International Congress Series 1288 (2006) 468–470

www.ics-elsevier.com

A graphical (Bayes net) molecular model of the entire DNA STR process to aid interpretation

Peter Gill [a,*], James Curran [b], Keith Elliot [a]

[a] *Forensic Science Service, Birmingham, UK*
[b] *Department of Statistics, University of Waikato, New Zealand*

Abstract. The use of expert systems to interpret short tandem repeat (STR) DNA profiles in forensic, medical and ancient DNA applications is becoming increasingly prevalent as high-throughput analytical systems generate large amounts of data that are time-consuming to process. With special reference to low copy number (LCN) applications we use a graphical model to simulate stochastic variation associated with the entire DNA process starting with extraction of sample, followed by the processing associated with preparation of a PCR reaction mix, and PCR itself. Each part of the process is modelled with input efficiency parameters. Then, the key output parameters that define the characteristics of a DNA profile are derived—namely heterozygous balance (Hb) and the probability of allelic dropout $p(D)$. © 2005 Published by Elsevier B.V.

Keywords: Simulation; Graphical model; Bayes net; Allele dropout; Heterozygous balance

1. Introduction

In forensic, ancient DNA and some medical diagnostic applications there may be limited, highly degraded DNA available (<100 pg) for analysis. To maximise the chance of a result sufficient PCR cycles must be used to ensure that a single template molecule will be visualised. The universally accepted preferred method to analyse crime samples is with short tandem repeat (STR) DNA. However, there are two main problems that result from stochastic events: one or more alleles of a heterozygous individual may be completely absent—this is known as allele dropout [1]; in addition, PCR generated slippage mutations or stutters [2] may be generated. Both events may compromise interpretation. In relation to analysis of ancient DNA such as museum specimens, or faeces

* Corresponding author. Trident Court, 2960 Solihull Parkway, Solihull, West Midlands B62 0LS, UK. Tel.: +44 121 329 5412; fax: +44 121 502 6021.
E-mail address: dnapgill@compuserve.com (P. Gill).

0531-5131/
doi:10.1016/j.ics.2005.08.040

of free-ranging animals, Taberlet et al. [3] used a computer simulation to address the question to determine the number of typing experiments needed to obtain a reliable result. These principles were adopted by Gill et al. [1], in conjunction with an extended statistical theory, to address similar issues that related to analysis of forensic samples. In forensic applications there is also the added complication that the sample itself may be a mixture from two or more individuals. A number of statistical methods have been devised to aid interpretation. The first methods were based on the binary absence/presence of alleles. Later methods subsequently incorporated electropherogram peak height and area into programmed expert systems [4,5]. In parallel to improvements of interpretation, the sensitivity of analysis has also improved to the extent that products of a single cell can be visualised, either by increased PCR cycle number [6], or by using novel individual cell-selection methods such as laser micro-dissection (LMD) [7].

To improve understanding of the dependencies of parameters associated with DNA analysis we provide a formal (graphical) statistical model along with a computer implementation (PCRSIM) that simulates the entire process starting from: extraction→aliquot into pre-PCR reaction mix→PCR amplification for t cycles→visualisation of alleles after electrophoresis. We use Monte Carlo simulation techniques to model the expected variation in PCR stutter artefacts, heterozygote balance, and to predict dropout rates. Wherever possible, we also provide a formal statistical model. We demonstrate that graphical models can be used to assess and measure unknown variables such as sample extraction efficiency, or to optimise parameters such as the amount of pre-PCR aliquot taken. By modelling 'what-if' scenarios, we can therefore improve entire DNA processes as a result, and this translates into improved success rates when real samples are analysed. The full model is described by Gill et al. [8].

