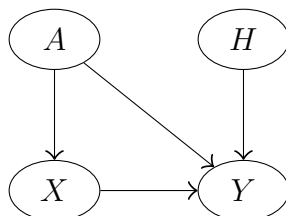
8.9   Lemma

Figure 8.4: Conditional dependencies of the random variables of the Lemma

Lemma 3. *Given four random variables A, H, X and Y, whose conditional dependencies are represented by the Bayesian network of Figure 8.4, the likelihood function for h, given $X = x$ and $Y = y$ satisfies*

$$\text{lik}(h \mid x, y) \;\propto\; \mathbb{E}(p(y \mid x, A, h) \mid X = x).$$

This Lemma, proven in Cereda (2016c), is very general: it applies to every group of random variables whose conditional dependencies are represented by the Bayesian network of Figure 8.4, and it is very useful due to the possibility of applying it to a very common forensic situation: the prosecution and the defence disagree on the distribution of part of the data (Y) but agree on the distribution of the other part (X), when the distribution of X and Y depends on some parameters (A). This Lemma can also be used for the DIP-STR model presented in Section 8.7. However, it is not straightforward to identify in the Bayesian network of Figure 8.3 the required structure shown in Figure 8.4. Luckily, the same model can be represented in several ways: we will propose a modification of the Bayesian network of Figure 8.3 into something which more clearly shows the required structure. This will be done in two steps: first, we will remove unnecessary nodes, and then we will group some of the others.

Step 1.   The Bayesian network presented in Figure 8.5 is obtained by removing from the Bayesian network of Figure 8.3 nodes U_1, U_2, C_1, and C_2. The conditional probability tables of nodes O_1 and O_2 can be directly expressed in terms of S_1, S_2, V_1, V_2, and H, in a way that makes the model equivalent to the previous one (Figure 8.5).
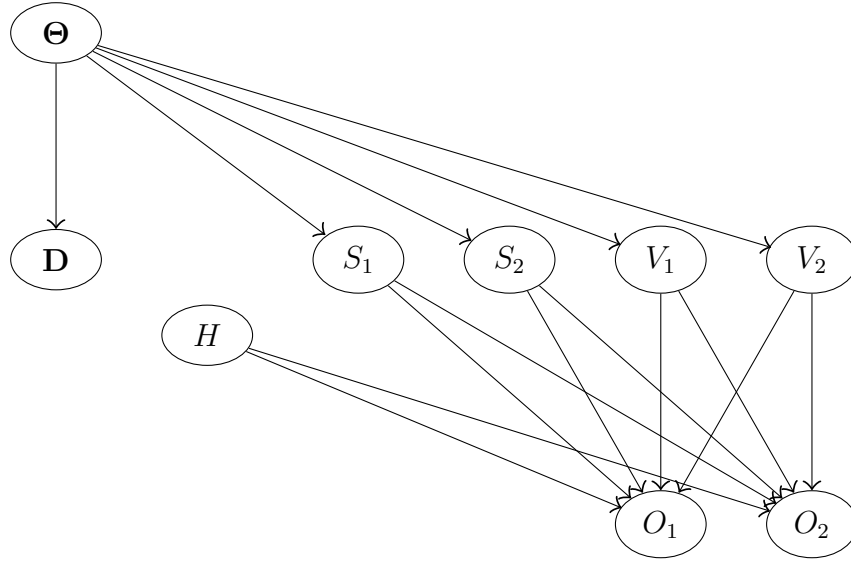
Figure 8.5: An alternative representation of the DIP-STR mixture model presented in Figure 8.3.

