USING IMPORTANCE SAMPLING TO IMPROVE ACCURACY AND REPEATABILITY OF CEESIt

By

SELINA HUI

A thesis submitted to the

Graduate School – Camden

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of Master of Science

Graduate Program in Computer Science

Written under the direction of

Dr. Desmond S. Lun

And Approved By

Dr. Desmond S. Lun

Dr. Jean-Camille Birget

Dr. Catherine M. Grgicak

Camden, New Jersey

October 2019

THESIS ABSTRACT

Using importance sampling to improve accuracy and repeatability of CEESIt

by SELINA HUI

Thesis Director:

Dr. Desmond S. Lun

CEESIt is a computational method for the analysis of short tandem repeats (STRs) in DNA for human identification. CEESIt computes the likelihood ratio (LR), the ratio of the probability of the evidence (the electropherogram obtained from the DNA sample) given a specific person of interest (POI) to the probability of the evidence given a random contributor from the background population. The DNA sample may be a mixture, comprised of multiple contributors at different ratios. With cases using low amounts of template DNA or cases with multiple contributors in the mixture, the results lacked consistency between computations. With 1-person mixtures, the tests ran with high repeatability and short runtime but with 2-people mixtures, the results had varying results and significantly longer runtime. The goal was to find the source of the discrepancies to improve repeatability and accuracy. CEESIt uses the Monte Carlo Method to generate the final probabilistic values. To improve repeatability and accuracy, importance sampling of the genotypes of the background population was implemented. By careful sampling and appropriate weighting to represent the background population, this improved the overall accuracy in the algorithm and allowed the algorithm to sample a smaller population, which decreases runtime.

CHAPTER 1

INTRODUCTION

The understanding of genetics has greatly improved over the years. With this increased understanding, more people are using genetics as evidence to prove their argument. Genetics is considered reliable and therefore an accepted method that identifies an individual. This is useful in forensics, especially in identifying a certain individual(s) from a population. There are many ways in identifying an individual such as qualitative properties like appearance but looks can easily be deceiving and altered. Genetics, however, are not easily altered. Due to the reliability of genetics, DNA comparison is used in forensic science to identify a person-of-interest (POI) in a DNA mixture.

1.1 DNA

Deoxyribonucleic acid (DNA) is a polymer in human cells whose sequence of nucleotide bases can be used to identify individuals. A gene is a genetic element at a specific location, or locus, in the DNA. Alleles are the variant forms of a gene and at each locus, humans can either have two identical alleles or two different alleles. If there are two different alleles at a locus, the organism is known as heterozygous with respect to that locus; if there are two identical alleles at a locus, the organism is known as homozygous with respect to that locus (Fincham).

An electropherogram is a signal that results from a person's DNA using a technique known as capillary electrophoresis where the DNA fragments are stained by fluorescent dyes and become visible under UV light (Westermeier). In an electropherogram, the alleles at each locus get represented as spikes with certain heights, which increase with the amount of DNA. During the translation of the alleles, the height can get disrupted which affects the spike in the diagram. These disruptions are classified into four categories: dropin, dropout, noise, reverse stutter and forward stutter.

CEESIt focuses on the human population so the algorithm focuses on specific DNA markers called STRs. STRs, or Short Tandem Repeats, are DNA regions that contain repeating components ranging in size from 2 to 7 base pairs (Swaminathan). The alleles over a number of STRs form a profile that can be used to identify an individual.

1.2 CEESIt

CEESIt is an algorithm that uses the POI's DNA profile to find the likelihood of the POI being a contributor to a DNA mixture and the probability of finding a random person having a higher likelihood than the POI. This algorithm uses a continuous model to analyze the DNA information instead of a binary representation. Binary representation only considers if the allele is absent or present. A continuous model includes quantitative information like the peak heights in the electropherogram which therefore includes noise, stutter, and drop-ins (Swaminathan, Garg, et al.).

