Stochastic Models in the Identification Process

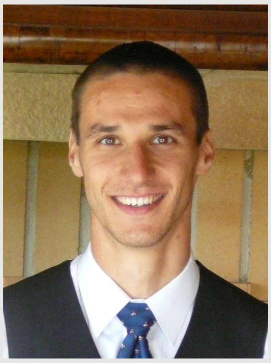
Dalibor Slovák[1], **Jana Zvárová**[1,2]

[1]Center of Biomedical Informatics, Institute of Computer Science AS CR, Prague, Czech Republic

[2]Institute of Hygiene and Epidemiology, First Faculty of Medicine, Charles University, Prague, Czech Republic

Abstract

Objectives: The DNA analysis is now accepted by the broad public as a completely standard and faultless procedure but in some circumstances its reliability can decrease substantially. This paper deals with the process of identifying and determining the weight of evidence against the suspect. Main stochastic approaches to identification are shown.
Methods: The weight-of-evidence formula was derived from Bayes theorem and its application in the model of the island problem was demonstrated. The beta-binomial formula derived from Dirichlet distribution was used for calculation of more complex situations.
Results: From many various complications in the model of the island problem there was shown how to work with uncertainty in a population size. The beta-binomial formula was used to include a subpopulation structure and in issues of DNA mixtures.
Conclusions: In particular, the influence of a population structure is now explored insufficiently. Using the results of H. Kubátová in this area, a new formula was derived.



Mgr. Dalibor Slovák

Keywords

Identification process, Weight-of-evidence formula, Coancestry coefficient, Beta-binomial sampling formula, DNA mixtures

Correspondence to:

Mgr. Dalibor Slovák
Center of Biomedical Informatics,
Institute of Computer Science AS CR
Address: Pod Vodarenskou vezi 2, Prague 8, Czech Republic
E–mail: slovak@euromise.cz

EJBI 2011; 7(1):44–50
recieved: September 20, 2011
accepted: October 24, 2011
published: November 20, 2011

1 Introduction

DNA profiling, discovered by Alec Jeffreys during the 1980s, has caused a revolution in criminology. DNA helps to convict the perpetrators of those crimes that once appeared irresolvable and also helps to prove the innocence of those who have already been convicted. The DNA analysis is now accepted by the broad public as a completely standard procedure, which reliably convicts the offender. Here, however, hides one of the main problems that results from using DNA, for even DNA evidence is not foolproof.

Several possibilities keep DNA from being completely reliable: for example there may be a false location of the trace (more specifically, the offender may have discarded a cigarette butt which had previously been smoked by someone else); the wrong take of a biological samples or damage to the samples could have occurred; or there may have been secondary transfer of a biological material. However, mathematicians do not deal with any of these things. Rather, they are faced with the following task: if all of the above options are excluded, what is the probability that a particular offender and a detained person are the same, given that the perpetrator's and the suspect's DNA profiles are available? As we will see, the answer depends mainly on the number of loci we use to DNA profiling, and the variability within each of them.

In forensic practice, genetic profiles consisting of the short tandem repeat polymorphisms (STRs) are currently used. STRs are known to vary widely between individuals by virtue of variation in their length and they are found only in the non-coding region of DNA, so they provide no information of medical or personal significance. Therefore, STRs are very useful and convenient for identification purposes.

The numerical representation of DNA profile consists of two numbers of alleles at each locus examined, one allele inherited from the mother and the other one from the father, along with two letters (XX or XY) which show the result of the gender test. The number of examined loci varies from country to country, with the smallest being seven used in Germany and a maximum of sixteen used in the Czech Republic.

For example, the system of DNA profiling used in the UK is known as SGM Plus. It examines ten loci plus a gender test and produces a numeric DNA profile which may look like this:

```
15,18; 6,9; 11,13; 22,22; 31,32.2; 14,17;
17,20; 11,12; 13,16.3; 15,16; XY.
```

The number provides information about a feature of DNA at each locus we examine. The number of complete repeat units observed is designated by an integer. Variant alleles that contain a partial repeat are designated by a decimal followed by the number of bases in the partial repeat. For example, an 32.2 allele contains 32 complete repeat units and a partial repeat unit of 2 bases ([10]).

