Analysis of Y-Chromosomal STR Population Data Using the Discrete Laplace Model

9th ICFIS
Leiden University, August 2014

Mikkel Meyer Andersen, Poul Svante Eriksen and Niels Morling
The discrete Laplace method and its applications

Comparing methods for calculating $LR$ for Y-STR data
Introduction
Evidential weight

\( H_p \) (prosecutor’s hypothesis): ’The suspect left the Y-chromosome DNA in the crime stain.’

\( H_d \) (defence attorney’s hypothesis): ’A random man left the Y-chromosome DNA in the crime stain.’

\( E \): Evidence (e.g. DNA profile from crime scene)

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

Non-match:

\[
LR = \frac{0}{P(E \mid H_d)}
\]

Match:

\[
LR = \frac{1}{P(E \mid H_d)}
\]

(Ideal situation, no errors, etc.)

Y-STR: Loci are not independent \( \Rightarrow \) No product rule
## Sparsity of Y-STRs

<table>
<thead>
<tr>
<th>19,630 samples</th>
<th>Forensic marker set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MHT 9 loci</td>
</tr>
<tr>
<td>$n = 1$ (singletons)</td>
<td>6,083 (31.0%)</td>
</tr>
<tr>
<td>$n = 2$ (doubletons)</td>
<td>1,131</td>
</tr>
<tr>
<td>$n = 3$</td>
<td>435</td>
</tr>
<tr>
<td>$n = 4$</td>
<td>226</td>
</tr>
<tr>
<td>$n = 5$</td>
<td>114</td>
</tr>
<tr>
<td>$n = 6$</td>
<td>86</td>
</tr>
<tr>
<td>$n = 7$</td>
<td>63</td>
</tr>
<tr>
<td>$n = 8$</td>
<td>43</td>
</tr>
<tr>
<td>$n = 9$</td>
<td>29</td>
</tr>
<tr>
<td>$n = 10$</td>
<td>31</td>
</tr>
<tr>
<td>$n = 11$</td>
<td>22</td>
</tr>
<tr>
<td>...</td>
<td>13</td>
</tr>
<tr>
<td>$n \in (30, 40]$</td>
<td>8</td>
</tr>
<tr>
<td>$n \in (100, 515]$</td>
<td>8</td>
</tr>
</tbody>
</table>

Estimators
Estimators

▶ Forensic ’conservatism’ (innocent suspect): For whom – what about paternity, immigration, etc.?
▶ Precise (low prediction error) – how do we measure this (more later)?
▶ Does it work for all datasets, also for those only consisting of singletons?
▶ Statistical model: Guaranteed behaviour (e.g. probabilities sum to 1)
  ▶ Assign probability to all possible haplotypes (e.g. for mixture LR)
  ▶ Probability mass 1 to be distributed among possible haplotypes
Estimators

- Match probability $\approx$ DNA profile population frequency
- Count method (works for any trait, e.g. blood type)
  - $n$: Dataset size
  - $n_x$: Number of times $x$ is observed in the dataset
  - $P(X = x) = \frac{n_x}{n}$
Estimators

- Include in dataset (new observation)
  - Additional information: Under $H_d$, suspect considered as a random (wrongly accused) individual from the population; the haplotype is just another random sample

- Old dataset: $D^-$ of size $n$

- New dataset: $D$ of size $n+1$

- $P(X = x) = (n_x + 1)/(n + 1)$
  - $n_x = 0$: $P(X = x) = \frac{1}{n+1}$

- $\sum_{x \in D} \frac{n_x}{n+1} = \frac{1}{n+1} \sum_{x \in D} n_x = \frac{n+1}{n+1} = 1$, hence $P(X = x) = 0$ for $x \notin D$

- Corrected count estimators:
  - Brenner’s $\kappa$ (CH Brenner (2010) / HE Robbins (1968))
  - Generalised Good (IJ Good (1953), G Cereda/R Gill)
The Discrete Laplace method
Motivation

