Evaluation of forensic evidence in DNA mixture using RMNE

Wing Kam Fung* and Yuk-Ka Chung
Department of Statistics and Actuarial Science, University of Hong Kong, Hong Kong
*Corresponding author: Wing Kam Fung, email: wingfung@hku.hk

Abstract

To establish the link between an arrested suspect and a crime case based on a DNA mixture, one of the two main statistical tools used by forensic scientists is the random man not excluded (RMNE) probability. The traditional RMNE approach omits any knowledge on the number of contributors and is commonly regarded as being less powerful than the likelihood (LR) approach. In view of the simplicity of interpretation of RMNE, which is the major advantage of using it to present DNA evidences in court, we present a new concept for the interpretation and calculation of the RMNE probability. A new approach for determining the non-exclusion of a random man is proposed, upon which a general formula for the calculation of RMNE probability is developed. By taking account of the number of contributors, the new RMNE probability can be much more powerful for evaluating the evidentiary value of non excluded suspects, compared to the traditional RMNE approach. As illustrated by an example based on a real rape case, our approach can be easily implemented and can shorten the gap between the two approaches by utilizing more information of the case.

Keywords: Deoxyribonucleic acid, Probability of exclusion, Random man not excluded

1 Introduction

When a crime is committed, it may happen that the biological trace found at the crime scene is a mixed blood stain or semen, from which a mixture of Deoxyribonucleic acid (DNA) from more than one person is obtained. In the context of forensic analysis on DNA mixtures, the probability of exclusion (PE) is the probability of excluding a random person as a possible contributor to the observed DNA mixture. If a suspect is not excluded as the unknown contributor to the DNA mixture according to the DNA profiles, a large value of the PE will indicate a strong evidence to convict the suspect. This method is often referred to as the "random man not excluded" ($RMNE$) approach whereas the $RMNE$ probability is the mathematical complement of PE (Buckleton *et al.*, 2005; Gill *et al.*, 2006; Fung and Hu, 2008; Nieuwerburgh *et al.*, 2009). The $RMNE$ probability evaluates the proportion of the population that would by chance have the DNA profiles present in the mixture. If a mixture contains alleles $\{a, b\}$ at a particular locus, any individual with genotype aa, ab, or bb is not excluded as the possible contributor and the corresponding $RMNE$ probability at this locus is $(p_a + p_b)^2$ where p_a, p_b are respectively the allele frequencies of a, b. In the calculation of such $RMNE$ probability, only the DNA profiles of the mixture and the known contributors are taken into account, while the DNA profile of the suspect and other quantitative data such as the number of contributors is not used. In contrast, the likelihood ratio (LR) approach that fully utilizes all the available information is commonly accepted as being more powerful. However, the result obtained from the $RMNE$ approach is still conceptually correct and the simplicity of the method makes it much easier than

the LR approach to present the DNA evidence in court (Buckleton *et al.*, 2005; Gill *et al.*, 2006). See Buckleton and Curran (2008) for a thorough discussion on the merits and drawbacks of the two approaches. It was concluded that "the $RMNE$ statistic wastes information that should be utilized" while the "LRs are more difficult to be presented in court". In fact, LRs are not only difficult to be presented in court, but also difficult to be understood by the jury. People without statistical training may easily fall into the trap of the Prosecutor's Fallacy by misinterpreting the LR as posterior odds (Evett and Weir, 1998). Nevertheless, the debates on whether the $RMNE$ approach or the LR approach should be used in court is still ongoing, despite the extensive reviews and discussions in the literature, see for example Gill *et al.* (2006) and Buckleton and Curran (2008) among others. In this article, we seek to provide a unified and generic approach for the calculation of $RMNE$ probability that also takes into account the number of contributors, thereby shortening the gap between the two approaches.

The remaining part of the paper is organized as follows. Firstly, we present a general formula for calculating the $RMNE$ probability for any known number of contributors. The formula is simple and generic, and therefore can be easily implemented into a computer program. The implementation of the formula is demonstrated by a real example, which shows that the proposed method is much more powerful than the traditional approach. Finally, some concluding remarks are provided.