When the mixture only contains one contributor, the testing is intuitive by comparing the POI's DNA to the mixture. However, when there is more than one contributor, the mixture contains an unknown ratio of the contributors' DNA, which complicates the comparison(Westermeier). CEESIt has to consider different ratios of the major and minor contributor of the DNA mixture in order correctly calculate the likelihood ratio.

1.3 Likelihood Ratio

The Likelihood Ratio is defined as:

$$LR = \frac{\Pr\left(E\middle|H_P, n_p\right)}{\Pr(E|H_d, n_d)}$$

where E is the evidence in the form of a electropherogram, H_p is the hypothesis from the prosecution, H_d is the hypothesis from defense, n_p is the number of contributors from prosecution and n_d is the number of contributors from defense. The numerator is the probability of the evidence knowing that the POI's genotype contributed to the evidence. The denominator is the probability of evidence from random contributors who may or may not have the POI's DNA profile. When the likelihood ratio is greater than one, the likelihood is in favor of the prosecutor's side and when the likelihood ratio is less than one, the likelihood ratio is in favor of the defense's side (Swaminathan, Qureshi, et al.). For this study, n_p and n_d are the same value. Both defense and prosecution assume the same number of contributors, which is the most common scenario.

This likelihood ratio is then expanded to:

$$LR \approx \frac{\Pr\left(E|R = s, N = n\right)}{\Pr(E|R = s, N = n)\Pr(R = s) + \Pr\left(R \in R_1 \backslash \{s\}\right) \sum_{i=1}^{M} \Pr(E|R = r^i, N = n)\,/M}$$

(Swaminathan, Qureshi, et al.). The numerator is the probability of the evidence given the person-of-interest's genotype and the denominator is the probability of the evidence given that the POI is a person randomly drawn from the background population. The variables are as follows: E is the evidence from an electropherogram; R is a genotype of a contributor; N is a number of contributors; s is the suspect's or person-of-interest's genotype; M is the number of sample genotypes in the sample population; r^i is a sequence of M genotypes sampled randomly from the background population; and R_1 is a set of genotypes containing all genotypes r such that $\Pr(E_l|\,R_l = r_l)$ does not evaluate to 0 within the limits of computational precision for all loci l (E_l is the evidence at locus l, R_l is genotype at locus l, r_l is the genotype r at locus l).

1.4 Monte Carlo Method

The CEESIt algorithm uses randomness to populate the sample population. Algorithms based on random choices are known as probabilistic algorithms (Rosen). The randomness in the sample population affects the final likelihood ratio, which causes different results between runs. Because the final results from a probabilistic algorithm differs between runs, there may be a percentage of runs where the results are drastically different from the correct results. In order to decrease the chances of generating incorrect results, the algorithm requires sufficient amounts of computation in order to generate the correct results.

The CEESIt algorithm is dependent on the population size to find the probability of finding the POI's template DNA. Running the algorithm over the entire population's genotype requires long runtime. To decrease runtime, CEESIt uses the Monte Carlo Method to select random samples from the population to create the sample population. A larger sample population has a better representation of the population but uses more runtime. This is one of the tradeoffs of this algorithm: precision with accuracy verses low runtime.

1.5 Importance Sampling

The Monte Carlo method is an effective way of representing the population without sacrificing runtime but because Monte Carlo sampling is random, random sampling does not prevent the sample population from over sampling a smaller distribution. By repeatedly sampling a subset of the background population, this skews the sample population towards that distribution. When the population is sampled over a similar subset, the results are a repeatable over several tests but the value is skewed towards that subset and not the background population.

One way to decrease the likelihood of data skewing towards a subset of the background population is to increase sample population size. By increasing sample population size, the sample population is able to include a range of genotypes from different distributions. This however comes with the caveat of increased runtime and does not guarantee that the sample population does not heavily sample from one subset of the population.

Another way to prevent data from being skewed towards a specific subset of the population is importance sampling. In addition to randomly sampling from a smaller population, each sample is weighted by its 'importance' to the population. By weighing samples according to their 'importance' allows for samples that are undervalued or overvalued in one population distribution to be appropriately represented in the sample population. The weight is defined as

$$w_r \equiv \frac{P(x^{(r)})}{Q(x^{(r)})}$$

where the weight for sample r is reweighted from the original population distribution P to the sample population distribution Q (MacKay).

