Current challenges in statistical DNA evidence evaluation

Cereda, G.

Chapter 2

Results

This Phd thesis consists of several challenging problems concerning statistical DNA evidence evaluation, investigated in a series of distinct studies, whose details are presented in Chapters 3 to 8. These studies, explained and summarized in the sections to come, are interrelated and follow a logical and chronological path.

2.1 Object-oriented Bayesian networks for evaluating DIP-STR profiling results from unbalanced DNA mixtures

This section is intended as a summary of the research contained in Cereda et al. (2014b), reproduced in full in Chapter 3.

DIP-STR markers, described in Section 1.1.5, were proposed as a novel type of genetic markers in Castella et al. (2013). The available kit for the DIP-STR analysis allows one to obtain the DIP-STR profile of a DNA mixture, to be compared with the DIP-STR profile of one or more questioned contributors, at a certain number of loci. This new set of markers is especially useful in case of extremely unbalanced mixtures, where the standard STR marker system fails to detect the alleles of the minor contributor. However, to exploit completely their potential, it is important to build an evaluative framework which helps in the assessment of observed profiling results in the light of the hypotheses of interest. This amounts to the calculation of the likelihood ratio, as described in Section 1.2.

To this extent, we decided to build an object-oriented Bayesian network to calculate the likelihood ratio for the two source level hypotheses 'the mixture contains the DNA of the victim and the suspect' (h_p), and 'the mixture contains the DNA of the victim and of an unknown person, unrelated to the suspect' (h_d). The data to evaluate is made of the DIP-STR results obtained from (i) DNA mixtures of two contributors, (ii) the known contributor (for instance, the victim of a sexual assault), (iii) the questioned contributor (the suspect). Figure 2.1 shows the OOBN whose details can be found in Chapter 3.

The probability tables of the class `DIP-STR` is filled with the so-called "Bayes estimates" of the allelic frequencies, based on a Dirichlet prior. They are obtained putting a Dirichlet prior

over the set of the allelic proportions, and updating this prior through the use of a database, seen as a multinomial observation. The posterior mean of each allelic proportion is used as estimate for the unknown allelic proportions. These estimates change from marker to marker, and from case to case. For this reason, an additional RHugin function is made available to automatise the procedure.

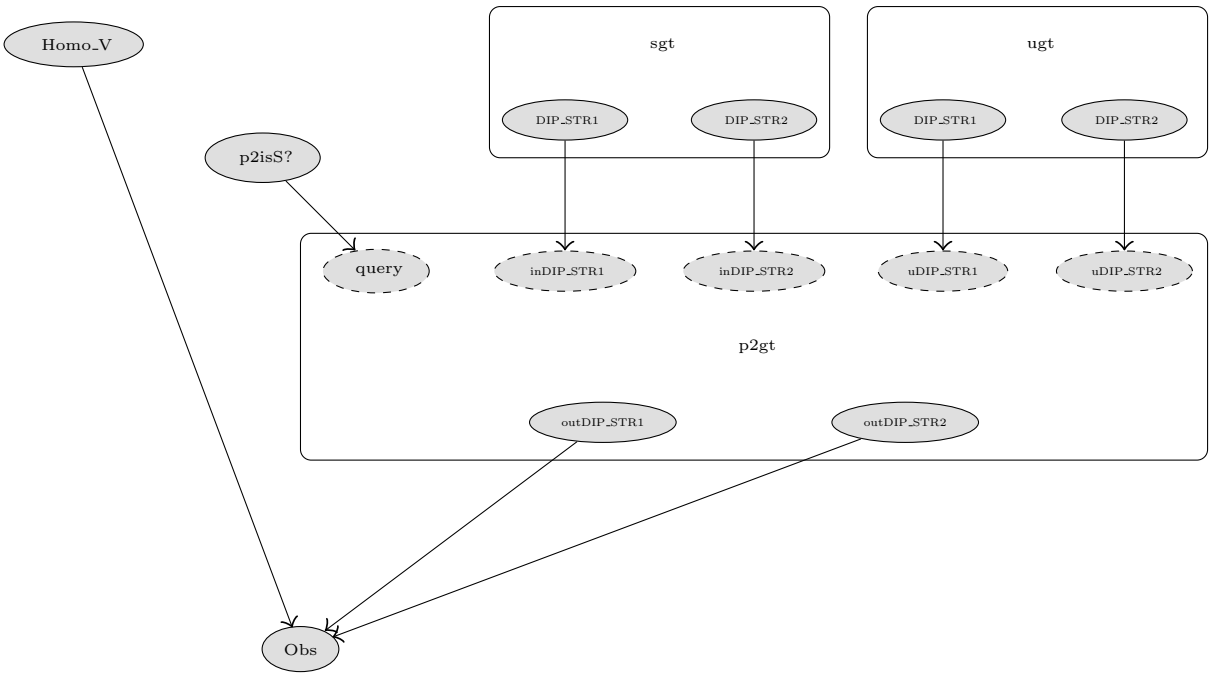

Figure 2.1: Expanded representation of the class `Marker`, modelling the mixture observation from a single locus. The victim is represented by the node **Homo_V**, with three states: *HomL*, *HomS* and *Hetero*, corresponding to his/her homozygosity or heterozygosity for the DIP allele. The right part of the network (i.e., all components other than **Homo_V** and **Obs**) models the minor contributor, that could be either the suspect, or an unknown person, whose alleles are represented by nodes **sgt** and **ugt**. The Boolean node **p2isS?** addresses the question of whether the second contributor is the suspect or an unknown person, while Node **p2gt** represents the genotype of the actual second contributor to the mixture. Node **Obs**, with states *La, Lb, Lab, Sa, Sb, Sab, X, nr*, represents the observed (minor contributor's) DIP-STR allele(s) in the trace. More details on the structure of each instance node, and on conditional probability tables can be found in Chapter 3.

Alternatively, several instances of the class `Marker` can be joined into the compound OOBN of Figure 2.2. The only modification required is to change node **p2isS?** inside the class `Marker` into an input node. Output node **H** is then linked to each node **p2isS?**.

Then, the compound likelihood ratio can be read as the ratio of the posterior probabilities of the state of node **H** when DIP-STR are entered, in virtue of the choice of equal prior probabilities over **H**.

A further step was that of modifying the interpretative model to be used when the suspect's DNA is not available for comparison, but that of a brother of the suspect is. The network `Marker for brother` riproduced in Figure 3.2 of Chapter 3, was built with that purpose.

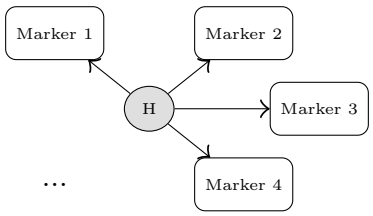

Figure 2.2: A compound object-oriented Bayesian network to be used for the evaluation of DIP-STR results from several markers. Each rectangular node is an instance of the class `Marker`, shown in Figure 2.1.

This is a good example of the flexibility of object-oriented Bayesian networks, which are readily adapted to different scenarios, in addition to provide a concise representation of the genotypic configuration of the various (assumed) contributors.

At the time when Cereda et al. (2014b) was written only 9 DIP-STR loci were available (Castella et al., 2013). However, the classes `Marker` and `Marker for brother` are usable with any marker. It is enough to adapt the meaning of the states and the probability tables of the input nodes. Hence, the same model can be used with the additional 9 loci that have recently been identified (Oldoni et al., 2015).

The paper contains also the application of the model to a casework example, where a relevant blood stain was collected on the body of a dead woman, and circumstantial evidence led to three suspects: a man and his two sons. The likelihood ratio for the hypotheses h_p and h_d defined above, was calculated marker by marker. Also, the likelihood ratio for the case in which the suspect's DIP-STR profile is not available, but that of a brother of the suspect is, is calculated, using the class `Marker for brother`. These examples allowed also to check that the likelihood ratio values obtained with the help of the probabilistic graphical model was the same as that obtained using a traditional formulaic approach. The advantage over the formulaic approach, which take different forms depending on the allelic configuration is clear, since except for the input values (i.e., the initial numerical specification), the structure of the OOBN remains constant. Moreover, formulae may become complicated when the scenario involves other members of the family of the suspect, depending on his degree of relatedness with the typed individuals. Thus, the use of an OOBN could also help to make evaluative procedures less prone to possible errors, since computations are entirely confined to the model. Note that the model presented in Figure 2.1 does not take into account extra variables of (potential) influence, such as typing errors and subpopulation effects.

2.2 An investigation of the potential of DIP-STR markers for DNA mixture analyses

This section is intended as a summary of the research contained in Cereda et al. (2014a), reproduced in full in Chapter 4.

The STR marker system is the most used set of markers for the analysis of DNA mixtures. It proved itself reliable for mixtures of any kind, except for extremely unbalanced mixtures,

for which it fails to detect the alleles of the minor contributor (see Section 1.1.5). This problem is usually overcome by the use of Y-STR markers, but they only work in case one contributor is female and the other is male. Also, they do not allow to distinguish between patrilineal relatives of the male contributor. The DIP-STR marker system hasn't got any of these limitations, and certainly can be preferred to STR markers and Y-STR markers in case of extremely unbalanced mixtures with a male major contributor (or a female minor contributor), or if it is necessary to distinguish between male relatives of the suspect, as in the casework example described in Section 2.1. However, an unconditional use of the DIP-STR technology, for any recovered trace, may not be the best option, given that available databases are sensibly smaller than the enormous amount of data already collected for STR and Y-STR. Also, the newness of this technology makes it more expensive. The question of which marker system is to be chosen is thus at the same time interesting and challenging.

In order to answer this interesting question, we decided to compare the DIP-STR marker system to the STR and Y-STR marker systems, from a statistical and forensic perspective. To do so, we contrasted the distribution of the likelihood ratio results obtained from 100,000 simulated cases using DIP-STR markers, with the distribution obtained (i) using traditional STR markers (assuming that we are in presence of moderately unbalanced mixture), and (ii) using Y-STR markers (assuming we are in presence of female-male mixtures). The comparison is performed under the prosecution's and the defence's case. For each marker system, each of the 100,000 simulated cases consists in the DNA profiles of the victim, of the suspect, and of another contributor. These are simulated by drawing locus by locus the alleles of the three individuals, independently and according to the allelic proportions in the population of interest. Under the prosecution's point of view, we "compose" a virtual mixture whose profile, at each locus, is made of the alleles of the victim and of the suspect. Under the defence's point of view, the profile contains alleles of the victim and of the other contributor. The virtual mixture obtained in this way is the one that could have been observed if we were in presence of real two-person mixtures, when the major contributor's genotype is available and under a set of assumptions, detailed in Chapter 4, concerning its quality.

For each simulated case, and each point of view, the likelihood ratio under the hypotheses "the mixture contains DNA material from the victim and the suspect" (h_p), "the mixture contains DNA material from the victim and an unknown contributor" (h_d) is calculated. All the likelihood ratios obtained for the prosecution's case are expected to be higher than 1, since we don't take into account the possibility of errors of laboratory. The higher the likelihood ratios obtained, the more the chosen marker system is interesting from the prosecution's point of view. On the other hand, in the defence's case, every time the genotype of the suspect is not compatible with the alleles in the mixture, a likelihood ratio of zero is obtained. The higher the number of zero values obtained, the more attractive is the chosen method from the defence's point of view.