Although each person's DNA is unique (apart from identical siblings), there is a very small, but finite chance (less than 1 in a billion in SGM Plus) that two unrelated people could share the same DNA profile. For this reason it is not possible to convict a person on DNA evidence alone and there must be additional corroborating evidence available.

DNA left at a crime scene may also decompose over time because of bacteria, UV light, environmental conditions etc. Due to the quality of biological material and/or its amount it is not always possible to investigate all of the polymorphisms. An incomplete DNA profile may look like

```
15,  ; 6,9; 11,13;   ,  ; 31,32.2; 14,17;
  ,20;   ,12; 13,16.3; 15,16; XY.
```

If an incomplete DNA profile is obtained, the probability of unique identification drops accordingly. However, even very incomplete profiles can still be used to conclusively eliminate a person from an investigation.

In the following text we will assume the examination of one locus only. Assuming independence of loci, generalization to a larger number of loci can be performed using a product rule (i.e. multiplying the individual marginal probabilities).

2 Methods

Denotation

- E - evidence or information about the crime (i.e. the circumstances, witness testimonies, crime scene evidence, etc.),
- G - an event at which the suspect is guilty,
- I - an event at which the suspect is innocent,
- C_i - an event at which the culprit is a person i,
- $\mathcal{I}$ - the population of alternative suspects.

Our goal is to determine the conditional probability of $\mathsf{P}(G|E)$ that, given circumstances E, the suspect is truly the culprit of the investigated crime. According to Bayes theorem

$$\mathsf{P}(G|E) = \frac{\mathsf{P}(E|G)\mathsf{P}(G)}{\mathsf{P}(E|G)\mathsf{P}(G) + \mathsf{P}(E|I)\mathsf{P}(I)}. \tag{1}$$

However, the expression $\mathsf{P}(E|I)$ cannot be counted directly. The suspect is innocent if and only if there exists an index $i \in \mathcal{I}$ in which the event C_i occurs. Then the event I is equivalent to the event $\cup_{i\in\mathcal{I}} C_i$ and thanks to the disjunction of events C_i holds:

$$\mathsf{P}(I) = \mathsf{P}\left(\cup_{i\in\mathcal{I}} C_i\right) = \sum_{i\in\mathcal{I}} \mathsf{P}(C_i).$$

Thus

$$\begin{aligned}
\mathsf{P}(E|I)\mathsf{P}(I) &= \mathsf{P}\left(E|\cup_{i\in\mathcal{I}} C_i\right)\mathsf{P}\left(\cup_{i\in\mathcal{I}} C_i\right) = \\
&= \frac{\mathsf{P}\left(E \cap \left(\cup_{i\in\mathcal{I}} C_i\right)\right)}{\mathsf{P}\left(\cup_{i\in\mathcal{I}} C_i\right)}\mathsf{P}\left(\cup_{i\in\mathcal{I}} C_i\right) = \\
&= \mathsf{P}\left(\cup_{i\in\mathcal{I}} (E \cap C_i)\right) = \sum_{i\in\mathcal{I}} \mathsf{P}\left(E \cap C_i\right) = \\
&= \sum_{i\in\mathcal{I}} \mathsf{P}(E|C_i)\mathsf{P}(C_i).
\end{aligned}$$

Let define **likelihood ratio**

$$R_i = \frac{\mathsf{P}(E|C_i)}{\mathsf{P}(E|G)} \tag{2}$$

which expresses how many times the probability of evidence E is greater under the condition that the culprit is a person i than under the condition that the culprit is the suspect.

Further we define **likelihood weights**

$$w_i = \frac{\mathsf{P}(C_i)}{\mathsf{P}(G)}$$

which expresses how many times the prior probability of committing a crime by a person i is greater than the prior probability of committing a crime by the suspect.

Then

$$\mathsf{P}(G|E) = \frac{1}{1 + \sum_{i\in\mathcal{I}} w_i R_i}. \tag{3}$$

The formula (3) is usually called **the weight-of-evidence formula**.

3 The Island Problem

The simplest application of the previous part is the "island problem". This is a model where a crime is committed on an inaccessible island which contains N people who are unrelated to each other. At the beginning, there is no information about the offender, so we assign to each of the islanders the same (prior) probability of committing a crime. Then the offender is found to possess a certain characteristic Υ (it can be an allele, or a pair of alleles respectively, in the appropriate locus) and the suspect is also found to have that characteristic, Υ. The question becomes, to what extent can we be sure that we have found the suspect who is truly the culprit.

First we calculate the likelihood ratio using the formula (2). Let p be the frequency of the Υ in the population. We suppose that the evidence consists only of the information that the suspect's and the culprit's DNA profiles are the same. If the hypothesis G holds, both profiles come from the same individual and thus the denominator equals 1. The numerator of R_i, $\mathsf{P}(E|C_i)$, can be estimated by p. Because $w_i = 1\ \forall i \in \mathcal{I}$, using the formula (3) we get

$$\mathsf{P}(G|E) = \frac{1}{1 + N \cdot p} . \tag{4}$$

For example if $p = 0.01$ and $N = 100$ then $\mathsf{P}(G|E) = 1/2$.

The previous result can be modified for more complex (and realistic) situations. Let's see which situations is this simple model inadequate for:

- *Typing and handling errors*
  As the test may give erroneous results in a small percentage of cases, errors caused by a human factor must also be considered: contamination or replacement of a sample from which the Υ-status is investigated; incorrect evaluation of the results, or even intentional misrepresentation.

- *The population size*
  Often the population size N is only estimated and furthermore, if there is migration in the population, then it is necessary to account for greater uncertainty within the population size.

- *The probability of occurrence Υ in the population*
  The value of p is usually unknown and is therefore estimated on the basis of relative frequency of the Υ in a smaller sample or in a similar population, about which we have more information. However, this auxiliary data may be outdated or may only partially describe the ivestigated population.

- *Suspect searching*
  The suspect is not usually chosen randomly from the population but on the basis of other circumstantial evidence which increase the probability of guilt. Another possibility is to choose the suspect by testing individuals from the population for the presence of Υ. In this way, people who are not Υ-bearers can be excluded and thus the population size of alternative suspects is reduced.

- *Relatives and population subdivision*
  If the suspect (or other individual being tested) is a Υ-bearer and some of his relatives are included in the population too, then in the case of DNA profile increases the probability of other individuals having Υ due to inheritance. Similarly, unusually high relative frequency of a rare character usually occurs within the same subpopulation due to its shared evolution history.

- *The same prior probability of committing a crime*
  Although this requirement intuitively corresponds with the general presumption of innocence, we can asses varying prior probability (i.e. based on the distance from the scene, time availability, or a possible alibi).

We will analyze some of these cases in detail in the following sections.

4 Uncertainty about the Population Size

The uncertainty about the size of the population of possible alternative suspects affects the prior probability of $\mathsf{P}(G)$. Consider the population size $\tilde{N}$ is a random variable with mean N. The prior probability of guilt, given value $\tilde{N}$, is

$$\mathsf{P}(G|\tilde{N}) = 1/(\tilde{N} + 1)$$

but since $\tilde{N}$ is not known, we use the expectation:

$$\mathsf{P}(G) = \mathsf{E}\left[G|\tilde{N}\right] = \mathsf{E}\left[\frac{1}{\tilde{N} + 1}\right].$$

The function $1/(\tilde{N} + 1)$ is not symmetric but it is at least convex on the interval $(0, \infty)$. Jensen's inequality for convex functions ($\mathsf{E}[f(x)] \geq f(\mathsf{E}[x])$) implies

$$\mathsf{P}(G) = \mathsf{E}\left[\frac{1}{\tilde{N} + 1}\right] \geq \frac{1}{N + 1}$$

because $\mathsf{E}[\tilde{N}] = N$.

Thus the failure to uncertainty about the value of N tends to favor defendant. Moreover, this effect is usually very small, let it show in a concrete example.