The sample population distribution was originally sampled upon the allele frequency. To decrease runtime and increase the accuracy of the sample population, the new distribution is based on the allele height. By sampling allele based on allele height instead of allele frequency, the strength of the allele is included in the calculation in addition to the frequency of the allele in a population. The tests with importance sampling samples the population based on allele height while the tests without importance sampling uses allele frequency to generate the sample population.

CHAPTER 2

FINDINGS

2.1 Testing

The algorithm was tested against known mixture samples with one or two contributors and template DNA between 0.0156ng to 0.25ng. Two variables were focused upon in this study: genotype tolerance and sample population size. The sample population size is the variable M referenced in the Likelihood Ratio. The genotype tolerance is the standard error tolerance for the probability of evidence at a given locus, quantification parameters, and number of contributors. The lower the genotype tolerance, the more samples are included in the sample population

For this series of testing, the known number of contributors was used. It is known from previous studies that the proper number of contributors highly impacts the final likelihood value(Benschop et al.). When the incorrect number of contributors is used, the probability of the evidence given that the POI is a contributor can be highly inaccurate. For example, when a 2-people DNA mixture gets mistaken for a 1-person mixture, electropherogram peaks resulting from the POI's alleles can get mistaken for stutter, thereby yielding an small likelihood ratio even though the POI's DNA is indeed in the mixture. The purpose of this study is to improve runtime and accuracy under ideal conditions so the known number of contributors was used in testing.

The final likelihood ratio is given in the logarithmic domain. The final likelihood ratio can either be extremely large or small depending on the strength of the likelihood ratio. Performing the calculations in the logarithmic domain allow very large of very small values to be represented. For this reason, the computation is performed in the logarithmic domain.

2.2 1-person Tests

1-person tests included mixtures with one contributor and tested against one known contributor. The runs without importance sampling used CEESIt default value: genotype tolerance of 0.5 with sample population size of one billion. The runs with importance sampling used a genotype tolerance of 0.5 with a sample population size of one million. These parameters with importance sampling were selected after running repeated trials under different parameter settings. Tests with large amounts of DNA showed little difference between importance sampling and without importance sampling. In many cases when there is a large amounts of DNA, sampling the population between allele height and allele frequency did not result in different likelihood ratio.

