Forensic Science International: Genetics 22 (2016) 149–160

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

CEESIt: A computational tool for the interpretation of STR mixtures

Harish Swaminathan[a], Abhishek Garg[b], Catherine M. Grgicak[c], Muriel Medard[d], Desmond S. Lun[a,b,e,f,*]

[a] Center for Computational and Integrative Biology, Rutgers University, Camden, NJ 08102, USA
[b] Department of Computer Science, Rutgers University, Camden, NJ 08102, USA
[c] Biomedical Forensic Sciences Program, Boston University School of Medicine, Boston, MA 02118, USA
[d] Research Laboratory for Electronics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[e] Department of Plant Biology and Pathology, Rutgers University, New Brunswick, NJ 08901, USA
[f] School of Information Technology and Mathematical Sciences, University of South Australia, Mawson Lakes, SA 5095, Australia

ARTICLE INFO

Article history:
Received 26 June 2015
Received in revised form 25 January 2016
Accepted 10 February 2016
Available online 23 February 2016

Keywords:
Mixture interpretation
Likelihood ratio
p-Value
DNA analysis
LR distribution

ABSTRACT

In forensic DNA interpretation, the likelihood ratio (LR) is often used to convey the strength of a match. Expanding on binary and semi-continuous methods that do not use all of the quantitative data contained in an electropherogram, fully continuous methods to calculate the LR have been created. These fully continuous methods utilize all of the information captured in the electropherogram, including the peak heights. Recently, methods that calculate the distribution of the LR using semi-continuous methods have also been developed. The LR distribution has been proposed as a way of studying the robustness of the LR, which varies depending on the probabilistic model used for its calculation. For example, the LR distribution can be used to calculate the *p*-value, which is the probability that a randomly chosen individual results in a LR greater than the LR obtained from the person-of-interest (POI). Hence, the *p*-value is a statistic that is different from, but related to, the LR; and it may be interpreted as the false positive rate resulting from a binary hypothesis test between the prosecution and defense hypotheses. Here, we present CEESIt, a method that combines the twin features of a fully continuous model to calculate the LR and its distribution, conditioned on the defense hypothesis, along with an associated *p*-value. CEESIt incorporates dropout, noise and stutter (reverse and forward) in its calculation. As calibration data, CEESIt uses single source samples with known genotypes and calculates a LR for a specified POI on a question sample, along with the LR distribution and a *p*-value. The method was tested on 303 files representing 1-, 2- and 3-person samples injected using three injection times containing between 0.016 and 1 ng of template DNA. Our data allows us to evaluate changes in the LR and *p*-value with respect to the complexity of the sample and to facilitate discussions regarding complex DNA mixture interpretation. We observed that the amount of template DNA from the contributor impacted the LR – small LRs resulted from contributors with low template masses. Moreover, as expected, we observed a decrease of *p*-values as the LR increased. A *p*-value of 10^{-9} or lower was achieved in all the cases where the LR was greater than 10^{8}. We tested the repeatability of CEESIt by running all samples in duplicate and found the results to be repeatable.

1. Introduction

Until recently, statements of inclusion or exclusion were exclusively used when reporting or presenting DNA comparisons to the trier-of-fact. If a suspect, or other known, is 'included' as a potential contributor to the item of evidence, then the inclusion statement must be accompanied by the calculation of a statistic that conveys the strength of the match [1]. Alternatives to inclusion/exclusion statements have, of late, been adopted, where a verbal scale is used to describe the number obtained [2].

Two protocols for calculating a match statistic are the random man not excluded (RMNE) approach, based on the combined probability of inclusion (CPI) statistic, and the likelihood ratio (LR) approach. The RMNE method seeks to determine the fraction of the population that would not be excluded as a contributor to the profile. During the calculation of the CPI statistic, some information like the genotype of the suspect, the peak heights and the number of contributors to the profile is not utilized [3].

* Corresponding author at: Department of Computer Science, Rutgers University, 227 Penn Street, Camden, NJ 08102, USA. Fax: +1 856 225 6624.
E-mail address: dslun@rutgers.edu (D.S. Lun).

http://dx.doi.org/10.1016/j.fsigen.2016.02.005
1872-4973/

Though RMNE is still employed in practice, this method of evaluation is being replaced with the LR approach [4]. The likelihood ratio is defined as:

$$\mathrm{LR} = \frac{Pr(E|H_p, n_p)}{Pr(E|H_d, n_d)},$$

where E is the evidence in the form of the electropherogram (epg); H_p and H_d are the hypotheses specified by the prosecution and the defense, respectively; and n_p and n_d are the number of contributors specified by the prosecution and the defense, respectively. The likelihood ratio can be expressed either as the ratio of probabilities or as the ratio of probability densities, depending on whether the evidence is treated as a discrete or as a continuous random variable. The numerator is the probability of observing the evidence given the prosecution's hypothesis and the denominator is the probability of observing the evidence given the defense's hypothesis. The evidence shows support for the prosecution's hypotheses if $LR > 1$; if $LR < 1$ the defense's hypothesis is supported by the evidence. Unlike the RMNE method, the LR can use information like the number of contributors to the sample, the heights of the peaks observed and the genotype of the suspect [3].

