Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2011/14

Evaluation of Kinship Identification Systems Based on STR DNA Profiles

Fabio Corradi, Federico Ricciardi

Università degli Studi
di Firenze

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

EVALUATION OF KINSHIP IDENTIFICATION SYSTEMS BASED ON STR DNA PROFILES

By Fabio Corradi and Federico Ricciardi

Università degli Studi di Firenze - Italy

In this paper we detail how to evaluate a kinship identification system, a probabilistic tool devoted to obtain the Likelihood Ratio required in deciding if a candidate is a specific member of a family given some genetic profiles observed in the familial pedigree. The paper considers the LR as a random variable, depending on the still unobserved genetic DNA evidence of a candidate to identification, posing attention to the familial expected possibilities to identify a specific member. In a decision theory perspective, we evaluate which system, among a set of possible alternatives, is the most suitable to fulfil the requirements of the parts involved. The proposed system evaluation, carried on before the identification trial is performed, is specific for each case and does not require any additional laboratory expenses, since it makes use of a subset of the employed data. Special attention is devoted to the computational aspects of the implied high dimensional space problem: matters concerning approximations are discussed. A case study illustrates how the approach proves to be especially helpful when the distance in the pedigree between the observed DNA donors and the unobserved relative possibly identifying a candidate is large.

1. Introduction. Nowadays, the identification of individuals trough DNA evidence can definitely be considered out of its infancy. Reliable kits of primers allow to determine the genotypes of biological traces typed on some short tandem repeat (STR) loci. The identification issue is addressed by calculating the ratio of the evidence's probabilities, usually named Likelihood Ratio (LR), expressed conditionally to a pair of competitive hypotheses meaningful for the case and evaluated following largely accepted theoretical developments as in Evett and Weir (1998), Aitken and Taroni (2004) and Balding (2006).

This paper focuses on kinship analysis (Brenner (1997)), a form of indirect identification comprising a wide class of problems including disputed paternities, searching missing persons, family reunifications for citizens of foreign birth, permanent resident aliens and so on. More precisely, we consider identification cases consisting in the attempt to identify a candidate

Keywords and phrases: Kinship Identification, DNA STR markers, Likelihood ratio distribution, Decision analysis approach, Utility functions.

as an unavailable and therefore unobserved specific member of a family, exploiting the knowledge of some of the family members' genetic profiles and the familial relationships.

It is commonly acknowledged that the ability of a System[1] to provide strong support to one of the hypotheses declines according on the "distance" between the persons who require the identification and provide their DNA, and the position in the pedigree of the searched individual. Nevertheless, the LR is almost invariably evaluated by practitioners using the set of loci included in the kit of primers adopted by their own laboratory and exclusively considering the family members offering themselves to provide their DNA profiles. Sensitivity analysis about the use of different population and segregation models has been rarely proposed. Moreover, in standard practice, no one provides information about the characteristics of the proposed System, concerning for instance the ability to produce high LR values when the candidate actually is the member the family, and to obtain small LR figures when the opposite is true.

In the literature the importance to pre-assess the expected distribution of the relevant LR has been recognised (Cook et al (1998)), but there is a limited number of contributions in the field of identification through DNA. Many of them focus on the LR obtained using sets of real or simulated cases for which the identification of the candidate was already ascertained, not providing an evaluation about the characteristic of the System if applied to a case with different DNA familial evidence. Evett and Buckleton (1996), using a data base of 1401 different individuals on 4 loci, evaluated the likelihood ratio's empirical distribution originated by criminal identification cases occurring when two traces are questioned to belong to the same person. On the same track, in a decision analysis perspective, Taroni et al (2007) obtained the LR distribution for the criminal identification issue by simulating 100.000 genetic profiles on 16 loci. All these proposals are confined to the realm of criminal cases, which are not kinship problems. Another contribution, dealing with kinship analyses, is due to Brenner and Staub (2003) who evaluated the LR distribution only under the identification hypothesis for 19 different pedigrees and identification issues simulating the genetic evidence of 100 familial groups. Results were synthesized by the LR geometric mean and standard deviation showing how these measures vary according to the distance between the donors and the searched person.

[1] The term Identification System or just System indicates all the elements required for the identification, including the familial pedigree, the DNA evidence provided by familial donors, the hypotheses concerning whether a candidate is or not a well specified family member, population and segregation models, these latter detailed in Section 2.

Also Lauritzen and Mazumder (2008) have studied the problem for different kinship identification issues proposing an information-theoretic measure to evaluate the informativeness of different loci. In these latter cases the authors do not refer to a specific identification case, but provide indication on how the System works with respect to cases invariably different from the one we are interested in. Finally in the documentation of DNA-VIEW[2], a commercial software for kinship identification, there are indication on how to evaluate the "value" of the a further additional relative introduced in a case.

Despite the previous experiences, our proposal consists in ascertaining the LR distributions specifically for a case of interest and to provide a synthesis of the results *before* carrying out the identification, proposing to calculate the LR including the candidate evidence only if the System shows satisfactory performance. Ignorance about the System characteristics exposes to the risk of producing misleading results with an unknown, possibly high, probability. Evaluating such probabilities and other measures of the System features allows to answer to very common questions posed by actors involved in the identification trial and concerning reliability.

The material is arranged as follows. Section 2 refers about some probability models for DNA evidence. Section 3 shows how, in kinship analysis, the LR can be expressed as a function of a candidate's genetic profile. Section 4 details how to obtain the LR distributions conditionally to the hypotheses and how to evaluate the System in a decision perspective. Section 5 describes computational issues. Section 6 proposes a real case concerning an indirect paternity case. Finally a discussion is given in Section 7.

2. Probability models for DNA evidence. To evaluate the probability of the 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 it is possible to 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 usually assumed known, $\mathcal{A} = \{a_1, \ldots, a_k\}$, and the probability to observe them is indicated by $Pr(A = a_i) = \theta_i$, $\theta = \{\theta_1, \ldots, \theta_k\}$. The random variable $X = (a_r, a_s)$, with $r \leq s$, represents the uncertainty about genotypes.