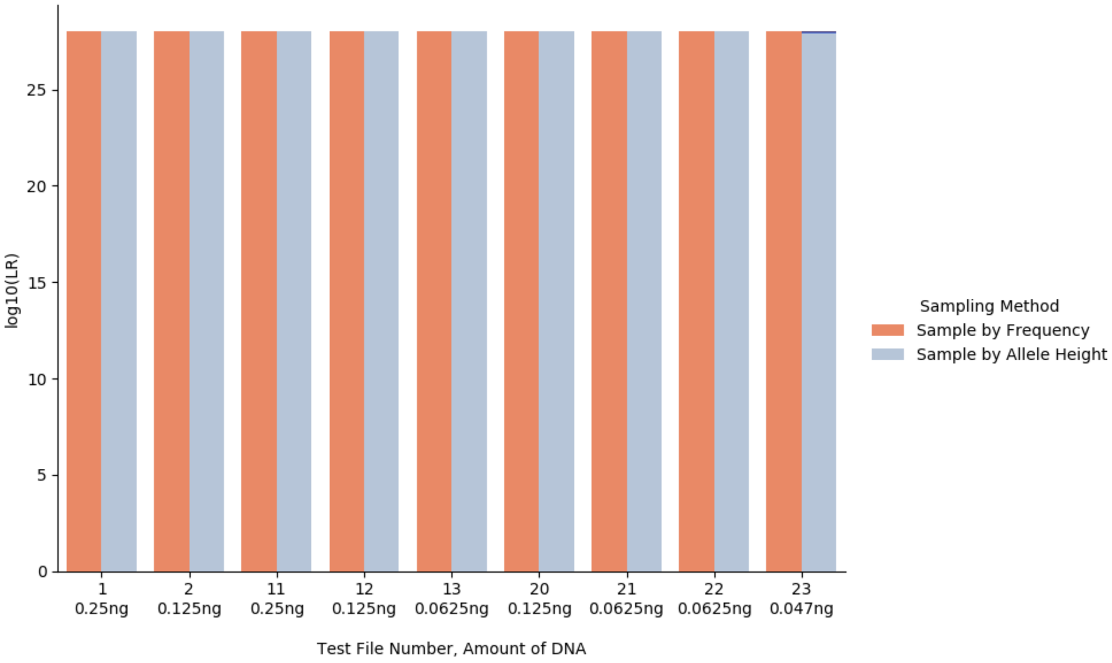

Fig. 2.1: Bar graph of a series of 1-person tests performed on known contributor #4 to test the different LR value by sampling from frequency and sampling by allele height. The mixture amount is from 0.047ng to 0.25ng. Each test ran at least five times. The light, medium, and dark shades in the bar represent the minimum, average, and maximum log likelihood ratio value for a series of runs of that given test, respectively. For some bars, there is no distinct color difference due to same log likelihood ratio value between runs.

For this series of test, the DNA mixture is tested against its contributor. With or without importance sampling produced the same results of maximum likelihood ratio against the POI. The varying amount of DNA, which ranged from 0.047ng to 0.25ng, did not affect the likelihood ratio.

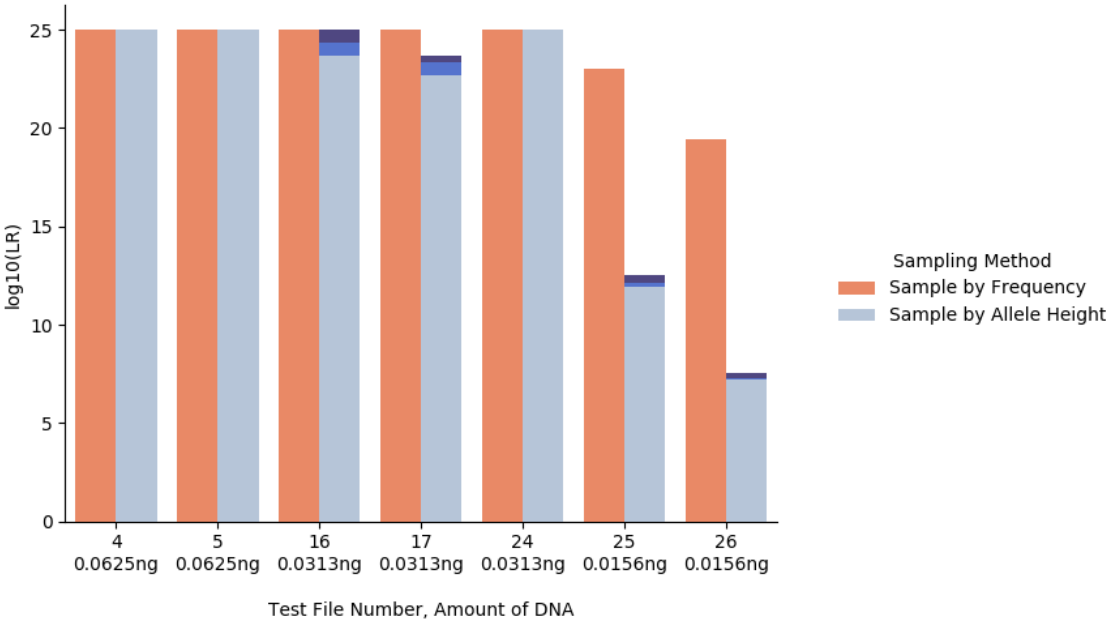

Fig. 2.2: Bar graph of a series of 1-person tests performed on known contributor #5 to test the different LR value by sampling from frequency and sampling by allele height. The amount of DNA ranged from 0.0156ng to 0.0625ng. Each test ran at least five times. The light, medium, and dark shades in the bar represent the minimum, average, and maximum log likelihood ratio value for a series of runs of that given test, respectively. For some bars, there is no distinct color difference due to same log likelihood ratio value between runs.

Figures 2.1 and 2.2 show the results for a series of tests with different amounts of template DNA in the mixtures and tested with against the actual contributor. The amount of template DNA ranged from 0.0625ng to 0.0156ng. For many tests, there was no difference between importance sampling and without importance sampling. These tests had a maximum log likelihood ratio of 25 and many of the tests evaluated to maximum likelihood ratio value. In other words, CEESIt had evaluated the POI's genotype in favor of the prosecutor's hypothesis that the POI indeed contributed to the evidence.

The tests where there are differences between importance sampling are DNA mixtures are where the amount of template DNA is 0.0313ng or less. At 0.0313ng, tests 16 and 17, the importance sampling tests showed a lower likelihood ratio value than without importance sampling. Without importance sampling still calculated maximum likelihood ratio. At 0.0156ng, both sampling methods did not reach maximum likelihood ratio and importance sampling still resulted in a likelihood ratio value less than without importance sampling.

At low amounts of template DNA, the sampling distribution influenced the final likelihood ratio value. When generating the sample population based on frequency, alleles that have low frequencies are less likely to appear in the sample population and alleles with high frequencies are more likely to appear in the sample population. With importance sampling, the sample population is based on the DNA mixture's allele height. By sampling on the allele height and then weighing to the frequency distribution, the sample population is distributed differently from the frequency sample population. The sample population is more likely to include the alleles of actual contributors because importance sampling generates the sample population from observed heights in the DNA mixture. Importance sampling is more likely to contain a sample that is potentially similar to the actual contributor, which lowers the final likelihood ratio of the person-of-interest

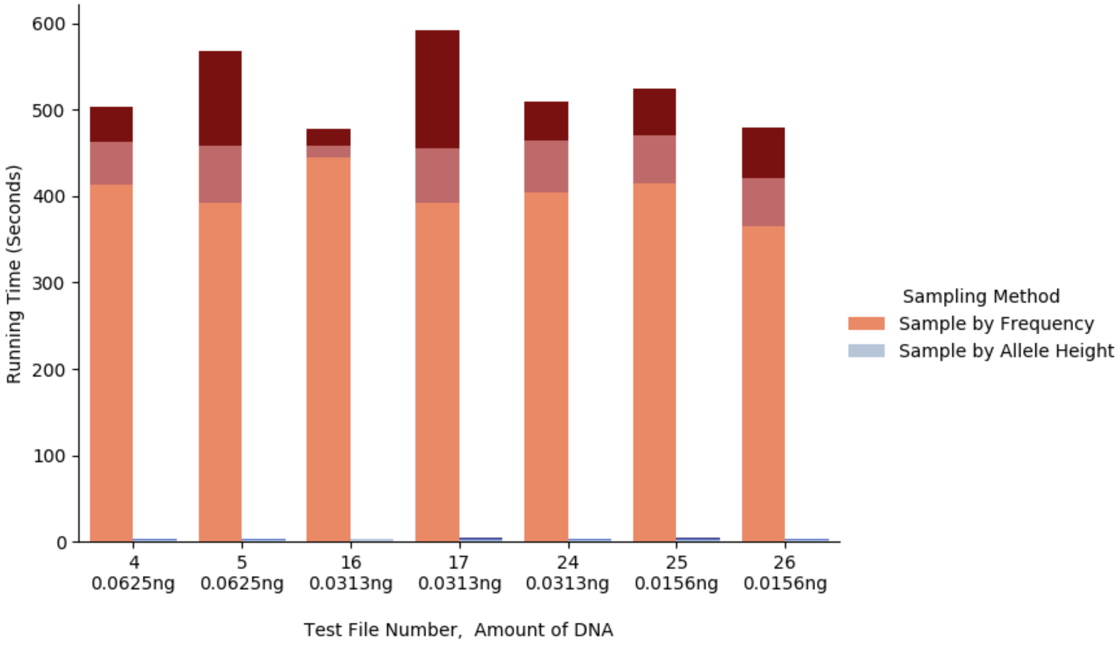

Fig. 2.3 Bar graph for a series of 1-person tests performed on known contributor focusing on runtime. Each test ran a minimum of five times. Each test ran at least five times. The light, medium, and dark shades in the bar represent the minimum, average, and maximum runtime for a series of runs of that given test, respectively.