Step 2. The Bayesian network of Figure 8.6 can be obtained by substituting some of the nodes of the Bayesian network of Figure 8.5 with a single node. Indeed, instead of having the random vector $\mathbf{D}$ and four additional random variables (S_1, S_2, V_1, and V_2), we can group all these together into a random vector $\mathbf{B}$, of length $n + 4$. The first n elements are the labels contained in $\mathbf{D}$, the fourth to last and third to last are the labels in S_1 and S_2, while the second to last and the last are the labels in V_1, and V_2, respectively.
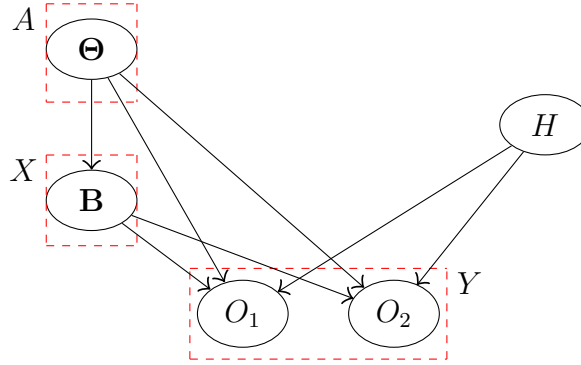

Figure 8.6: A simpler structure for the Bayesian network, suitable to be used for the Lemma. Dashed lines show the choice for the corresponding variables A, X, and Y of Figure 8.4.

The Bayesian network of Figure 8.6 can be used to represent the same model as that represented by Figure 8.1, by carefully adapting the conditional distribution of O_1, and O_2. We can apply the Lemma to our model by defining $Y = (O_1, O_2)$, $X = \mathbf{B}$, and $A = \Theta$. This leads to

$$\text{LR} = \frac{p(o_1, o_2, \mathbf{b} \mid h_p)}{p(o_1, o_2, \mathbf{b} \mid h_d)} = \frac{\text{lik}(h_p \mid o_1, o_2, \mathbf{b})}{\text{lik}(h_d \mid o_1, o_2, \mathbf{b})} = \frac{\mathbb{E}(p(o_1, o_2 \mid \mathbf{b}, \Theta, h_p) \mid \mathbf{B} = \mathbf{b})}{\mathbb{E}(p(o_1, o_2 \mid \mathbf{b}, \Theta, h_d) \mid \mathbf{B} = \mathbf{b})}.$$

Notice that we assume that under the prosecution's hypothesis, $p(o_1, o_2 \mid \mathbf{b}, \Theta, h_p) = 1$.

171

Therefore, the likelihood ratio can be simplified:

$$\text{LR} = \frac{1}{\mathbb{E}(p(o_1, o_2 \mid \mathbf{b}, \boldsymbol{\Theta}, h_d) \mid \mathbf{B} = \mathbf{b})}. \tag{8.4}$$

Victim's DIP alleles	o_1	o_2	$p(o_1, o_2 \mid \mathbf{b}, \boldsymbol{\Theta}, h_d)$
	(S, i)	(S, j)	$2\Theta_i^S \Theta_j^S$
L-L	(S, i)	0	$(\Theta_i^S)^2 + 2\Theta_i^S \Psi$
	0	0	Ψ^2
	(L, i)	(L, j)	$2\Theta_i^L \Theta_j^L$
S-S	(L, i)	0	$(\Theta_i^L)^2 + 2\Theta_i^L(1 - \Psi)$
	0	0	$(1 - \Psi)^2$

Table 8.2: Different forms that $p(o_1, o_2 \mid \mathbf{b}, \boldsymbol{\Theta}, h_d)$ can take, based on the DIP-STR alleles observed from the trace and on the victim's DIP alleles. The case in which the victim is heterozygous is not of interest.

$p(o_1, o_2 \mid \mathbf{b}, \boldsymbol{\Theta}, h_d)$ is a function of some components of the vector $\boldsymbol{\Theta}$. The form of this function depends on the combination of the DIP-STR alleles of the victim and of the trace (see Table 8.2). The expectation in the denominator of (8.4) is to be taken using the posterior distribution $\boldsymbol{\Theta} \mid \mathbf{B} = \mathbf{b}$. This is developed in detail in the Appendix.

8.10 Choice of priors

The Appendix shows the form of the denominator of the likelihood ratio for the different cases which one may encounter (for any m, any parameter $\alpha > 0$ for the Dirichlet distribution, and any prior $p(k)$ over k^L and k^S). The choice of a value for α, m, and of a prior over k^L is very delicate. If the expert has strong opinions about the number of L-STR (S-STR) alleles potentially present in nature (m) and in the population of interest (k^L and k^S), he can choose a prior which reflects his beliefs. Otherwise, he can try to use classical priors such as the Poisson distribution, the Negative binomial distribution (both of them truncated so as to have support only over $\{1, ..., m\}$), or the uniform prior over $\{1, ..., m\}$.

8.10.1 Alternative solutions

The most natural choice is to give a uniform prior (over $\{1, ..., m\}$) to k^L and k^S, combined with that of having all the $k^L + k^S$ hyperparameters of the Dirichlet priors over $\boldsymbol{\phi}_\mathbf{b}^L$ and $\boldsymbol{\phi}_\mathbf{b}^S$ equal one another. These choices represent the lack of knowledge on the number of categories and make the computations tractable.

One of the limitations of having all the hyperparameters α equal one another is that the posterior for k^L, given $\mathbf{b}$ uses only the number of distinct alleles of type L as information, and

ignores other useful information contained in $\mathbf{b}$. An alternative solution, which compensates for this undesired feature, consists of estimating k^L through the database, instead of putting a prior on it. This can be called an empirical Bayesian approach. Notice that such an undesired situation does not appear if personal beliefs are used to specify the prior distribution. Let us define the vector $\boldsymbol{\phi_b}^L$ made of the allelic proportions of the L-STR alleles observed in the augmented database, and of a last component $\bar{\phi}_{\mathbf{b}}^L$ which is the sum of the allelic proportions of all the L-STR alleles not observed in $\mathbf{b}$. $\bar{\phi}_{\mathbf{b}}^L$ is the probability of observing a new L-STR allele in the $n + 1$th draw from the population. $\boldsymbol{\phi_b}^L$ given k^L and $\mathbf{b}$ is Dirichlet distributed, hence we can obtain the posterior expected values of $\bar{\phi}_{\mathbf{b}}^L$:

$$\mathbb{E}(\bar{\phi}_{\mathbf{b}}^L \mid k^L, \mathbf{b}) = \frac{(k^L - k_{\mathbf{b}}^L)\alpha}{k^L \alpha + n^L}. \tag{8.5}$$

The so-called *Good-Turing estimator* (Good, 1953) says that the expected value for the probability of the unobserved types can be approximated by the proportion of L-STR singletons (i.e, alleles observed only once) in the database. Stated otherwise,

$$\mathbb{E}(\bar{\phi}_{\mathbf{b}}^L | k^L, \mathbf{b}) \approx \frac{n_1^L}{n^L}. \tag{8.6}$$

where n_1^L is the number of DIP-STR alleles observed only once in the augmented database. The two quantities (8.5) and (8.6) can be equated in order to obtain an empirical Bayesian estimate of k^L as

$$\hat{k^L} = \frac{n_1^L n^L + k_{\mathbf{b}}^L \alpha n^L}{\alpha n^L - \alpha n_1^L}.$$

The likelihood ratio for this choice can be obtained using the same formulas developed in the Appendix by using prior over k^L the degenerate prior which gives a probability of one to value $\widehat{k^L}$. This solution allows one to use more information (n_1^L and n_1^S) from $\mathbf{b}$.

The third option is to use the plug-in approximation proposed by some literature, which estimates the allelic frequencies by their posterior expectation, after the observation of a database. One of the aims of this paper is to investigate the goodness of this approximation, in the case of a rare type match problem.

We did some experiments using marker MID1950-D20S473 (Castella et al., 2013), and considering the two cases described in Table 8.3.