The probability distribution for the genotype of an individual is provided by two kind of models depending on whether their parents are explicitly included in the model (Segregation models) or not (Population models). Here two versions of these models are considered to also perform a sensitivity

[2] `http://dna-view.com/simulate.htm`

analysis with respect to the implied assumptions.

2.1. *Segregation Models.* The baseline segregation model (ML) strictly follows the first Mendelian law allowing for the transmission of each parental allele with probability 0.5. The genotype's probability of a child, c, given the genotypes of their parents, m and f, if $x_c = (a_t, a_z)$, $x_m = (a_i, a_j)$ and $x_f = (a_r, a_s)$, is:

$$Pr(x_c \mid x_m, x_f) = \frac{1}{4}(\mathbb{1}_{\{a_i,a_r\}}(x_c) + \mathbb{1}_{\{a_i,a_s\}}(x_c) + \mathbb{1}_{\{a_j,a_r\}}(x_c) + \mathbb{1}_{\{a_j,a_s\}}(x_c)).$$

A more realistic approach includes in the segregation process a mutation mechanism. **?** discuss a number of such models. Here we consider the Mixed Mutation Model (MMM) obtained by mixing a single step model - any mutation can only happen in the strict neighbor of the parental allele - and the proportional model - whenever a mutation takes place the new allele value is generated at random from the population allelic distribution.

Following MMM the child allele received from the father (A_c^f) could be different, since a mutation occurred, from the father transmitted allele, A_f, which is one of the father's alleles chosen with probability 0.5. For example, the probability that the A_c^f takes value a_r, $Pr(A_c^f = a_r)$, is given by

$$\underbrace{(1-\mu)\cdot Pr(A_f = a_r)}_{no\ mutations} \quad + \quad \mu \cdot \underbrace{\Big\{\Big(h + h\mathbb{1}_{\{1\cup k\}}(r)\Big)}_{single\ step\ MM\ weight} \times$$

$$\underbrace{\Big[\Big(\frac{1}{2} - \frac{1}{2}\mathbb{1}_{\{k\}}(r)\Big)\cdot Pr(A_f = a_{r+1}) + \Big(\frac{1}{2} - \frac{1}{2}\mathbb{1}_{\{1\}}(r)\Big)\cdot Pr(A_f = a_{r-1})\Big]}_{single\ step\ MM\ probability} +$$

$$+ \quad \underbrace{(1-h)\theta_r\Big\}}_{proportional\ MM} \quad .$$

The mixing proportion is specified according to the idea that a mutation of more than one repetition in the STR sequence characterizing a locus is uncommon, so that a $h \simeq 0.9$ is largely accepted. The overall mutation rate μ is estimated over a large number of meiosis as in Brinkmann et al (1998).

2.2. *Population Models.* The baseline model (HW) derives from the conditions introduced by Hardy-Weinberg for a population in equilibrium. The genotype probability is calculated from the assumed known probabilities of the alleles in the population. For a generic individual g, the genotype probability $x_g = (a_i, a_j)$ is:

$$Pr(x_g \mid \theta) = \theta_i \cdot \theta_j \cdot (2 - \mathbb{1}_{\{a_i\}}(a_j)), \tag{2.1}$$

which implies that the maternal and paternal alleles are independent conditionally to the population parameters θ.

A more realistic model relaxes the assumption of known alleles' probabilities and considers them uncertain. If a database of individuals available for forensic inference can be assumed as a random sample from a reference population, for a locus the observed alleles frequencies, $n = \{n_1, \ldots, n_k\}$, with $N = \sum_{i=1}^{k} n_i$, follow a multinomial distribution expressed conditionally to the parameters θ. If prior probabilities on $\theta \sim Dir(\delta)$, $\delta = \{\delta_1, \ldots, \delta_k\}$, then the posterior distribution is $\theta|n, \delta \sim Dir(\delta_1 + n_1, \ldots, \delta_k + n_k)$. If in the pedigree involved in the identification trial two or more founders' alleles are not observed and their probabilities are uncertain, alleles become dependent. The Uncertainty in Allele Frequencies model (UAF), proposed by Green and Mortera (2009), states that, if S founders' alleles are considered, the marginal distribution of the Sth allele probability assumes value a_j is a mixture formed by the marginal of $\theta|n, \delta$, i.e. a Beta, and a probability mass proportional to the number of a_js observed on the previous $S-1$ founder alleles, i.e.:

$$\theta_j^S|\delta, n \sim \frac{\sum_i^k \delta_i + N}{M} Beta(\delta_j + n_j, \sum_{i \neq j}^{k} \delta_i + n_i) + \frac{1}{M} \sum_{s=1}^{S-1} \mathbb{I}_{\{a^s\}}(a_j),$$

so that

$$Pr(A = a_j^S|\delta, n) = \frac{\delta_j + n_j}{M} + \frac{1}{M} \sum_{s=1}^{S-1} \mathbb{I}_{\{a^s\}}(a_j), \tag{2.2}$$

where $M = \sum_{i=1}^{k} \delta_i + N + S - 1$. Including (2.2) into (2.1) as one of the θs produces the required genotype probability.

The UAF model is derived by miming the allele probability in an individual which could appear either because it arises from the reference population or because it captures the ambient degree relatedness often named coancestry. The same model can be also derived by the Pòlya urn scheme.

3. Kinship identification.

3.1. *Generalities.* Consider the relative support to the hypothesis H_1, against H_0, provided by the entire observed genetic evidence, generically indicated by $E = e$ and measured by the likelihood ratio:

$$LR = \frac{Pr(E = e|H_1)}{Pr(E = e|H_0)}.$$

Computing the LR does not require the assessment of the prior probabilities for the hypotheses, which simply appear as conditioning circumstances.

