Outline

Y-STR: Haplotype Frequency Estimation and Evidence Calculation

by Mikkel Meyer Andersen

Introduction

Outline

1. Short biological recap
2. Motivation and aims of using Y-STR
3. Frequency estimation
   3.1 Methods and comments to the methods
   3.2 Model control
4. Evidence calculation
5. Further work
Y-STR: Haplotype Frequency Estimation and Evidence Calculation by Mikkel Meyer Andersen

Introduction Outline Biological framework Motivation Aims Estimating frequencies Comparing models Calculating evidence Further work Questions

Example of (autosomal) STR DNA-type: 
\[
(\{14, 13\}, \{22\}, \{15, 16\}, \ldots, \{22\})
\]

Example of Y-STR DNA-type: 
\[
(17, 14, 22, \ldots, 15)
\]

Motivation

- In some situations Y-STR is more sensible to use than A-STR (autosomal STR), e.g. to avoid noise in the trace from a rape victim.
- Y-STR and A-STR differs in several areas, e.g. the number of alleles at each locus and statistical dependence between loci.
- The statistical methods developed to handle A-STR cannot be applied on Y-STR directly, so reformulation is required (e.g. for calculating evidence) or new methods must be developed (to estimate Y-STR haplotype frequencies).
Aims

Be able to calculate statistical evidence in trials

Estimate frequencies for Y-STR haplotypes (also unobserved ones) is required to do this

First, methods for estimating frequencies for Y-STR haplotypes will be discussed and afterwards calculation of evidence will be introduced.
Estimating frequencies

Y-STR: Haplotype Frequency Estimation and Evidence Calculation

by Mikkel Meyer Andersen

Introduction

Estimating frequencies

Dimension reduction
Existing methods
New methods
Frequency surveying
Ancestral awareness
Classification models
Kernel smoothing

Comparing models
Calculating evidence
Dimension reduction

Neither principal component analysis nor factor analysis yields good results, so that path has not been followed any further.
Existing methods

- Simple count estimates
  - Not precise enough
  - Several problems exist; some will be presented in this presentation (some also presented in a talk at 7th International Y Chromosome User Workshop in Berlin, Germany, from April 22 to April 24, 2010)
New methods

- Graphical models would be an obvious choice
  - Structure based learning (e.g. PC algorithm) or score based learning (e.g. AIC and BIC)
  - Standard tests for conditional independence (e.g. $G^2 = 2nCE(A, B \mid \{S_i\}_{i \in I})$, where $n$ is the sample size and $CE$ is the cross entropy, which is $\chi^2_\nu$ distributed when $A$ and $B$ are independent given $\{S_i\}_{i \in I}$) do not exploit the ordering in the data nor does it incorporate prior knowledge such as the single step mutation model
  - Better independence tests are required

- Ancestral awareness

- Classification models (e.g. classification trees, ordered logistic regression, and support vector machines)

- Kernel smoothing and model-based clustering
Y-STR: Haplotype Frequency Estimation and Evidence Calculation by Mikkel Meyer Andersen

Introduction

Estimating frequencies
Dimension reduction
Existing methods
New methods

Frequency surveying
Ancestral awareness
Classification models
Kernel smoothing

Comparing models
Calculating evidence
The idea: Bayesian approach

Notation:
- \( N \): number of observations
- \( M \): number of haplotypes (i.e. unique observations)
- \( N_i \): the number of times the \( i \)'th haplotype has been observed
- \( f_i = \frac{N_i - 1}{N - M} \): the frequency for the \( i \)'th haplotype

Model using Bayesian inference:
1. A priori: assume \( f_i \) is Beta distributed with parameters non-stochastic parameters \( u_i \) and \( v_i \)
2. Likelihood: Given \( f_i \), then \( N_i \) is Binomial distributed
3. Posterior: Given \( N_i \), then \( f_i \) is (still) Beta distributed (Beta distribution is a conjugate prior for the Binomial distribution)

Model expressed in densities using generic notation:
\[
p(f_i|N_i) \propto p(N_i|f_i) p(f_i)
\]
Estimating $u_i$ and $v_i$

1. Calculate $W_i = \frac{1}{N-N_i} \sum_{i \neq j} \frac{N_j}{d_{ij}}$ for $i = 1, 2, \ldots, M$, where $d_{ij}$ denotes the Manhattan distance/$L^1$ norm.

2. Order the $W_i$'s by size and divide into 15 (?) groups and calculate the mean and variance of the $f_i$'s in each group.

3. Fit regression models $\mu(W) = \beta_1 + \exp(\beta_2 W + \beta_3)$ and $\sigma^2(W) = \beta_4 + \exp(\beta_5 W + \beta_6)$ based on the 15 estimates.

4. Calculate $\mu_i = \mu(W_i)$ and $\sigma_i^2 = \sigma^2(W_i)$ and use these to calculate the prior parameters $u_i$ and $v_i$.