	Victim's alleles	Suspect's alleles	Observed alleles
Case 1	S11 - S11	L2 - L13	L2 - L13
Case 2	S11 - S11	L2 - S12	L2

Table 8.3: Allelic configurations of the victim and of the suspect in the two cases of interest, at marker MID1950-D20S473.

Allele L2 was not contained in the database of reference, hence we are in presence of the rare type match case. The available database, augmented with the victim's and the suspect's DIP-STR alleles, contains 11 different DIP-STR alleles, for a total number of 210 observations (from 105 individuals).
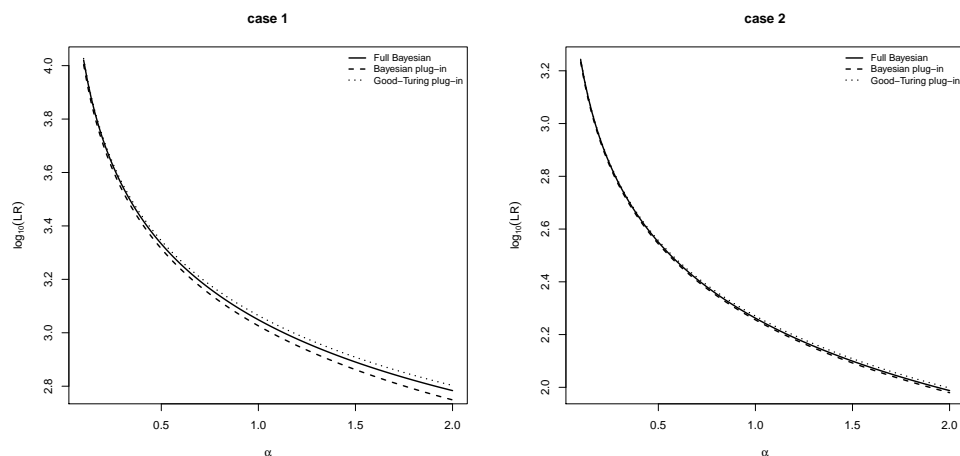
Figure 8.7: Sensitivity analysis for the $\log_{10}(\text{LR})$ obtained with (i) the full Bayesian approach, (ii) the hybrid Good-Turing plug-in (iii) classical Bayesian plug-in, for the two cases described in Table 8.3 when the prior over k^L and k^S is uniform over $\{1, ..., m\}$.

The sensitivity analysis for the $\log_{10}(\text{LR})$, shown in Figure 8.7, has been conducted for different loci and different combinations of alleles, without showing substantial differences (in terms of sensitivity). Moreover, it tells us that the two plug-in approaches represent acceptable solutions in terms of quantification. Varying m does not change Figure 8.7 much.

8.11   Conclusion

Mostly due to the limited size of the available database (about one hundred people in a given relevant population), the rare type match situation is very likely to be encountered when DIP-STR data is used. The recipients of this new technology should be prepared for such an eventuality, which was not taken into account in the OOBN proposed in Cereda et al. (2014b). This paper provides a methodology that allows one to obtain the full Bayesian likelihood ratio also when there are DIP-STR alleles which are not present in the reference database among the alleles of the known contributor and of the suspect. This is done by extending the OOBN, and introducing a more complex prior over the allelic frequencies (a mixture of Dirichlet and uniform distribution) based on a previously developed solution for Y-STR data (Cereda, 2016a). Notice that this issue also represents an opportunity to discuss the use of plug-in approximations which are compared with the full Bayesian likelihood ratio. They proved to be valid approximations.

The sensitivity analysis of the hyperparameters of the prior is also studied. The results show that the likelihood ratio moderately depends on the choices of the parameters α of the Dirichlet prior. Hence, there is the need for further investigations to find better priors, either less sensitive to hyperparameters, or more realistic, such as it was done for Y-STR data in Cereda (2016c).

Appendix. Full Bayesian likelihood ratio development

In Table 8.4, a summary of the relevant symbols used is reported. The aim of this Appendix is to develop the conditional expectation of the functions reported in Table 8.2, which constitute the denominator of the likelihood ratio (8.4). Those conditional expectations can be rewritten in terms of $\boldsymbol{\phi}^L$, $\boldsymbol{\phi}^S$, and ψ, in the following way:

$$\mathbb{E}(2\Theta_i^L \Theta_j^L \mid \mathbf{b}) = \mathbb{E}(2\Phi_i^L \Phi_j^L \Psi^2 \mid \mathbf{b}) = 2\mathbb{E}(\Phi_i^L \Phi_j^L \mid \mathbf{b})\mathbb{E}(\Psi^2 \mid \mathbf{b}),$$

$$\mathbb{E}((\Theta_i^L)^2 + 2\Theta_i^L(1-\Psi) \mid \mathbf{b}) = \mathbb{E}((\Phi_i^L)^2 \mid \mathbf{b})\mathbb{E}(\Psi^2 \mid \mathbf{b}) + 2\mathbb{E}(\Phi_i^L \mid \mathbf{b})\mathbb{E}(\Psi \mid \mathbf{b})\mathbb{E}(1-\Psi \mid \mathbf{b}).$$

The distribution of Ψ given B.

As explained in Section 8.7, $\boldsymbol{\theta}$ can be represented through a set of three independent variables $(\boldsymbol{\phi}^L$, $\boldsymbol{\phi}^S$, $\psi)$. The vector $\mathbf{b}$ can also be reduced by sufficiency to three random variables: $(n^L, \mathbf{n}^L, \mathbf{n}^S)$, where n^L is the total number of observed L-STR alleles in the enlarged database, $\mathbf{n}^L$ is the vector of length m containing the counts in the augmented database of each of the m L-STR alleles, in an order that corresponds to that of $\boldsymbol{\phi}^L$, $\mathbf{n}^S$ is the vector of counts of each of the m S-STR alleles. n^L is binomial distributed with parameters $(n+4, \boldsymbol{\psi})$, while $\mathbf{n}^L$ is multinomial distributed with parameters $(n^L, \boldsymbol{\phi}^L)$. Similarly, $\mathbf{n}^S$ is multinomial distributed with parameters $(n^S, \boldsymbol{\phi}^S)$, where $n^S = n + 4 - n^L$ is the number of S-STR alleles in the augmented database.

It holds that the likelihood for $\boldsymbol{\phi}^L, \boldsymbol{\phi}^S$, and ψ factors:

$$p(n^L, \mathbf{n}^L, \mathbf{n}^S \mid \boldsymbol{\phi}^L, \boldsymbol{\phi}^S, \psi) = p(n^L \mid \psi)p(\mathbf{n}^L \mid n^L, \boldsymbol{\phi}^L)p(\mathbf{n}^S \mid n^S, \boldsymbol{\phi}^S).$$

The priors for $\boldsymbol{\phi}^L, \boldsymbol{\phi}^S$, and ψ factors as well, since they are independent. Therefore, the posteriors for $\boldsymbol{\phi}^L, \boldsymbol{\phi}^S$, and for ψ given $\mathbf{b}$ factors as the product of three independent posteriors. Thus, it holds that

$$p(\psi \mid \mathbf{b}) \propto p(n^L \mid \psi)p(\psi),$$

which is a product of the density of a binomial distribution and of a beta prior. By conjugacy,

$$\Psi \mid \mathbf{B} = \mathbf{b} \sim \text{Beta}(1 + n^L, 1 + n^S).$$

In conclusion, by using properties of the Beta distribution, it holds that

$$\mathbb{E}(\Psi^2 \mid \mathbf{b}) = \frac{(n^L + 1)(n^L + 2)}{(n + 6)(n + 7)}, \tag{8.7}$$

and

$$\mathbb{E}((1-\Psi)^2 \mid \mathbf{b}) = \frac{(n^S + 1)(n^S + 2)}{(n + 6)(n + 7)}. \tag{8.8}$$

Name	Description	Type
m	number of theoretically possible L-STR (and S-STR) alleles	fixed
k^L	number of L-STR allele types present in the population	random
k^S	number of S-STR allele types present in the population	random
$\mathbf{t}^L$	positions (from 1 to m) of the k^L L-STR allele types in the population	random
$\mathbf{t}^S$	positions (from 1 to m) of the k^S S-STR allele types in the population	random
$\boldsymbol{\theta}$	population proportions of the $2m$ possible DIP-STR alleles	random
ψ	sum of the relative frequencies of the L-STR alleles	random
$\boldsymbol{\phi}^L$	normalised vector of the relative frequencies of L-STR alleles	random
$\boldsymbol{\phi}^S$	normalised vector of the relative frequencies of S-STR alleles	random
n	size of the available database	observed
n^L	total number of L-STR alleles in the augmented database	observed
n^S	total number of S-STR alleles in the augmented database	observed
$\mathbf{b}$	labels (j,i) corresponding to each of the $n+4$ DIP-STR alleles in the augmented database	observed
$\mathbf{b}^L$	labels (L,i) corresponding to each of the n^L L-STR alleles in the augmented database	observed
$\mathbf{b}^S$	labels (S,i) corresponding to each of the n^S S-STR alleles in the augmented database	observed
$k_{\mathbf{b}}^L$	number of distinct L-STR alleles in the augmented database	observed
$k_{\mathbf{b}}^S$	number of distinct S-STR alleles in the augmented database	observed
$\mathbf{n}^L$	counts of all m L-STR alleles in the augmented database	observed
$\mathbf{n}^S$	counts of all m S-STR alleles in the augmented database	observed

Table 8.4: Some relevant symbols used in the paper.

The distribution of ϕ^L and ϕ^S given B.

Let $p(k)$ be the prior distribution over k^L and k^S. In this section we will omit superscripts L and S from k, $\mathbf{t}$, ϕ, and $\mathbf{n}$, in order to obtain general results valid for both cases. Notice that n will stand for n^L or n^S, and $\mathbf{b}$ will stand for $\mathbf{b}^L$ or $\mathbf{b}^S$ as described in Table 8.4 (so temporarily, the meaning of n, and $\mathbf{b}$ is different from its meaning in the rest of the paper).

Given k, $\mathbf{t}$ is uniformly distributed over the ordered vectors containing k indexes from 1 to m. Let us denote with $k_{\mathbf{b}}$ the number of distinct L-STR (or S-STR) alleles observed in the augmented database, and with $\phi_{\mathbf{b}}$ the vector of length $k_{\mathbf{b}}$ containing only the frequencies of the L-STR alleles observed in the augmented database in the order in which they appear in ϕ. $\phi_{\mathbf{b}}$ does not sum to one, since there are L-STR alleles of positive frequency, which are not observed: the total probability mass of the unobserved alleles is $\bar{\phi}_{\mathbf{b}} = 1 - \sum_{i=1}^{k_{\mathbf{b}}} \phi_{\mathbf{b}i}$. The vector $\phi_{\mathbf{b}}^* = (\phi_{\mathbf{b}}, \bar{\phi}_{\mathbf{b}})$ sums up to one.

We can look for the posterior distribution of $\phi_{\mathbf{b}}^*$ given the vector $\mathbf{b}$.

$$p(\phi_{\mathbf{b}}^* \mid \mathbf{b}) = \sum_k \sum_{\mathbf{t}} p(\phi_{\mathbf{b}}^* \mid \mathbf{b}, \mathbf{t}) p(\mathbf{t} \mid k, \mathbf{b}) p(k \mid \mathbf{b}) \tag{8.9}$$

It can be proved that

- the posterior density $p(\phi_{\mathbf{b}}^* \mid \mathbf{b}, \mathbf{t})$ depends on $\mathbf{t}$ only through k. Hence, we can denote it as $p(\phi_{\mathbf{b}}^* \mid \mathbf{b}, k)$

- if k is less than $k_{\mathbf{b}}$, then $p(k \mid \mathbf{b}) = 0$.