2 Methods

For a particular locus l, $l = 1, 2, \cdots, L$, denote M_l as the set of alleles found in the mixed stain. A person would not be excluded as a possible contributor to the mixture if all the alleles of his/her genotype at this locus are present in M_l. Under Hardy-Weinberg equilibrium, traditionally, the $RMNE$ probability at this locus can be evaluated as

$$RMNE_l = \left(\sum_{A_i \in M_l} p_i \right)^2, \tag{1}$$

where p_i is the population allele frequency of allele A_i at locus l. Assuming linkage equilibrium, i.e. independence of alleles across all loci, the overall $RMNE$ probability is calculated as

$$RMNE = \prod_{l=1}^{L} RMNE_i. \tag{2}$$

The calculation of the $RMNE$ probability, and hence the exclusion probability $PE = 1 - RMNE$, does not take account of the DNA profile of any individual involved in the case and also the number of contributors to the mixture. In some practical crime cases such as rape cases, the mixed stain found at the body of the victim may be observed as the mixing of body fluids that contain DNA from both the victim and the perpetrator. Therefore some of the alleles in the mixture are known to be contributed by the victim. For instance, suppose in a rape case the mixture $M = \{A_1, A_2\}$ is found to be contributed by a victim with genotype A_1A_1 and an unknown contributor. The allele A_1 in the mixture is explained by the victim's genotype while allele A_2 remains unexplained. The genotype of the perpetrator involved in this case must be either A_1A_2 or A_2A_2.

Under Hardy-Weinberg equilibrium assumption, the $RMNE$ probability should be calculated as $2p_1p_2 + p_2^2$. The traditional approach ignores the information of the victim, and evaluates the $RMNE$ probability as $p_1^2 + 2p_1p_2 + p_2^2$, which is larger and less discriminative. Therefore the allele information from the known contributors, as well as the number of contributors, can enhance the power of the $RMNE$ probability as a measure of the evidentiary value of an non-excluded suspect, using the following formulation of $RMNE$.

Suppose there are x unknown contributors of the mixture M beside the known contributors (e.g. the victim). For simplicity, here we consider only a particular locus l and drop the subscript l from the symbols of the DNA profiles. Denote U as the set of alleles present in M but not in the DNA profiles of the known contributors, S as the set containing all the alleles of k ($k < x$) suspects, and G as the set containing all the alleles of $x - k$ arbitrary individuals. These k suspects will not be excluded as the possible contributors if $U \subset (S \cup G) \subset M$ for some G, subject to the constraint that $1 \leq |G| \leq 2(x - k)$ where $| \bullet |$ is the cardinality of a set.

Proposition 1

$U \subset (S \cup G) \subset M$ for some G if and only if $|U \setminus S| \leq 2(x - k)$ and $S \subset M$.

Proof:

We only consider the case with $k < x$ as it is trivial for $k = x$. Suppose that there exists G such that $U \subset (S \cup G) \subset M$. Then for this G, we have

$$U \subset (S \cup G) \Rightarrow (U \setminus S) \subset (G \setminus S) \subset G \Rightarrow |U \setminus S| \leq |G| \leq 2(x - k)$$

and

$$S \cup G \subset M \Rightarrow S \subset M.$$

On the other hand, suppose $|U \setminus S| \leq 2(x - k)$ and $S \subset M$. If $U \setminus S = \phi$, then $U \subset S$ and hence $U \subset (S \cup G)$ for any arbitrary G. Since S is not empty, we can let G be the set containing only one of the alleles in S. For this G we have $(S \cup G) = S \subset M$. If $U \setminus S \neq \phi$, we can let $G = U \setminus S$ as it fulfills the constraint that $1 \leq |G| \leq 2(x - k)$. Therefore there exists G such that $S \cup G = U$ and hence $U \subset (S \cup G) \subset M$ as $U \subset M$. □

