Legal Medicine 19 (2016) 16–21

Contents lists available at ScienceDirect

Legal Medicine

journal homepage: www.elsevier.com/locate/legalmed

New stutter ratio distribution for DNA mixture interpretation based on a continuous model

Sho Manabe [a,b], Yuya Hamano [a,c], Chie Morimoto [a], Chihiro Kawai [a], Shuntaro Fujimoto [a], Keiji Tamaki [a,*]

[a] Department of Forensic Medicine, Kyoto University Graduate School of Medicine, Yoshida-Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan
[b] Research Fellow of Japan Society for the Promotion of Science, Japan
[c] Forensic Science Laboratory, Kyoto Prefectural Police Headquarters, 85-3, 85-4, Yabunouchi-cho, Kamigyo-ku, Kyoto 602-8550, Japan

ARTICLE INFO

Article history:
Received 16 October 2015
Received in revised form 4 January 2016
Accepted 13 January 2016
Available online 14 January 2016

Keywords:
Forensic DNA typing
Mixture interpretation
Low-template DNA
Continuous model
Stutter ratio

ABSTRACT

In forensic science, DNA mixture interpretation is traditionally based on a binary model, which does not account for peak-height information in DNA profiles. In recent years, some countries have adopted a continuous model in which peak heights are used and stochastic effects are considered to enable rigorous calculation of likelihood ratios. However, this model requires certain biological parameters which affect the expected allelic and stutter peak heights. In this paper, we focused on estimating the distribution of the stutter ratio (*SR*) in 15 short tandem repeat loci in relation to the allele repeat number. We estimated the *SR* values of 234 single-source DNA samples by using a commercially available kit. In all loci except for D8S1179, D21S11, and D2S1338, a simple log-normal distribution model was fitted to the variability of *SR*. For D21S11, we developed a new distribution model in which distinct log-normal distributions between complete and incomplete repeat units are used (a separate log-normal distribution model). For D8S1179 and D2S1338, we developed another new distribution model that mixes two log-normal distributions to explain two types of repeat structures appearing within the same number of allele repeats. These two models were fitted to the observed *SR* values more accurately than the simple log-normal distribution model. We expected these new *SR* models to be applied to DNA mixture interpretation based on a continuous model.

1. Introduction

In Japanese forensic casework, DNA mixture interpretation using short tandem repeat (STR) loci is based on a binary model, which describes the fact that the probability of the evidence is assigned as 0 or 1. This model is known to make no use of information such as peak heights. In recent years, some countries have begun to use a continuous model to calculate rigorous likelihood ratios [1–3]. In this method, peak heights are used and stochastic effects (e.g., heterozygote imbalance, and allele drop-out) are considered in order to calculate the probability of the peak heights given all the possible genotype combinations of the contributors; therefore, it is unnecessary to use a fixed stutter threshold to determine whether a peak in the stutter position represents a stutter or an allele. Thus, this method can avoid some of the criticism regarding subjectivity [4].

Continuous models incorporate biological parameters such as locus-specific amplification efficiency, heterozygote balance, mixture ratio, stutter ratio (*SR*), and degradation, all of which can affect the expected allelic and stutter peak heights [2]. The distributions of these parameters must be estimated using experimental or empirical data. Here, we describe new models of *SR* distribution for mixture interpretation that are based on a continuous model.

In theory, *SR* is positively correlated with the allele repeat number [5]. However, this is not always the case with complex repeat structures, such as the D21S11 locus, where the alleles have internal variations in repeat sequences. Brookes et al. investigated the underlying behavior of *SR* and showed that the longest uninterrupted stretch (LUS) was a more reliable predictor of *SR* than the repeat number was [6]. For example, TH01 allele 9.3 possesses the structure $[AATG]_6$ ATG $[AATG]_3$ and its *SR* values are close to those of allele 6 because the LUS value of allele 9.3 is 6 [7]. However, in forensic practice, we only type the allele repeat number, not the length of the LUS. Therefore, the models of *SR*

Abbreviations: AIC, Akaike information criterion; LUS, longest uninterrupted stretch; MLE, maximum likelihood estimation; RFU, relative fluorescence unit; SR, stutter ratio; STR, short tandem repeat.

* Corresponding author.
E-mail address: ktamaki@fp.med.kyoto-u.ac.jp (K. Tamaki).

http://dx.doi.org/10.1016/j.legalmed.2016.01.007
1344-6223/

distribution that do not consider the LUS values are desired for mixture interpretation in actual casework.

In this study, we experimentally investigated the *SR* distribution in 15 commercially available STR loci by using single-source DNA samples. We then developed *SR* distribution models without considering the LUS values especially for loci with complex repeat structures.

2. Materials and methods

2.1. STR typing

We used 276 single-source DNA profile analyzed by an AmpFℓSTR® Identifiler® Plus PCR Amplification Kit (Life Technologies, Carlsbad, CA). The PCR products were analyzed using an ABI 3130xl Genetic Analyzer (Life Technologies) and data were analyzed using GeneMapper™ *ID* version 3.2.1 (Life Technologies) with 30 relative fluorescence units (RFU) being used as the limit of detection threshold. This study was approved by the ethics committee of Kyoto University Graduate School and Faculty of Medicine.