Even though many of the tests share the same likelihood ratio with or without importance sampling, there is a substantial difference in runtime. There is over 300-second difference between runs sampling by frequency and sampling by allele height. The above figure shows that without importance sampling, runs require more than 300 seconds to complete while with importance sampling takes less than 10 seconds. For each test that samples according to frequency, there are about 20-100 second differences between each run. With tests that sample by allele height, there is minimal difference between each run as denoted by a line in the bar graph. Sampling by frequency requires a sample population of one billion to accurately report the final value, which significantly increases runtime. If the algorithm uses a population size less than one billion, the sample population does not accurately represent the background population. Importance sampling only requires a sample population of one million to yield the same final result value.

2.3 2-people Tests

2-people tests included tests with two contributors in the mixture ratio and the person of interest is one of the known contributors. The other contributor acts as interference in the DNA mixture. Each test consists of a mixture in ratios of 1:1, 1:2, 1:4, or 1:9 with different amounts of total template DNA, ranging from 0.25ng to 0.0625ng.

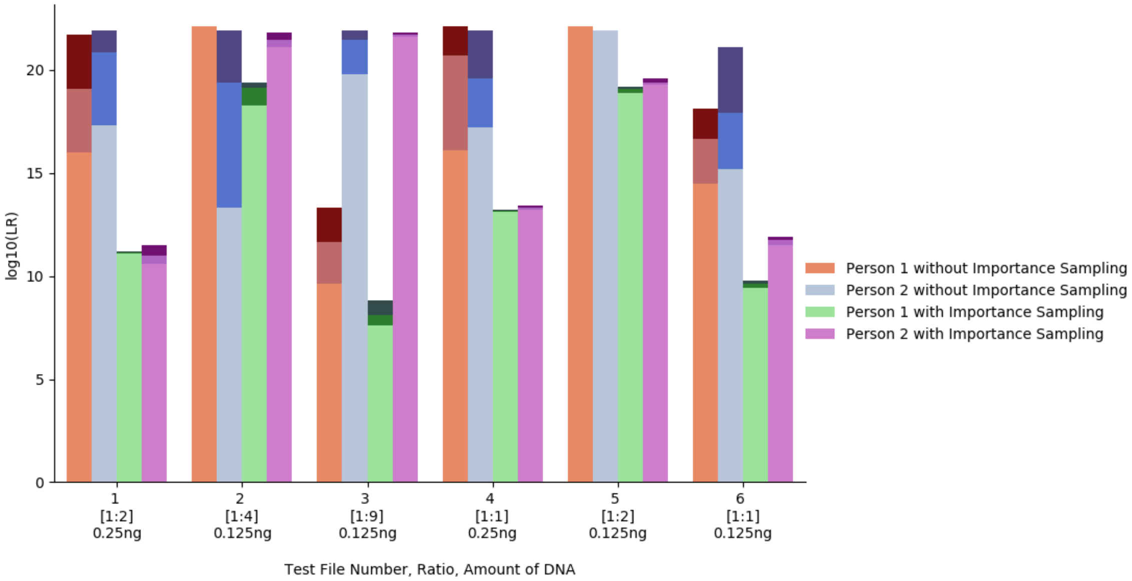

Fig. 2.4: Bar graph of a series of 2-people tests comparing major and minor contributor with and without importance sampling and focusing on the likelihood ratio. In samples with uneven contribution from the two contributors, person 2 is the major contributor (the contributor with a greater amount of template DNA) and person 1 is the minor contributor (the contributor with a lesser amount of template DNA). Each test ran at least five times. The light, medium, and dark shades in the bar represent the minimum, average, and maximum log likelihood ratio value for a series of runs of that given test, respectively. These tests focus on 0.125ng and 0.25ng of DNA mixture.