In summary, the following conclusions can be made:

- For cases of, at worst, moderately unbalanced mixtures, the distributions of the likelihood ratio values both from the prosecution's and the defence's point of view suggest that the traditional STR marker system should be preferred.
- In case of extremely unbalanced mixtures STR markers are not reliable, but Y-STR

markers and DIP-STR markers can be used, the DIP-STR method should be preferred from the prosecution's point of view. From the the defence's point of view, preferences depend on whether one is more concerned with the strength of support obtained in case of a wrong association (i.e., when the likelihood ratio supports the wrong proposition), or on the number of times in which such a wrong indication is obtained.

The research went on with a discussion about the proportion of cases in which each of the three methods cannot be used: this has clearly a great influence on the choice of an analytical methodology.

The STR method is generally not useful to detect the minor contributor of extremely unbalanced mixtures, but in current practice, many (or most) extremely unbalanced mixtures probably go undetected (Oldoni et al., 2015), hence it is difficult to assess the proportion of cases in which such mixtures are encountered. In turn, it is easier to circumscribe the proportion of cases in which Y-STR markers are not usable: this happens whenever the major and the minor contributors are not female and male, respectively.

With regards to DIP-STR markers, the probability that an actual two-person mixture will not be recognised as such (i.e., the presence of a second contributor cannot be pointed out because the mayor is heterozygous or both contributors have the same DIP alleles) has been calculated. It turned out that about 4% of recovered stains, which are actually mixtures, will leave one with the uncertainty about the presence of a second contributor.

This allowed to conclude that the use of DIP-STR markers can be desirable for all those kind of traces that appear as a single stain with the use of STR and Y-STR markers, but for which one suspects the presence of a second contributor. In these cases, DIP-STR markers can also complement Y-STR results to discriminate paternally related individuals. Also, the use of DIP-STR markers could be of interest for all kinds of DNA stains, since one can establish in advance if they could be used, given that one starts by determining the DIP-STR genotype of the assumed known major contributor. In the case of a favourable outset, DIP-STR profiling can provide information about the second contributor in terms of one, two or no alleles. Although the likelihood ratio distributions obtained under the defence's and the prosecution's point of view are not as marked as those that can be obtained with traditional STR markers, they can still be regarded as practically useful. Moreover, new DIP-STR markers have been recently located (Oldoni et al., 2015). This will favourably improve the likelihood ratio distributions that could be obtained under the various competing points of view in a near future. However, analysts should also remind that the definition of the practical procedures will also encompass additional factors such as time and monetary constraints.

2.3 Some methodological issues

The research explained in the sections to come combines the attempt of finding a solution to the rare type match problem described in Section 1.2.4, and the treatment of some methodological issues encountered while working with DIP-STR data and, more generally, studying the state of the art of forensic DNA evaluation. There are four main methodological issues

discussed:

Plug-in likelihood ratio assessment. While building the interpretative framework for DIP-STR results, we were confronted with the necessity of quantifying DIP-STR allelic proportions having only a small database to represent the population of interest. A "full Bayesian" procedure would consist in choosing a prior distribution for this nuisance parameter, and integrating it out. However, in the researches summarized in Sections 2.1 and 2.2, we decided to follow the common forensic procedure, which estimates the allelic proportions using the posterior mean, after the observation of the database. In fact, this is only (at best) an approximation to the full Bayesian procedure. In Chapter 6 we call this approach 'hybrid' since it is reminiscent of the frequentist approach. The use of these plug-in methods, seen as an approximation of the exact Bayesian likelihood ratio, is not a problem in itself, and may present advantages in terms of computability, but the accuracy of these approximations should be carefully investigated. Moreover, we will show that often the full Bayesian procedure is not more difficult to use.

Evidence and background This issue is deeply related to the discussion about the use of hybrid plug-in methods described here above. In Section 1.2, the likelihood ratio is defined as the ratio of the probabilities of observing data D under the competing hypotheses of interest. It is important to discuss what to consider as D, since there are diverging options in literature. In our opinion, the data available to the expert can be divided into *evidence*

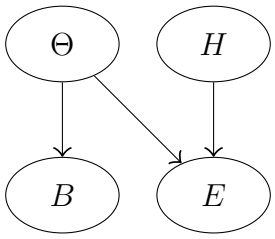

Figure 2.3: Bayesian network representing the dependency relations between E (evidence of the case), B (background data in the form of a database), Θ (population parameter), and H (hypotheses of interest).

and *background*. The first refers to data related to the crime, directly useful to discriminate between hypotheses of interest. The second refers to data which is not related to the crime, but is made available to the expert in order to help him to deal with the nuisance parameter(s) of the model. For instance, the evidence may consist of the DNA stain found at the crime scene, and of a suspect with the same DNA profile. The nuisance parameter of this model is θ, the population proportion of this DNA profile, and background data is a database from the population of possible perpetrators. Let us denote with E, B and Θ the random variables that correspond to evidence, background data, and nuisance parameter, respectively. The conditional dependency relation between these variables (together with H) are represented by the Bayesian network of Figure 2.3.

We believe that a Bayesian statistician or a Bayesian forensic scientist would use all available data to assign the value of the evidence. Hence, the Bayesian likelihood ratio should be

defined as

$$\mathrm{LR} = \frac{\Pr(E = e, B = b \mid H = h_p)}{\Pr(E = e, B = b \mid H = h_d)}. \tag{2.1}$$

This is in discrepancy with the definition of the likelihood ratio which can be found in much literature where only evidence data E is considered, with no mention for B.

It is true that B is independent of H, and that under the frequentist approach B and E are independent given H. This allows to simplify B from the formula 2.1. However, as can be deduced from the Bayesian network of Figure 2.3, this is not valid in a Bayesian framework, unless Θ is known.

Levels of uncertainty For a frequentist statistician, the likelihood ratio is a ratio of probabilities usually based on a model which is at best only a good approximation to the truth, and whose parameters have to be estimated using a limited sample. Thus, in a frequentist framework, along with the first basic initial uncertainty about the hypotheses, two more levels of uncertainty arise in the attempt of calculating the likelihood ratio. Some forensic literature (Morrison, 2010; Stoel and Sjerps, 2012; Curran et al., 2002; Curran, 2005) already pointed out the necessity for uncertainty assessment in the estimation of the likelihood ratio, even though they don't differentiate among levels.

