Available online at www.sciencedirect.com
ScienceDirect
Forensic Science International 162 (2006) 17–23
Forensic Science International
www.elsevier.com/locate/forsciint

On statistical analysis of forensic DNA: Theory, methods and computer programs

Wing K. Fung [a,*], Yue-Qing Hu [a,b], Yuk-Ka Chung [a]

[a] Department of Statistics and Actuarial Science, The University of Hong Kong, Pokfulam Road, Hong Kong, China
[b] Department of Mathematics, Southeast University, Nanjing 210096, China

Available online 25 July 2006

Abstract

Statistics plays an important role in evaluating the evidential weight of forensic DNA. In this paper, general statistical principles for forensic DNA analysis are presented. We introduce the theory and methods for the statistical assessment in kinship determination and DNA mixture evaluation. In particular, analytical formulas for testing for biological relationship among three individuals and for assessing the DNA mixture evidence in the case of multiple subdivided ethnic groups are developed. Two user-friendly computer programs are demonstrated to exhibit their wide applicability in tackling with complex kinship/paternity and mixture problems. The EasyDNA program can solve a complicated paternity case in 1 min.

Keywords: Forensic DNA; Statistics; Computer programs; EasyDNA

1. Introduction

Statistics has been playing an important role in forensic science. Recently, several books were published on the use of statistics in forensic science and in the courtroom [1–4]. In particular, the use of statistics on DNA forensics has attracted special attention [5–7]. In the Second National Research Council Report (USA) on the evaluation of forensic DNA evidence [8], there were a lot of discussions on the statistical aspects of forensic DNA, and a number of recommendations related to the proper use of statistics were suggested.

In this paper, we are giving a review on the general theory and methods of statistical DNA forensic. We will concentrate more on the statistical aspects of paternity/kinship testing and DNA mixed stains analysis. A formula is newly established to determine if three individuals are ''partially related''; and it generalizes the well known result of Li and Sacks [9] on testing the biological relationship of two individuals. In the analysis of DNA mixtures, the general match probability formula for multiple subdivided ethnic groups is explicitly provided. Besides analytical formula, we will also discuss on the available software which is particularly useful for the analysis of complex kinship and DNA mixture problems.

2. The use of statistics

Why statistics has attracted (much) more attention in DNA forensics than in other areas of forensic science? Several possible reasons are given as follows. First, DNA profiling is generally scientifically unambiguous and very powerful. Since the DNA evidence is repeatable, statistical evaluation would then be possible and in most situations objective. Second, when there is a match to the DNA evidence, people would like to know how likely there is a random match. Third, extremely small probabilities are commonly encountered in DNA profiling, and people are curious about their derivations and interpretations (note: these probabilities are sometimes interpreted incorrectly, e.g., prosecutor's fallacy). Fourth, many forensic scientists are not that familiar with statistics, particularly on different approaches of the subject. Fifth, some problems such as kinship determinations and DNA mixtures need complex statistical analysis. In the following, we are going to discuss about various statistical approaches, and the

* Corresponding author. Tel.: 852 2859 1988; fax: +852 2858 9041.
E-mail address: wingfung@hku.hk (W.K. Fung).

0379-0738/$ – see front matter © 2006 Elsevier Ireland Ltd. All rights reserved.
doi:10.1016/j.forsciint.2006.06.025

principles that have been adopted in DNA forensics. For simplicity, Hardy–Weinberg (HW) and linkage equilibrium are assumed, unless otherwise stated.

The first approach that is commonly used is the match probability (MP) approach. We consider a criminal case that a blood specimen (of the perpetrator) was found at the crime scene and a suspect had been identified. Suppose the DNA profiles of the suspect at three loci (for illustration purpose only) are typed as D3S1358: $\{16, 16\}$, vWA: $\{17, 19\}$, FGA: $\{24, 25\}$, and they are found to be matched with the crime scene DNA. The match probability approach evaluates the probability of having a match under the following defense explanation/hypothesis,

H_d : A random man (not the suspect) is the contributor of the DNA sample,

which is obtained as $(0.289)^2(2 \times 0.245 \times 0.109)(2 \times 0.202 \times 0.116) = 2.09 \times 10^{-4}$ for Chinese population [10]. Such an approach is commonly used in single source sample situations.