With two-people DNA mixtures, importance sampling showed a more drastic difference. The above figure shows the bar graph of the likelihood ratio with and without importance sampling. The runs with importance sampling had a genotype tolerance of 0.0625 with sample population size of two million. The runs without importance sampling used CEESIt default value: genotype tolerance of 0.5 with sample population size of one billion. These parameters with importance sampling were selected after running repeated trials under different parameter settings.

Following a similar trend from 1-person tests, 2-people tests showed higher variability between runs without importance sampling on tests that do not result in a maximum likelihood ratio, as denoted by the bar graph's confidence interval. In other words, sampling by frequency is not as repeatable as sampling by allele height. Importance sampling's sampling population is based on sampling by allele height. Because this is a two-people mixture, the amount of template DNA per contributor is reduced even further by the mixture ratio. Unlike 1-person tests where the difference between sampling by frequency and allele height differed starting at 0.0313ng, 2-people tests showed different likelihood ratio values starting at 0.25ng of DNA. When 0.25ng of DNA is split in a mixture ratio of 1:1, the expected amount of DNA per contributor is 0.125ng. At 0.125ng 1-person tests yielded same results regardless of the method of sampling, but 2-people tests show a lower likelihood ratio with importance sampling.

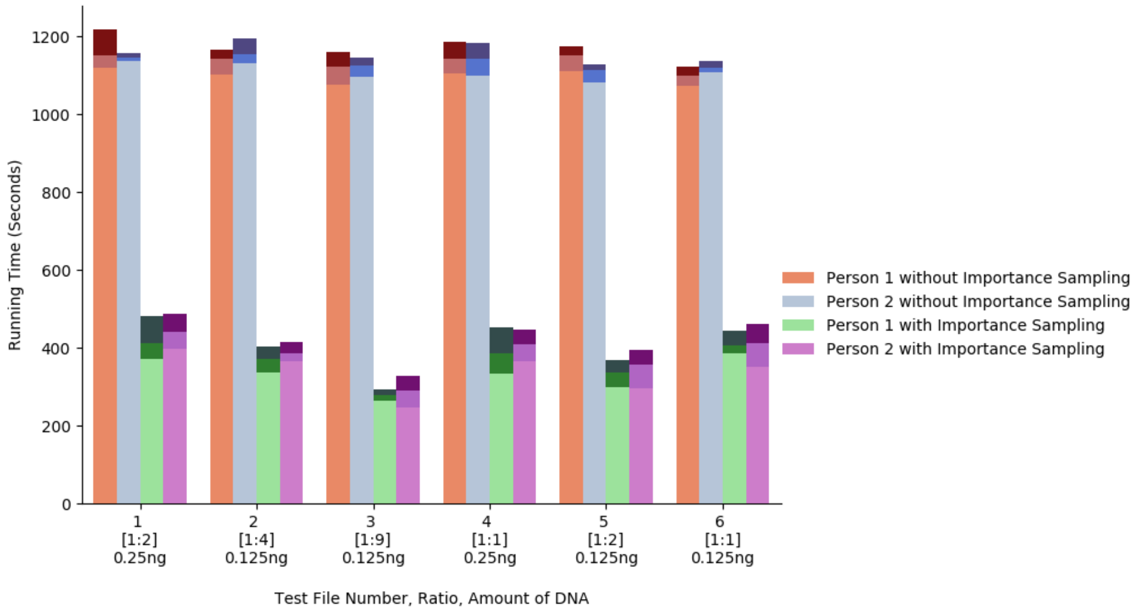

Fig. 2.6: Bar graph of a series of 2-people tests comparing major and minor contributor with and without importance sampling and focusing on runtime. In samples with uneven contribution from the two contributors, person 2 is the major contributor (the contributor with a greater amount of template DNA) and person 1 is the minor contributor (the contributor with a lesser amount of template DNA). Each test ran at least five times. The light, medium, and dark shades in the bar represent the minimum, average, and maximum runtime for a series of runs of that given test, respectively. These tests focus on 0.125ng and 0.25ng of DNA mixture.

