Dipartimento di Statistica
"Giuseppe Parenti"

W O R K I N G P A P E R 2009/13

Evaluation of indirect identification systems based on STR DNA Profiles

Fabio Corradi

Università degli Studi
di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Evaluation of indirect identification systems based on STR DNA Profiles

Fabio Corradi

University of Florence - Department of Statistics,
59, Viale Morgagni, Florence 50134 Italy;
email: corradi@ds.unifi.it

Abstract

In this paper we describe an approach to assess the potentiality of a probabilistic system for indirect identification through DNA evidence. The aim is, before an identification is attempted, to provide, the expected performances of the system and help all the parts involved to judge if the system fits their requirements. The analysis provided considers the probabilities associated to the weight of evidence related to the case, also suggesting possible strategies to improve the system. A detailed case study is illustrated as well as some possible information theoretic measures able to summarize the information capabilities of the DNA markers employed in a specific indirect identification.

Keywords: Indirect Identification, DNA STR markers, Weight of Evidence distribution, Information decay, Information theoretic measures.

1 Introduction

Nowadays, the identification of individuals obtained trough DNA evidence has definitely to be considered out of its infancy. Actually, reliable kits of primers [1] allow to determine inexpensively the genotypes of biological traces typed on some STR loci. Moreover, in many circumstances, the probabilities required to evaluate the hypotheses relevant for a case can be obtained by means of software available for free (*Familias* , http://www.math.chalmers.se/ mostad/familias/), or at a reasonable cost (*DNA-View*, http://dna-view.com/).

The identification issue is addressed by comparing the evidence probability expressed conditionally to a pair of competitive identification hypotheses meaningful for the case. In forensic science their ratio is often called the Weight of evidence (WE), Balding (2006) whereas, in Statistics, according to different inferential approaches, the same quantity is named the Likelihood ratio or the Bayes factor.

[1] In this work we have used results obtained through the *PrepFiler (TM) Forensic DNA extraction Kit* produced by the Applied Biosystem

This paper considers the potential identification capabilities of probabilistic models built to cope with specific identification cases. The focus is on indirect identifications, a wide class of problems including issues such as disputed paternities, searching missing persons, the family reunification of citizens of foreign birth and permanent resident aliens and many others.

It is commonly acknowledged that an identification model reduces is ability to provide strong support to one of the hypotheses according to the "distance" between the persons who require the identification and the position in the pedigree of the searched individual. Common sense also suggests that more conclusive results are expected if the family donors' characteristics are rare.

Despite the expected differences in the system's performances according to the specific case, the WE is almost invariably evaluated making use of those loci included in the kit of primers adopted by each laboratory, only considering the family members more easily available to provide their DNA profile, but not taking into account the expected performances of the resulting identification tool.

The aim of this paper is to formalize how to measure the ability of a system to investigate on the specific identification, to give suggestions on how to improve its performances, and to illustrate the matter trough an articulated case study.

2 Probability models for DNA evidence

To evaluate the probability of observed DNA evidence, consider the individuals implied in the analysis with respect to some nuclear STR DNA loci, among those commonly used for forensic identification.

In a locus we observe a genotype, i.e. two alleles inherited from each of the parents. In a population and in a specific locus the possible observable alleles are generally known, $\mathcal{A} = \{a_1, ..., a_k\}$ and so is the probability to observe them, $Pr(A = a_i = \theta_i, \forall i$. The random variable X represents the uncertainty about genotypes and a determination of X is simply indicated by x. The X sample space $\mathcal{X}$ is constituted by $k(k+1) \cdot 0.5$ different genotypes. The genotype's probability distribution is provided by two kinds of models.

Segregation models: for a locus, they evaluate the probability of an offspring's genotypes conditionally to their parents. The first Mendelian law is the basic model to specify the genotype' s probability of a child, c, given the genotypes of their parents, m and f. If $x_c = (t, z)$, $x_m = (i, j)$ and $x_f = (r, s)$, we have:

$$Pr(x_c \mid x_m, x_f) = \frac{1}{4}(I_{\{i,r\}}(x_c) + I_{\{i,s\}}(x_c) + I_{\{j,r\}}(x_c) + I_{\{j,s\}}(x_c)). \qquad (1)$$

If mutations or laboratory errors are involved, other more sophisticated models are required to describe the segregation process as described in Dawid et al. (2007).