If, instead, H is considered a random variable, with $\mathcal{H} = \{H_z : z \in \{0,1\}\}$, and $Pr(H_0)$ and $Pr(H_1)$ are available, we can use the LR to easily derive interpretable posterior probabilities:

$$Pr(H_1|E=e) = LR\frac{Pr(H_1)}{Pr(H_0)}\Big(1 + LR\frac{Pr(H_1)}{Pr(H_0)}\Big)^{-1}.$$

On the other hand we can compute the LR able to update a given prior to a specified posterior:

$$LR = \frac{Pr(H_1|E=e)}{Pr(H_0|E=e)} \times \frac{Pr(H_0)}{Pr(H_1)}. \tag{3.1}$$

3.2. *LR computations based on STR DNA evidence.* In kinship identifications we consider an individual, the Candidate (C), and an unobserved person (U) posed in the pedigree of a certain family in a well defined position. Conventionally, H_0 is the no-identification hypothesis, which assumes C not to be related to the family, being for instance a generic member of the reference population, whereas H_1 recognizes C to be the family member U. Let the set $\mathcal{F} = \{\mathcal{F}^+, \mathcal{F}^-, U\}$ contain the family members involved in the analysis: $\mathcal{F}^+$ is the set of relatives providing their DNA profiles while $\mathcal{F}^-$ considers the unobserved relatives possibly required to link the members in $\mathcal{F}^+$ to U.

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

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

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

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

$$\Pr(x_C|x_{\mathcal{F}^+}, x_U, H_1) = \begin{cases} 1, & \text{if } x_C \equiv x_U, \\ 0, & \text{otherwise.} \end{cases}$$

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

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

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

$$
\begin{aligned}
&LR(X_C = x_C) = \\
&\quad = \frac{\Pr(x_C, x_{\mathcal{F}^+}|H_1)}{\Pr(x_C, x_{\mathcal{F}^+}|H_0)} = \frac{\Pr(x_C|x_{\mathcal{F}^+}, H_1)\Pr(x_{\mathcal{F}^+}|H_1)}{\Pr(x_C|x_{\mathcal{F}^+}, H_0)\Pr(x_{\mathcal{F}^+}|H_0)} \\
&\quad = \frac{\displaystyle\sum_{x_{\mathcal{F}^-}, x_U \in \mathcal{X}} \Pr(x_C|x_{\mathcal{F}^+}, x_{\mathcal{F}^-}, x_U, H_1)\Pr(x_U|x_{\mathcal{F}^+}, x_{\mathcal{F}^-}, H_1)\Pr(x_{\mathcal{F}^-}|x_{\mathcal{F}^+}H_1)}{\Pr(x_C|, x_{\mathcal{F}^+}, H_0)} \\
&\quad = \frac{\displaystyle\sum_{x_U \in \mathcal{X}} \Pr(x_C|x_{\mathcal{F}^+}, x_U, H_1)\Pr(x_U|x_{\mathcal{F}^+}, H_1)}{\Pr(x_C|H_0)} \\
&\quad = \frac{\Pr(x_C \equiv x_U|x_{\mathcal{F}^+}, H_1)}{\Pr(x_C|H_0)}.
\end{aligned}
\tag{3.2}
$$

As a result, in kinship identifications based on STR loci, the LR can be evaluated by assessing two probabilities for $X_C = x_C$, conditionally to two different states of information.

4. The evaluation of the identification System. To evaluate a System, the first activity consists in deriving the $LR(X_C)$ distributions conditionally to the states of H; then the decision approach is employed to produce an evaluation of the System appropriate for the parts.

4.1. *The LR distributions.* To derive the LR distributions we can take advantage of the circumstance that loci commonly used in forensic identification are located at large genetic distance and therefore are considered independent[3].

a) For a generic locus i with k_i different allele values, the possible LRs are determined evaluating (3.2) for all the $k_i(k_i+1)/2$ genotypes. Considering n loci in the set $\mathcal{L} = \{l_i : i \in \{1, \ldots, n\}\}$, the number of the possible genetic profiles observable on C is $|\mathcal{LR}| = \prod_{i=1}^{n} k_i(k_i+1)/2$ and the $LR(X_C)$ support is given by the set:

$$
\mathcal{LR} = \Big\{ \prod_{i=1}^{n} LR(x_{C,l_i}) : x_{C,l_1} \in \mathcal{X}_{l_1}, \cdots, x_{C,l_n} \in \mathcal{X}_{l_n} \Big\}, \tag{4.1}
$$

where $\mathcal{X}_{l_i}$ is the sample space for the genotype of a generic locus i.

[3]If segregation or population models would include some parameters shared among all the loci this form of independence cannot be assumed: here this possibility is not taken into account.

b) By (4.1), $\mathcal{LR}$ merely depends on the genetic profiles of C and the LR distributions follow their distributions conditionally to H_0 and H_1. If H_0 holds, the probability of a genetic profile is obtained by factorizing the genotypes' probabilities over the loci through the assumed population model.

$$Pr(LR(x_{C,l_1},\ldots,x_{C,l_n})|H_0)=\prod_{i=1}^{n} Pr(x_{C,l_i}|H_0) \quad \forall x_{C,l_i}\in\mathcal{X}_{C,l_i}. \tag{4.2}$$

If H_1 holds, i.e. $C\equiv U$, then the genetic profiles probabilities is obtained by factorizing the loci's probabilities derived by $X_C|x_{\mathcal{F}^+},H_1$:

$$\begin{aligned} &Pr\big(LR(x_{C,l_1},\ldots,x_{C,l_n})|H_1\big)= \\ &\quad=\prod_{i=1}^{n} Pr(x_{C,l_i}\equiv x_{U,l_i}|x_{\mathcal{F}^+,l_i},H_1) \quad \forall x_{C,l_i}\in\mathcal{X}_{C,l_i}. \end{aligned} \tag{4.3}$$

Hereafter we consider the likelihood ratio distributions with regard to all the available loci altogether and for simplicity we refer to LR instead of $LR(x_{C,l_1},\ldots,x_{C,l_n})$.

4.2. *The decision approach.* To evaluate the System by a decision analysis we need to define the following quantities.

Decisions. Consider the possibility to choose among n identification Systems, differing for some characteristics, and devoted to cope with a specific identification problem. The decision consists in choosing among the alternatives $\mathcal{D}=\{d_1,\ldots,d_n\}$ indicating the System to use.

Outcomes. The LR distributions, one for each identification hypothesis, are the uncertain outcomes. They vary according to the System employed.

