Research paper

A fully continuous system of DNA profile evidence evaluation that can utilise STR profile data produced under different conditions within a single analysis

Duncan Taylor\textsuperscript{a,b,*}, Jo-Anne Bright\textsuperscript{c}, Hannah Kelly\textsuperscript{c}, Meng-Han Lin\textsuperscript{c}, John Buckleton\textsuperscript{c,d}

\textsuperscript{a} Forensic Science SA, 21 Divett Place, Adelaide, SA 5000, Australia
\textsuperscript{b} School of Biological Sciences, Flinders University, GPO Box 2100, Adelaide, SA 5001, Australia
\textsuperscript{c} Institute of Environmental Science and Research Limited, Private Bag 92021, Auckland, 1142, New Zealand
\textsuperscript{d} University of Washington, Department of Biostatistics, Seattle, WA 98195, United States

\textbf{ARTICLE INFO}

\textbf{Keywords:}
Multiple DNA profiling kits
Continuous DNA interpretation
STRmix
Modelling
DNA mixtures

\textbf{ABSTRACT}

The introduction of probabilistic DNA interpretation systems has made it possible to evaluate many profiles that previously (under a manual interpretation system) were not. These probabilistic systems have been around for a number of years and it is becoming more common that their use within a laboratory has spanned at least one technology change. This may be a change in laboratory hardware, the DNA profiling kit used, or the manner in which the profile is generated. Up until this point, when replicates DNA profiles are generated, that span a technological change, the ability to utilise all the information in all replicates has been limited or non-existent. In this work we explain and derive the models required to evaluate (what we term) multi-kit analysis problems. We demonstrate the use of the multi-kit feature on a number of scenarios where such an analysis would be desired within a laboratory. Allowing the combination of profiling data that spans a technological change will further increase the amount of DNA profile information produced in a laboratory that can be evaluated.

1. Introduction

DNA profile interpretation has substantially changed towards probabilistic genotyping. Testament to this are the number of probabilistic DNA profile interpretation systems now available [1–6], the recommendations that have been released by multiple advisory bodies on their validation [7,8], and the vast array of published works on various models that can be used to describe different elements of DNA profile generation or behaviour [9–26].

Most of the available systems have features that carry out two standard tasks:

1) Deconvolution of a DNA profile into sets of genotypes that could describe the profile, weighted by some system of probabilistic modelling (whether this be in a semi-continuous manner that uses probabilities of drop-in and drop-out or a fully continuous system that applies probabilities for various mass parameters within the model).

2) The calculation of a likelihood ratio (LR) when comparing one or more references to the deconvolution and considering two competing propositions.

These two tasks fulfil the vast majority of functions required by most forensic laboratories. However, as our knowledge and comfort with these systems increases, extensions to the above elements are starting to be explored. These have included extensions to the calculation of the LR to database searching [27], familial searching [28], the calculation of an LR comparing two mixed profiles where the propositions are whether one or more common contributors exist [29], or seeking information on potential familial relationships between donors of a mixture [30]. The deconvolution has also been extended to carry out deconvolutions looking for one or more common contributors to multiple evidence items [31].

The use of multiple PCR replicates within a single interpretation has been undertaken for as long as probabilistic genotyping has been considered [32]. For semi-continuous systems this involves the accumulation of probabilities of drop-in and drop-out for each peak in each profile in the analysis. For fully continuous systems the use of replicates involves the comparison of multiple sets of observed peak heights to their expectations, built up from the models in the system.

We present here an extension to this idea of using replicates, by considering replicates that have been produced under different conditions. This would be of particular use under two broad conditions:
1) The forensic laboratory uses complementary DNA profiling kits that have the same loci, but ordered differently with respect to molecular weight i.e. as per the ESI and ESX profiling system.

2) Data has been produced in one system that has then been super-seded by another before the replicate was produced. This may be due to a change in technology such as the move of a laboratory to a different profiling kit, or an update in the model of electrophoretic instrumentation, or the decision to trial an enhancement technique, such as increased PCR cycles, when standard profiling has produced insufficient data.

We term these replicates a ‘multi-kit’ analysis noting that it may be the same DNA profiling kit in use, just being used under two conditions (for example, a replicate PCR amplification after having moved from a 3130 capillary electrophoresis instrument to a 3500). The difference in the consideration of a multi-kit analysis to a standard replicate analysis is that the profiles have been produced under differing conditions, meaning that they are behaving differently from one another. While the same general models should apply (e.g. for stutter, peak height variability or drop-in, etc.) the parameter values (or their prior probabilities) within the models will differ between replicates. Multi-kit analyses pose additional challenges to standard continuous DNA interpretation systems as different parameters within the models may be correlated across replicates while others will not. An example of this would be to consider a mixed and degraded DNA sample that is profiled in two different kits. We would expect the relative DNA amounts of each contributor to be the same between replicates (they in fact must be as this is a property of the DNA extract, which precedes DNA profiling) where the amplification efficiencies of the loci would not be (as this is a kit related parameter).

We outline here an extension to the theory presented in [1], that considers multi-kit analyses. We then demonstrate the performance of the proposed model on a number of scenarios that laboratories could encounter, where multi-kit analysis capabilities would prove beneficial.