Population models: they determine the probability of an individual's genotype conditionally to their belonging to a specified population in which the

alleles' probabilities, θ, are assumed known. The most popular of such models derives by the conditions introduced by Hardy-Weinberg for a population in equilibrium and the genotypic probability is calculated from the probabilities of the alleles in the population. For a generic individual g, the genotype probability is:

$$Pr(x_g = (i,j) \mid \theta) = \theta_i \cdot \theta_j \cdot (1 + I_{\{i,j:i \neq j\}}\{i,j\}), \tag{2}$$

According to the specific structure of the population related to a case, to take into account the characteristics of inbreeding and co-ancestry, it is possible to make use of more realistic models, as described in Balding (2006).

Hereafter we refer to generic segregation and population models since our proposal is not influenced by such choices.

3 Indirect identification

3.1 Generalities

In indirect identifications we consider the possibility that someone, the candidate, (C), is the unobserved individual (I) posed in the pedigree of a certain family in a well defined position. In many circumstances, as in case of disputed paternity, identification does not mean to name the candidate, whose identity is known, but it consists in confirming that C and I are the same person.

More formally, consider two alternative conjectures, $\mathcal{H} = \{H_0, H_1\}$, where, conventionally, H_0 is the no-identification hypothesis, which assumes C to be a generic member of a population, and H_1 recognizes C to be the family member I.

The relative support to H_1 provided by some generic observed characteristics, $E = e$, is measured by the weight of evidence, defined by:

$$WE = \frac{Pr(E = e|H_1)}{Pr(E = e|H_0)}. \tag{3}$$

Computing the WE does not require to probabilize the hypotheses, which simply appear as conditioning circumstances but, since a $WE \in [0, \infty]$, it is difficult to appreciate how much considerable the WE must be to provide strong support to H_1 or H_0.

If otherwise prior probabilities for H_0 and H_1 are available, we can directly derive their posteriors. For H_1:

$$Pr(H_1|E = e) = \frac{WE\frac{Pr(H_1)}{Pr(H_0)}}{1 + WE\frac{Pr(H_1)}{Pr(H_0)}}, \tag{4}$$

or, equivalently, in odds form:

$$\frac{Pr(H_1|E = e)}{Pr(H_0|E = e)} = WE\frac{Pr(H_1)}{Pr(H_0)}. \tag{5}$$

3.2 WE computations based on STR DNA evidence

To specify the WE computation in indirect identifications based on DNA STR loci, let the set $\mathcal{F} = \{\mathcal{F}^+, \mathcal{F}^-, I\}$ to contain the family members involved in the analysis. $\mathcal{F}^+$ is the set of relatives providing their DNA profiles; $\mathcal{F}^-$ considers the unobserved relatives required to link the members in $\mathcal{F}^+$ to I.

Once the candidate C and the donors in $\mathcal{F}^+$ have been typed, the required WE can be easily computed since the following assertions of conditional independence hold:

a) H only affects the probability to observe x_C, i.e. $X_{\mathcal{F}^+} \perp\!\!\!\perp H$, so that:

$$\Pr(x_{\mathcal{F}^+}|H_1) = \Pr(x_{\mathcal{F}^+}|H_0); \tag{6}$$

b) If H_1 holds, $C \equiv I$, so that:

$$\Pr(x_C|x_{\mathcal{F}}, x_I, H_1) = \begin{cases} 1 \text{ se } x_C = x_I \\ 0 \text{ otherwise} \end{cases}$$

c) $X_C \perp\!\!\!\perp X_{\mathcal{F}}|H_0$, so that:

$$\Pr(x_C|x_{\mathcal{F}^+}, H_0) = \Pr(x_C|H_0). \tag{7}$$

Considering a), b) and c) we have:

$$\begin{aligned} WE(x_C) &= \frac{\Pr(x_C, x_{\mathcal{F}^+}|H_1)}{\Pr(x_C, x_{\mathcal{F}^+}|H_0)} = \frac{\Pr(x_C|x_{\mathcal{F}^+}, H_1)\Pr(x_{\mathcal{F}^+}|H_1)}{\Pr(x_C|x_{\mathcal{F}^+}, H_0)\Pr(x_{\mathcal{F}^+}|H_0)} \\ &= \frac{\sum_{x_{\mathcal{F}^-}, x_I \in \mathcal{X}} \Pr(x_C|x_{\mathcal{F}^+}, x_{\mathcal{F}^-}, x_I, H_1)\Pr(x_I|x_{\mathcal{F}^+}, x_{\mathcal{F}^-}, H_1)\Pr(x_{\mathcal{F}^-}|x_{\mathcal{F}^+}H_1)}{\Pr(x_C|, x_{\mathcal{F}^+}, H_0))} \\ &= \frac{\sum_{x_I \in \mathcal{X}} \Pr(x_C|x_{\mathcal{F}^+}, x_I, H_1)\Pr(x_I|x_{\mathcal{F}^+}, H_1)}{\Pr(x_C|H_0)} \\ &= \frac{\Pr(x_I = x_C|x_{\mathcal{F}^+}, H_1)}{\Pr(x_C|H_0)}. \end{aligned} \tag{8}$$

As a result, in indirect identification based on STR loci, the WE can be evaluated by assessing two probabilities for each considered locus. If H_1 holds, we only need to evaluate the probability to observe the C's genotypes conditionally to the observations in $\mathcal{F}^+$. To derive $Pr(x_I = x_C|x_{\mathcal{F}^+}, H_1)$ is possible algebraically (Evett and Weir (1998), Balding (2006)) or numerically, by means of some specialized software, as illustrated in (Egeland and Mostad (2002), and (Brenner (1997)). Conditionally to H_0, the probability to observe the C's genotypes is evaluated according to the population model considered adequate to the specific identification.

4 The evaluation of the identification system

4.1 Theory

Now we concentrate on the evaluation of a indirect identification system characterized by a familial structure and by the genetic evidence observed on the family's members in $\mathcal{F}^+$ who want to identify someone as their relative I. Note that, now, we do *not* consider a specific candidate, i.e. this activity is realized before the evidence x_C is observed.

The evaluation of the system is achieved by considering WE as a random variable whose uncertainty is induced by the unobserved random variable X_C. Since the X_C distribution varies according to H_0 and H_1, also the uncertainty on WE varies conditionally to the hypotheses.

More specifically we need to define:

a) $\mathcal{WE}(X_C)$, the support for WE. For a generic locus with k alleles, the set of possible WEs is determined evaluating (8) for each of the possible $k(k+1)/2$ genotypes for each of the n considered loci posed in the set $\mathcal{L} = \{l_i : i \in \{1, \ldots, n\}\}$. The cardinality of $\mathcal{WE}(X_{C,l_1}, \ldots X_{C,l_n})$ is equal to $\prod_{i=1}^{n} k_i(k_i+1)/2$, corresponding to the number of the possible genetic profiles observable on C for the loci in $\mathcal{L}$, so that:

$$\mathcal{WE}(X_{C,l_1}, \ldots X_{C,l_n}) = \otimes_{l_i \in \mathcal{L}} \mathcal{WE}(x_{C,l_i}) \quad \forall x_{C,l_i} \in \mathcal{X}_{l_i}, \forall l_i \in \mathcal{L}; \tag{9}$$

b) two probability distributions for the WE random variable, specified according to the conditioning hypotheses. Since loci commonly used in forensic identification are located at large genetic distance they can be considered independent and their joint probability obtained by factorization.

If H_0 holds, $\forall x_{C,l_i} \in \mathcal{X}_{C,l_i}$:

$$Pr(WE(x_{C,l_1}, \ldots, x_{C,l_n}|H_0) = \prod_{l_i \in \mathcal{L}} Pr(x_{C,l_i}|H_0) \tag{10}$$

where each term which can be evaluated by means of the assumed population model.