- let us denote with $\mathcal{T}_{k,\mathbf{b}}$ the set of ordered vectors $\mathbf{t}$ of length k and compatible with $\mathbf{b}$ (i.e., which contain among others the positions corresponding to the elements in $\mathbf{b}$). For all the $\mathbf{t}$ which are not in $\mathcal{T}_{k,\mathbf{b}}$, then $p(\mathbf{t} \mid k, \mathbf{b}) = 0$.

We can change the summation indexes in (8.9) to obtain:

$$p(\phi_{\mathbf{b}}^* \mid \mathbf{b}) = \sum_{k=k_{\mathbf{b}}}^{m} p(k \mid \mathbf{b}) p(\phi_{\mathbf{b}}^* \mid k, \mathbf{b}) \sum_{\mathbf{t} \in \mathcal{T}_{k,\mathbf{b}}} p(\mathbf{t} \mid k, \mathbf{b}).$$

For any of the $\binom{m-k_{\mathbf{b}}}{k-k_{\mathbf{b}}}$ vectors $\mathbf{t}$ in $\mathcal{T}_{k,\mathbf{b}}$, $p(\mathbf{t} \mid k, \mathbf{b})$ has the same value $\frac{1}{\binom{m-k_{\mathbf{b}}}{k-k_{\mathbf{b}}}}$. Thus, in the end we have that

$$p(\phi_{\mathbf{b}}^* \mid \mathbf{b}) = \sum_{k=k_{\mathbf{b}}}^{m} p(k \mid \mathbf{b}) p(\phi_{\mathbf{b}}^* \mid k, \mathbf{b}).$$

The distribution $p(k \mid \mathbf{b})$ can be obtained in the following way.

$$p(k, \mathbf{t}, \phi, \mathbf{b}) = p(k) p(\mathbf{t} \mid k) p(\phi \mid \mathbf{t}) p(\mathbf{b} \mid \phi).$$

Integrating out ϕ, we obtain

$$p(k, \mathbf{b}, \mathbf{t}) = p(k) p(\mathbf{t} \mid k) \int_{\phi} p(\phi \mid \mathbf{t}) p(\mathbf{b} \mid \phi) \mathrm{d}\phi, \tag{8.10}$$

where the integral contains a Dirichlet density and the categorical density defined in (8.2). They are conjugate, thus we obtain

$$p(k, \mathbf{t} \mid \mathbf{b}) \propto p(k)p(\mathbf{t} \mid k)\frac{\Gamma(k\alpha)}{\Gamma(n + k\alpha)}.$$

Now we can sum over the $\mathbf{t}$ compatible with $\mathbf{b}$, to get to

$$p(k \mid \mathbf{b}) \propto \binom{k}{k_{\mathbf{b}}}p(k)\frac{\Gamma(k\alpha)}{\Gamma(n + k\alpha)}. \tag{8.11}$$

In conclusion,

$$p(\boldsymbol{\phi_b}^* \mid \mathbf{b}) \propto \sum_{k=k_{\mathbf{b}}}^{m} \binom{k}{k_{\mathbf{b}}}p(k)\frac{\Gamma(k\alpha)}{\Gamma(n + k\alpha)}p(\boldsymbol{\phi_b}^* \mid k, \mathbf{b}), \tag{8.12}$$

where $\boldsymbol{\Phi_b}^* \mid K = k, \mathbf{B} = \mathbf{b} \sim \text{Dir}^{k_{\mathbf{b}}+1}(\alpha + \tilde{n}_1, ..., \alpha + \tilde{n}_{k_{\mathbf{b}}}, (k - k_{\mathbf{b}})\alpha)$, and $\tilde{\mathbf{n}}$ is the vector of length $k_{\mathbf{b}}$ with the positive elements of $\mathbf{n}$.