2. Theory

Starting with the standard theory of [1,32] we seek the probability of the observed evidence given a proposition. Nuisance parameters must be considered in order to enumerate the probability and we term these the mass parameters (\( \mathbf{M} \)). We do not specifically care what the values of these parameters are to calculate the probability (hence ‘nuisance’), although it may be of interest to know their values for either diagnostic purposes or for comments on higher level evidential considerations that rely on extrinsic properties of the DNA profile such as the biological source of the DNA or the activities that could have led to the presence of the DNA. We also do not know the values of these mass parameters and so integrate across their prior probability distributions, which can be achieved through the use of Markov chain Monte Carlo (MCMC) analysis. The mass parameters considered in [1] are template DNA (t), degradation (d), locus amplification efficiency (A) and replicate amplification efficiency (R). To extend the theory to multiple kits we introduce a fifth mass parameter, a kit multiplier (B). We now have:

1. A variable \( t_n \) for each of the \( n \) contributors that may usefully be thought of as template amount,
2. A variable \( d_{a,k} \), which models the decay with respect to molecular weight (m) in the template for each of the contributors to genotype set, \( S_j \), where \( j \) is one of the \( J \) possible genotype sets the contributors can take. This may usefully be thought of as a measure of degradation. Note that this term is now dependent on kit, \( k \), although this dependence on kit is optional and can be fixed within the analysis so that \( d_{a,k} = d_{a,k'} \) under certain circumstances (explained later),
3. Locus efficiencies at each locus, \( A^l_i \) to allow for the observed amplification levels of each locus. Note that this term is now dependent on kit,
4. Replicate multipliers \( R_{y,k} \). This effectively scales all peaks up or down between replicates. Note that this term is now dependent on kit, and
5. Kit multipliers, \( B_k \). This scales all peaks in all replicates up or down between kits.

We write the mass variables, \( D = \{d_{a,k}: n = 1,..., N; k = 1,..., K\} \), \( T = \{t_n: n = 1,..., N\} \), \( A = \{A^l_i: l = 1,..., L; k = 1,..., K\} \), \( R = \{R_k: k = 1,..., K\} \) as (defining \( Y_k \) as the number of PCR replicates used for kit \( k \)) and \( B = \{B_k: k = 1,..., K\} \) as \( \mathbf{B} \). The variables \( D, A, R, T \) and \( B \) are written collectively as \( \mathbf{M} \).

Note that the degradation term is optionally kit dependant, despite it being a property of the DNA extract and not the manner in which the profile has been generated. We make the choice to introduce an optional kit dependence on degradation as a result of the type of samples for which this analysis will commonly be used. That is, replicates that span technology (e.g. profiling kit) are likely to be separated by some time, during which degradation will have been acting on the sample. Therefore, the degradation term should be free between kits. For samples where this is not the case, the degradation values for contributors should be similar between kits and hence the term would be fixed in those situations (this property could be used as an analysis diagnostic).

Knowing the values of the parameters in \( \mathbf{M} \) allows the calculation of Total Allelic Product (\( T \)), the total amount of fluorescence expected resulting from an allele in a DNA extract, which will ultimately get broken into components of fluorescence in an allelic position and its stutter positions on the electropherogram (EPG). Calculation of \( T \), for a combination of contributor, kit, locus, replicate and allele, is achieved formulaically by:

\[
T^l_i = t_n \times A^l_i \times R_{y,k} \times B_k \times X_l^i \times e^{-d_{a,k}\left(m^l_i - m_{offset}\right)}
\]  
(1)

The \( X_l^i \) term in Eq. (1) represents a ‘dose’ and takes values of 1 or 2. The dose considers that if contributor \( n \) is homozygous for allele \( a \) at locus \( l \) (\( X_l^i = 2 \)), then the expected value for \( T \) will be twice as high than if allele \( a \) was one in a heterozygous pair. The offset marks the molecular weight at which degradation starts to be applied, i.e. at the offset (and technically before it), degradation is not acting to reduce fluorescence. This offset is usually set to be the lowest molecular weight peak observed in one or more electropherograms (or some value below it), and so only depends on the profiling kit being used. As the PCR occurs, some of the fluorescence that was destined for the allele will shift to stutter positions on the EPG. There are a number of stutter types that can occur (back stutter, forward stutter, half stutter, double stutter, etc.) and we will define the number of types of stutter as \( I \), the stutter ratio of stutter type \( i \), for locus \( l \) in kit \( k \) for allele \( a \) as \( n^l_{i,k} \) and the position of stutter type \( i \) relative to the parent peak, \( a_i \). We can now split the total allelic product into components of, respectively, allele and stutter by:

\[
E^l_i = \frac{T^l_i}{1 + \sum_l T^l_i}
\]  
(2)

and

\[
E^l_{i+a,k} = \frac{T^l_i \times n^l_{i,k}}{1 + \sum_l T^l_i \times n^l_{i,k}}
\]  
(3)

where \( i, \) indicates ‘not \( i \)’ and hence the allelic component. The total expected height of a peak at a locus, replicate and kit combination is then the sum of the stutter and allelic components of all individuals that fall on that allelic position:

150
دریافت فوری متن کامل مقاله

امکان دانلود نسخه تمام متن مقالات انگلیسی
امکان دانلود نسخه ترجمه شده مقالات
پذیرش سفارش ترجمه تخصصی
امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
امکان دانلود رایگان ۲ صفحه اول هر مقاله
امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
دانلود فوری مقاله پس از پرداخت آنلاین
پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات