Simulation of the PCR-CE process

Khang Ee Pang

University College Dublin

September 30, 2019

1 Introduction

The polymerase chain reaction capillary electrophoresis or PCR-CE process is the basis of profiling short tandem repeat (STR), which in turns used to identify contributor(s) to unknown biological specimens [1]. Challenges of the inferring of the contributor(s) remains, especially within forensic applications where samples might be degraded or came from multiple sources.

In this work, we compare (a simplified version of) the proposed model of the PCR-CE process from [2] with observation data, focusing on the stutter productions, in hoping to uncover the nature of said process. We achieve this by looking only at single parent alleles (homozygous observations) so that stutter interactions between different parent alleles could be ignored.

2 Methods

2.1 Model

We re-state the model from [2]. Let $E(N_a)$ be the copying efficiency of PCR, π_1 be the probability of stuttering, and π_2 be the probability of the stutter product further stuttering. Then for each allele a, the number of true amplicon start with $T_0 = T_2 \sim \mathrm{Binomial}(T, p_{\mathsf{sample}})$ where $T \sim \mathrm{Normal}(\eta\mu_{\mathsf{DNA}}, (\eta\sigma_{\mathsf{DNA}})^2)$, $\eta = 6.3/(10^3 V_{\mathsf{tot}})$ and number of stutter amplicons is $S_0 = S_2 = 0$. For $n \geq 3$, the recursion relation is given

by

$$T_{n+1} = T_n + \sum_{i=1}^{T_n} C_i^{\beta_n}(1 - D_i^{\beta_n}), \tag{1}$$

$$S_{n+1} = S_n + \sum_{i=1}^{S_n} C_i^{\alpha_n}(1 - D_i^{\alpha_n}) + \sum_{i=1}^{T_n} C_i^{\beta_n} D_i^{\beta_n}, \tag{2}$$

where $C^\beta \sim \mathrm{Ber}(E(T_n))$, $C^\alpha \sim \mathrm{Ber}(E(S_n))$, $D^\beta \sim \mathrm{Ber}(\pi_1)$, and $D^\alpha \sim \mathrm{Ber}(\pi_2)$.

The observed fluorescence at allele a is given by

$$F = (T^a_{N_{PCR}} + S^{a+1}_{N_{PCR}})\alpha + bN \tag{3}$$

where α is the CE sensitivity, $b \sim \mathrm{Ber}(p_{\mathsf{noise}})$ and $N \sim \text{Log-normal}(\mu_{\mathsf{noise}}, \sigma^2_{\mathsf{noise}})$.

The goal is to infer T from F.

2.2 Expectation

The expectation value for the stutter product from one parent allele (homozygous) of eq(1) and eq(2), for $n \geq 3$, is given by

$$\mathcal{T}_{n+1} = \mathcal{T}_n(1 + E(\mathcal{T}_n)(1 - \pi_1)), \tag{4}$$

$$\mathcal{S}_{n+1} = \mathcal{S}_n(1 + E(\mathcal{S}_n)(1 - \pi_2)) + \mathcal{T}_n E(\mathcal{T}_n)\pi_1. \tag{5}$$

If we assume that $E(N_a) = E$ and $\pi_1 = \pi_2 = \pi$, one can formulate the following:

Lemma 2.1 *Let $\mathcal{T}_i^n$ be the expected number of amplicons of allele $a-i$ (i.e. distance i form the parent allele a) after the nth PCR cycle, eq(4) and eq(5) become*

$$\mathcal{T}_0^{n+1} = p\mathcal{T}_0^n, \tag{6}$$

$$\mathcal{T}_i^{n+1} = p\mathcal{T}_i^n + q\mathcal{T}_{i-1}^n + \mathrm{O}(\pi^2), \qquad \textit{for } i = 1\ldots, n-2 \tag{7}$$

where $p = 1 + E(1 - \pi)$ and $q = E\pi$. $\mathrm{O}(\pi^2)$ includes the contribution from forward stutter and multiple stutter within one PCR cycle.

By induction,

$$\mathcal{T}_i^n = T_0^0 \binom{k}{i} p^{k-i} q^i, \qquad \textit{for } i = 1, \ldots, k \tag{8}$$

where $k = n - 2$.

Corollary 2.2 *Fix $n = N_{PCR}$. As $q \ll p$, eq(8) has the form*

$$\mathcal{T}_i = Ce^{-\lambda i} \tag{9}$$

where the decay factor λ, independent of T_0^0, is approximately constant.

