Calculating forensic trace-suspect match probabilities for Y-STRs using coalescent theory

Mikkel Meyer Andersen\textsuperscript{a}, Amke Caliebe\textsuperscript{b}, Arne Jochens\textsuperscript{b}, Sascha Willuweit\textsuperscript{c}, Michael Krawczak\textsuperscript{b}

\textsuperscript{a) Department of Mathematical Sciences, Aalborg University, Denmark}
\textsuperscript{b) Institute of Medical Informatics and Statistics, Christian-Albrechts University, Kiel, Germany}
\textsuperscript{c) Institute of Legal Medicine, Charité, Universitätsmedizin, Berlin, Germany}

DNA in Forensics 2012
Match probabilities are needed

\[ LR = \frac{P(E|H_p)}{P(E|H_d)} \]

- \( H_p \): The suspect is the donor of the genetic data (prosecutor’s hypothesis)
- \( H_d \): The suspect unconnected to the crime (defense attorney’s hypothesis)
- \( P(E|H_p) = 1 \) is often assumed
- \( P(E|H_d) \): 'match probability', the probability that the suspect matches the haplotype found at the crime scene given that the suspect is unconnected to the crime (how probable is it that some random man’s haplotype matches the haplotype found at the crime scene)
Aim and terminology

- **Aim**: Obtain the match probability for a haplotype (If we knew the haplotypes of the entire population, the population frequency of the haplotype in question would be the match probability.)

- **Terminology**:
  - The match probability is $p$ assuming the $X$ model
  - Using the $X$-based estimator, the match probability is $p$

- A way to estimate the match probability is called a match probability estimator
Most haplotypes are rare

- German Y-STR database (15 loci, ignoring DYS385a/b): 1,757 of 1,469 haplotypes (84%) were singletons (haplotypes only observed once)

- We focus on singletons because they are difficult to treat and there are a lot of them
Notation

- Let $n - 1$ be the size of the reference database.
- The haplotype $h$ found at the crime scene has not previously been observed.
- Adding $h$ to the database, we get a database of size $n$ and $h$ is a singleton.
Existing estimators

- **Binomial estimator**: Match probability is $1/n$, thus $LR = n$
  - Adding the haplotype that has not previously been observed to a database of size $n-1$, we get 1 observation in $n$ haplotypes
  - Too conservative

- **Surveying estimator (implemented on [http://www.yhrd.org](http://www.yhrd.org))**: Match probability depends on the genetic information of the haplotype in question

- **Brenner’s $\kappa$**
  - Inspired by a clever argument by Robbins (1968) about unobserved probability mass in an experiment
  - $\kappa$ is the singleton proportion ($\kappa = \alpha/n$, where $\alpha$ is the number of singletons observed)
  - Assume $0 < \kappa < 1$, thus $0 < 1 - \kappa < 1$ making $\frac{1}{1 - \kappa} > 1$
  - Match probability is $\frac{1 - \kappa}{n} = \frac{1}{n} - \frac{\kappa}{n} < \frac{1}{n}$
  - Inflates the $LR$ from the binomial $LR = n$ to

  $$LR = \frac{n}{1 - \kappa} = n \times \frac{1}{1 - \kappa}$$
Existing estimators

- Binomial estimator $1/n$ and Brenner’s estimator $(1 - \kappa)/n$ does NOT depend on the genetics of the singleton in question:
  - Same match probability to all singletons
- Surveying estimator depends on the genetics of the haplotype in question
  - Different estimates depending on the singleton’s weighted inverse distance – how much alleles are alike – to the other haplotypes in the database
Fisher-Wright model: basics

- $N$: the constant number of individuals in the population
- Generations are discrete and non-overlapping
- Selectively neutral mutation process (single-step mutation model)
- Evolution forwards in time
- Y-STR: assume that one individual can get children (as opposed to when two individuals are required to get a child)
Fisher-Wright model: evolution

- Initial population: Each individual in the initial population of size $N$ is assigned a haplotype (this can be the same for all individuals).
- A new generation is obtained as follows:
  1. Create $N$ individuals:
     For individual $i = 1, 2, \ldots, N$, choose the $i$’th individual’s parent (father) at random from the previous generation (each with probability $1/N$) and inherit this parent’s haplotype.
  2. For every locus at every individual in the new generation, determine if a mutation (and its direction) will happen.