By Proposition 1, the k suspects will not be excluded as the possible contributors if $|U \setminus S| \leq 2(x - k)$ and $S \subset M$. The calculating formula of the corresponding $RMNE$ probability is

$$RMNE = \sum_{R_1} \cdots \sum_{R_k} \left(I_{|U \setminus R| \leq 2(x-k)} I_{R \subset M} \right) P(R_1) \cdots P(R_k) \tag{3}$$

where I_A represents the indicator function that takes the value of 1 if A is satisfied and 0 otherwise; R_i, $i = 1, 2, \ldots, k$ are the set of alleles running over all the possible genotypes; and $R = \cup_{i=1}^{k} R_i$. Assuming Hardy-Weinberg equilibrium, the genotype probabilities $P(R_i)$ can be simply computed by

$$P(A_iA_j) = \begin{cases} p_i^2 & \text{if } i = j \\ 2p_ip_j & \text{if } i \neq j \end{cases}.$$

The general equation (3) can be used for calculating the $RMNE$ probability, and hence the PE, for any number of contributors as well as any number of random persons considered. It can be easily implemented into a computer program by brute force evaluation on all possible genotypes.

3 Implementation

We consider a case reported by Robert Goetz in New South Wales, Australia (Buckleton and Curran, 2008). A woman was sexually assaulted by a number of men. As told by the woman, there were allegedly three or four men involved. The DNA profile of the mixed stain found at the crime scene revealed that there were at least three contributors. However, the victim was excluded as being one of the contributors as her DNA profile did not match the mixture. Three suspects were arrested, with their DNA profiles determined. Table 1 shows the DNA profiles of the mixture and the suspects at three STR loci, as well as the allele frequencies of the Caucasian in New South Wales given in Ayres et al. (2002). Since there was no known contributor, all the alleles in the mixture were unexplained and hence $U = M$. For simplicity, here we assume Hardy-Weinberg equilibrium.

From the testimony of the victim, the prosecution alleged that the three suspects were the contributors of the mixture and set up the following prosecution and defense hypotheses:

H_p : the mixture contains the DNA of suspects 1-3 and no others;
H_d : the mixture contains the DNA of three unknown persons.

As can be seen from Table 1, the three suspects have matched DNA profiles with the mixture and were not excluded. According to the traditional method and using equations (1) and (2), the overall $RMNE$ probability is calculated as 0.3362, i.e. one out of three random men would not be excluded as being the contributor of the mixture. Therefore the weight of evidence is weak if the number of contributors is not taken into account. However, based on the proposed method, using equation (3) with $k = 3$ and $x = 3$, the overall $RMNE$ probability is calculated as 4.787×10^{-6} which is approximately 1 in 200,000, providing a much stronger evidence to convict the suspects.

For illustrative purpose, consider the following hypothetical scenario. Suppose suspects 1 and 2 had confessed to the crime. However, the defense attorney representing suspect 3 argues against the setting of the hypotheses by claiming that suspect 3 was not involved in the sexual assault despite the fact that suspects 1 and 2 did perpetrate this crime. According to this claim, the prosecution and defense hypotheses would become

H_p : the mixture contains the DNA of suspects 1-3 and no others;
H_d : the mixture contains the DNA of suspects 1 and 2 and one unknown person.

Based on the traditional approach, the overall $RMNE$ probability is still 0.3362, despite having a new set of hypotheses. Using the proposed $RMNE$ method by equation (3) with $k = 1$ and $x = 1$, the overall $RMNE$ probability is calculated as 0.0009 which is approximately 1 in 1000. The $RMNE$ probability becomes larger, but is still much smaller than that obtained by using equation (1) according to the traditional method. The power of the $RMNE$ probability as a measure of the weight of evidence is therefore enhanced, by taking account of

Table 1: DNA profiles of the mixture and three suspects in a group rape case in New South Wales, Australia reported in Buckleton and Curran (2008). The allele frequencies are estimated from a sample of Caucasians in New South Wales Ayres et al. (2002)

Locus	*Mixture*	*Suspect 1*	*Suspect 2*	*Suspect 3*	*Frequency*
D3	14	14			0.0978
	15		15		0.2640
	16		16	16	0.2696
	17	17			0.1983
	18			18	0.1522
vWA	16	16			0.2277
	17	17	17		0.2849
	19			19	0.0698
	20		20	20	0.0140
FGA	20		20		0.1327
	21		21	21	0.1872
	24	24			0.1355
	25	25			0.0810
	26			26	0.0377

the number of contributors and the observed DNA profiles. In fact, taking the reciprocals of the $RMNE$ probabilities give exactly the LRs of the evidence under the prosecution and defense hypotheses in both the original setting and the hypothetical scenario. With the aid of our proposed formula in calculating the $RMNE$ probability, the difference between the $RMNE$ and LR approaches becomes less significant than they used to be according to the common claim.