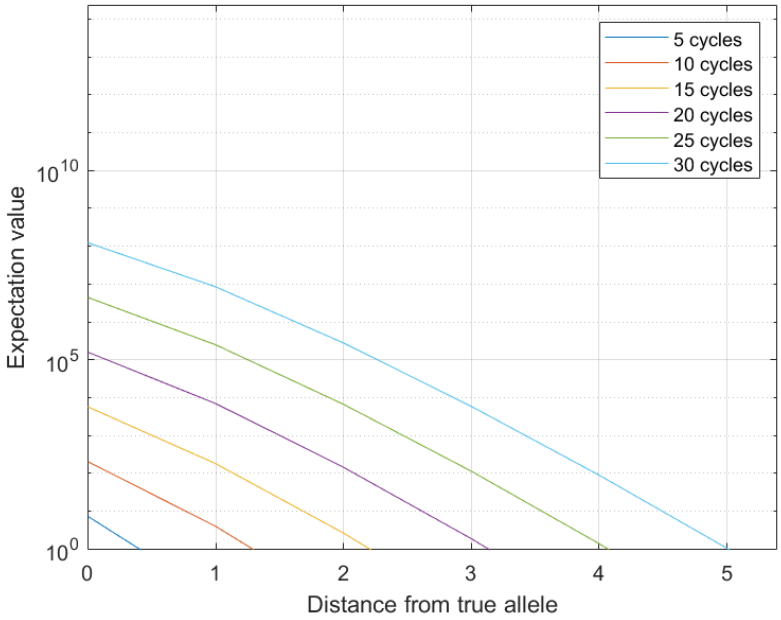

Figure 1: Expected amplicon and stutter count for $T_0^0 = 1$, $E = 0.95$, $\pi = 0.005$ for different N_{PCR}. The decay factor λ is approximately 3.68 for $N_{\text{PCR}} = 30$.

Corollary 2.3 *If $p_{sample} = 1$, then the distribution of $\mathcal{T}_i^n, i = 1, \ldots, k$ can be shown to be*

$$\mathcal{T}_i^n \sim \text{Normal}(\xi_i^n \mu_{DNA}, (\xi_i^n \sigma_{DNA})^2), \qquad \xi_i^n = \eta \binom{k}{i} p^{k-i} q^i. \tag{10}$$

3 Result

We use $N_{\text{PCR}} = 29$, $p_{\text{noise}} = 0.1$, $\mu_{\text{noise}} = 1$, $\sigma_{\text{noise}} = \sqrt{0.5}$, $E(N_a) = 0.95$, $\pi_1 = \pi_2 = 0.005$, and $\alpha = 5 \times 10^{-7}$ for the simulations as specified by the assignment. We further assume $p_{\text{sample}} = 1$, $V_{\text{tot}} = 50\mu L$, $\mu_{\text{DNA}} = 0.00078ng/\mu L$, and $\sigma_{\text{DNA}} = 0$ as taken from [2]. Simulations of the model within one loci is shown in Figure 2.

4 Discussion

4.1 Inference

Let $n = N_{\text{PCR}}$ and let $T_i := F_i/\alpha, i = 1, \ldots, N$ be the observed amplicon count of stutter products with parent allele a, then it is possible to estimate $m := \ln \widehat{T}$

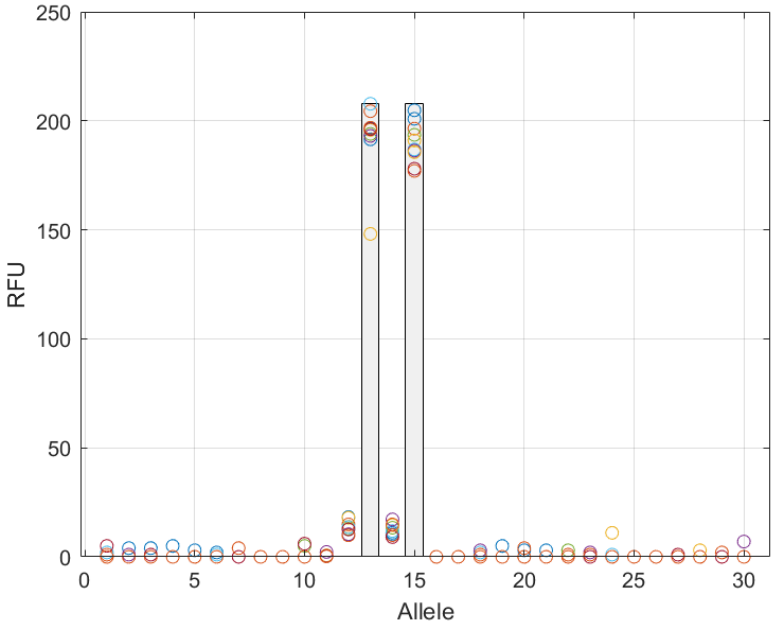

Figure 2: 10 simulations (marked in circles) of the PCR-CE process of D8S1179 loci of a 13,15 heterozygous contributor.

for the stutter product from single allele interactions (homozygous observations) by minimizing the sum of squared residuals

$$S(m) = \sum_{i=1}^{N} (\ln \mathcal{T}_i - \ln T_i + m)^2. \tag{11}$$

$S(m)$ attain the global minimum when

$$m = \frac{1}{N} \sum_{i=1}^{N} \ln \frac{T_i}{\mathcal{T}_i}, \qquad \widehat{T} = \mathrm{e}^m. \tag{12}$$

Finally, the decay factor λ of the observation is given by the slope of $\{\ln T_i\}_{i=1,\ldots,N}$.