The LR framework can be applied using a binary model that uses the set of alleles observed in a DNA profile [5]. This method assigns a probability of 0 or 1 to genotypes based on the presence or absence of alleles. Alternatives to the binary model have been proposed that allow for drop-in and/or dropout of alleles [6,7]. These 'semi-continuous' methods use the peak heights to establish probabilities of dropout and drop-in. Unlike binary methods, they can be used to interpret profiles in which one or more of the suspect's alleles are not observed, or when there are incidences of drop-in. Fully continuous methods that employ probabilistic genotyping by modeling the peak heights have also been created, resulting in the ability to incorporate stutter and noise, or drop-in, into the calculation of the statistic [8–12]. Fully continuous methods make use of the entire data obtained in the epg, including the qualitative (alleles observed) and the quantitative (peak heights) data. The TrueAllele system [8,9] uses an MCMC sampler to compute a probability for every possible genotype combination based on how well it explains the observed data. The peak heights are linearly modeled with respect to the mixture weights using a multivariate normal distribution. Degradation is modeled as an exponential decay with respect to the allele product length and stutter as a linear function of the allele. Cowell et al. [10] model the peak heights using a Gamma distribution and employ a Bayesian network for analyzing mixtures that incorporate dropout and a stutter model that is independent of DNA mass and the marker. Puch-Solis et al. [11] also use a gamma distribution for stutter and allele heights but differ from [10] by using the total peak height at a locus as a proxy for the DNA mass, estimating parameters conditional on peak heights and jointly modeling stutter and allelic peaks. Both these methods do not take into account drop-in. Taylor et al. [12] implement MCMC using the Metropolis–Hastings algorithm to compute the genotype probabilities. Allele peak heights are modeled using an exponential decay with respect to the molecular weight of the allele and stutter peak heights are modeled as a linear function of the allele height. Drop-in is modeled as an exponential decay with respect to the peak height.

In addition to the methods and models that evaluate LRs, computational methods that compute the LR distribution have recently garnered attention [7,13–16]. The LR distribution can be used to evaluate the robustness of the model by performing Tippett tests – "*what is the probability that a non-contributor will give rise to an LR greater than 1 (Type I error)?*" [13]. Another statistic that can be obtained from the LR distribution is the p-value [7,14–16]. The p-value is a summary statistic that provides the probability that a randomly picked person from the population would give rise to an LR at least as large as the one observed for the person of interest. It can be interpreted as the false positive rate and may be useful when the analyst wants to compute the probability of a random non-contributor giving rise to an LR greater than the one observed for the suspect. While there is controversy surrounding the use of p-values [17], several authors have shown that it is a useful statistic that assists in the interpretation of LRs and has other applications like database searching [18] and kinship analysis [19].

In this work, we seek to combine the twin features of a fully continuous method to calculate the LR and the calculation of the LR distribution and a p-value. To this end, we developed a computational method called 'CEESIt' to calculate the LR for a person of interest, given an STR profile. CEESIt (CEES: computational evaluation of evidentiary signal) is a fully continuous method that works by modeling the peak heights observed in a calibration data set consisting of single source samples with known genotypes. CEESIt accounts for dropout, noise and stutter (both reverse and forward), artifacts observed regularly in low template samples [19]. Additionally, CEESIt also computes a p-value for the LR by sampling a large number of random genotypes from the population. The method was tested on 303, 1-, 2- and 3-person experimental sample files with template masses ranging from 0.016 to 1 ng, and represents the largest empirical study to evaluate the p-value and LR that we know of. We found that the amount of template DNA from the contributor had an impact on the LR—small LRs were associated with low template masses. Since we used 10^9 samples to calculate the p-value, the lowest p-value that CEESIt reports is 10^{-9}, and this was obtained in all the cases where the LR was greater than 10^8. We also tested the system's repeatability and found that the results were repeatable.

2. Materials and methods

2.1. Calibration set

CEESIt employs a continuous method to calculate the LR and hence uses the peak heights in the signal to calculate probabilities. Characterization of the peak heights is accomplished by using single source calibration profiles with known genotypes obtained from samples amplified from a wide range of input DNA masses. For a detailed description of how the calibration samples (Calibration Set – Supplementary Table 1) were created, refer to [20]. Briefly, DNA was extracted from 28 individuals. Absolute DNA quantification was performed using real-time PCR and the Quantifiler® Duo™ Quantification kit according to the manufacturer's recommended protocol and one external calibration curve [21,22]. The extracted DNA was amplified using the manufacturer's recommended protocol for AmpFℓSTR® Identifiler® Plus Amplification Kit (Life Technologies, Inc.) [23]. Separation of the STR fragments was accomplished with a 3130 Genetic Analyzer using an injection voltage of 3 kV and injection times of 5, 10 and 20 s. Analysis was performed using GeneMapper IDX v1.1.1 (Life Technologies, Inc.) and an RFU threshold of 1. A threshold of 1 RFU was used in order to capture all peak height information, i.e. the allelic peaks, baseline noise and stutter peaks, in the signal. Known artifacts such as pull-up, spikes, -A, and artifacts due to dye dissociation were manually removed, as previously detailed in [20].

2.2. Testing set

A total of 303 1-, 2- and 3-person sample files were used to test CEESIt (Testing set – Supplementary Tables 2–4). These 1-person test samples were created using the same protocol described for the single source samples in the calibration set. The mixtures were