Forensic Science International: Genetics 35 (2018) 113–122

Contents lists available at ScienceDirect

Forensic Science International: Genetics

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

Research paper

A sensitivity analysis to determine the robustness of STRmix™ with respect to laboratory calibration

Hannah Kelly[a,*], Jo-Anne Bright[a], Maarten Kruijver[a,b], Stuart Cooper[a], Duncan Taylor[c,d], Kyle Duke[e], Melissa Strong[f], Vickie Beamer[g], Christina Buettner[h], John Buckleton[a,i]

[a] Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland, 1142, New Zealand
[b] Department of Mathematics, VU University, Amsterdam, Netherlands
[c] Forensic Science South Australia, 21 Divett Place, Adelaide, SA 5000, Australia
[d] School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia
[e] California Department of Justice Bureau of Forensic Services, United States
[f] San Diego Police Department Crime Laboratory, CA, United States
[g] Scottsdale Police Department Crime Laboratory, AZ, United States
[h] Wyoming State Crime Laboratory, United States
[i] University of Washington, Department of Biostatistics, Seattle, WA 98195, United States

ARTICLE INFO

Keywords:
Interpretation
Mixtures
Likelihood ratio
PCAST report
Probabilistic genotyping
Validation

ABSTRACT

STRmix™ uses several laboratory specific parameters to calibrate the stochastic model for peak heights. These are modelled on empirical observations specific to the instruments and protocol used in the analysis. The extent to which these parameters can be borrowed from laboratories with similar technology and protocols without affecting the accuracy of the system is investigated using a sensitivity analysis. Parameters are first calibrated to a publicly available dataset, after which a large number of likelihood ratios are computed for true contributors and non-contributors using both the calibrated parameters and several borrowed parameters. Differences in the *LR* caused by using different sets of parameter values are found to be negligible.

1. Introduction

Probabilistic genotyping (PG) is now widely used to interpret DNA profiles for forensic purposes. PG refers to the use of biological modelling to describe the behaviour of DNA encompassed within a mathematical model [1]. A first generation of PG software reduced the electrophoretic data (peak designations, heights, and size) to the presence or absence of alleles and is often referred to as a semi continuous or drop model. A second generation of PG software explicitly models the peak heights in the electropherogram (epg) as random variables and is referred to as a fully continuous model.

The validation of any new technique is important prior to its use in a forensic laboratory. Validation studies help define the scope or range of conditions under which reliable results may be obtained [2]. With respect to PG, validation is used to verify the functionality of the system, accuracy of the calculations, and the limits of the software. Developmental validation is typically undertaken by the developer of the software, whereas internal validation is undertaken by a laboratory prior to its use in casework. The behaviour of peak heights depends on the instrument, protocol, DNA quantity, degradation of the mixture components, dye, locus, and even the allele that is amplified. Capturing these effects is essential when validating a fully continuous system and determining whether it is fit for purpose.

Current ISFG (International Society of Forensic Genetics) guidelines for the internal validation of probabilistic genotyping software recommend empirical studies to ensure that the software runs as expected. ISFG recommend that internal validation of PG should cover a wide range of functionality of the software and all relevant parameter settings [3]. The SWGDAM (Scientific Working Group on DNA Analysis Methods) guidelines for the validation of PG systems state that internal validation is the accumulation of test data within the laboratory to establish that the parameters, software settings, formulae, algorithms, and functions perform as expected [1].

STRmix™ is one example of a PG system that implements a fully continuous model for the interpretation of forensic DNA profiles [4,5]. Developmental validation of STRmix™ following the SWGDAM guidelines has previously been published [6]. Internal validation studies have also been published [7,8].

STRmix™ uses the quantitative information from an electropherogram such as peak heights (O), to calculate the probability of the

* Corresponding author.
E-mail address: Hannah.kelly@esr.cri.nz (H. Kelly).

https://doi.org/10.1016/j.fsigen.2018.04.009
Received 20 March 2018; Received in revised form 25 April 2018; Accepted 27 April 2018
Available online 27 April 2018
1872-4973/

profile given all possible genotype combinations, S_j. The observed height of each allele and stutter peak is a random variable, drawn from a distribution that varies about some expected value modelled from the profile. If the expected value is obtained correctly, then the difference from expectation of each peak is independent. The model used to compare an observed peak height (O_a) and expected peak height (E_a) is given by [5]:

$$\log_{10}\left(\frac{O_a}{E_a}\right) \sim N\left(0, \frac{c^2}{E_a}\right) \text{for allelic peak, } a, \quad (1)$$

and

$$\log_{10}\left(\frac{O_{a-1}}{E_{a-1}}\right) \sim N\left(0, \frac{k^2}{O_a}\right) \text{for stutter peak } a\text{-}1 \quad (2)$$

Where a is an allelic peak and a-1 a stutter peak. The random variables c^2 and k^2 describe the variability of a profile, the prior distributions of which are determined empirically by modelling the variability over a range of single source profiles using component-wise Markov chain Monte Carlo (MCMC). This process is described in Taylor et al. [9]. In addition to the allele and stutter variance prior parameters, STRmix™ also has a prior distribution for the amplification efficiency at a locus (LSAE). The locus specific amplification efficiency recognises that not all loci amplify equally well and allows the expected peak heights to be above or below those modelled by per contributor template and degradation alone.