2.2. Calculation and modeling of stutter ratio

SR values are calculated as follows:

$$SR = \frac{h_{a-1}}{h_a}$$

where h_a denotes the height (RFU value) of the allele ***a*** (repeat number) and h_{a-1} denotes the stutter peak height of allele ***a***. If the stutter position of an allele was the same as another allelic position in a heterozygote locus, we did not use either allele data for calculating the *SR* values. Next, we investigated the correlation between *SR* and allele repeat numbers in order to determine the probability distribution of the *SR* in each locus. We first adopted the following log-normal distribution model (*LN*) proposed by Bright et al. [8]:

$$SR_{al} \sim LN(\mu_{al}, \sigma_{al}^2).$$

The parameters μ_{al} and σ_{al}^2 are written as

$$\mu_{al} = \log_e(\beta_{0,l} + \beta_{1,l}a)$$

and

$$\sigma_{al}^2 = \frac{\sigma_l^2}{h_a}$$

where μ_{al} is the mean value of the logarithmic *SR* at allele ***a*** in locus *l*. Bright et al. proposed a linear model to describe the relationship between the mean of the *SR* and the LUS values [8]. We replaced the LUS values with allele repeat numbers in our linear model. Thus, according to the linear model, the mean *SR* values in each locus exhibit a linear relationship with allele repeat numbers. The parameters $\beta_{0,l}$ and $\beta_{1,l}$ are the intercept and the slope of the linear model for locus *l*, respectively, and σ_{al}^2 is the variance of the logarithmic *SR* at allele ***a*** with peak height h_a. The variability of *SR* typically increases due to stochastic effects when amplifying low levels of a DNA template [9]. According to the model, the variability of $\log_e(SR)$ increases as h_a decreases. The parameter σ_l^2 is the variance parameter of $\log_e(SR)$ for locus *l*.

We determined the parameters $\beta_{0,l}$, $\beta_{1,l}$, and σ_l^2 by means of a maximum likelihood estimation (MLE). For loci that showed a poor correlation between *SR* and allele repeat numbers, we developed an alternative model instead of the *LN* model. The accuracy of these models was compared by calculating the difference in the Akaike information criterion (AIC):

$$\mathrm{AIC} = 2k - 2\log_e L$$

where k is the number of parameters and $\log_e L$ is the maximum log-likelihood in each model. A model featuring a low AIC is considered a superior fit for the experimental *SR* data as compared with a model featuring a high AIC. All programs used for data interpretation and model fitting were written using the statistical software R (version 3.1.2) [10].

3. Results

3.1. PCR artifacts other than −1 backward stutter peaks

We detected some artifacts other than the typical −1 backward stutter peaks (i.e., one repeat shorter than the allele). Forward stutter peaks or −2 backward stutter peaks were observed in 42 experimental DNA samples. These peaks were excluded from this analysis. Pull-up peaks located in allelic positions were also observed in 79 experimental DNA samples; 1–5 pull-up peaks per sample were observed and the heights of these peaks ranged from 32 to 160 RFU. We excluded 33 DNA samples for estimating the distribution of *SR* because their pull-up peaks were stacked on the stutter peaks. The pull-up peaks in the remaining 46 samples did not affect any allelic and stutter peak heights.

We observed an extreme heterozygote imbalance that was probably caused by a mutation of the primer-binding site ($n = 4$), a tri-allelic pattern ($n = 1$), or an off-ladder allele ($n = 5$, including one sample in which we also detected pull-up peaks stacked on stutter peaks). We also removed these samples for estimating the *SR* distribution because these samples might not be accurately genotyped. Finally, we used 234 DNA samples with a total of 3510 stutter peaks.

3.2. Stutter ratio of loci following simple log-normal model

The *SR* values demonstrated only one trend of positive correlation with allele repeat numbers in all 15 loci except for D8S1179, D21S11, TH01, and D2S1338. Spearman's rank correlation values between the *SR* and the allele repeat numbers in the 11 loci ranged from 0.813 (in TPOX) to 0.929 (in D18S51). Although D3S1358, D19S433, vWA, and FGA were considered to contain variable repeat blocks based on previous sequence analyses [11], our data revealed only one trend of positive correlation between the *SR* values and allele repeat numbers without any LUS influence. Thus, we considered the 11 loci to follow the *LN* model and estimated the parameters $\beta_{0,l}$, $\beta_{1,l}$, and σ_l^2 of each locus by using the MLE method (Table 1). These parameters are different in each locus.

On the other hand, the relationship between the *SR* and the allele repeat numbers could not be explained by only one positive

Table 1
Estimated values of $\beta_{0,l}$, $\beta_{1,l}$, and σ_l^2 in loci following a simple log-normal model.

Locus	$\beta_{0,l}$ ($\times 10^{-2}$)	$\beta_{1,l}$ ($\times 10^{-3}$)	σ_l^2
D7S820	−4.27	8.82	19.2
CSF1PO	−5.16	9.74	13.2
D3S1358	−7.14	9.50	11.0
TH01[a]	−1.96	5.82	62.5
D13S317	−7.11	11.4	42.9
D16S539	−5.91	10.6	14.3
D19S433	−7.39	10.7	9.02
vWA	−17.2	14.5	20.9
TPOX	−2.59	5.84	51.7
D18S51	−3.70	7.50	8.03
D5S818	−4.67	9.52	16.6
FGA	−8.32	7.16	13.2

[a] Regarding the allele repeat numbers of allele 9.3 as 6.