- Haplotype probability distribution (statistical model)
- Enables a wide range of inferences using one model:
  - Haplotype frequency estimation (observed and unobserved)
  - Mixtures (e.g. separation and LR)
  - Cluster analysis
  - ... 
- Not a new ad-hoc tool for each task
- A statistical model gives desirable properties:
  - $P(x)$: Probability mass function
  - Consistent:
    \[
    \sum_{x \in \mathcal{H}} P(x) = 1
    \]
  - $P(x) > 0$ for all $x \in \mathcal{H}$
Model

- Y-STR: Loci not statistically independent
- Our approach: Condition on [something] to obtain independency between loci
Discrete Laplace distribution

Discrete Laplace distributed \( X \sim DL(p, \mu) \):
- Dispersion parameter \( 0 < p < 1 \) and
- Location parameter \( \mu \in \mathbb{Z} = \{ \ldots, -2, -1, 0, 1, 2, \ldots \} \)

Probability mass function:

\[
f(X = x; p, \mu) = \frac{1 - p}{1 + p} \cdot p^{|x-\mu|} \quad \text{for } x \in \mathbb{Z}
\]

Perfectly homogeneous population with 1-locus haplotypes:

\[
P(X = x) = f(X = x; p, \mu)
\]

The image shows a bar graph illustrating the probability distribution of \( X \) with parameters \( p = 0.3 \) and \( \mu = 13 \), with bars indicating the probability mass function for each allele value from 8 to 18.
Statistical model for Y-STR haplotypes

Perfectly homogeneous population with \( r \)-locus haplotypes:

\[
P(X = (x_1, x_2, \ldots, x_r)) = \prod_{k=1}^{r} f(x_k; p_k, \mu_k)
\]

- \( \bar{\mu} = (\mu_1, \mu_2, \ldots, \mu_r) \): Central haplotype
- \( \bar{\rho} = (p_1, p_2, \ldots, p_r) \): Discrete Laplace parameters (one for each locus)
- Mutations happen independently across loci (relative to \( \bar{\mu} \))
Statistical model for Y-STR haplotypes

Non-homogeneous population with \( c \) subpopulations and \( r \)-locus haplotypes:

\[
P(X = (x_1, x_2, \ldots, x_r)) = \sum_{j=1}^{c} \tau_j \prod_{k=1}^{r} f(x_k; \mu_{jk}, \phi_{jk})
\]

- \( \tau_j \): A priori probability for originating from the \( j \)'th subpopulation (\( \sum_{j=1}^{c} \tau_j = 1 \))
- \( \vec{\mu}_j = (\mu_{j1}, \mu_{j2}, \ldots, \mu_{jr}) \): Central haplotype for the \( j \)'th subpopulation
- \( \vec{\phi}_j = (\phi_{j1}, \phi_{j2}, \ldots, \phi_{jr}) \): Parameters for all loci at the \( j \)'th subpopulation
- Parameter estimation from observations using \( \text{R} \) library disclapmix
Data and fit

c: Number of subpopulations

\[ P(X = x) = \sum_{j=1}^{c} \tau_j f(x; p_j, \mu_j) \]
Data and fit

$c$: Number of subpopulations

\[ P(X = x) = \sum_{j=1}^{c} \tau_j f(x; p_j, \mu_j) \]

\[ P(DYS392 = x) = 1 \cdot f(x; p = 0.41, \mu = 11) \]
Data and fit

\[ c: \text{Number of subpopulations} \]
\[ P(X = x) = \sum_{j=1}^{c} \tau_j f(x; p_j, \mu_j) \]
\[ P(\text{DYS392} = x) = 0.519 \cdot f(x; p = 0.004, \mu = 11) + 0.481 \cdot f(x; p = 0.179, \mu = 13) \]
Data and fit

\[ P(X = x) = \sum_{j=1}^{c} \tau_j f(x; p_j, \mu_j) \]

c: Number of subpopulations

- Observations
- Estimated (c = 3)
Data and fit

$c$: Number of subpopulations

\[ P(X = x) = \sum_{j=1}^{c} \tau_j f(x; p_j, \mu_j) \]