For $\varepsilon \in (0, 0.5)$ we put

$$\tilde{N} = \begin{cases} N - 1 & \text{with probability } \varepsilon \\ N & \text{with probability } 1 - 2\varepsilon \\ N + 1 & \text{with probability } \varepsilon. \end{cases}$$

Then

$$\mathsf{P}(G) \quad = \quad \mathsf{E}\left[\frac{1}{\tilde{N} + 1}\right] = \frac{\varepsilon}{N} + \frac{1 - 2\varepsilon}{N + 1} + \frac{\varepsilon}{N + 2} =$$

$$= \frac{1}{N+1} + \frac{2\varepsilon}{N(N+1)(N+2)} \geq \frac{1}{N+1}$$

and if we put $\varepsilon = 0.25$ and $N = 100$ then $\mathsf{P}(G)$ is greater than $1/(N+1)$ by only 0.000000485.

Let's see what the population size uncertainty causes in formula (4):

$$\begin{aligned}
\mathsf{P}(G|E) &= \frac{1}{1 + \sum_i R_i \frac{\mathsf{P}(C_i)}{\mathsf{P}(G)}} = \frac{1}{1 + p\frac{1}{\mathsf{P}(G)} \underbrace{\sum_i \mathsf{P}(C_i)}_{=1-\mathsf{P}(G)}} = \\
&= \frac{1}{1 + p\frac{N(N+1)(N+2)}{N^2+2N+2\varepsilon}\left(1 - \frac{N^2+2N+2\varepsilon}{N(N+1)(N+2)}\right)} = \\
&= \frac{1}{1 + Np\frac{N^3+2N^2-2\varepsilon}{N^3+2N^2+2N\varepsilon}} = \\
&= \frac{1}{1 + Np\left(1 - 2\varepsilon\frac{N+1}{N^3+2N^2+2N\varepsilon}\right)}.
\end{aligned}$$

Substituting again $\varepsilon = 0.25$ and $N = 100$ we receive $\mathsf{P}(G|E) = 0.5000124$ which value, despite the high value of ε, differs from the original result of 50 %, at which we calculate with N fixed, in an order of just one thousandth of a percent. If we want to still count with uncertainty about N,

$$\mathsf{P}(G|E) \approx \frac{1}{1 + Np\,(1 - 2\varepsilon/N^2)}$$

is very good approximation to take. In our example this approximation gives $\mathsf{P}(G|E) = 0.5000125$, i.e. 50.00125 %.

Balding in [1] uses an approximation order of magnitude worse than

$$\mathsf{P}(G|E) \approx \frac{1}{1 + Np\,(1 - 4\varepsilon/N^3)}$$

which gives in our example the value $\mathsf{P}(G|E) = 0.5000003$, i.e. 50.00003 %.

5 DNA Database

DNA profiles as a sequence of alphanumeric data allow relatively easy storage in the database, therefore national databases are created from late 1990's. Currently there are three major forensic DNA databases: CODIS (Combined DNA Indexing System), which is maintained by the United States FBI; the ENFSI (European Network of Forensic Science Institutes) database; and the ISSOL (Interpol Standard Set of Loci) database maintained by Interpol.

All systems mentioned above divide the DNA database into two subdatabases. In *the crime scene database* biological samples collected at the scene are stored, in *the convicted offender database* figure genetic profiles of individuals convicted in the past. These two databases are compared with each other and eventual match of profiles is examined by qualified professionals.

The type of offenses for which DNA is stored differs among countries and states. Initially, these databases contained only samples from violent offenders, those convicted of aggravated assault, rape, or murder. However, the value of obtaining DNA from offenders of less severe crimes has been recognized, as many small time criminals become repeat offenders and also more violent offenders. The power of a large bank of DNA samples extends to the possibility of it acting as a deterrent. A match of DNA evidence from a crime scene (which would then be logged in the crime scene database) to one in the convicted offender database rapidly solves the crime, saving time, effort, and money ([3]).

The absolutely largest national database is the US National DNA Index System (NDIS). It contains almost ten million offender profiles and over 380 000 forensic profiles as of July 2011 ([7]). The oldest and relatively largest database is the national DNA database of UK (NDNAD) which currently consists of over six and a half million profiles.