Since according to H_1, $C \equiv I$, $\forall x_{C,l_i} \in \mathcal{X}_{C,l_i}$:

$$Pr(WE(x_{C,l_1}, \ldots, x_{C,l_n}|H_1) = \prod_{l_i \in \mathcal{L}} Pr(x_{I,l_i} = x_{C,l_i}|x_{\mathcal{F}^+,l_i}, H_1) \tag{11}$$

where each term is obtainable analytically or computationally.

For the case-study analysed in section 5, we developed an allele Bayesian network (Lauritzen and Sheehan (2003)) making use of the Matlab BN toolbox (Murphy (2001)). Note that, since the X_C probability distribution is required, forensic specialized software devoted to evaluate the WE for

a specific x_C should be repetitively used a very large number of times, becoming almost useless.

The possibility to get a large variety of different WEs, able to discriminate between the hypotheses, depends on how much (10) and (11) diverge. Some information theoretic measures of information are considered in Appendix B.

4.2 Thresholds for the system evaluation

Considering a WE defined as in (3), the unity is the natural threshold producing a meaningful partition of the WE support. As a matter of fact:

$$Pr(WE(X_{c,l_1}, \ldots, X_{c,l_n}|H_1)) < 1 \tag{12}$$

is the probability to get support *against* the identification hypothesis if it is actually true. At the same time, if a prior probability on H_1 is available, (12) provides the probability to get a posterior probability of identification *smaller* than the prior, even if the identification hypothesis is true.

Also:

$$Pr(WE(X_{c,l_1}, \ldots, X_{c,l_n}|H_0)) > 1, \tag{13}$$

is the probability to get support *against* the no-identification hypothesis when the latter is true or to get a posterior probability of no-identification *smaller* than the prior, even if H_0 is true.

As it is apparent, this form of evaluation can be used either the pure likelihood or the subjective Bayesian paradigms are employed. Whether (12) and (13) are small enough to make the system performances acceptable depends on an utility function, to be specified according to the case and the decision maker.

An implicit form of decision rule is embedded in the classification of likelihood ratios proposed by Royal (2000) and based on an ancillary experiment.

The author considers two urns: H_1 contains only white balls; H_0 has the same number of white and red balls. The experiment consists in drawing balls with replacement from one unspecified urn: conditionally to the results we are required to evaluate the support provided to the hypotheses that H_0 or H_1 is the employed urn. Royal imagines that, after n draws, only white balls have been drawn so that the WE favouring H_1 is equal to 2^n. The aim of the experiment is to give the opportunity to determine a value for WE assumed to give support to H_1. For instance, after 8 consecutive white balls you might believe H_1 has received strong support. This would imply that, in the indirect identification of interest, all the genetic profiles implying a $WE \geq 256$ or, generally speaking WE greater than a threshold, τ_2, must be considered strong evidence for H_1. Obviously an analogous experiment could be arranged to determine a value of WE, τ_1, such that all the possible observations implying a smaller WE, must be considered strong evidence favouring H_0. Although Royal is convinced of the *objectivity* of the proposed experiment, the role of the individual called to decide on what is the number of drawn balls required to classify the experimental results into "strong evidence" is of paramount importance.

More explicitly, according to the subjective Bayesian approach, we could introduce the thresholds τ_1 and τ_2 as the WE values able to convert the elicited prior probabilities on H_0 and H_1 into reputed high identification posteriors probabilities. For instance, if we specify the no-informative prior $Pr(H_0) = P(H_1) = 0.5$, and we believe that posteriors must be, for both hypotheses, close to the certainty, e.g. 0.9973, to be considered conclusive, then, by (5), for each possible value of WE less than or equal to $\tau_1 = \frac{0.0027}{0.9973} = 0,002707$, or greater than or equal to $\tau_2 = \frac{0.9973}{0.0027} = 369,37$, the hypotheses H_0 and H_1 could be considered highly supported.

According to which distribution of WE is employed and using terms introduced by Royal, the WE support is partitioned into the subsets described in Table 1 and the evaluation of the system can be achieved computing the corresponding probabilities.