On the other hand, for a true Bayesian individual these additional levels of uncertainty are part of the model, so that the definition of the Bayesian Pr includes not only beliefs about chances when picking people from the population, but also beliefs about parameters of the model, and beliefs about model. Hence, these levels could in principle be taken care of within the same framework.

There is a debate in literature (e.g. Taroni et al., 2016; Sjerps et al., 2016; Berger and Slooten, 2016; Curran, 2016) as to whether it makes sense to talk about 'estimation' and 'uncertainty assessment' regarding the likelihood ratio. Both the points of view are valid, depending on the context: if a frequentist approach to probability definition is chosen, it is pertinent to talk about estimations and uncertainty assessment. On the other hand, in a full Bayesian context, with Bayesian probabilities subjectively defined, they are misplaced. Moreover, most of the time, the Bayesian procedure consists of choosing priors (and hyperpriors) which are a compromise between personal beliefs and mathematical convenience. Additionally, Bayesian forensic statistics makes use of approximations, which may be seen as frequentist. It is thus interesting to investigate how good the choice of such priors is. Hence, whether Bayesian or frequentist approaches are chosen, the attempt at producing the likelihood ratio may lead to several levels of uncertainty which should be accounted for.

This subject matter is studied and discussed in the research summarized in Section 2.4.

Different data different likelihood ratios The evaluation of the totality of the data at the expert's disposal is often of difficult fulfilment. This is why often statisticians resort to reducing data to something less informative, but of more feasible evaluation. Especially in presence of many nuisance parameters, it can be wise to discard the part of data which primarily regards the nuisance parameters, and only indirectly regards which hypothesis is

more likely to be true. The more the data is reduced, the easier and more precisely the likelihood ratio for that reduction is estimated. However, each reduction comes with a cost: the stronger the reduction, the less the corresponding likelihood ratio value is helpful to discriminate between the two hypotheses. A compromise has to be made, between the gain in terms of precision and the loss in terms of strength of the evidence. Thus, one has to strike a balance between weakening the likelihood ratio and gaining in the precision of the estimate.

Many different methods to calculate the likelihood ratio are proposed in the literature. They are divided into Bayesian and frequentist, and most of the time they use different reductions of the data. In practice, the different methods are not suggesting different ways to obtain the same likelihood ratio. On the contrary, they are providing different methods to obtain different likelihood ratios: each reduction corresponds to a different likelihood ratio to be estimated. The choice of the reduction is often only implicit and one of the aim of this research is to make explicit the reduction corresponding to the proposed methods.

In the research described in the sections to come, several models for the likelihood ratio assessment in the rare type match case for Y-STR data are proposed, each using a different reduction of the data. The entirety of data to evaluate would be $D = (E, B)$, where E is composed by E_s (suspect's Y-STR hayplotype), E_t (crime stain's Y-STR haplotype, matching with the suspect), and B is a reference database of size n, containing a sample of n Y-STR haplotypes from the population of possible perpetrators. Each method corresponds to a different reduction of the data D, to be evaluated in the light of the two hypotheses h_p ('The crime stain was left by the suspect'), and h_d ('The crime stain was left by someone else').

2.4 Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach)

This section is intended as a summary of the research published in Cereda (2016b), reproduced in full in Chapter 5.

Even though the likelihood ratio – seen as a way to update prior beliefs in the light of new observations – is motivated by Bayes' theorem, it may be of interest for frequentist statisticians as well, as a tool to measure the evidential value of data. With the aim of discussing the last two methodological aspects listed in Section 2.3, we studied and compared two frequentist methods to estimate the likelihood ratio in the rare type match case: the discrete Laplace method (see Section 1.2.5), and a modification of the nonparametric Good-Turing estimator (Good, 1953). Beside representing two interesting solutions to the rare type match problem with Y-STR data, they are also useful to show that:

(i) when the likelihood ratio is defined and estimated in a frequentist way, there are different levels of uncertainty that come into play, which have to be investigated and carefully discussed.

(ii) in a frequentist framework, the data to evaluate can be reduced in several ways, each

entailing the definition of a different likelihood ratio. The reduction is usually performed in order to reduce the error, since likelihood ratios for reduced data are more precisely estimated than likelihood ratios for more data, but it has to be clearly discussed, and justified. It is also important to understand that any reduction comes with a cost.

The discrete Laplace method First proposed in Andersen et al. (2013b), the discrete Laplace method is based on the model assumption that a single locus Y-STR allele is distributed according to the discrete Laplace distribution described in Section 1.2.5. The R package `disclapmix` (Andersen, 2013) allows one to estimate the frequency of any Y-STR haplotype, performing some statistical inference on a limited database. The method estimates the frequencies of unobserved haplotypes as well, hence we decided to apply it to the Y-STR rare type match problem. The data to evaluate is the Y-STR haplotype of the suspect and of the recovered stain, along with the list of Y-STR haplotypes in the available database. The data is not reduced, and the likelihood ratio that we want to estimate is equal to $1/f$, where f is the unknown frequency of the suspect's Y-STR haplotype in the population, to be estimated with the use of the `disclapmix` package. The reciprocal of this estimate is used as an estimate of the likelihood ratio.