4 Discussion

In this article we have proposed simple representation of the criteria for checking the non-exclusion of arbitrary number of individuals as being the contributors to a DNA mixture, given the known number of contributors and the DNA profiles of the known contributors. Based on the criteria presented in Proposition 1, we have derived a computational formula of the $RMNE$ probability that can be applied in general situations, including the scenario when there are more than one suspect involved. As illustrated by a numerical example, our approach can utilize more information and is shown to be more powerful than the traditional approach as a measure of the evidentiary value of non-excluded suspects.

It should be pointed out that we are not suggesting the use of $RMNE$ probability as a replacement of the LR approach. Both approaches have their own merits. In practice, we may not rely on just one statistical approach in evaluating and interpreting the DNA evidence. In statistics, there is always more than one way to tackle a problem. We concur with the suggestion of Budowle et al. (2009) that forensic scientists should acquire knowledge and skills for both statistical approaches. Here we strongly recommend incorporating as much available information as possible in the calculation if the $RMNE$ probability is chosen as a measure of evidentiary value of the DNA evidence to be presented in the court.

The calculation of the $RMNE$ probability illustrated in the previous section

is based on the assumption of Hardy-Weinberg equilibrium. In reality, Hardy-Weinberg equilibrium may not be satisfied. To correct for the possible deviation from this assumption, we can make use of the coancestry coefficient which is the chance that random alleles taken from two individuals are identical by descent (*ibd*) (Fung and Hu, 2008; Evett and Weir, 1998), and the formula suggested by the NRC II Recommendation 4.1 (National Research Council, 1996). Such correction will have little effects on the complexity of the $RMNE$ probability calculation using equation (3). Nevertheless, one key assumption of equation (3) is the independence among the genotypes of the random men involved. In practical crime cases the involved persons may come from a structured population in which the independence assumption would be violated and NRC II Recommendation 4.2 may be adopted (National Research Council, 1996). Therefore seeking for a general procedure for calculating joint probabilities of genotypes under structured populations constitutes a possible direction of our future work.

References

Ayres, K.L., Chaseling, J., Balding, D.J. (2002) Implications for DNA identification arising from an analysis of Australian forensic databases, *Forensic Sci Int* **129**, 90–98.

Buckleton, J., Triggs, C., Walsh, S.J. (2005) *Forensic DNA Evidence Interpretation.* CRC Press, London.

Buckleton, J., Curran, J. (2008) A discussion of the merits of random man not excluded and likelihood ratios, *Forensic Sci Int: Genetics*, **2**, 343–348.

Budowle, B., Onorato, A.J., Callaghan, T.F., Manna, A.D., Gross, A.M., Guerrieri, R.A., Luttman, J.C., McClure, D.L. (2009) Mixture interpretation: defining the relevant features for guidelines for the assessment of mixed DNA profiles in forensic casework, *J Forensic Sci* **54**, 810–821.

Evett, I.W. and Weir B.S. (1998) *Interpreting DNA Evidence.* Sinauer Associates, Massachusetts.

Fung, W.K., Hu, Y.Q. (2008) *Statistical DNA Forensics: Theory, Methods and Computation.* Wiley, Chichester.

Gill, P., Brenner, C.H., Buckleton, J.S., Carracedo, A., Krawczak, M., Mayr, W.R., Morling, N., Prinz, M., Schneider, P.M., Weir, B.S. (2006) DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures, *Forensic Sci Int*, **160**, 90–101.

National Research Council (1996) *The evaluation of forensic DNA evidence.* :National Academy Press, Washington, D.C.

Nieuwerburgh, F.V., Goetghebeur, E., Vandewoestyne, M., Deforce, D. (2009) Impact of allelic dropout on evidential value of forensic DNA profiles using RMNE, *Bioinformatics*, **25**, 225–229.