- 3 subpopulations: 
  \[ \begin{array}{c|c|c|c} 
  \hat{\mu}_j & 11 & 13 & 14 \\
  \hat{\tau}_j & 52\% & 46\% & 2\% 
  \end{array} \]

- Observed vs expected:

<table>
<thead>
<tr>
<th>Allele</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>Observed</td>
<td>0.5248</td>
<td>0.0567</td>
<td>0.3322</td>
<td>0.0714</td>
<td>0.0083</td>
</tr>
<tr>
<td>Expected</td>
<td>0.5248</td>
<td>0.0567</td>
<td>0.3315</td>
<td>0.0715</td>
<td>0.0089</td>
</tr>
</tbody>
</table>
The Discrete Laplace Method
MM Andersen
mikl@math.aau.dk

Introduction
Estimators
Discrete Laplace
Match probability
Mixture analysis
Cluster analysis
Conclusion

Match probability
Simulation study:

- Simulate populations (each 7 loci and 20 mio individuals)
- Draw random datasets
- Estimate haplotype frequencies of all singletons and compare with the true values
- Result: Smaller prediction error than those with count estimator and Brenner’s $\kappa$ method
Estimate match probability
Real data (Y23 dataset)

<table>
<thead>
<tr>
<th>Population</th>
<th>Size, $n$</th>
<th>5 loci</th>
<th>7 loci</th>
</tr>
</thead>
<tbody>
<tr>
<td>World</td>
<td>18,925</td>
<td>0.026</td>
<td>0.108</td>
</tr>
<tr>
<td>Europe</td>
<td>11,664</td>
<td>0.029</td>
<td>0.101</td>
</tr>
</tbody>
</table>

- Dataset sizes: 200 and 500
- Sampling cases with singleton haplotype:
  1. Draw dataset, $D$, from population
  2. Draw an extra observation, $h$
  3. If $h \in D$, skip and go to next sample
  4. If $h \notin D$: Estimate frequency and compare to $n_h/n$ (‘true’)

- 100 cases for each dataset size, population and locus count
- Compare to Brenner’s $\kappa$ method and Generalised Good
Estimate match probability

![Box plots](image)

- **Estimator**
  - Brenner
  - Good
  - Disclap

- **LR**
  - $10^2$
  - $10^4$
  - $10^6$
  - $10^8$

- **Loci**
  - 5 loci
  - 7 loci
## Prediction error

<table>
<thead>
<tr>
<th>Case</th>
<th>Probability</th>
<th>LR</th>
<th>LR inflation</th>
</tr>
</thead>
</table>
| Case 1 | $p_1 = 0.01$  
$\hat{p}_1 = 0.00995$ | $LR = 100.0$  
$\hat{LR} = 100.5$ | $\hat{LR}/LR = 1.005$ |
| Case 2 | $p_2 = 0.0001$  
$\hat{p}_2 = 0.00015$ | $LR = 10,000$  
$\hat{LR} = 6,667$ | $\hat{LR}/LR = 0.667$ |
### Prediction error