After the creation of DNA databases the number of solved crimes in the UK has increased from 24 % to 43 %. The success of this approach is also confirmed by the fact that a new crime scene DNA profile being loaded to the DNA database had a 45 % chance of matching a persons DNA profile in 2002/03 against 60 % in 2008/09 ([8]). Thus the database system has the support of public. On the other hand, from DNA very sensitive personal information can be obtained and therefore it is necessary to ensure a thorough protection of databases against abusing.

The Czech national database was created in 2002. Then there was a rapid development of the database and it currently contains approximately 90 000 genetic profiles.

6 Relatives and Population Structure

Alleles, which are identical and come from a common ancestor, are called *ibd* (identical by descent). A common recent evolution history of two individuals, whether relatives or members of the same subpopulation, increases the probability of occurrence of ibd alleles. Therefore, as the degree of relatedness within subpopulations is used, the *coancestry coefficient* θ indicated the probability that two randomly selected alleles on fixed locus are ibd. Neglecting the influence of kinship and population structure leads to overestimation of posterior probability of the suspect's guilt. Ignoring this tends to suspect's disfavour, so this topic is given considerable attention.

Balding and Nichols in [2] proposed a method which allows to calculate probability of observing considered genotype in structure population via coancestry coefficient. More detailed mathematical derivation of method including several corrections was provided by Kubátová in [6].

Let's denote p_A and p_B frequencies of alleles A a B in the whole population, k proportion of the subpopulation in the general population and θ coancestry coefficient in the subpopulation. The probability of observing homozygous genotype can be calculated as

$$\mathsf{P}(AA) = p_A \left(\theta + (1-\theta)\,\frac{p_A - \theta k}{1-\theta k} \right) \qquad (5)$$

and similarly heterozygous genotype:

$$\mathsf{P}(AB) = 2 p_A p_B \frac{1-\theta}{1-\theta k}\,. \qquad (6)$$

Balding and Nichols do not use variable k in their derivation, we get their results by putting $k = 1$. Thus, probabilities of homozygous genotypes decreased and conversely, probabilities of heterozygous genotypes increased.

7 Beta-binomial Formula

To get formulas (5) and (6), we can use also a more general approach proposed by Wright ([11]). Consider on given locus J alleles $A_1, \ldots, A_J$ having probability of occurrence in the population $p_1, \ldots, p_J$, $\sum_{i=1}^{J} p_i = 1$. Allele proportions in the subpopulation can be modelled by the Dirichlet distribution with parametres $\lambda p_i, \lambda = \frac{1-\theta}{\theta(1-k)}$. Thus the probability of observing m_i alleles A_i ($\sum_i m_i = n$) is given by

$$\mathsf{P}(m_1, \ldots, m_J) = \frac{\Gamma(\lambda)}{\Gamma(\lambda+n)} \prod_{i=1}^{J} \frac{\Gamma(\lambda p_i + m_i)}{\Gamma(\lambda p_i)}. \qquad (7)$$

Putting $m = (m_1, \ldots, m_J)$ we can adjust formula (7) to

$$\mathsf{P}(m) = \frac{\prod\limits_{j=1}^{J} \prod\limits_{i=0}^{m_j-1} \left[(1-\theta)\,p_j + \theta i\,(1-k) \right]}{\prod\limits_{i=0}^{n-1} \left[1-\theta+\theta i\,(1-k) \right]}. \qquad (8)$$

The formula (8) is usually called **beta-binomial sampling formula** and applies to ordered samples. If we want to work with unordered samples, it is necessary to multiply the result by $\frac{n!}{m_1! \cdots m_J!}$.