The generalized Good method The Good-Turing estimator, first described in the famous paper Good (1953), is a nonparametric nearly unbiased estimator for the total probability of the species which are unseen in a sample of size n, obtained by drawing species independently from the population. This estimator is equal to $\frac{N_1}{n}$, where N_1 is the number of singletons in the sample. We decided to build an estimator of the likelihood ratio for the rare type match problem, based on the Good-Turing estimator. Indeed, by reducing the data to the simple fact that the Y-STR haplotype of the crime stain matches the Y-STR haplotype of the suspect, but they are not in the database (hence, discarding information about the specific Y-STR haplotype of the suspect, and those in the database), the numerator of the likelihood ratio is precisely the probability of observing an unseen species at the $n+1$st observation. The denominator is the probability of observing the same unseen species twice (both in the $n+1$st and $n+2$nd observations). We proved that, as N_1/n is nearly unbiased for the numerator, $\frac{2N_2}{n(n-1)}$ is nearly unbiased for the denominator (at least when n is big enough). Thus, $\frac{N_1 n}{2N_2}$ can be used as estimate for the likelihood ratio. We call this estimator the "generalized Good estimator".

These two methods are a good example of different reductions of the data. The discrete Laplace method allows one to estimate a likelihood ratio that evaluates almost[1] the entirety of the data at disposal, while the generalized Good method is suitable to estimate likelihood ratios that only evaluate part of the data: information about the particular Y-STR haplotypes of the suspect and those in the database are discarded. Hence, the likelihood ratio values obtained with the generalized Good method were expected to be smaller than those obtained with the discrete Laplace method. Thus, before directly comparing the likelihood ratio values obtained with the two estimators, we compared them with the values they aim at estimating.

[1] the database is reduced to count so we loose the information about the order in which data is listed.

Moreover, we discussed and quantified the uncertainty that is involved in each of the two estimates. To do so, for each method we simulated many cases, for which we were able to calculate the true likelihood ratio and the estimates, and we studied the difference of their logarithm. Notice that to actually know the true likelihood ratio it would be necessary to know the entire population. However, the available Y-STR database contains approximately 19,000 haplotypes from 129 different locations in the world (Purps et al., 2014). We consider only 7 loci out of the 23 available, and we pretend that the database contains the entire population of interest for our case.

In the discrete Laplace case, the error is both due to the fact that a specific distribution was chosen to model the Y-STR alleles data (the discrete Laplace), and to the fact that parameters for that distributions are estimated using databases of limited size. For the generalized Good method the entire error is due to the approximation step.

Comparing the distributions of the error is not enough. Indeed, one has to take into account the fact that the generalized Good method discards a lot of data and thus is less useful for the final aim of the evaluation (being able to distinguish the correct hypothesis). One can say that the hallmark desired likelihood ratio is one that evaluates the entirety of the data at disposal, thus $1/f$, hence a comparison with it is also proposed (see Chapter 4, for the details).

2.5 A useful Lemma

In a forensic casework, it is very common that prosecution and defence agree on part of the available information, and disagree on other. For instance, the prosecution may consider the correspondence between the profiles of two DNA traces a sure event (since they came from the same source), while the defence may believe that this correspondence is due to coincidence. Nevertheless, both parts may accept a database as representative of the population of possible perpetrators. Mathematically, one can say that prosecution and defence disagree on the distribution of the evidence, while they agree on the distribution of the background data. This situation is represented by the Bayesian network of Figure 2.4, where node Y stands for the entirety of data at disposal (say the DNA profile of the crime stain, of the suspect and a list of DNA profiles in the form of a database), while node X is the part of data whose distribution is agreed on by prosecution and defence (only the suspect DNA profile and the database). The distributions of X and Y depend on the nuisance parameter(s) represented by A (such as the vector containing the population proportions of all DNA profiles in the population of possible perpetrators).

The Lemma is very useful to simplify the development of the likelihood ratio for the evaluation of X and Y in this common forensic situation, since it is enough to calculate two posterior expectations. In most of the cases of interest X represents everything except the observation of the crime stain (hence, the reference database and the suspect's DNA profile). The probability of the entirety of data given X and the parameter(s) is 1 according to the prosecution, while is a function of the parameter according to the defence.

Lemma. *Given four random variables A, H, X and Y, whose conditional dependencies are represented by the Bayesian network of Figure 2.4, the likelihood function for h, given $X = x$ and $Y = y$ satisfies*

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

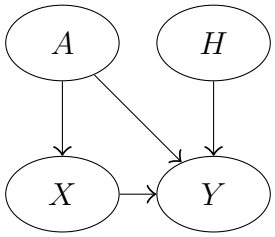

Figure 2.4: Conditional dependencies of the random variables of the Lemma.

Hence the likelihood ratio can be developed as

$$\text{LR} = \frac{p(x, y \mid h_p)}{p(x, y \mid h_d)} = \frac{\mathbb{E}(p(y \mid x, A, h_p) \mid X = x)}{\mathbb{E}(p(y \mid x, A, h_p) \mid X = x)}.$$

This Lemma is used for the development of complex likelihood ratio formulae in Cereda (2016a), Cereda et al. (2016), and Cereda (2016c), where a proof is also given. These publication are reported in full length in Chapters 7, 2.8, and 2.7, respectively.

2.6 Bayesian approach to LR for the rare match problem

This section is intended as a summary of the research published in Cereda (2016a), reproduced in full in Chapter 7.

The first two methodological issues detailed in Section 2.3, prompted us to study the two most common Bayesian solutions used in forensic science, the beta-binomial and Dirichlet-multinomial model. Again, we have a double aim: to customise these models and make them suitable as solution for the rare type match problem, and to show the difference between the widespread plug-in solutions and a proper full Bayesian approach to likelihood ratio assessment, which is obtained by evaluating also background data.

The beta-binomial model for the rare Y-STR haplotype match problem If the data to be evaluated is made of the suspect's and the crime stain's Y-STR haplotypes (E_s and E_c) which correspond, and of a database containing n Y-STR haplotypes from the population of possible perpetrators, the beta-binomial model can be adopted. This amounts to put a beta prior over θ (the unknown population proportion of the suspect's Y-STR haplotype) and to represent the database as a binomial variable B with parameters n and θ. The use of a full Bayesian approach which evaluates both $E = (E_s, E_c)$ and B leads to the following likelihood ratio (details can be found in Chapter 6):

$$\text{LR} = \frac{\alpha + \beta + n + 1}{\alpha + b + 1}. \tag{2.2}$$

where α and β are the shape parameters of the beta prior, and b is the number of times the suspect's Y-STR haplotype is observed in the database.

On the other hand, the plug-in approach would lead to the following estimate for the likelihood ratio:

$$\widehat{\mathrm{LR}} = \frac{\alpha + \beta + n}{\alpha + b}.$$

Sensitivity analysis of the logarithms of these two quantities, when $b = 0$ (i.e., in the case of a rare type match problem) and of their difference, shows that the two quantities depend a lot on α while they do not depend much on β. On the whole, the plug-in estimate $\widehat{\mathrm{LR}}$ is more sensitive than LR to changes in α, and is more anti-conservative (it always exceeds the full Bayesian likelihood ratio). The difference, however, is important only for small values of α. Otherwise, the two methods would lead essentially to the same conclusions, so that the plug-in can be seen as a good approximation of the proper Bayesian procedure.

The Dirichlet-multinomial model for the rare Y-STR haplotype match problem With the same data to evaluate, another possibility is to treat the database as a multinomial sample of size n. For a population with k different Y-STR haplotypes, the conventional choice for the prior over the parameter vector $\boldsymbol{\theta}$, containing the population frequencies of all the different haplotypes in Nature, is the Dirichlet distribution with all the hyperparameters equal to the same value α. The reason for that, is that our prior knowledge about the probabilities of the different categories is invariant to permutations of the categories. Green and Mortera (2009) proposed a similar model, along with an implementation, using an equivalence with a Polya urn scheme to avoid dealing with continuous parameter θ. Their implementation, though, is not directly exploitable for our model, since they assume that k is known. For several applications, k is chosen equal to the number of types in the database, but this is not appropriate for Y-STR haplotypes, since the database usually does not offer a good coverage. We don't even have an idea of a maximal number of categories, so we proposed a modification of the classical Dirichlet-multinomial model, where k is now modelled as a random variable, with a prior $p(k)$.

The full Bayesian likelihood ratio for this model is developed (for α=1) and is equal to

$$\mathrm{LR} = \frac{1}{2}\frac{\displaystyle\sum_{k=k_{obs}+1}^{m} \binom{k}{k_{obs}+1} \frac{\Gamma(k)}{\Gamma(k+N+1)} p(k)}{\displaystyle\sum_{k=k_{obs}+1}^{m} \binom{k}{k_{obs}+1} \frac{\Gamma(k)}{\Gamma(k+N+2)} p(k)}, \tag{2.3}$$

where k_{obs} is the number of distinct Y-STR haplotypes observed in the database. In our research two prior distributions over k are studied: the Poisson distribution and the negative binomial distribution. The full Bayesian likelihood ratio obtained with these priors can be compared to likelihood ratio estimated through the classical Bayesian plug-in method. This, for $\alpha = 1$, is equal to

$$\widehat{\mathrm{LR}} = \bar{k} + N, \tag{2.4}$$

where the number of haplotypes is a fixed value $\bar{k}$, to be chosen (or estimated) in advance. Both using a Poisson prior and a negative binomial prior over k, we decided to perform a

sensitivity analysis of the logarithms of the two likelihood ratios (2.4) and (2.3), and of their difference, when $\bar{k}$ is chosen equal to $\mathbf{E}(K)$.

Results show that, using the Poisson prior, the plug-in approach can be seen as a rather good approximation of the full Bayesian likelihood ratio, while with the negative binomial prior over K the difference between the two can be quite significant, especially if the mean value for K is big. Moreover, for both the priors, the plug-in and the full Bayesian likelihood ratios strongly depend on the hyperparameters, and not much on the data. In particular, they depend on the data only through k_{obs}, the number of distinct observed types, and not on their frequencies. On the other hand, both these quantities depend much on the mean value of K.

This research pointed out the inadequacy of both models (beta-binomial, and Dirichlet-multinomial) for the rare type match problem. Indeed, for both models, the prior over the parameter is chosen for mathematical convenience, rather than because it does represent the expert's belief. This procedure would be sensible only if, at the end, the data overrules the prior choice, whilst this is not the case. This is the reason why we decided to investigate different priors, more realistic for Y-STR data, such as those proposed in the study described in Chapter 7.

2.7 Nonparametric Bayesian approach to LR assessment in case of rare haplotype match

This section is intended as a summary of the research published in Cereda (2016c), reproduced in full in Chapter 7.

The need of priors which better represent Y-STR haplotypes frequencies, prompted us to study new kinds of distributions. In particular, we realized that one of the problems of the method used in Cereda (2016a) (see Section 2.6), is the choice of equal hyperparameters for the Dirichlet prior, since it generates a posterior distribution for θ (after the observation of the multinomial sample) such that the probability of the non observed types is almost uniformly distributed over these types. In fact, looking at the distribution of the frequencies of thousands of Y-STR haplotypes collected all over the world (shown in Figure 2.5), one can realize that there are many types with very rapidly decreasing probabilities, showing a sort of power-law behavior, described in Section 1.4.5. A tempting solution would be to use an asymmetric Dirichlet prior. However, this would lead to extremely difficult computations. The density of the two-parameter Poisson Dirichlet distribution (described in Section 1.4.3) shows a behavior similar to that of Figure 2.5, at least for the smaller probabilities (those of interest in the rare type match case). For this reason the use of this distribution as a prior over the nuisance parameter was investigated. The necessary assumption for this solution is that there are infinitely many Y-STR haplotypes in nature. This assumption, already used by Kimura (1964), is acceptable given the huge number (almost four thousand) of distinct 7 loci Y-STR haplotypes observed in the available database. To use this model, the additional assumption that the specific sequence of STR alleles forming each Y-STR haplotype carries no information was made. This is not realistic, since the distance of Y-STR numbers between