	$WE \leq \tau_1$	$\tau_1 < WE < \tau_2$	$WE \geq \tau_2$
H_0	strong evidence	weak evidence	misleading evidence
H_1	misleading evidence	weak evidence	strong evidence

Table 1: WE classification according to τ_1 and τ_2 and H_0 and H_1

Nevertheless the probability to observe misleading evidence under H_0 reminds the Neyman-Pearson (N-P) type I error, differences should be clear. In N-P the WE (or LR) is simply the way of finding the set of profiles $\mathcal{A} = \{\bigwedge_{l_i \in \mathcal{L}} x_{C,l_i} : Pr(WE(x_{C,l_1}, \ldots, x_{C,l_n}|H_0 \geq \tau_2)$ with probability α, such that $Pr(WE(\mathcal{A}|H_1))$ is maximized. In N-P the subjective choice of α dominates the analysis and the WE threshold is determined consequently, regardless of its meaning.

Here the result of the analysis is the probability to get a bigger or smaller WEs determined at a reputed meaningful value, conditionally to H_0 or H_1.

5 Case study

In this section we consider a real case which seems to take advantage of the proposed method.

Mr. A. R. would would like to assess his father identity. For a number of reasons he believes to be the son of Mr. G.M., who died some years ago.

The indirect identification is considered from three perspectives, with special regard to the systems' identification potentialities.

Genetic data of the individuals related to the case are in Tab. 2; genetic population data are in Brisighelli et al. (2009).

Results are gathered in Tab. 3 and 4 where we display the probabilities to get the WE values according to the dichotomous and trichotomous partition of the WE support, along the lines of Section 4. The values of τ_1 and τ_2 are derived having in mind a prior for the hypotheses such that: $Pr(H_0) = P(H_1) = 0.5$ and posteriors probabilities equal to 0.9933. For this choice, the thresholds become: $\tau_1 = 0.0067$ and $\tau_2 = 148.41$.

Loci	Miss M.P.M.	Mrs A.L.P.	Mr A.R.
D8S1179	14-14	13-14	13-13
D21S11	31.2-32.2	30-32.2	28-28
D7S820	10-10	10-10	8-11
CSF1PO	12-12	12-12	11-12
D3S1358	18-18	15-18	15-18
TH01	6-9	9-9.3	6-7
D13S317	12-12	10-12	8-8
D16S539	10-10	9-10	11-11
VWA	17-17	17-19	16-16
TPOX	8-9	8-9	8-11
D18S51	15-17	16-17	15-15
D5S818	10-12	9-12	10-13
FGA	20-22	22-23	23-25
PENTAD	13-13	12-13	9-10
PENTAE	12-14	12-17	17-17

Table 2: Genetic Data on 15 loci for the individuals considered in the case

			Probabilities		
$\mathcal{F}^+$	I	H	$WE \leq 0.0067$	$0.0067 < WE < 148.41$	$WE \geq 148.41$
Miss M.P.M.	Half Brother	H_0	0.1647	0.8345	0.0008
		H_1	0.0005	0.7936	0.2019
Miss M.P.M. & Mrs. A.L.P.		H_0	0.3079	0.6911	0.0010
		H_1	0.0008	0.5961	0.4031
	Father	H_0	0.9998	0.0002	0.0000
		H_1	≈ 0	≈ 0	≈ 1
Mr. A.R.	Father	H_0	0.9997	0.0001	0.0002
		H_1	0.000005	0.000001	0.999994

Table 3: Probability of *Strong, Weak* and *Misleading* evidence in the considered identification perspectives

			Probabilities	
$\mathcal{F}^+$	I	H	$WE \leq 1$	$WE > 1$
Miss M.P.M.	Half Brother	H_0	0.8890	0.1110
		H_1	0.1151	0.8849
Miss M.P.M. & Mrs. A.L.P.		H_0	0.9262	0.0738
		H_1	0.0804	0.9196
	Father	H_0	0.9998	0.0002
		H_1	0.0067	0.9933
Mr. A.R.	Father	H_0	0.9980	0.0020
		H_1	0.000006	0.999994

Table 4: Probability for the hypotheses H_0 and H_1 to receive or not to receive support in the considered identification perspectives

5.1 The first attempt

The first attempt to assess the Mr. A.R.'s paternity started from the fact that his alleged father, Mr. G.M., had a daughter, Miss M.P.M. with his wife, Mrs. A.L.P.. To avoid the exhumation of Mr. G.M' corpse, Miss M.P.M requested that her own genetic profile be used. Since her mother, Mrs. A.L.P. did not provide her DNA, we evaluated the WE according to the hypotheses specified below:

- H_1: Mr. A.R. is Miss M.P.M.'s half brother;
- H_0 Mr. A.R and Miss M.P.M. do not share recent relatives.