<table>
<thead>
<tr>
<th></th>
<th>Probability</th>
<th>LR</th>
<th>LR inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_1 = 0.01 )</td>
<td>( \hat{p}_1 = 0.00995 )</td>
<td>( LR = 100.0 )</td>
<td>( \hat{LR}/LR = 1.005 )</td>
</tr>
<tr>
<td><strong>Case 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>( p_2 = 0.0001 )</td>
<td>( \hat{p}_2 = 0.00015 )</td>
<td>( LR = 10,000 )</td>
<td>( \hat{LR}/LR = 0.667 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Error type</th>
<th>( \hat{p}_i - p_i )</th>
<th>( \hat{p}_i - p_i )</th>
<th>( (\hat{p}_i - p_i)^2 )</th>
<th>( (\hat{p}_i - p_i)^2 )</th>
<th>( \log_{10} \left( \frac{\hat{p}_i}{p_i} \right) )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Case 1</strong></td>
<td>(-0.00005)</td>
<td>(-0.005)</td>
<td>(2.5 \cdot 10^{-9})</td>
<td>(2.5 \cdot 10^{-7})</td>
<td>(-0.002)</td>
</tr>
<tr>
<td><strong>Case 2</strong></td>
<td>(0.00005)</td>
<td>(0.5)</td>
<td>(2.5 \cdot 10^{-9})</td>
<td>(2.5 \cdot 10^{-5})</td>
<td>(0.176)</td>
</tr>
</tbody>
</table>

Taking summary (sum, mean, median, ...): What is 0?
Estimate match probability

The Discrete Laplace Method

Introduction
Estimators
Discrete Laplace
Match probability
Mixture analysis
Cluster analysis
Conclusion
The discrete Laplace method and Brenner’s $\kappa$ is implemented in upcoming version of http://www.yhrd.org

The discrete Laplace method helped finding haplotypes with wrong metapopulation assignments
Mixture separation
Mixture separation

Yfiler trace, 15 loci (DYS385a/b removed):

<table>
<thead>
<tr>
<th>Locus</th>
<th>Alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>DYS19</td>
<td>14, 15</td>
</tr>
<tr>
<td>DYS389I</td>
<td>13, 14</td>
</tr>
<tr>
<td>DYS389II'</td>
<td>16, 17</td>
</tr>
<tr>
<td>DYS390</td>
<td>24, 26</td>
</tr>
<tr>
<td>DYS391</td>
<td>10, 11</td>
</tr>
<tr>
<td>DYS392</td>
<td>11, 13</td>
</tr>
<tr>
<td>DYS393</td>
<td>13</td>
</tr>
<tr>
<td>DYS438</td>
<td>11, 12</td>
</tr>
<tr>
<td>DYS439</td>
<td>10, 11</td>
</tr>
<tr>
<td>DYS437</td>
<td>14, 15</td>
</tr>
<tr>
<td>DYS448</td>
<td>19, 20</td>
</tr>
<tr>
<td>DYS456</td>
<td>15, 16</td>
</tr>
<tr>
<td>DYS458</td>
<td>14, 18</td>
</tr>
<tr>
<td>DYS635</td>
<td>23</td>
</tr>
<tr>
<td>Y GATA H4</td>
<td>12, 13</td>
</tr>
</tbody>
</table>

\[2^{13-1} = 4,096\] possible contributor pairs
### Mixture separation

<table>
<thead>
<tr>
<th>Loci</th>
<th>Danish</th>
<th>Somali</th>
<th>German</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DEN (21)</td>
<td>DEN (15)</td>
<td>DEN (10)</td>
</tr>
<tr>
<td>n</td>
<td>181</td>
<td>181</td>
<td>181</td>
</tr>
<tr>
<td>Singletons</td>
<td>181</td>
<td>164</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>(100%)</td>
<td>(90.6%)</td>
<td>(61.9%)</td>
</tr>
</tbody>
</table>

- For each dataset, 550 mixtures were simulated
- \(i^{th}\) contributor pair \(c_i = \{h_{i,1}, h_{i,2}\}\), find \(\hat{p}_i = \hat{P}(h_{i,1})\hat{P}(h_{i,2})\)
- Order all pairs according to the \(\hat{p}_i\) values (highest to lowest)
## Mixture separation