2. Results and discussion

Heterozygote balance (Fig. 1) is a particularly important parameter in mixture interpretation. For a heterozygote locus with alleles A and B, for each allele we simulate 1000 times the number of post-

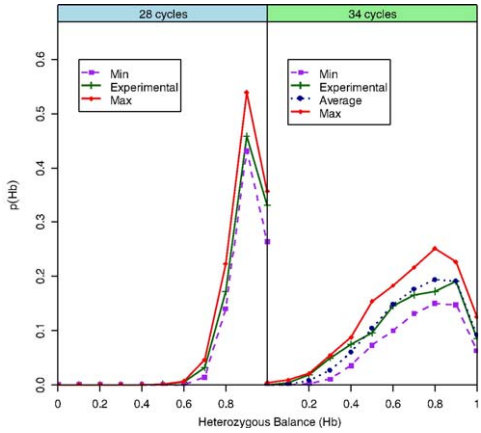

Fig. 1. Simulations of Hb (1000×) of 500 pg (83 diploid cells), 28 PCR cycles and 25 pg (4 diploid cells), 34 PCR cycles, compared to experimental observations.

PCR molecules $n_A(t)$ and $n_B(t)$. Given the two parameters $\pi_{aliquot}$ and π_{PCReff} we obtain 1000 estimates of $Hb = \min(n_A(t), n_B(t))/\max(n_A(t), n_B(t))$. Simulation results were compared to experimental data from 1692 heterozygotes from 500 pg and 25 pg of genomic DNA (Fig. 1). The results also provide a strong theoretical basis for the widely used guideline for the acceptable range of Hb, $Hb \geq 0.6$ [9,10], which is used to assist interpretation of mixtures when optimal amounts of DNA are analysed. Note that allele dropout is simply an extreme form of heterozygous balance.

Finally, we can also use the PCRSIM model to generate random DNA profiles from allelic frequency databases [11]. Given the parameters that describe quantity and PCR efficiency, it is possible to simulate entire SGM plus profiles comprising 11 loci. At low quantities of DNA, stochastic effects result in partial DNA profiles. Consequently, each time a different PCR is carried out, each will give a different result. Either dropout occurs, or samples are very unbalanced within and between loci. Some researchers have attempted to improve systems by using alternative amplification methods. In particular, there is much interest in Whole Genome Amplification [12]. However, we have demonstrated that the reasons for imbalance are predominantly stochastic, and not related to biochemistry.

References

[1] P. Gill, et al., An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA, Forensic Sci. Int. 112 (2000) 17–40.
[2] P.S. Walsh, N.J. Fildes, R. Reynolds, Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA, Nucleic Acids Res. 24 (1996) 2807–2812.
[3] P. Taberlet, et al., Reliable genotyping of samples with very low DNA quantities using PCR, Nucleic Acids Res. 24 (1996) 3189–3194.
[4] M. Bill, et al., PENDULUM—a guideline based approach to the interpretation of STR mixtures, Forensic Sci. Int. 148 (2004) 181–189.
[5] J. Curran, P. Gill, M.R. Bill, Interpretation of repeat measurement DNA evidence allowing for multiple contributors and population substructure, Forensic Sci. Int. 148 (2004) 47–53.
[6] I. Findlay, et al., DNA fingerprinting from single cells, Nature 389 (1997) 555–556.
[7] K. Elliott, et al., Use of laser microdissection greatly improves the recovery of DNA from sperm on microscope slides, Forensic Sci. Int. 137 (2003) 28–36.
[8] P. Gill, J. Curran, K. Elliot, A graphical simulation model of the entire DNA process associated with the analysis of short tandem repeat loci, Nucleic Acids Res. 33 (2005) 632–643.
[9] P. Gill, R. Sparkes, C. Kimpton, Development of guidelines to designate alleles using an STR multiplex system, Forensic Sci. Int. 89 (1997) 185–197.
[10] J.P. Whitaker, E.A. Cotton, P. Gill, A comparison of the characteristics of profiles produced with the AMPFlSTR SGM Plus multiplex system for both standard and low copy number (LCN) STR DNA analysis, Forensic Sci. Int. 123 (2001) 215–223.
[11] P. Gill, et al., A comparison of adjustment methods to test the robustness of an STR DNA database comprised of 24 European populations, Forensic Sci. Int. 131 (2003) 184–196.
[12] P.M. Schneider, et al., Whole genome amplification—the solution for a common problem in forensic casework? in: C. Dutremepuich, N. Morling (Eds.), Progress in Forensic Genetics—International Congress Series, vol. 1261, Elsevier, 2004, pp. 24–26.