Stated in this way the identification procedure clearly deals with one person, Miss M.P.M., who wants to identify her half brother. This circumstance only indirectly implies they share the father. In other words, if the case were simplistic labelled as a "paternity test", the very indirect nature of the identification would be obscured.

According to the traditional approach, the case was addressed through evaluating the WE, making use of the Miss M.P.M.'s and Mr. A.R.'s genetic traces, obtaining a $WE = 0.017$, a figure which definitely does not support H_1. Then some doubts arose about the ability of the system to identify her half brother, using exclusively Miss M.P.M.'s genetic data.

After an evaluation of the identification system it became clear that the suspicion was well founded. Looking at Tab. 3, whatever hypothesis is assumed true, the system only rarely can achieve a definite result, since the probability of weak evidence is about 0.80. Moreover, if we consider the partition of the WE support into the sets $[0, 1), (1, \infty]$, an even more embarrassing result arises. Now the probability to observe evidence *not* supporting H_0 and H_1 when they are actually true turns to be around 0.11, again an unacceptably high value for this delicate matter.

5.2 The second attempt

Later, the analysis was replicated since Mrs. A.L.P., Miss M.P.M.'s mother, was convinced to provide her genetic profile. The results, provided in Tab. 3 - lines 3-4, make clear the benefits of such additional evidence: the probability to observe weak evidence now reduces to 0.70 if H_0 holds, or to less than 0.60 if the identification hypothesis is assumed. Nevertheless, lines 3-4 in Tab. 4 show that the probability to observe evidence *against* the hypotheses when they are respectively true, is around 0.08, definitely too high a value.

After these attempts there seem to be only two ways to cope with the case.

The first one is to include, among the evidence, the genetic profiles of some further family's members, as, for instance, Mr. A.R.'s mother, or to extend the analysis to more loci.

If the previous remedies cannot be overtaken or if some doubts arise about the assumed familial relationships, the only chance is to re-state the identification hypotheses.

5.3 The third attempt

Since the evaluation of the identification system has given poor results and since Mr. A.R. questioned about the assumption that Miss M.P.M.'s father was Mr.G.M., a further possibility was to evaluate two separate identification systems devoted to:

1. the identification of Miss M.P.M.'s father, starting from Miss M.P.M and her mother's genetic evidence;

2. the identification of Mr. A.R.'s father, starting from Mr. A.R.'s genetic profile.

The performances of the two identification systems are summarized in the last four lines of Tab. 3 and 4 . It clearly shows that, now, Mr. A.R. has the opportunity to carry on the identification of his father quite safely, and the same happens for Miss M.P.M.. The only drawback is that, unfortunately, the Mr. G.M.'s corpse must be exhumed.

The results obtained can be considered an example of satisfactory performances. These empathize how the identification systems' performances depend on the distance between the family's DNA donors and the individual providing the identification of possible candidates. In Appendix A we formalize such idea in a simplified setting.

6 Discussion

In this paper we described a methodology to deal with the assessment of the identification abilities of a probabilistic identification system, devoted to indirect identification.

The goal of the analysis is to provide, *before* the identification process is undertaken, probabilistic information on the possible misleading results possibly deriving from the analysis. We also provide the probability for the system to produce strong or weak support to the hypotheses of interest. In a design-of-experiment perspective we consider how to improve the model performances. In the considered framework, the traditional requirement of additional observations for certain values of the design variables becomes the request of additional genetic profiles from family's members and/or increasing the number of typed loci.

The main contribution of the paper is to recognize a large variety of the indirect identification system behaviours according to specific cases, which implies the use of different amounts of information to maintain a satisfactory standard of performances. This is in our opinion a very important issue since, up to now, the capabilities of proposed identification systems have not been revealed to the interested parts, including those called to express the final judgement on the identification trial.

For sake of simplicity we did not consider segregation models implying mutation or population models including co-ancestry but the effect of such reasonable refinements are under study.

A Information decay on the direct lineage

Now we formally illustrate how much the information provided by a DNA donor to identify a relative on their direct lineage decays according to the distance between the DNA donor and the considered unobserved relative.

Let $X = (a_r, a_s)$ the ancestor genotype and assume the alleles' probabilities $\theta_i, i = 1, \ldots, k$ to be known in the population. For sake of simplicity let the HW conditions and the first Mendelian law hold.

On the ancestor lineage consider the probability distribution of the transmitted allele. At the first generation, $n = 1$, the transmitted allele can assume only two values, r and s, with probability 0.5. For $n > 1$, the ancestor alleles have its probability to be IBD equal to 0.5^n plus the probability to come from the no-ancestor lineage.

Simple probability computations allow to express the allele distribution transmitted on the ancestor lineage according to the number of generations, n:

$$\begin{aligned} Pr(A^n = a_i | X^0 = (a_r, a_s) &= (\frac{1}{2})^n + (1 - (\frac{1}{2})^{n-1})\theta_i \quad i = r, s \\ &= (1 - (\frac{1}{2})^{n-1})\theta_i \quad i \neq r, s. \end{aligned}$$

Since the allele coming from the no-ancestor lineage still has probability ruled by the population parameters, the genotype probability along the generations results to be:

$$
\begin{aligned}
Pr(X^n = (a_i, a_j)|X^0 = (a_r, a_s)) &= (\frac{1}{2})^n(\theta_r + \theta_s) + (1 - (\frac{1}{2})^{n-1})2\theta_r\theta_s \quad i = r, j = s \\
&= (\frac{1}{2})^n(\theta_j) + (1 - (\frac{1}{2})^{n-1})2\theta_r\theta_j \quad i = r, j \neq s \\
&= (1 - (\frac{1}{2})^{n-1})2\theta_i\theta_j \quad i \neq r, j \neq s
\end{aligned}
$$

so that the usual WE, can easily be evaluated:

$$
\begin{aligned}
WE(X^n = (a_i, a_j)|X^0 = (a_r, a_s)) &= (\frac{1}{2})^{n+1}\frac{(\theta_r + \theta_s)}{\theta_r\theta_s} + (1 - (\frac{1}{2})^{n-1}) \quad i = r, j = s \\
&= (\frac{1}{2})^{n+1}\frac{1}{\theta_r} + (1 - (\frac{1}{2})^{n-1}) \quad i = r, j \neq s \\
&= 1 - (\frac{1}{2})^{n-1} \quad i \neq r, j \neq s,
\end{aligned}
$$

As a result, the WE approaches the unity from above at the $(\frac{1}{2})^n$ rate if the candidate shares one or two alleles with the donor, from below if the ancestor and the candidate have no common alleles.

B Some useful information-theoretic measures

Lauritzen and Mazumder (2008) have recently proposed the mutual information between X_I and $X_{\mathcal{F}^+}$ as a measure of the expected reduction in uncertainty regarding X_I achievable exploiting the dependence between X_I and $X_{\mathcal{F}^+}$.

The mutual information proposed by the authors, expressed in our notation. is:

$$
\begin{aligned}
I(X_I, X_{\mathcal{F}^+}) &= \sum_{x_I \in \mathcal{X}_I} \sum_{x_{\mathcal{F}^+} \in \mathcal{X}_{\mathcal{F}^+}} \log(Pr(x_I|x_{\mathcal{F}^+}, H_1)Pr(x_I|x_{\mathcal{F}^+}, H_1) \\
&\quad - \sum_{x_I \in \mathcal{X}_I} \log(Pr(x_I|H_1)Pr(x_I|H_1), \\
&= H(X_I|H_1) - H(X_I|x_{\mathcal{F}^+}, H_1)
\end{aligned}
$$

where $H(X)$ is the entropy of the random variable X.

As it appears, the proposed measure is not related to a specific case, but it concerns the evaluation of a marker informativeness for a specific identification issue and for some specified segregation and population models.