It is recommended by the STRmix™ developers that these variables are calibrated as part of the software implementation. In addition, stutter ratios, the drop-in rate, and the CE instrument analytical threshold and saturation limit are also variables within the STRmix™ model that are recommended to be determined empirically prior to use of the software [10]. Stutter ratios have previously been shown to not differ significantly between different laboratories using the same kit and cycle number [11]. Camera saturation is likely CE model dependent and not influenced by the technology or laboratory. Drop-in is likely to be PCR cycle dependent but there may also be a laboratory component in regards to prevention. PCR cycle number also affects stutter ratios (increasing stutter ratios as cycle number increases) and profile variance (heterozygote imbalance increases with elevated cycle number) [9,12].

Traditionally, a 'bottom up' approach to setting these variables is taken by modelling the observed variability in a range of single source profiles (with respect to peak height variances and stutter ratios) and negative control data (with respect to drop-in rates). The appropriateness of these values is then tested by interpreting a range of DNA profiles of varying quality and quantity (template) as part of a laboratory's internal validation of the software.

Empirical observations suggest however that the parameters are not too different between laboratories with the same kit chemistry, PCR cycle number, and CE machine (collectively termed technology in this paper) and the resulting likelihood ratios do not depend much on these parameters. This suggests one may 'borrow' parameters from a laboratory using the same technology to calculate a likelihood ratio (*LR*) without having to establish laboratory specific parameters; a 'top down' approach to setting parameters. The current study aims to rigorously address the question to what extent parameters can be borrowed from laboratories with similar technologies without rendering the results unreliable. This study demonstrates that STRmix™ parameters used for profile interpretation are portable between different laboratories. As with many organisations, we advocate internal laboratory validation prior to use in casework for any interpretation method, irrespective of how the parameters were determined.[1]

[1] We acknowledge standard 8.3.1.1 of the Quality Assurance Standards for Forensic DNA Testing Laboratories which states that "Internal validation data may be shared by all locations in a multi-laboratory system."

In the first part of this paper, STRmix™ parameters are determined using a publicly available dataset. In the second part of this paper, sets of data created in different laboratories were interpreted using their validated kit settings and then re-interpreted using the kit validated for the publically available data set. The resulting *LR*s were compared.

In part one, the publicly available Boston University PROVEDIt Initiative (Project Research Openness for Validation with Experimental Data) dataset was used. PROVEDIt is a freely available dataset which contains 25,000 DNA profiles generated by Boston University using different multiplex kits [13]. The intent of the PROVEDIt initiative is to provide an empirical data set to facilitate laboratories to "efficiently compare, contrast, and validate forensic DNA computational systems". A selection of single source GlobalFiler™ (Applied Biosystems), 29 cycle, 3500 CE samples was used from this data set to calibrate the PG software. Subsequently a series of mixtures of two and three contributors was used for validation purposes. STRmix™ [6] was calibrated to the PROVEDIt dataset according to the standard 'bottom up' implementation protocol. In order to determine the suitability of the STRmix™ parameters to interpret different mixture profiles, sensitivity and specificity studies were undertaken. With respect to interpretation methods, sensitivity is defined as the ability of the software to reliably resolve the DNA profile of known contributors within a mixed DNA profile for a range of starting DNA templates. The *LR* for known contributors (H_p true tests) with adequate template should be high and should trend to 1 as less information is present within the profile, i.e. sensitivity is dependent on the DNA input amount.

Specificity is defined as the ability of the software to reliably exclude known non-contributors (H_d true tests) within a mixed DNA profile for a range of starting DNA templates. The *LR* should trend upwards to 1 as less information is present within the profile [14].

In part two, the authors have considered whether, once calibrated to a given technology, such settings could be used more broadly by other laboratories/organisations employing the same technology. This involves applying the calibrated settings for Boston University to mixtures developed in another laboratory or organisation (which we term non-cognate experiments). For completeness, a selection of the publicly-available mixtures created by the Boston University were interpreted using settings calibrated within other laboratories/organisations around the world. This work is used to demonstrate the robustness of STRmix™ to variation in the input run parameters

2. Methods

2.1. Setup of the PROVEDIt parameters

2.1.1. Analytical threshold

The purpose of an analytical threshold (AT) is to reliably assign peaks above the threshold as signal (either truly allelic or stutter) and discard measurements below the threshold as potentially noise [15]. There is a trade-off, since setting this threshold too high can result in a loss of profile data, while setting the threshold lower increases the risk of interpreting noise spuriously as allelic. An analytical threshold is usually based on a signal to noise analysis determined from a laboratory's empirical data and should be generated including positive control data [16–18].

20 profiles were selected from the PROVEDIt 1-Person Profiles 3500 15 s GF 29 cycles, RD14-0003(01216ADG_15sec) data (a full list of samples used can be found in Supplementary material). The profiles selected had been amplified with 0 ng (negative controls) to 0.5 ng of template using GlobalFiler™ and 29 PCR cycles. Raw data files (.hid files) were analysed in GeneMapper® ID-X v1.5 with a 1 rfu AT for all dyes. Allelic peaks, stutter peaks, pull-up, and other artefactual peaks were removed from the profiles. The remaining peaks were attributed to noise and were exported to Excel for analysis.

These data were collated per dye colour (for all DNA quantities) and the average signal was determined. The signal per locus was also