Match probabilities for Y-STRs using coalescent theory

Mikkel Meyer Andersen - mikl@math.aau.dk
Fisher-Wright model: evolution

Image based on one from http://www.csbio.unc.edu/mcmillan

Match probabilities for Y-STRs using coalescent theory

Mikkel Meyer Andersen - mkl@math.aau.dk
Fisher-Wright model: evolution

Image based on one from http://www.csbio.unc.edu/mcmillan
Fisher-Wright model: evolution

Image based on one from http://www.csbio.unc.edu/mcmillan

Match probabilities for Y-STRs using coalescent theory

Mikkel Meyer Andersen - mikl@math.aau.dk
Fisher-Wright model: evolution

Image based on one from http://www.csbio.unc.edu/mcmillan
Match probabilities for Y-STRs using coalescent theory

Mikkel Meyer Andersen - mikl@math.aau.dk

Image based on one from http://www.csbio.unc.edu/mcmillan
Most recent common ancestor (MRCA)

Image based on one from http://www.csbio.unc.edu/mcmillan
Sample from population

Image based on one from http://www.csbio.unc.edu/mcmillan

Match probabilities for Y-STRs using coalescent theory

Mikkel Meyer Andersen - mikl@math.aau.dk
Coalescent theory

- Given a sample from a population, coalescent theory aims to make inference about the population and its evolution (e.g. time to most recent common ancestor, TMRCA)
- Invented by J. F. C. Kingman in the 1970s-1980s
- Lineages are said to coalesce when they have the same father
- Uses approximations of properties of the Fisher-Wright model
- Widely used for a lot of population genetics applications
Inference:

- Sample random evolutionary histories giving rise to the sample (e.g., by using MCMC as Wilson and others (1998, 2003) or importance sampling) and record the relevant quantities in each evolutionary history (for example the time to the MRCA)
The match probability of \( h_S \), \( P(h_X = h_S \mid H, h_S) \), under this database and evolutionary history is the probability that \( h_Z \) mutates into \( h_S \) during the time span indicated by the dotted line, thereby creating a match between the suspect and trace haplotype.
Match probability under an unknown evolutionary history

- We do not know the actual evolutionary history (tree)
- Sample a large number of histories at random according to their probabilities and averaging out to get match probability under the coalescent model
Convergence

Two different types of convergence:

- For a given reference database $H$ and a given suspect haplotype $h_S$, estimates $\hat{p}_{H,h_S,m}$ converge to $P(h_X = h_S \mid H, h_S)$ as the number of sampled trees $m$ increases.

- $P(h_X = h_S \mid H, h_S)$ converges to the true match probability $P(h_X = h_S)$ when the reference database $H$ expands to comprise the whole population.
Simulation study

- Sample a population of 50,000,000 individuals (7 loci and mutation rate 0.003)
- From this, sample a database
- For each singleton in the database, assume it belongs to the suspect
- Estimate the match probability of this singleton and compare estimate with (the now known) population frequency
- We did this for 5 databases of size 100 and 5 databases of size 200
Comparison instruments

- $h_{S_j}$: $j$th singleton out of $v$ singletons in the database
- $H_j$: the database with the $j$th singleton excluded
- $\hat{p}_{H_j,h_{S_j}}$: an estimate of the match probability (e.g., coalescent-based, surveying, ...)
- $p_{h_{S_j}}$: the population frequency of $h_{S_j}$
- Bias

\[ \frac{1}{v} \sum_{j=1}^{v} (\hat{p}_{H_j,h_{S_j}} - p_{h_{S_j}}) \]

- Mean squared error

\[ \frac{1}{v} \sum_{j=1}^{v} (\hat{p}_{H_j,h_{S_j}} - p_{h_{S_j}})^2 \]

- Spearman correlation coefficient between $\hat{p}_{H_j,h_{S_j}}$ and $p_{h_{S_j}}$
Plot of biases

Solid lines: databases of size 100 (500,000 simulated coalescent trees per singleton)

Dashed lines: databases of size 200 (200,000 simulated coalescent trees per singleton)

Each color correspond to a database
Introduction

Existing estimators

Coalescent-based estimator

Simulation study

Results

Plot of mean squared error

Solid lines: databases of size 100 (500,000 simulated coalescent trees per singleton)

Dashed lines: databases of size 200 (200,000 simulated coalescent trees per singleton)

Each color correspond to a database
## Comparison with other estimators

Databases of size 100 (500,000 simulated coalescent trees per singleton)  
Databases of size 200 (200,000 simulated coalescent trees per singleton)