Since this paper focuses on a specific indirect identification case, i.e. having observed the family's DNA donors and studying the uncertainty on WE on

which randomness is induced by the unobserved candidate's genetic profile, the most suitable information measure seems to be Kullback-Leibler measure of divergence.

In fact, Kullback and Leibler (1951) (K-L) defined this measure as the expected value of the log of a likelihood ratio, explicitly considered as the updating factor of the Bayes formula (4) expressed in log form. Since it is not known which one of the hypotheses is true, Kullback and Leibler consider two different measures.

If H_1 is true, the $\log(WE) = \log(\frac{Pr(x_C|x_{\mathcal{F}^+}, H_1)}{Pr(x_C|x_{\mathcal{F}^+}, H_0)})$, which is equal, by (7), to $\log(\frac{Pr(x_C|x_{\mathcal{F}^+}, H_1)}{Pr(x_C|H_0)})$ is averaged with respect to $Pr(x_C|x_{\mathcal{F}^+}, H_1)$ and the K-L measure of divergence is:

$$\begin{aligned} K_{H_1}(X_C) &= \sum_{x_C \in \mathcal{X}_C} \log(\frac{Pr(x_C|x_{\mathcal{F}^+}, H_1)}{Pr(x_C|H_0)}) Pr(x_C|x_{\mathcal{F}^+}, H_1) \\ &= \sum_{x_C \in \mathcal{X}_C} \log(Pr(x_C|x_{\mathcal{F}^+}, H_1)) Pr(x_C|x_{\mathcal{F}^+}, H_1) - \\ &\quad \sum_{x_C \in \mathcal{X}_C} \log(Pr(x_C|H_0) Pr(x_C|x_{F^+}, H_1). \end{aligned}$$

If, otherwise, H_0 holds, X_C distribution depends on the population model and related parameters and the expected value of the $log(WE) = \log(\frac{Pr(x_C|H_0)}{Pr(x_C|x_{\mathcal{F}^+}, H_1)})$ is:

$$\begin{aligned} K_{H_0}(X_C) &= \sum_{x_C \in \mathcal{X}_C} \log(\frac{Pr(x_C|H_0)}{Pr(X_C|X_{\mathcal{F}^+}, H_1)}) Pr(x_C|H_0) \\ &= \sum_{x_C \in \mathcal{X}_C} \log(Pr(x_C|H_0) Pr(x_C|H_0) \\ &\quad - \sum_{x_C \in \mathcal{X}_C} \log(Pr(x_C|x_{\mathcal{F}^+}, H_1)) Pr(x_C|H_0) \end{aligned}$$

The result is that not only we have a measure of the expected performances of a marker for the specific identification problem at hand but we also get information on the expected ability of the system to support each of the hypotheses when they are true.

References

Balding, D. J. (2006). *Weight-of-evidence for forensic DNA profiles.* Wiley, New York.

Brenner, C. (1997). Symbolic kinship program. *Genetics 145*, 533–542.

Brisighelli, F., C. Capelli, I. Boschi, P. Garagnani, M. Lareu, V. Pascali, and A. Carracedo (2009). Allele frequencies of fifteen strs in a representative sample of the italian population. *Forensic Science International: Genetics 3*(2), 2, e29–e30.

Dawid, A., J. Morter, and P. Vicard (2007). Object-oriented bayesian networks for complex forensic dna profiling problems. *Forensic Science International 169*, 195–205.

Egeland, T. and P. Mostad (2002). Statistical genetics and genetical statistics: a forensic perspective. *Scand. Journal of Statistics 29*, 297–307.

Evett, I. and B. Weir (1998). *Interpreting DNA evidence*. Sinauer Associates, Sunderland.

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics 22*, 79–86.

Lauritzen, S. and A. Mazumder (2008). Informativeness of genetic markers for forensic inference-an information theoretic approach. *Forensic Science International: Genetics supplement Series 1*, 652–653.

Lauritzen, S. and N. Sheehan (2003). Graphical models for genetic analyses. *Statistical Science 18*, 489–514.

Murphy, K. (2001). The bayes net toolbox for matlab. *Computing Science and Statistics 33*, 1024–1034.

Royal, R. (2000). *Statistical Evidence*. Chapman&Hall London - New York.