From the formula (8) we can also deduce the probability of observing certain combination of alleles. For $J = 2$, $m_A = 2$ and $m_B = 0$ we have

$$\begin{aligned}
\mathsf{P}(AA) &= \frac{(1-\theta)\,p_A \left[(1-\theta)\,p_A + \theta\,(1-k) \right]}{(1-\theta) \left[1-\theta+\theta\,(1-k) \right]} = \\
&= p_A \left[\frac{(1-\theta)\,p_A + \theta - \theta k}{1-\theta k} + \theta - \frac{\theta - \theta^2 k}{1-\theta k} \right] = \\
&= p_A \left[\theta + \frac{(1-\theta)\,p_A + \theta - \theta k - \theta + \theta^2 k}{1-\theta k} \right] = \\
&= p_A \left[\theta + \frac{(1-\theta)\,p_A - \theta k\,(1-\theta)}{1-\theta k} \right] = \\
&= p_A \left[\theta + (1-\theta)\,\frac{p_A - \theta k}{1-\theta k} \right],
\end{aligned}$$

which is in agreement with (5).

Similarly putting $J = 2$, $m_A = 1$ and $m_B = 1$ in the formula (8) we get

$$\mathsf{P}(AB) = 2 \frac{(1-\theta)\,p_A\,(1-\theta)\,p_B}{(1-\theta)\,(1-\theta+\theta\,(1-k))} = 2 p_A p_B \frac{1-\theta}{1-\theta k},$$

which agrees with formula (6).

8 Aplication of Beta-binomial Formula

Using the formula (8) we can also deduce the probability of observing certain allele given by our knowledge of previous alleles observing:

$$\mathsf{P}(m_j{+}1 | m_1, \ldots, m_j, \ldots, m_J) = \frac{(1-\theta)\,p_j + m_j \theta\,(1-k)}{1-\theta+n\theta\,(1-k)}. \qquad (9)$$

Denote G_C and G_S genotype of culprit and suspect respectively, and generally G_i genotype of person i. Likelihood ratio (2) can be rewritten as

$$R_i = \frac{\mathsf{P}\,(G_C = G_S = D | C_i)}{\mathsf{P}\,(G_C = G_S = D | G)} = \frac{\mathsf{P}\,(G_i = G_S = D)}{\mathsf{P}\,(G_S = D)} =$$

$$= \mathrm{P}\,(G_i = D | G_S = D)\,.$$

Suppose first that the suspect has a homozygous profile $A_j A_j$ and with this knowledge calculate the probability that the suspect has the same homozygous profile:

$$\begin{aligned}
R_i &= \mathsf{P}(G_i = A_j A_j | G_S = A_j A_j) \equiv \mathsf{P}(A_j^2 | A_j^2) = \\
&= \mathsf{P}(A_j | A_j^3) \cdot \mathsf{P}(A_j | A_j^2)
\end{aligned}$$

We know how to calculate these conditional probabilities by using (9). First we put $m_j = n = 2$ and then $m_j = n = 3$, that is

$$R_i = \frac{\left[(1-\theta)\,p_j + 2\theta\,(1-k) \right] \left[(1-\theta)\,p_j + 3\theta\,(1-k) \right]}{\left[1-\theta+2\theta\,(1-k) \right] \left[1-\theta+3\theta\,(1-k) \right]}.$$

Similarly we proceed for a heterozygous profile $A_j A_k$:

$$\begin{aligned}
R_i &= \mathsf{P}(G_i = A_j A_k | G_S = A_j A_k) \equiv \mathsf{P}(A_j A_k | A_j A_k) = \\
&= \mathsf{P}(A_k | A_j^2 A_k^1) \mathsf{P}(A_j | A_j^1 A_k^1) + \\
&\qquad + \mathsf{P}(A_j | A_j^1 A_k^2) \mathsf{P}(A_k | A_j^1 A_k^1).
\end{aligned}$$

To quantify both expressions on the bottom line we put $m_j = 1, n = 2$ and $m_k = 1, n = 3$; $m_k = 1, n = 2$ and $m_j = 1, n = 3$ respectively. In total

$$R_i = 2 \frac{\left[(1-\theta)\,p_j + \theta\,(1-k) \right] \left[(1-\theta)\,p_k + \theta\,(1-k) \right]}{\left[1-\theta+2\theta\,(1-k) \right] \left[1-\theta+3\theta\,(1-k) \right]}.$$

9 DNA Mixtures

If the DNA sample is found to have more than two alleles at one locus, it is clear to be a mixture. The number of contributors to the mixture can be known or estimated, usually as $\left\lceil \frac{n}{2} \right\rceil$ where n is the maximum number of alleles detected. Because of the large number of situations that may have arised we show for illustration only the case when the victim (V) and one other individual contribute to the mixture.