Therefore, (8.12) is a mixture of Dirichlet distributions with weights $w(k) = \binom{k}{k_{\mathbf{b}}}p(k)\frac{\Gamma(k\alpha)}{\Gamma(n+k\alpha)}$. Using properties of the Dirichlet distribution we obtain that, $\forall i, j$ corresponding to different observed DIP-STR alleles:

$$\mathbb{E}(\Phi_i\Phi_j \mid \mathbf{b}) = (\alpha + n_i)(\alpha + n_j)\frac{\sum_{k=k_{\mathbf{b}}}^{m} w(k)g(k)}{\sum_{k=k_{\mathbf{b}}}^{m} w(k)}, \tag{8.13}$$

$$\mathbb{E}(\Phi_i^2 \mid \mathbf{b}) = (\alpha + n_i)(\alpha + n_i + 1)\frac{\sum_{k=k_{\mathbf{b}}}^{m} w(k)g(k)}{\sum_{k=k_{\mathbf{b}}}^{m} w(k)}, \tag{8.14}$$

where $g(k) = \frac{1}{(k\alpha+n)(k\alpha+n+1)}$.

The conditional expectations in Table 8.2

Using (8.7), (8.8), (8.13) and (8.14), we obtain that, $\forall i, j$ corresponding to different observed alleles,

$$\mathbb{E}(\Theta_i^L\Theta_j^L \mid \mathbf{b}) = \mathbb{E}(\Phi_i^L\Phi_j^L\Psi^2 \mid \mathbf{b}) = \mathbb{E}(\Phi_i^L\Phi_j^L \mid \mathbf{b})\mathbb{E}(\Psi^2 \mid \mathbf{b}) \tag{8.15}$$

$$= (\alpha + n_i^L)(\alpha + n_j^L)\frac{\sum_{k=k_{\mathbf{b}}^L}^{m} w^L(k)g^L(k)}{\sum_{k=k_{\mathbf{b}}^L}^{m} w^L(k)}\frac{(n^L + 1)(n^L + 2)}{(n + 6)(n + 7)}, \tag{8.16}$$

$$\mathbb{E}((\Theta_i^L)^2 \mid \mathbf{b}) = \mathbb{E}((\Phi_i^L)^2\Psi^2 \mid \mathbf{b}) = \mathbb{E}((\Phi_i^L)^2 \mid \mathbf{b})\mathbb{E}(\Psi^2 \mid \mathbf{b}) = \tag{8.17}$$

$$= (\alpha + n_i^L)(\alpha + n_i^L + 1)\frac{\sum_{k=k_{\mathbf{b}}^L}^{m} w^L(k)g^L(k)}{\sum_{k=k_{\mathbf{b}}^L}^{m} w^L(k)}\frac{(n^L + 1)(n^L + 2)}{(n + 6)(n + 7)}, \tag{8.18}$$

$$\mathbb{E}(\Theta_i^L(1-\Psi) \mid \mathbf{b}) = \mathbb{E}(\Phi_i^L \mid \mathbf{b})(\mathbb{E}(\Psi \mid \mathbf{b}) - \mathbb{E}(\Psi^2 \mid \mathbf{b})) = \tag{8.19}$$

$$= (\alpha + n_i^L)\frac{\sum_{k=k_{\mathbf{b}}^L}^m \frac{w^L(k)}{k\alpha + n^L}}{\sum_{k=k_{\mathbf{b}}^L}^m w^L(k)}\frac{(n^L+1)(n^S+1)}{(n+6)(n+7)}, \tag{8.20}$$

where n and $\mathbf{b}$ have now their original meaning, and $w^L(k) = \binom{k}{k_{\mathbf{b}}^L}p(k)\frac{\Gamma(k\alpha)}{\Gamma(n^L+k\alpha)}$, and $g^L(k) = \frac{1}{(k\alpha+n^L)(k\alpha+n^L+1)}$.

These formulas can be directly used to obtain the conditional expectations of the last three rows of Table 8.2. In a very similar way one can easily obtain the conditional expectations contained in the first three rows.