4.2 Validation

From Figure 3, we see that the decay factor λ is far from being a constant, which Corollary 2.2 predicts. The tapering of λ for increasing sample count N suggests a double exponential decay relation instead of a (single) exponential decay relation in eq(9). The physical relevance of such relation could come from the reduced PCR efficiency E for large amplicon number as detailed in [3], however the precise reason is unsure.

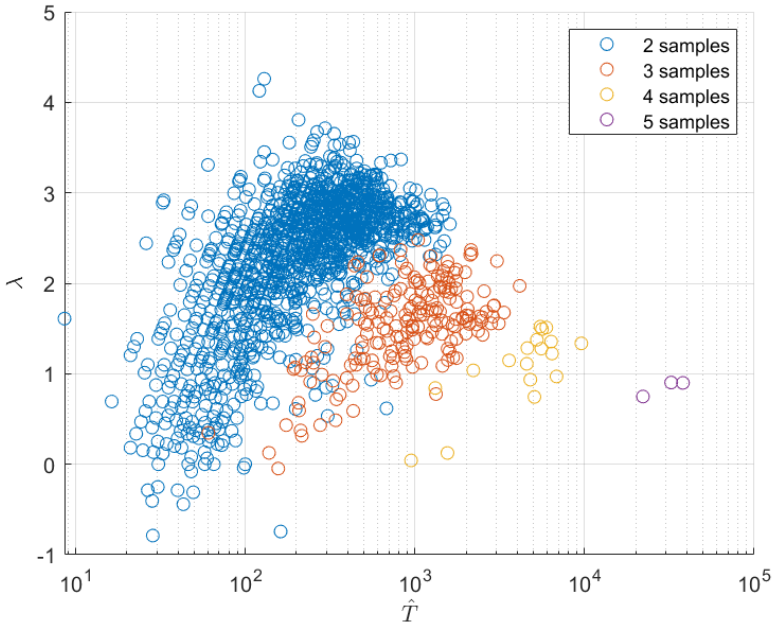

Figure 3: Predicted initial amplicon count $\widehat{T}$, and decay factor λ of 1646 homozygous observations across 21 core STR loci, with different number of samples N.

4.3 Noise

Consistent with the finding of [2], we observed that the small (and more varied) decay factors at the low-template region ($\widehat{T} \lesssim 10^2$) are likely due to the noise interfering with the underlying stutter count.

The noise is model as log-normal-ly distributed with a probability of occurrence of p_{noise} at each allelic position, superimposed to the underlying signal. This produced similar looking signal noise as ones that were observed. However, as mentioned in [4], the model only capture CE induced noise and fails to describe noise induced by other source such as the PCR process, where one would expect the noise to increase as N_{PCR} increases.

5 Conclusion

Given the simplicity of the model, even with the assumption that $E(N_a)$ is constant, the model is able to reproduced many of the feature observed in real EPG including the decay in stutter count as the distance between the stutter allele and the parent allele increases. This is not surprising as the model has a theoretical basis that is based on the mechanism of the PCR process.

However, we observed a deviation of the decay relation between the (simplified) model and the observational data which behaves closer to a double exponential function. Investigating the effect of the inclusion of an amplicon number dependent PCR efficiency on the decay relation would be of interest in understanding the stutter production.

References

[1] L.E. Alfonse, A.D. Garrett, S.L. Desmond, K.R. Duffy. A large-scale dataset of single and mixed-source short tandem repeat profiles to inform human identification strategies: PROVEIt. *Forensic Science International: Genetics*, 32:62-70, 2018.

[2] K.R. Duffy, N. Gurram, K.C. Peters, G. Wellner, C.M. Grgicak. Exploring STR signal in the single- and multicopy number regimes: Deductions from an in silico model of the entire DNA laboratory process. *Electrophoresis*, 38:855-868, 2017.

[3] J.M. Butler. Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers. *Elsevier*, 2005.

[4] N. Gurram. A mathematical model of polymerase chain reaction induced stutter. *Massachusetts Institute of Technology*, 2015.