The exclusion probability (EP) approach instead considers the probability that

$\bar{H}$: A random man is excluded as the contributor of the DNA sample.

In the above example, the hypotheses H_d and $\bar{H}$ are complementary to each other, with $\text{EP} = 1 - \text{MP}$, and so the two approaches are equivalent.

The third approach is the use of likelihood ratio (LR). This approach, however, considers two hypotheses instead of one, namely the prosecution and defense hypotheses:

H_p : The suspect is the contributor of the DNA sample,
H_d : A random man is the contributor of the sample.

The LR is generally defined as

$$\text{LR} = \frac{P(\text{DNA Evidence} E|H_p)}{P(\text{DNA Evidence} E|H_d)}. \quad (1)$$

In the above particular example, $\text{LR} = 1/\text{MP}$, and so the match probability, exclusion probability and likelihood ratio approaches are all equivalent.

The fourth approach is the Bayesian approach which computes the posterior odds as follows:

$$\text{posterior odds} = \text{LR} \times \text{prior odds}.$$

The prior odds are often determined by non-DNA evidence, which can be subjective in some situations. The LR is based on scientific DNA evidence and still has to be computed in the Bayesian approach.

Next, we consider a more complex DNA mixture problem. Suppose that the mixture M (a mixed blood stain from two perpetrators) is found in the crime scene. A suspect is arrested and his genotype at locus TPOX is typed as $\{8, 9\}$, which matches with the mixture $M = \{8, 9, 11\}$. Two explanations/hypotheses are considered as follows:

H_p : The suspect and an unknown person are contributors of the mixture,
H_d : Two unknown persons (not the suspect) are contributors (2)

The likelihood ratio approach, as in the single source case, evaluates the weight of DNA evidence according to Eq. (1) under both H_p and H_d, though the probability under H_p (the numerator) is no longer equal to 1. The match probability approach, however, evaluates the probability under H_d only. The exclusion probability approach essentially does in a similar way. Thus, the likelihood ratio, match probability and exclusion probability approaches are no longer equivalent. The latter two are not very appropriate since they do not consider the weight of DNA evidence under the prosecution explanation H_p. Moreover, the LR approach evaluates the DNA evidence by considering explanations from both the prosecution and defense sides, which seems to be in line with the situation in common court case.

3. Paternity and kinship testing

3.1. Paternity testing

For a standard trio paternity case, suppose we have obtained the following genotypes at locus D3S1358: Alleged Father $\text{AF} = \{14, 17\}$, Child $C = \{14, 15\}$ and Mother $M = \{15, 16\}$. Two alternative explanations are posed as follows:

H_p : Alleged father is the true biological father of the child,
H_d : A random man is the true father. (3)

In assessing the weight of DNA evidence the paternity index is often considered:

$$\text{PI} = \frac{P(\text{DNA Evidence} E|H_p)}{P(\text{DNA Evidence} E|H_d)},$$

which is in fact a LR. The LR or PI in this case can be evaluated as

$$\begin{aligned}\text{LR} &= \frac{P(E|H_p)}{P(E|H_d)} = \frac{P(C, M, \text{AF}|H_p)}{P(C, M, \text{AF}|H_d)} \\ &= \frac{P(C|M, \text{AF}, H_p)}{P(C|M, \text{AF}, H_d)} \cdot \frac{P(M, \text{AF}|H_p)}{P(M, \text{AF}|H_d)} = \frac{P(C|M, \text{AF}, H_p)}{P(C|M, \text{AF}, H_d)} \\ &= \frac{1/4}{p_{14}(1/2)} = \frac{1}{2p_{14}}.\end{aligned}$$

Refer to [5] for the PI's for other possible genotypes of the trios.

There are some other paternity testing problems similar to the standard trio case, e.g., paternity testing without mother being typed, determination of both parents, standard trio case but having H_d: true father is a brother of the alleged father. Interested readers may refer to [5] for details.