Figure 2.6 is a bar graph that shows the runtime for each test with and without importance sampling. Just like the tests with one contributor, due to the sample population size, sampling by frequency has a higher runtime compared to sampling by allele height.

2.4 Summary

Starting with 1-persons test, DNA mixtures with high amounts of template DNA yield similar results between importance sampling and without importance sampling. Sampling by frequency and sampling by allele height did not affect the likelihood ratio because high amounts of DNA mixture yielded strong alleles peaks; therefore giving strong likelihood value of the POI having contributing to the DNA mixture since the POI did contribute to the DNA.

As the amount of DNA decreases, with and without importance sampling computed different likelihood ratio. While in theory both sampling methods should produce the same likelihood ratio, it appears that, without importance sampling, computing the correct likelihood ratio is unlikely even with the large number of samples (one billion) that was used.

One consistent difference between importance sampling and without importance sampling is the runtime. For sampling by frequency, a sample population size of one billion was used while for sampling by allele height, a sample population size of one million was used. Even with far fewer samples, importance sampling showed low run-to-run variability. This difference in sample population size contributes to the amount of iterations used in the calculations and therefore directly affects the runtime.

With two-people tests, there is a larger deviation for each test without importance sampling. A wider deviation between runs means that when a test is repeatedly run under the same parameters, the results are not repeatable. Because the final likelihood ratio value is a logarithmic value with base 10, each increment or decrement by 1 is a tenfold change in the likelihood ratio. Importance sampling shows a smaller deviation between each run, which means that this method not only reduces runtime but also produces repeatable results.

CHAPTER 3

CONCLUSION

Forensic DNA analysis uses loci and alleles to identify an individual. The goal of this study is to improve the accuracy and repeatability of calculating the likelihood ratio of an individual in a population while using reasonable runtime.

CEESIt uses the Monte Carlo method to repeatedly and randomly sample from a population. When sampling by the frequency of an allele, alleles that are more likely to appear in the population are just as likely to appear in the sample population. Another method to represent the sample population is to implement importance sampling. With importance sampling, each allele is sampled based on the height of an allele. The frequency of the height of allele is determined by the observed peak heights distribution. Because the distribution is different from the target population, weights are implemented to counteract the distribution differences. After running a series of runs on different tests, sampling by allele height yielded smaller variability between runs than sampling by frequency. Sampling by allele height also allows for a smaller sample population to be used, which decreases runtime.

BIBLIOGRAPHY

Benschop, Corina C. G., et al. "The Effect of Varying the Number of Contributors on Likelihood Ratios for Complex DNA Mixtures." *Forensic Science International: Genetics* 19 (2015): 92-99. Print.

Fincham, J. R. S. *Genetic Analysis: Principles, Scope, and Abjectives*. Oxford ; Boston: Blackwell Scientific Publications, 1994. Print.

MacKay, David J.C. . "Information Theory, Inference, and Learning Algorithms." Print.

Rosen, Kenneth H. *Discrete Mathematics and Its Applications*. 7th ed ed. New York: McGraw-Hill, 2012. Print.

Swaminathan, Harish. "Computational Methods for the Interpretation of Forensic DNA Samples." (2015). Print.

Swaminathan, Harish, et al. "Ceesit: A Computational Tool for the Interpretation of Str Mixtures." *Forensic Science International: Genetics* 22 (2016): 149-60. Print.

Swaminathan, Harish, et al. "Four Model Variants within a Continuous Forensic DNA Mixture Interpretation Framework: Effects on Evidential Inference and Reporting." *PLOS ONE* 13.11 (2018): e0207599. Print.

Westermeier, Reiner *Electrophoresis in Practice*
a Guide to Methods and Applications of DNA and Protein Separations. Fifth Edition ed: Wiley-VCH, 2016. Print.