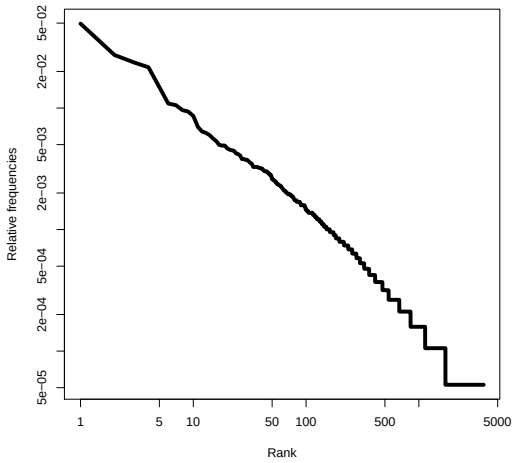

Figure 2.5: Power-law behaviour (logarithmic scale) observed for the ranked frequencies of 7 loci Y-STR haplotypes of Purps et al. (2014).

two haplotypes is an indication of the number of mutation drifts that separate them, but we considered this information too complex to be exploited. Thanks to this assumption, one could reduce by sufficiency the database of size n to a partition of the set $[n]$, obtained by grouping in the same subsets the indexes corresponding to the same haplotypes. According to this model, prosecution and defence agree that this partition is distributed according Pitman's sampling formula (see Section 1.4.2). The entirety of data is made of $n+2$ observed hyplotypes (those in the database and two additional haplotypes from the suspect and from the crime scene trace), and can be reduced to a partition of the set $[n+2]$, where $n+1$ and $n+2$ form a subset by themselves in the rare type match case. Prosecution and defence disagree on the distribution of this random partition since, according to prosecution, elements $n+1$ and $n+2$ are in the same class with probability one.

Using the Chinese Restaurant representation described in Section 1.4.3, and the Lemma presented in Section 2.5, and modelling the hyperparameters α and θ as random variables A and Θ, the likelihood ratio can be written as

$$\mathrm{LR} = \frac{1}{\mathbb{E}(\frac{1-A}{n+1+\Theta} \mid \Pi_{[n+1]} = \pi_{[n+1]})},$$

where $\pi_{[n+1]}$ is the partition obtained enlarging the database with the Y-STR haplotype of the suspect.

Empirical validation showed that choosing a flat hyperprior for the parameters α and θ, the likelihood ratio can be accurately approximated by

$$\mathrm{LR} \approx \frac{n+1+\theta_{MLE}}{1-\alpha_{MLE}}, \tag{2.5}$$

where α_{MLE} and θ_{MLE} are the maximum likelihood estimates of α and θ, respectively. This is equivalent to a plug-in approach where the parameters are estimated through data, and where estimates were plugged into the likelihood ratio. Here this approach is validated by

empirical studies, whose details can be found in Chapter 7. This result was unexpected, but very useful, since it makes the method very practical to be used, despite the complex theoretical background. The paper offers also a comparison between the values obtained with (2.5), and the 'true' likelihood ratio one would obtain knowing the vector $\mathbf{p}$ with the sorted population frequencies of all Y-STR haplotypes in Nature, both using the same reduction of data, and not reducing the data at all. To make this comparison, we built a Metropolis-Hastings algorithm similar to the one proposed in Anevski et al. (2013). This comparison led to the conclusion that the use of the two-parameter Poisson Dirichlet prior is very convenient, and allows one to obtain accurate approximation of the likelihood ratio one would obtain knowing the nuisance parameter $\mathbf{p}$. Additional details can be found in Chapter 7.

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

This section is intended as a summary of the research published in Cereda et al. (2016), reproduced in full in Chapter 8.

The available database of DIP-STR alleles contains only observations from about 100 Swiss individuals. Hence, whenever a new trace is analysed, it is very likely to observe alleles not contained in the database. Not all the solutions proposed and summarized in the previous sections for the Y-STR rare type match problem are appropriate for the DIP-STR case. For instance, the generalized Good method requires a large sample size, while the discrete Laplace and two-parameter Poisson Dirichlet model requires trust in the prior: one would need a sort of empirical validation that cannot be achieved given the small data available. On the other hand, the method proposed in Cereda (2016a), which uses a Dirichlet prior with a random number of categories can be successfully applied, and the model can be developed in a way that does not require the ad hoc plug-in approximations used in Cereda et al. (2014b) for the DIP-STR allelic proportions, and allows one to obtain the full Bayesian likelihood ratio.

The Bayesian network of Figure 2.6, is developed starting from the structure of the object-oriented Bayesian network presented in Cereda et al. (2014b), with two additional nodes, representing the database ($\mathbf{D}$) and the nuisance parameter ($\mathbf{\Theta}$).

The data to evaluate is made of a database containing n DIP-STR alleles from a relevant population, along with the two DIP-STR alleles of the victim, the two DIP-STR alleles of the suspect, and one or two alleles observed from the mixed trace. The hypotheses of interest are the same as those in Section 2.1 regarding whether the second contributor of the stain is the suspect or an unknown (unrelated) individual from the relevant population. We assume that there may be at most $2m$ possible DIP-STR alleles, but only some of them are actually present in the population and even less are those observed in the available database. The nuisance parameter $\boldsymbol{\theta} = (\theta_1^L, ..., \theta_m^L, \theta_1^S, ..., \theta_m^S)$, contains the population proportions of the $2m$ potential DIP-STR alleles.[2] The values corresponding to alleles which are not present

[2] L and S represent "long" and "short" as defined in Section 1.1.5.

