CSE 494/594 Algorithms in Computational Biology
Project Proposal

Alexandra Dent, Juan Garcia Mesa, Matthew Huff, Raj Shah

March 1, 2020

1 Introduction

Topic description, background information and what's been done

- Forensic identification based on DNA
- mtDNA (Amorim et al.), HLA, DIP-STR (Kuffel et al.)

2 Motivation

Why is this project important:

- Need for improvements in the field of forensic identification

3 Methods

Different statistical approaches and workflows exist to determine whether the DNA an individual is present in a genomic DNA mixture (e.g. [3], [2], [4]). However, there is a common pattern that all methods follow and that we intend to replicate.

First, the development of a robust theoretical framework for detecting the presence of an individual in a mixture sample is needed. Current probabilistic approaches are generally divided into deterministic versus Bayesian analysis to deconvolute a set with multiple contributors [8]. We intend to develop a novel framework by combining the power of single nucleotide polymorphisms (SNPs) from mtDNA, well stablished and widely used in the filed of forensic identification, together with HLA, both highly polymorphic genomic regions. Kuffel et al. [1] conclude that the application of HLA together with any standard short tandem repeat (STR) based analysis (e.g. deletion/insertion polymorphism (DIP-STR)) can show a significant increase in the probability of positive identification. We do not discard the use of DIP-STR to complement and strengthen our analysis.

Then, the following step is to test the limits of differentiating power of our framework through computational simulations. Hu et al. (2014) [8] provide a thorough review of software, including dynamic and web-based, that have been applied to produce accurate analysis of complex DNA profiles. These include LoComatioN [5], targeting low-copy DNA profiling in mixed DNA samples, an open-source R package developed by Forensim [6] that interprets and weights forensic DNA evidence, and LRmix software [7], which builds upon the previous Forensim package. Other approaches based on coalescent simulation of pairs of alleles have also been done and remain a possibility, although perhaps slightly out of the scope of this project.

Finally, if possible, demonstrate the validity of the simulation results with data from real world samples. Fortunately, there exist many online data bases that provide public data fitted for this validation. An thorough search is yet to be done to determine what database will be used, given all previous milestones are completed successfully and we are to perform validation with real data.

The innovation of the workflow of the project will be combining genomic polymorphism from both mtDNA and HLA. The latter has been understudied and has recently shown promising results when combined with other standard methods [1].

References

[1] Agnieszka Kuffel et al. "Human Leukocyte Antigen alleles as an aid to STR in complex forensic DNA samples". In: *Science & Justice* (2019).

[2] Nils Homer et al. "Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays". In: *PLoS genetics* 4.8 (2008).

[3] RG Cowell et al. "Analysis of forensic DNA mixtures with artefacts". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 64.1 (2015), pp. 1–48. DOI: https://doi.org/10.1111/rssc.12071.

[4] Samuel H. Vohr et al. "A method for positive forensic identification of samples from extremely low-coverage sequence data". In: *BMC Genomics* 16 (2015). DOI: 10.1186/s12864-015-2241-6.

[5] Peter Gill, Amanda Kirkham, and James Curran. "LoComatioN: a software tool for the analysis of low copy number DNA profiles". In: *Forensic Science International* 166.2-3 (2007), pp. 128–138.

[6] Hinda Haned. "Forensim: an open-source initiative for the evaluation of statistical methods in forensic genetics". In: *Forensic Science International: Genetics* 5.4 (2011), pp. 265–268.

[7] Hinda Haned and Peter Gill. "Analysis of complex DNA mixtures using the Forensim package". In: *Forensic Science International: Genetics Supplement Series* 3.1 (2011), e79–e80.

[8] Na Hu et al. "Current developments in forensic interpretation of mixed DNA samples". In: *Biomedical reports* 2.3 (2014), pp. 309–316.