<table>
<thead>
<tr>
<th>Probability</th>
<th>DEN (21)</th>
<th>DEN (15)</th>
<th>DEN (10)</th>
<th>SOM (10)</th>
<th>GER (7)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rank ≤ 1</td>
<td>13%</td>
<td>26%</td>
<td>45%</td>
<td>72%</td>
<td>53%</td>
</tr>
<tr>
<td>Rank ≤ 5</td>
<td>33%</td>
<td>55%</td>
<td>84%</td>
<td>94%</td>
<td>89%</td>
</tr>
<tr>
<td>Rank ≤ 10</td>
<td>42%</td>
<td>69%</td>
<td>93%</td>
<td>98%</td>
<td>97%</td>
</tr>
<tr>
<td>Random ≤ 10</td>
<td>0.03%</td>
<td>0.78%</td>
<td>12.15%</td>
<td>26.79%</td>
<td>53.93%</td>
</tr>
</tbody>
</table>

![Graph showing probability of true rank ≤ x for different methods](image)

**Ranking**
- **Discrete Laplace**
- **Random**

---

**The Discrete Laplace Method**

MM Andersen

mikl@math.aau.dk

- Introduction
- Estimators
- Discrete Laplace
- Match probability
- Mixture analysis
- Cluster analysis
- Conclusion
Mixture \( LR \)

\[
H_p : S + U \\
H_d : U_1 + U_2
\]

\[
LR = \frac{P(H_U)}{\sum_{(H_{U_1}, H_{U_2})} P(H_{U_1})P(H_{U_2})}
\]

![Graphs showing log10(LR) vs log10(1/P(H_s)) for different methods: DEN (21), DEN (15), DEN (10), SOM (10), GER (7). Predicted ranks for each range of LR values are indicated.](image-url)
Cluster analysis
Cluster analysis

- \( \tau_j = P(\text{From subpopulation } j) \)
- Haplotype frequency by summing the contributions from each subpopulation:

\[
P(Haplotype = x) = \sum_{j=1}^{c} \tau_j \cdot P(Haplotype = x \mid \text{From subpopulation } j).
\]

- Discrete Laplace model

- Bayes theorem:

\[
P(\text{From subpopulation } j \mid Haplotype = x) = \frac{\tau_j \cdot P(Haplotype = x \mid \text{From subpopulation } j)}{P(Haplotype = x)}
\]
Cluster analysis of European data
7 loci

First analysed in 'Signature of recent historical events in the European Y-chromosomal STR haplotype distribution' by Roewer et al. in 2005
Cluster analysis of Y23
21 loci (from Purps J, Siegert S, et al. (2014))
Pairwise population distances:

- 7-locus, 12,727 European males (91 locations): Correlation(AMOVA, discrete Laplace) = 0.90
- 10-locus, 2,736 African males (26 locations): Correlation(AMOVA, discrete Laplace) = 0.82
- 21-locus (Y23), 18,925 males (129 locations): Correlation(AMOVA, discrete Laplace) = 0.78
Concluding remarks
The discrete Laplace method

- Sound statistical properties
- Applications
  - Estimation of Y-STR haplotype population frequencies
  - Mixture analysis
  - Cluster analysis
- Computationally feasible
- Open source software: R libraries disclap and disclapmix (and fwsim for simulating populations)
- Criticism
  - Intermediate alleles (e.g. 10.2)
  - Duplications (e.g. DYS385a/b)
  - Copy number variation (e.g. Yfiler Plus)
  - ̂μ’s difficult to estimate (curse of dimensionality)
Conclusion

- Match probability is of great interest and is difficult
- Validation of methods
  - Open source software (e.g. R library, C++ program): Compare to your own method
  - Availability of real data (PPY23)
    - R-object: http://people.math.aau.dk/~mikl/?p=y23
  - Battery of simulated populations
  - Measure of prediction error
- For a matching profile (e.g. Y23 or Yfiler Plus), use only subset (e.g. 7 or 10 loci) for LR calculations?
  - Easier to validate
Thank you for your attention