<table>
<thead>
<tr>
<th>Size</th>
<th>Database</th>
<th>Bias (Brenner)</th>
<th>Bias (Surveying)</th>
<th>Bias (Coalescent)</th>
<th>MSE (Brenner)</th>
<th>MSE (Surveying)</th>
<th>MSE (Coalescent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1</td>
<td>$-9.4 \cdot 10^{-5}$</td>
<td>$9.1 \cdot 10^{-3}$</td>
<td>$2.9 \cdot 10^{-4}$</td>
<td>$4.3 \cdot 10^{-6}$</td>
<td>$1.5 \cdot 10^{-4}$</td>
<td>$2.7 \cdot 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>$-3.3 \cdot 10^{-4}$</td>
<td>$8.9 \cdot 10^{-3}$</td>
<td>$2.7 \cdot 10^{-4}$</td>
<td>$4.7 \cdot 10^{-6}$</td>
<td>$8.3 \cdot 10^{-5}$</td>
<td>$2.4 \cdot 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>$4.3 \cdot 10^{-4}$</td>
<td>$9.6 \cdot 10^{-3}$</td>
<td>$4.2 \cdot 10^{-4}$</td>
<td>$2.4 \cdot 10^{-6}$</td>
<td>$9.5 \cdot 10^{-5}$</td>
<td>$2.0 \cdot 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>$5.4 \cdot 10^{-5}$</td>
<td>$8.8 \cdot 10^{-3}$</td>
<td>$-4.0 \cdot 10^{-5}$</td>
<td>$3.4 \cdot 10^{-6}$</td>
<td>$9.8 \cdot 10^{-5}$</td>
<td>$2.3 \cdot 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>$-6.4 \cdot 10^{-4}$</td>
<td>$8.1 \cdot 10^{-3}$</td>
<td>$2.5 \cdot 10^{-4}$</td>
<td>$2.5 \cdot 10^{-6}$</td>
<td>$9.6 \cdot 10^{-5}$</td>
<td>$1.9 \cdot 10^{-6}$</td>
</tr>
<tr>
<td>200</td>
<td>1</td>
<td>$7.7 \cdot 10^{-5}$</td>
<td>$4.1 \cdot 10^{-3}$</td>
<td>$5.6 \cdot 10^{-5}$</td>
<td>$1.8 \cdot 10^{-6}$</td>
<td>$2.7 \cdot 10^{-5}$</td>
<td>$2.3 \cdot 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>$3.5 \cdot 10^{-4}$</td>
<td>$4.9 \cdot 10^{-3}$</td>
<td>$3.3 \cdot 10^{-4}$</td>
<td>$2.6 \cdot 10^{-6}$</td>
<td>$2.6 \cdot 10^{-5}$</td>
<td>$3.8 \cdot 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>$6.1 \cdot 10^{-4}$</td>
<td>$4.2 \cdot 10^{-3}$</td>
<td>$1.4 \cdot 10^{-4}$</td>
<td>$1.8 \cdot 10^{-6}$</td>
<td>$2.5 \cdot 10^{-5}$</td>
<td>$1.9 \cdot 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>$4.9 \cdot 10^{-4}$</td>
<td>$4.7 \cdot 10^{-3}$</td>
<td>$1.2 \cdot 10^{-4}$</td>
<td>$1.7 \cdot 10^{-6}$</td>
<td>$2.7 \cdot 10^{-5}$</td>
<td>$3.3 \cdot 10^{-6}$</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>$-8.4 \cdot 10^{-5}$</td>
<td>$4.3 \cdot 10^{-3}$</td>
<td>$9.8 \cdot 10^{-5}$</td>
<td>$2.0 \cdot 10^{-6}$</td>
<td>$2.2 \cdot 10^{-5}$</td>
<td>$2.8 \cdot 10^{-6}$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Size</th>
<th>Database</th>
<th>Correlation (Brenner)</th>
<th>Correlation (Surveying)</th>
<th>Correlation (Coalescent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>1</td>
<td>0</td>
<td>0.528</td>
<td>0.446</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0.566</td>
<td>0.509</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0.413</td>
<td>0.327</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0.401</td>
<td>0.274</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0.389</td>
<td>0.266</td>
</tr>
<tr>
<td>200</td>
<td>1</td>
<td>0</td>
<td>0.309</td>
<td>0.154</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0</td>
<td>0.490</td>
<td>0.267</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0.283</td>
<td>0.343</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0</td>
<td>0.381</td>
<td>0.184</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0</td>
<td>0.389</td>
<td>0.250</td>
</tr>
</tbody>
</table>
Comparison with other estimators

In-depth analysis for 10 randomly selected singletons per database of size 200 of the coalescent-based estimator of match probabilities, using different numbers of simulated coalescent trees ($2 \times 10^5$ and $10^6$ per singleton):

<table>
<thead>
<tr>
<th>Sample</th>
<th>Bias</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$2 \times 10^5$</td>
<td>$10^6$</td>
</tr>
<tr>
<td>1</td>
<td>$-4.1 \cdot 10^{-4}$</td>
<td>$-1.8 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>2</td>
<td>$-6.7 \cdot 10^{-4}$</td>
<td>$-4.6 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>3</td>
<td>$-6.9 \cdot 10^{-4}$</td>
<td>$-5.6 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>4</td>
<td>$1.4 \cdot 10^{-3}$</td>
<td>$4.6 \cdot 10^{-4}$</td>
</tr>
<tr>
<td>5</td>
<td>$-3.6 \cdot 10^{-4}$</td>
<td>$-3.2 \cdot 10^{-4}$</td>
</tr>
</tbody>
</table>
Results

- Coalescent-based estimation seems promising (works for non-singletons, too)
- Computationally difficult (at the moment): Requires many trees (many simulations of the evolutionary history)
  - More loci and more samples lead to better estimates of the match probabilities (in principle)
  - More loci and more samples increase the number of possible evolutionary histories (more simulations)
- Research on improving speed must be done in order to make it an everyday tool
The modified BATWING program with the forensic match probability module included can be downloaded from the 'Software' page at http://people.math.aau.dk/~mikl

References:


$W_i$: weighted inverse molecular distance (like used in the surveying method); a measure of how much alleles are alike.