Consequences. Each possible value of the ou. Ftcome, $LR=j$ or simply LR_j, jointly with a decision d_i produces a consequence C_{ij}. For instance, different Systems may require different laboratory activities and costs, leading to different consequences for the same LR_j. Here costs related to different decisions are considered negligible with respect to the matter implied in an identification. For this reason consequences simply coincide with LRs.

Utility or loss. Consequences, conditionally to the hypothesis assumed to hold, can be measured by using an utility, $u(LR_j|H_z)$, or a loss function $l(LR_j|H_z)$, $z\in\{0,1\}$. In the following we describe two aptitudes, relevant in identification and concerning the evaluation of consequences. We define as *Problem-solver* aptitude that of an actor who eminently appreciates Systems strongly supporting the identification hypothesis assumed to hold. The same value of utility is attributed to all the LR_js strongly supporting the holding

hypothesis. On the opposite no utility is attributed to the other LR_js. The proposed utility function, also represented in (Fig. 1), is

$$u(LR_j|H_0) = \begin{cases} 1, & \text{if } LR_j \leq \tau_0, \\ 0, & \text{if } LR_j > \tau_0, \end{cases} \tag{4.4}$$

$$u(LR_j|H_1) = \begin{cases} 0, & \text{if } LR_j < \tau_1, \\ 1, & \text{if } LR_j \geq \tau_1. \end{cases} \tag{4.5}$$

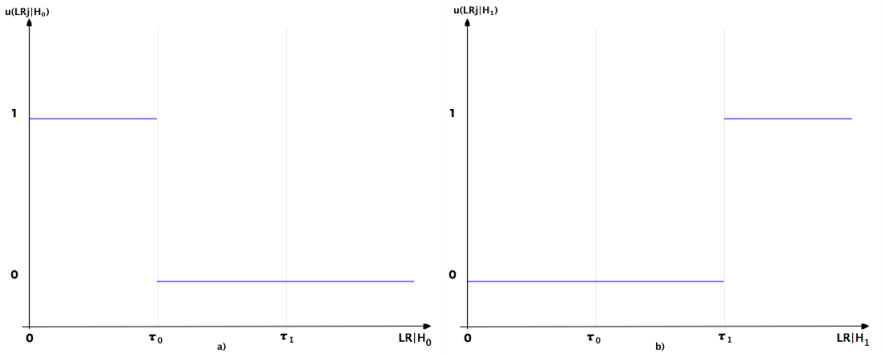

FIG 1. *Utility functions under H_0 a) and H_1 b), related to a problem solver aptitude.*

Alternatively, the *Conservative* aptitude is that of individuals mostly alarmed by the possibility the System produces false identifications. Since the pessimistic aptitude, consequences are measured by a loss function. The proposal, also represented in (Fig. 2), is

$$l(LR_j|H_0) = \begin{cases} 0, & \text{if } LR_j < \tau_1, \\ 1, & \text{if } LR_j \geq \tau_1, \end{cases} \tag{4.6}$$

$$l(LR_j|H_1) = \begin{cases} 1, & \text{if } LR_j \leq \tau_0, \\ 0, & \text{if } LR_j > \tau_0. \end{cases} \tag{4.7}$$

To define the thresholds the most natural way is to use a single $\tau_0 = \tau_1 = \tau = 1$, i.e. the value which splits the LR support in two regions: one favouring H_0 ($\tau < 1$), the other H_1 ($\tau > 1$). Other, perhaps more meaningful, thresholds can be specified considering the behaviour of the actors involved in the System evaluation.

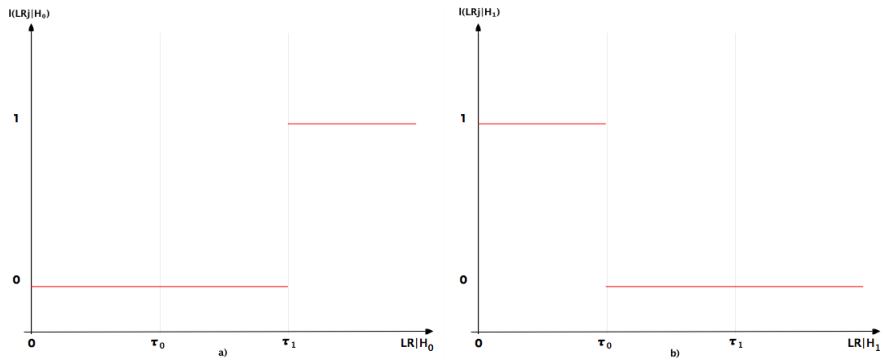

FIG 2. *Loss functions under* H_0 *a) and* H_1 *b), related to a conservative aptitude.*

More specifically, consider the *Judge*, the person required to decide about the identification controversy: their prior probabilities on the identification hypotheses, and the posteriors required to assume a decision about the identification, can be used to define the thresholds. A not informative, and possibly fair, position for the *Judge* could be $Pr(H_0) = Pr(H_1) = 0.5$. Moreover the Judge (and common law) could indicate a posterior probability leading to an identification decision "beyond any reasonable doubt". This implies, by (3.1), the evaluation of τ_0 and τ_1 in (4.4 - 4.7). In Section 6 we consider a real case making use of both the proposed approaches to define the thresholds. Finally, two other parts have interest in assessing the value of the System: those favouring identification (*pro-id*) and those against (*con-id*). Their role suggests reasonable prior probabilities on H. *The pro-id*s believe that $C \equiv U$, so that their prior probabilities could be close to $Pr(H_0) = 0$ and $Pr(H_1) = 1$. *The con-id*s strongly believe that $C \not\equiv U$, so their prior probabilities could tentatively be $Pr(H_0) = 1$ and $Pr(H_1) = 0$.

The aim of a decision analysis is to compute the expected utility and/or loss conditionally to each d_i through:

$$E(u|d_i) = \sum_{z\in\{0,1\}} \sum_{j\in\mathcal{LR}} u(LR_j|H_z) \cdot Pr(LR_j|d_i, H_z) \cdot Pr(H_z), \tag{4.8}$$

$$E(l|d_i) = \sum_{z\in\{0,1\}} \sum_{j\in\mathcal{LR}} l(LR_j|H_z) \cdot Pr(LR_j|d_i, H_z) \cdot Pr(H_z), \tag{4.9}$$

and to maximize (4.8) or to minimize (4.9) choosing among the available decisions.

The expected loss and utility vary according to the parts since their different prior probabilities on H. The *pro-id*s actually consider only the case

in which the LR distribution is expressed conditionally on H_1 (Fig. 1b and Fig. 2b). The *con-id*s only take account of the distribution of $LR|H_0$, as shown in Fig. 1a and Fig. 2a.

In this proposal the expected loss and utility evaluated for the parts are easily interpretable. If we choose $\tau = 1$ and the problem-solver attitude, the expected utility amounts to the probability the System supports H_1 (*pro-id*), H_0 (*con-id*) and average of them (*Judge*) when the hypotheses hold. Conversely, if the conservative attitude is assumed, the expected loss is the probability the System supports the hypotheses when they are not actually true. Alternatively, if τ_0 and τ_1 are specified, the expected utility and loss are, respectively, equal to the probability of false and correct identification according to the decision rule on which τ_0 and τ_1 are chosen.

The proposed loss and utility functions are an extreme version of more realistic and smooth alternatives, but they have the merit to produce expected values of the loss and utility functions interpretable in term of the probability of some System's features. Attempts to combine them inevitably would obscure important characteristics of the System.

5. Computational strategies. The main computational issue consists in obtaining the LR distributions for the loci considered altogether, when H_0 and H_1 respectively hold. The task can be usefully pursued in two steps. First, for each locus we need to efficiently derive the X_C distribution conditionally to each hypothesis following the population and the segregation models appropriate to the case. By the ratio of the probabilities of each possible state of X_C, conditionally to H_0 and H_1, we can derive the possible values assumed by the LR. Since $Pr(X_C|H_0)$ and $Pr(X_C|X_{\mathcal{F}^+}, H_1)$ are obtained, the LR distributions, given H_0 and H_1, arise as a by-product. The second step consists in deriving the LR distributions for each possible arrangement of the genotypes on different loci.

5.1. *LR single locus computations.* The best way to derive the $X_C|\mathcal{F}^+, H_z$ distributions is probably by using a Bayesian Network (BN), a representation of a stochastic system by a Directed Acyclic Graph (DAG), as in Cowell et al (1999). Stochastic nodes in the graph are opportunely linked by arrows; marginal and conditional dependence-independence relationships can be derived from the graph and exploited to efficiently obtain the marginal posterior distributions of every unobserved random variables. All the probabilistic information required to describe the joint distribution of the stochastic system represented by a BN is carried on by the probability distribution of each random variable conditionally to its incident nodes. In identification problems, the pedigree structure follows the BN requirements since the genotype

probability distributions of the founders are expressed by a population model and the genotype probability distributions of the other members of the family, whose parents are explicitly represented, can be derived by the segregation models and parameters as described in Section 2. The using of BN for identification issues by DNA traits has been introduced by Dawid et al (2002), Lauritzen and Sheehan (2002). Green and Mortera (2009) provide details about the use of the BN in modeling the UAF population model. Segregation allowing mutations has been implemented following the BN approach detailed in Vicard and Dawid (2004).

5.2. *LR multi locus computations.* The most immediate but naïve way to obtain the LR distributions is to achieve the result by exact computations. Making use of the single locus results, by **a)** in Section 4.1, the LR support is derived applying (4.1). Similarly, exploiting independence among loci, we obtained the LR distributions under H_0 and H_1 using (4.2) and (4.3), since they depends on X_C only.

Even if very simple, this procedure produces a number of LR values exponentially increasing according to the number of genotypes for every locus, so that the LR sample space rapidly becomes intractable.

LR equivalence classes. Fortunately, not all the LRs induced by the possible genetic traces assumed by the Candidate are different. Consider for instance the case provided in Appendix A, where a DNA donor posed on the direct lineage n generations far from U, is attempting the identification of a Candidate. In this case, by (A.1), the number of different values the LR can assume in a locus results to be 3 or 6, depending if the donor in $\mathcal{F}^+$ is homozygous or heterozygous, so it does not depend neither on k nor on n. This implies that we are not required to evaluate the LR for all the possible different candidate's profiles, but the LR distribution can be obtained considering only the different LR values, summing up the probabilities of the profiles producing the same LR and for this reason they constitute an equivalence class. For identification cases differing from that one detailed in the Appendix, these classes of equivalence could be found computationally, by aggregating identical LR obtained for each locus.

LR quasi-equivalence classes. The strategy outlined above does not introduce approximations of the LR distributions but it could not suffice to downsize $|\mathcal{LR}|$ to a tractable dimension. A possibility to go further is represented by a form of approximation treating some profiles producing very close LR values as belonging to the same equivalence class. The idea is

twofold and deals with the issue of obtaining a reduced LR size and with the evaluation of the LR distributions' approximation accuracy.

The first issue is addressed for the locus i by collapsing the jth and $j+1$th LR if

$$\Delta(i)_{(j)} = \frac{LR(l_i)_{(j+1)} - LR(l_i)_{(j)}}{LR(l_i)_{(j)}} < \varepsilon$$

algebraically or and decide what LR should be used instead of the two neighboring LRs:

$$LR^*(l_i)_{(j)} = f\big(LR(l_i)_{(j+1)}, LR(l_i)_{(j)}\big), \tag{5.1}$$

then derive the LR^* probabilities summing up all the probabilities of the profiles included in what result to be a quasi-equivalence class.

Note that the value of ε is purely instrumental to downsize $|\mathcal{LR}|$, but it is not functionally related to a desired approximation accuracy. To achieve this further goal consider that: if $f(\cdot,\cdot) = max\big(LR(l_i)_{(j+1)}, LR(l_i)_{(j)}\big)$, then:

$$Pr(LR^*_{\texttt{max}} > \tau) \geq Pr(LR > \tau) \quad \forall\tau.$$

If, alternatively, $f(\cdot,\cdot) = min\big(LR(l_i)_{(j+1)}, LR(l_i)_{(j)}\big)$:

$$Pr(LR^*_{\texttt{min}} < \tau) \geq Pr(LR < \tau) \quad \forall\tau.$$

This implies that, using both the approximations, we get:

$$Pr(LR^*_{\texttt{max}} < \tau) \leq Pr(LR < \tau) \leq Pr(LR^*_{\texttt{min}} < \tau)$$

and

$$Pr(LR^*_{\texttt{min}} > \tau) \leq Pr(LR > \tau) \leq Pr(LR^*_{\texttt{max}} > \tau),$$

i.e. the $LR^*_{\texttt{min}}$ and $LR^*_{\texttt{max}}$ distributions can be used as upper and lower bound to determine an interval surrounding the probability of the required subset of LR values. The approximation's accuracy can possibly be measured by the relative difference between the two approximations and the reduction in complexity could be evaluated through the relative difference between the size of LR and LR^*.

6. Case study. In this section we detail a case which seems to take advantage of the proposed method. Genetic data related to the case, obtained with a 15 STR loci kit of primers, are in a supplement table available on line. Here allelic frequencies referring to the Italian population, has been

provided by Brisighelli et al (2009) and mutation rates are taken from the 2008 Report of the American Association of Blood Banks[4].

A man, B, would like to assess his father's identity. He has serious reasons to believe he is the son of AF (the *Alleged Father*), who died some years ago. AF had a daughter (S) with his wife M, who is not the mother of B.

To follow our decision perspective, we define the following quantities.

Decisions. We decided to consider three Systems, so a decision among the set $\mathcal{D} = \{d_1, d_2, d_3\}$ has to be, hopefully, taken. The considered Systems are here shown as a whole, but they had been actually proposed at successive steps in time: at first we tried the less invasive choice; then more evidence has been collected; finally we proposed a completely new third System.

The first two Systems (associated to decisions d_1 and d_2) deal with one person, S, who wants to identify her half brother (HB). Their relatedness, if proved, implies they share the father. The first System (d_1) is the one in which, as in Fig. 3a, B is the still unobserved candidate (in blue) and the only person providing evidence (in green) is S. Decision d_2 implies a System differing from d_1 only because the inclusion of M in the set $\mathcal{F}^+$ of the observed evidence, as in Fig. 3b. For both Systems the considered hypotheses are:

- H_0: B and S do not share recent relatives;
- H_1: B is S's half brother.

Note that these two Systems don't require the exhumation of AF's body.

The third System (d_3) deals with a motherless paternity case: B, who provides DNA evidence, would like to identify AF, now the Candidate, as his *Father* (F), as in Fig. 3c. For this reason d_3 allows for the possibility to exhume AF and hypotheses are modified as follow:

- H_0: AF and B do not share recent relatives;
- H_1: AF is B's father.

Outcomes. The LR distributions have been obtained using the quasi-equivalence classes approach, as described in Section 5. Values for ε have been iteratively employed to obtain a relative difference between the approximations for all the LR subsets of interest equal or less to 10^{-3}.

Loss and Utility functions. We have used the loss and utility functions defined in (4.4 - 4.7). Thresholds τ_0 and τ_1 are defined by means of the Judge preferences: $Pr(H_0) = Pr(H_1) = 0.5$ are the prior probabilities and the posteriors probabilities are equal to 0.9933, as a consequence, by (3.1), $\tau_0 = 0.00675$ and $\tau_1 = 148.254$. Also $\tau = 1$ is considered.

[4] `http://www.aabb.org/sa/facilities/Documents/rtannrpt08.pdf`

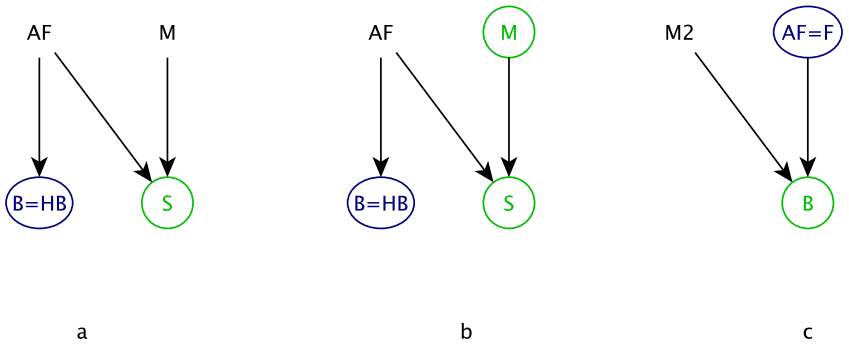

FIG 3. *Pedigree structures for the three Systems in the case study, under H_1 (blue indicates the unobserved candidate, green for the observed evidence).*

Results are gathered in Tab. 1 and 2 in which Systems are evaluated under three possible combinations of population and segregation models: HW+ML, UAF+ML and HW+MMM, realizing a sensitivity analysis with respect to the models' assumptions. For every actor involved two values are provided: the expected loss if the conservative aptitude is assumed and the expected utility if the problem-solver aptitude holds. Obviously a satisfactory System should have small expected loss and high expected utility values. Furthermore, in order to give an idea of the LR sample space reduction, Tab. 1 shows the size of $|\mathcal{LR}^*|$ to be compared with $|\mathcal{LR}| = 1.6243 \times 10^{29}$.

Results in Tab. 1 referring to the τ_0 and τ_1 thresholds, show that the three Systems produce very small values of expected loss for all the actors involved. Because these values can be interpreted as the probability to obtain LR values which strongly support the hypothesis not supposed to hold leading to a wrong decision, the conservative aptitude seems to be satisfied in all the circumstances.

If instead we focus our attention on the values of the expected utility for the baseline models, neither decision d_1 nor d_2 produce acceptable performance irrespectively of the actor. Results are even worse if the more realistic UAF and MMM models are introduced. On the opposite, looking at the System d_3, the value of the utility approaches its maximum, 1, for all the considered population and segregation models. Obviously d_3 implies the exhumation of the AF's body which is unfortunately an invasive activity. Finally, each system shows a substantial reduction of the LR size.

To conclude we consider the simplified setting in which $\tau_0 = \tau_1 = \tau = 1$. Now the expected loss corresponds to the probability to get LR values sup-

			d_1	d_2	d_3
		$\mathcal{F}^+$	S	S & M	B
		C	B	B	AF
		$\mathcal{U}$	HB	HB	F
HW+ML	pro-id	$E(u\|d_i)$	0.2018	0.4006	$\simeq 1$
		$E(l\|d_i)$	5.054×10^{-4}	7.9052×10^{-4}	$\simeq 0$
	con-id	$E(u\|d_i)$	0.1636	0.3094	0.9999954
		$E(l\|d_i)$	5.57×10^{-4}	7.55×10^{-4}	4.6×10^{-6}
	Judge	$E(u\|d_i)$	0.1827	0.355	0.9999977
		$E(l\|d_i)$	5.312×10^{-4}	7.7276×10^{-4}	2.3×10^{-6}
		$\|\mathcal{LR}^*\|$	394347	323315	56624
UAF+ML	pro-id	$E(u\|d_i)$	0.1049	0.3747	$\simeq 1$
		$E(l\|d_i)$	3.4×10^{-4}	7.62×10^{-4}	$\simeq 0$
	con-id	$E(u\|d_i)$	0.0999	0.2917	0.9999939
		$E(l\|d_i)$	3.57×10^{-4}	7.45×10^{-4}	6.07×10^{-4}
	Judge	$E(u\|d_i)$	0.1024	0.3332	0.999997
		$E(l\|d_i)$	3.49×10^{-4}	7.54×10^{-4}	3.04×10^{-4}
		$\|\mathcal{LR}^*\|$	341337	554167	200615
HW+MMM	pro-id	$E(u\|d_i)$	0.199	0.397	0.9982
		$E(l\|d_i)$	4.98×10^{-4}	7.85×10^{-4}	9×10^{-7}
	con-id	$E(u\|d_i)$	0.1602	0.304	0.9996
		$E(l\|d_i)$	5.5×10^{-4}	7.5×10^{-4}	1.69×10^{-5}
	Judge	$E(u\|d_i)$	0.1796	0.3505	0.9989
		$E(l\|d_i)$	5.24×10^{-4}	7.675×10^{-4}	8.9×10^{-6}
		$\|\mathcal{LR}^*\|$	539301	576362	297709

TABLE 1

Results of the System evaluation for the case study $\tau_0 = 0.00675$ *and* $\tau_1 = 148.254$.

porting the hypothesis not assumed to hold, whereas the expected utility represent the probability to obtain LR values supporting the assumed hypothesis. As displayed in Tab. 2, losses implied by d_1 and d_2 reach high figures that cannot be ignored and make even more important to switch to decision d_3, leaving unaltered the necessity of getting DNA evidence from AF.

7. Discussion. In this paper we propose a methodology to deal with the evaluation of probabilistic tools devoted to perform kinship analyses.

The goal is to provide information on the results deriving from the analysis *before* the identification process is undertaken. The analysis moves considering the DNA evidence belonging to the individuals promoting the identi-

			d_1	d_2	d_3
		$\mathcal{F}^+$	S	S & M	B
		C	B	B	AF
		$\mathcal{U}$	HB	HB	F
HW+ML	pro-id	$E(u\|d_i)$	0.8846	0.9193	$\simeq 1$
		$E(l\|d_i)$	0.1154	0.0807	$\simeq 0$
	con-id	$E(u\|d_i)$	0.8895	0.9268	0.999995
		$E(l\|d_i)$	0.1105	0.0732	4.63×10^{-6}
	Judge	$E(u\|d_i)$	0.8871	0.9231	0.999998
		$E(l\|d_i)$	0.1129	0.0769	2.32×10^{-6}
UAF+ML	pro-id	$E(u\|d_i)$	0.8622	0.916	$\simeq 1$
		$E(l\|d_i)$	0.1378	0.084	$\simeq 0$
	con-id	$E(u\|d_i)$	0.8635	0.9228	0.999994
		$E(l\|d_i)$	0.1365	0.0772	6.07×10^{-6}
	Judge	$E(u\|d_i)$	0.8629	0.9194	0.999997
		$E(l\|d_i)$	0.1371	0.0806	3.04×10^{-6}
HW+MMM	pro-id	$E(u\|d_i)$	0.884	0.9185	0.999968
		$E(l\|d_i)$	0.116	0.0815	3.2×10^{-5}
	con-id	$E(u\|d_i)$	0.889	0.9261	0.999906
		$E(l\|d_i)$	0.111	0.0739	9.4×10^{-5}
	Judge	$E(u\|d_i)$	0.887	0.9223	0.999937
		$E(l\|d_i)$	0.113	0.0777	6.3×10^{-6}

TABLE 2

Results of the system evaluation for the case study with $\tau = 1$.

fication trial, but not that of the candidate. Obviously this procedure can be performed without any additional costs nor laboratory work since it uses only a subset of the data required for a kinship analysis.

This implies to perform the traditional LR computation only *after* the identification System has been certified to achieve the characteristics required for the case of interest according to measures of loss and/or utility.

The main contribution of the paper is to acknowledge the large variety of behaviours of kinship identification Systems related to specific cases. This implies the need to use different amounts of information to reach satisfactory and well-specified standards of performance. The traditional requirement of additional observations for design variables, here becomes the request of additional genetic profiles from family members and/or the increase of the number of typed loci or an entirely different identification System to achieve an acceptable level of expected utility or loss. In our opinion the whole matter is relevant since, up to now, the capabilities of a proposed

identification System have not been revealed to the parts, including those called to express the final judgement on the identification trial. Moreover the proposal represents a way to take into account the third criteria of the sentence *Daubert v. Merree Dow Pharmaceutical Inc.* of the Supreme Court[5]. Specifically, to determine whether science is reliable and admissible, it wonders: "In the case of a particular technique, does it have a known error rate and standards controlling the techniques operations?"

APPENDIX A: LR EQUIVALENCE CLASSES FOR IDENTIFICATION ON THE DIRECT LINEAGE

Here we consider a case in which an individual, the DNA donor, is trying to identify a candidate as the family member U posed n generations far on the direct lineage. If $n = 1$ this is a motherless paternity case, if $n = 2$ it is the case of a grandparent trying to identify a candidate as the nephew, and so on. We illustrate how the number of *different* LRs arising in this circumstances, is not equal to the number of possible genotypes the candidate can assume, $k(k+1)/2$, but it is a number independent of k and n, being k the number of allele in the locus.

Let $X = (a_r, a_s)$ be the genotype of the donor and assume the population alleles' probabilities θ are known. For the sake of simplicity make use of the HW and the ML models.

On the donor lineage, consider the probability distribution of the transmitted allele. At first generation, $n = 1$, it can assume only two values, a_r and a_s, with probability 0.5. For $n > 1$, the probability to observe the a_r or a_s is equal to 0.5^n (if they are IBD) plus the probability to come from the no-donor lineage.

Let A^n be the distribution of the allele n generations after the donor had provided $X^0 = (a_r, a_s)$, then:

$$Pr\Big(A^n = i | X^0 = (a_r, a_s)\Big) = \begin{cases} (0.5)^n + (1 - (0.5)^{n-1})\theta_i, & \text{if } i \in \{r, s\}, \\ (1 - (0.5)^{n-1})\theta_i, & \text{if } i \notin \{r, s\}, \end{cases}$$

for $n > 1$.

Since the allele coming from the no-donor lineage still has a probability ruled by the population parameters, the genotype probability along the

[5] *Daubert v. Merrell Dow Pharmaceuticals Inc.*, 113 S. Ct. 2786, 1993.

generations, X^n, $Pr\Big(X^n = (a_i, a_j)|X^0 = (a_r, a_s)\Big)$ results to be:

$$\begin{cases} (0.5)^n(\theta_r + \theta_s) + (1 - (0.5)^{n-1})2\theta_r\theta_s, & \text{if } i = r, j = s, \\ (0.5)^n(\theta_j) + (1 - (0.5)^{n-1})2\theta_r\theta_j, & \text{if } i = r, j \neq s, \\ (0.5)^n(\theta_i) + (1 - (0.5)^{n-1})2\theta_s\theta_i, & \text{if } i \neq r, j = s, \\ (0.5)^n(\theta_r) + (1 - (0.5)^{n-1})\theta_r^2, & \text{if } i = r, j = r, \\ (0.5)^n(\theta_s) + (1 - (0.5)^{n-1})\theta_s^2, & \text{if } i = s, j = s, \\ (1 - (0.5)^{n-1})2\theta_i\theta_j, & \text{if } i \neq r, j \neq s. \end{cases}$$

For this reason the $LR = \dfrac{Pr(X^n = (a_i, a_j)|X^0 = (a_r, a_s))}{Pr(X^n = (a_i, a_j)|\theta)}$ is:

$$\begin{cases} (0.5)^{n+1}\dfrac{(\theta_r + \theta_s)}{\theta_r\theta_s} + (1 - (0.5)^{n-1}), & \text{if } i = r, j = s, \\ (0.5)^{n+1}\theta_r^{-1} + (1 - (0.5)^{n-1}), & \text{if } i = r, j \neq s, \\ (0.5)^{n+1}\theta_s^{-1} + (1 - (0.5)^{n-1}), & \text{if } i \neq r, j = s, \\ (0.5)^{n}\theta_r^{-1} + (1 - (0.5)^{n-1}), & \text{if } i = r, j = r, \\ (0.5)^{n}\theta_s^{-1} + (1 - (0.5)^{n-1}), & \text{if } i = s, j = s, \\ 1 - (0.5)^{n-1}, & \text{if } i \neq r, j \neq s. \end{cases} \tag{A.1}$$

The last line shows that for descendant's genotypes with alleles different from a_r and a_s the LR always assumes the value of $1 - (0.5)^{n-1}$. This circumstance reduces the LR sample space to 6 or 3 possible states, depending if the donor is heterozygous or homozygous, respectively.

REFERENCES

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

Aitken, C. and Taroni, F. (2004). *Statistics and the valuation of evidence for forensic scientist*. John Wiley and Sons, Chichester, UK.

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

Brenner, C. (1997). Symbolic Kinship Program. *Genetics.* **145** 533–542.

Cook, R. et al. (1998). A model for case assessment and interpretation. *Science & Justice.* **38**(3) 151–156.

Evett, I. and Buckleton, J. (1996). Statistical Analysis of STR Data. In *Advances in Forensic Haemogenetics.* Springer-Verlag Heilderberg.

Taroni, F. et al. (2007). Value of DNA Tests: A Decision Perspective. *Journal of Forensic Sciences.* **52**(1) 31–39.

Brenner, C. and Staub, C. (2003). Can DNA solve this? *Poster presented at the 14th International Symposium on Human Identification. Phoenix, AZ.*

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

Vicard, P. and Dawid, A.P. (2004). A statistical treatment of biases affecting the estimation of mutation rates. *Mutation Research.* **547** 19–33.

Green, P.J. and Mortera, J. (2009). Sensitivity of Inference in Forensic Genetics to Assumptions About Founding Genes. *The Annals of Applied Statistics.* **3**(2) 731–763.

Cowell, R. et al (1999). *Probabilistic Expert Systems and Bayesian Networks.* Springer Verlag, New York, NY

Dawid, A.P. et al. (2002). Probabilistic expert systems for forensic inference from genetic markers. *Scandinavian Journal of Statistics.* **29** 577–595.

Lauritzen, S. and Sheehan, N. (2002). Graphical Models for Genetic Analyses. *Statistical Science.* **18** 489–514.

Murphy, K.P. (2001). The Bayes Net Toolbox for Matlab *Computing Science and Statistics.* **33** 1024–1034.

Brisighelli, A.P. et al. (2009). Allele frequencies of fifteen STRs in a representative *Forensic Science International: Genetics.* **3**(2) 29–30.

Brinkmann, B. et al. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *American Journal of Human Genetics.* **62** 1408–1415.

Address of the First and Second authors:
Dipartimento di Statistica "G. Parenti"
Viale Morgagni, 59
50134 Firenze
Italy
E-mail: corradi@ds.unifi.it
ricciardi@ds.unifi.it