The likelihood ratio R_i defined by formula (2) can be rewritten as

$$\begin{aligned} R_i &= \frac{\mathsf{P}\left(E_C, G_S, G_V | C_i\right)}{\mathsf{P}\left(E_C, G_S, G_V | G\right)} = \\ &= \frac{\mathsf{P}\left(E_C | G_S, G_V, C_i\right)}{\mathsf{P}\left(E_C | G_S, G_V, G\right)} \cdot \frac{\mathsf{P}\left(G_S, G_V | C_i\right)}{\mathsf{P}\left(G_S, G_V | G\right)} = \\ &= \frac{\mathsf{P}\left(E_C | G_S, G_V, C_i\right)}{\mathsf{P}\left(E_C | G_S, G_V, G\right)} = \frac{\mathsf{P}\left(E_C | G_V, C_i\right)}{\mathsf{P}\left(E_C | G_S, G_V, G\right)} . \end{aligned} \tag{10}$$

Four alleles mixture

First we look at the case where the mixture consists of four alleles.

Suppose the following conditions apply:

1. None of the individuals are considered relatives to each other.
2. The population is homogeneous (i.e. $\theta = 0$).
3. The population follows Hardy-Weinberg equilibrium.

Let the mixture be made up of alleles A, B, C, D with known probabilities of occurrence in the total population p_A, p_B, p_C, p_D and let the suspect have alleles A, B and the victim C, D respectively. The denominator in the formula (10) is equal to one, the numerator is equal to the probability of observing the individual with alleles A, B which is under the above assumptions $2p_A p_B$. Therefore, the likelihood ratio equals to

$$R_i = 2p_A p_B .$$

Suppose now that all considered individuals have the same degree of relatedness to each other expressed by coancestry coefficient θ. Then according to (9)

$$\begin{aligned} R_i &= \mathsf{P}\left(AB | ABCD\right) = \\ &= \frac{2\left[\left(1-\theta\right)p_A + \theta\left(1-k\right)\right]\left[\left(1-\theta\right)p_B + \theta\left(1-k\right)\right]}{\left[1-\theta+4\theta\left(1-k\right)\right]\left[1-\theta+5\theta\left(1-k\right)\right]} . \end{aligned}$$

Three alleles mixture

In the case of three alleles in the sample, assuming at least two contributors to the mixture is also necessary. Consider alleles A, B, C with probabilities p_A, p_B, p_C. If the victim is homozygous for allele C, we get the same results as in the case of a mixture of four alleles.

Let's assume that the victim is heterozygous with alleles A, B. Let the suspect be homozygous for allele C and conditions 1 to 3 are fulfilled. The denominator of the formula (10) is again equal to one, the numerator equals to the probability of observing an individual who has the allele C and does not have a different allele than A, B or C. Therefore

$$\begin{aligned} R_i &= \mathsf{P}(AC) + \mathsf{P}(BC) + \mathsf{P}(CC) = \\ &= 2p_A p_C + 2p_B p_C + p_C^2 . \end{aligned} \tag{11}$$

To include the population structure we use the formula (9) again:

$$\begin{aligned} R_i &= \mathsf{P}\left(AC | ABCC\right) + \mathsf{P}\left(BC | ABCC\right) + \\ &\quad + \mathsf{P}\left(CC | ABCC\right) = \\ &= \frac{2\left[\left(1-\theta\right)p_A + \theta\left(1-k\right)\right]\left[\left(1-\theta\right)p_C + 2\theta\left(1-k\right)\right]}{\left[1-\theta+4\theta\left(1-k\right)\right]\left[1-\theta+5\theta\left(1-k\right)\right]} \\ &+ \frac{2\left[\left(1-\theta\right)p_B + \theta\left(1-k\right)\right]\left[\left(1-\theta\right)p_C + 2\theta\left(1-k\right)\right]}{\left[1-\theta+4\theta\left(1-k\right)\right]\left[1-\theta+5\theta\left(1-k\right)\right]} \\ &+ \frac{\left[\left(1-\theta\right)p_C + 3\theta\left(1-k\right)\right]\left[\left(1-\theta\right)p_C + 2\theta\left(1-k\right)\right]}{\left[1-\theta+4\theta\left(1-k\right)\right]\left[1-\theta+5\theta\left(1-k\right)\right]} \\ &= \frac{\left[\left(1-\theta\right)p_C + 2\theta\left(1-k\right)\right]}{\left[1-\theta+4\theta\left(1-k\right)\right]} \times \\ &\quad \times \frac{\left[\left(1-\theta\right)\left(2p_A + 2p_B + p_C\right) + 7\theta\left(1-k\right)\right]}{\left[1-\theta+5\theta\left(1-k\right)\right]} . \end{aligned}$$

We assumed in the previous calculation that the suspect is homozygous for allele C. If he is heterozygote with alleles A and C, or B and C respectively, formula (11) remains unchanged under conditions 1 to 3. If the population structure is included, we get in both cases the likelihood ratio

$$\begin{aligned} R_i &= \frac{\left[\left(1-\theta\right)p_C + \theta\left(1-k\right)\right]}{\left[1-\theta+4\theta\left(1-k\right)\right]} \times \\ &\times \frac{\left[\left(1-\theta\right)\left(2p_A + 2p_B + p_C\right) + 8\theta\left(1-k\right)\right]}{\left[1-\theta+5\theta\left(1-k\right)\right]} . \end{aligned}$$

10 Conclusion

We derived the weight-of-evidence formula and its simplest applications. To include the uncertainty about the population size we proposed a better approximation than Balding in ([1]). We showed how to include the subpopulation structure into the model. Here we used new results from ([6]) which we plan to investigate in more detail in future.

Acknowledgements

This work was partially supported by the project 1M06014 of the Ministry of Education, Youth and Sports of the Czech Republic and by the project SVV-2011-262514 of Charles University in Prague.

References

[1] Balding D.J.: *Weight-of-evidence for forensic DNA profiles*, John Wiley & Sons, Ltd, 2005, pp. 15-63

[2] Balding D.J., Nichols R.A.: *DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands*, Forensic Science International **64**, 1994, pp. 125-140

[3] eNotes. World of Forensic Science. *DNA Evidence, Social Issues* [online]. 2011 [cit. 2011-9-15]. Available at www.enotes.com/forensic-science/dna-evidence-social-issues.

[4] Slovák Dalibor: *Stochastic Approaches to Identification Process in Forensic Medicine and Criminalistics*, in Doktorandské dny '11, Matfyzpress, Praha, 2011

[5] The office for personal data protection. *Otevřete ústa, prosím... & Databáze DNA* [online, in czech]. February 2007 [cit. 2011-9-15]. Available at www.uoou.cz/uoou.aspx?menu=287 &submenu=288.

[6] Kubátová H., Zvárová J. (supervisor): *Statistical methods for interpreting forensic DNA mixtures*, MFF UK, Praha 2010, pp. 20-26

[7] The Federal Bureau of Investigation. *CODISNDIS Statistics* [online]. July 2011 [cit. 2011-9-15]. Available at www.fbi.gov /about-us/lab/codis/ndis-statistics.

[8] The National Policing Improvement Agency. *The National DNA atabase* [online]. 2010 [cit. 2011-9-15]. Available at www.npia.police.uk/en/8934.htm.

[9] Slovák D., Zvárová J. (supervisor): *Statistické metody stanovení váhy evidence v procesu identifikace jedince*, MFF UK, Praha, 2009

[10] The Applied Biosystems. *AmpFℓSTR SGM Plus. PCR Amplification Kit. User's Manual* [online]. 2011 [cit. 2011-9-15]. Available at www3.appliedbiosystems.com/cms/groups /applied_ markets_support/documents/generaldocuments /cms_041049.pdf, pp. 178.

[11] Wright S.: *The genetical structure of populations*, Ann. Eugen. **15**, 1951, pp. 323-354