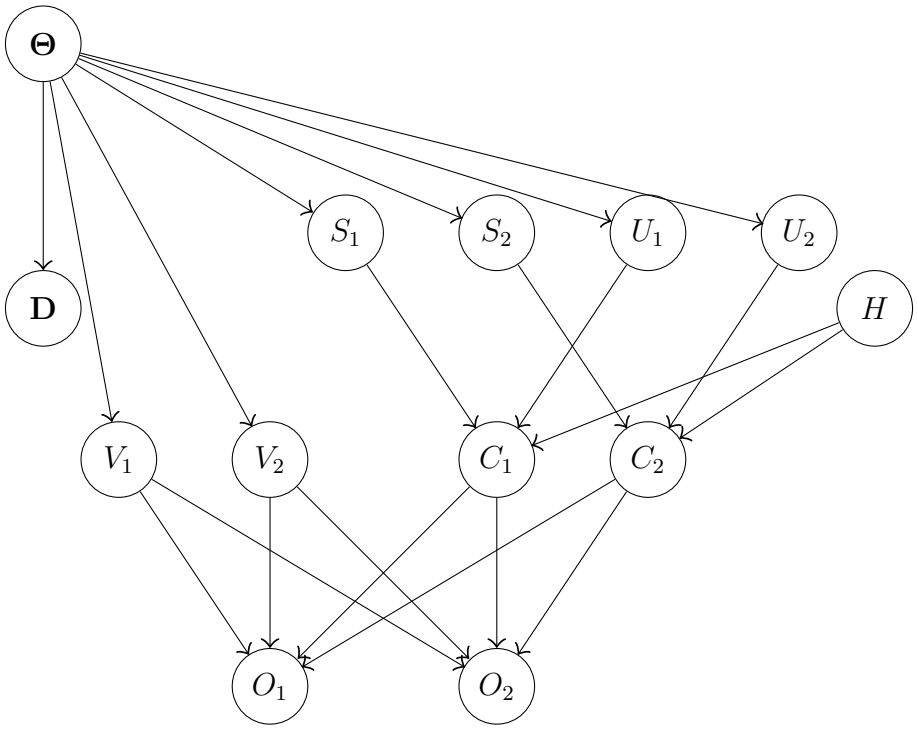

Figure 2.6: Bayesian network for the Dirichlet-multinomial model with a random number of types, to be used for single loci DIP-STR results from mixtures of two contributors. V represents the victim's (or the major contributor's) DIP-homozigozity at the represented locus, with three possible states: $HomoL$, $HomoS$ and $Hetero$. Node H represents the hypotheses h_p and h_d. Nodes S_1 and S_2 represent the two DIP-STR alleles of the suspect, while nodes U_1 and U_2 represent the two DIP-STR alleles of the alternative (unknown) minor contributor. Nodes C_1 and C_2 represent the DIP-STR alleles of the actual minor contributor. Depending on H they can be a copy of S_1 and S_2 (under h_p), or of U_1 and U_2 (under h_d). Nodes O_1 and O_2 contain the DIP-STR alleles observed from the mixed trace. Nodes S_1, S_2, U_1, U_2, C_1, C_2, O_1, and O_2 have states (L,i) (or (S,i)) where $i \in \{1, ..., m\}$ represents the STR part of the DIP-STR allele. Node $\boldsymbol{\theta}$ contains the population proportions of all the potential $2m$ DIP-STR alleles at that locus, (for instance, alphabetically ordered). For more details on the conditional probability distribution of each node, see Chapter 8.

in the population are zero. The prior for $\boldsymbol{\theta}$ is articulated and can be described as follows. The random variable ψ representing the sum of the frequencies of the DIP-STR alleles of type L has a uniform prior, while the normalized vector containing the frequencies of the DIP-STR alleles of type L (obtained dividing each θ_i^L by ψ) has a Dirichlet distribution with all hyperparameters equal to α, with a random number of categories (k^L). The same holds for the normalized vector containing the frequencies of the DIP-STR alleles of type S. k^L and k^S have uniform priors as well.

The Lemma discussed in Section 2.5, and proved in Cereda (2016c), is used to simplify the development of the full Bayesian likelihood ratio. Details can be found in Chapter 8.

The sensitivity analysis conducted on the likelihood ratio values, using data from different DIP-STR markers and different contributors, shows that these values moderately depend on α, at least when a uniform prior is given to k^L, k^S, and ψ. This shows the need of introducing better priors, either less sensitive to changes in α or more realistic, such as the prior used for Y-STR frequencies in Cereda (2016c) (see Section 2.7). Moreover, the full Bayesian likelihood ratio values are compared to those obtained with the plug-in approximation adopted in Cereda et al. (2014b). The values are practically identical.

The drawback of this model is, again, the fact that few information from the database are used in the likelihood ratio. This can be due to the choice of having all hyperparameters equal to α. In order to compensate for this undesired feature, another solution can be adopted. It consists of an hybrid combination with the Good-Turing estimator. Given k^L, and the observation from the database augmented with suspect's and victim's alleles, the posterior expectation of the total probability of the unobserved DIP-STR alleles of type L, is equal to

$$\frac{(k^L - k^L_{\text{obs}})\alpha}{k^L\alpha + n^L}, \tag{2.6}$$

where k^L_{obs} is the number of dinstinct DIP-STR alleles of type L in the augmented database. According to the Good-Turing estimator, the total probability of the unobserved DIP-STR alleles of type L can be estimated by

$$\frac{N_1^L}{n^L}, \tag{2.7}$$

where N_1^L is the number of singletons of type L, while n^L is the total number of alleles of type L in the augmented database. By equating (2.6) and (2.7), we obtain an empirical Bayes estimate of k^L, of the form

$$\hat{k^L} = \frac{N_1^L n^L + k^L_{\text{obs}}\alpha n^L}{\alpha n^L - \alpha N_1^L}.$$

This methods uses more information from the augmented database (namely n_1^L and n_1^S). Additional investigations led to the conclusion that it is a very good approximation to the full Bayesian likelihood ratio.