5. Apply the Bayesian approach to obtain the posterior distribution, e.g. to estimate $f_i$ using the posterior mean.

Only dane could fit the full models – and the fit is not too comforting – the others resulted in $\mu_i < 0$ or $\sigma_i^2 < 0$ for some $i$'s.
Y-STR: Haplotype Frequency Estimation and Evidence Calculation

by Mikkel Meyer Andersen

Introduction

Estimating frequencies

Dimension reduction

Existing methods

New methods

Frequency surveying

Ancestral awareness

Classification models

Kernel smoothing

Comparing models

Calculating evidence

\[
\mu(W) = 0.002 + \exp(81.389 \times W - 21.454)
\]
Introduction

Estimating frequencies

Dimension reduction
Existing methods
New methods

Frequency surveying
Ancestral awareness
Classification models
Kernel smoothing

Comparing models
Calculating evidence

Y-STR:
Haplotype
Frequency
Estimation
and Evidence Calculation

by Mikkel Meyer Andersen

\[ \sigma^2(W) = 0 + \exp(114.043 \times W - 31.492) \]
Modified regression models

■ Set $\beta_1 = \beta_4 = 0$ in the regression models yielding

$$\mu(W) = \exp(\beta_2 W + \beta_3)$$

and

$$\sigma^2(W) = \exp(\beta_5 W + \beta_6)$$

■ Now *berlin* makes the best fits, which seems quite reasonable for $\mu(W)$, but more doubtful for $\sigma^2(W)$

■ *dane* fits almost as before
Y-STR: Haplotype Frequency Estimation and Evidence Calculation by Mikkel Meyer Andersen

Introduction
Estimating frequencies

Dimension reduction
Existing methods
New methods

Frequency surveying
Ancestral awareness
Classification models
Kernel smoothing

Comparing models
Calculating evidence

Plot of the regression

\[ \mu(W) = 0 + \exp(34.442 \times W - 12.972) \]

berlin divided into 16 groups
Y-STR: Haplotype Frequency Estimation and Evidence Calculation

by Mikkel Meyer Andersen

Introduction

Estimating frequencies

Dimension reduction

Existing methods

New methods

Frequency surveying

Ancestral awareness

Classification models

Kernel smoothing

Comparing models

Calculating evidence

Plot of the regression

\[ \sigma^2(W) = 0 + \exp(21.217 \times W - 13.662) \]
Y-STR: Haplotype Frequency Estimation and Evidence Calculation by Mikkel Meyer Andersen

Introduction
Estimating frequencies
Dimension reduction
Existing methods
New methods

Frequency surveying
Ancestral awareness
Classification models
Kernel smoothing

Comparing models
Calculating evidence

somali divided into 17 groups

\[ \mu(W) = 0 + \exp(14.328 \times W - 9.286) \]
Y-STR: Haplotype Frequency Estimation and Evidence Calculation by Mikkel Meyer Andersen

Introduction

Estimating frequencies

Dimension reduction
Existing methods
New methods
Frequency surveying
Ancestral awareness
Classification models
Kernel smoothing

Comparing models

Calculating evidence

somali divided into 17 groups

\[ \sigma^2(W) = 0 + \exp(10.29 \times W - 9.328) \]
At the 7th International Y Chromosome User Workshop in Berlin, 2010, Sascha Willuweit (one of the persons behind http://www.yhrd.org) mentioned a couple of changes between their implementation at http://www.yhrd.org and the original article:

- Using the reduced regression models, i.e. without intercepts $\beta_1$ and $\beta_4$
- The number of groups are determined by fitting several regressions and choosing the best one (details for selecting the minimum number of groups was not mentioned)
Comments and proposals for changes

- Not a statistical model, more an ad-hoc method
- \( \mu_i = \mu(W_i) = \exp(aW_i + b) \) is not bounded above such that \( \mu(W) \geq 1 \) for \( W \geq -\frac{b}{a} \): for \textit{berlin} \( a = 34.44 \) and \( b = -12.97 \) so \( \mu_i \geq 1 \) for \( W_i \geq 0.377 \) (\( 0 \leq W_i \leq 1 \) and \( 0 < \mu_i < 1 \) by definition)
- Fitting \( u_i \) and \( v_i \): only \( W_i \) to fit two exponential regression models
- The model is not consistent: generalisation to a Dirichlet prior and a multinomial likelihood might solve this
Fitting two parameters ($u_i$ and $v_i$) using only one ($W_i$)

Y-STR:
Haplotype
Frequency
Estimation
and Evidence
Calculation

by Mikkel
Meyer
Andersen

Introduction
Estimating frequencies
Dimension reduction
Existing methods
New methods
Frequency surveying
Ancestral awareness
Classification models
Kernel smoothing

Comparing models
Calculating evidence

Using regression without beta1 and beta4
Fitting two parameters \((u_i \text{ and } v_i)\) using only one \((W_i)\)

Y-STR:
Haplotype
Frequency
Estimation
and Evidence
Calculation

by Mikkel
Meyer
Andersen

Introduction

Estimating frequencies

Dimension reduction
Existing methods
New methods

Frequency surveying
Ancestral awareness
Classification models
Kernel smoothing

Comparing models

Calculating evidence
Fitting two parameters ($u_i$ and $v_i$) using only one ($W_i$)
Fitting two parameters \((u_i \text{ and } v_i)\) using only one \((W_i)\)

Y-STR:
Haplotype
Frequency
Estimation
and Evidence
Calculation

by Mikkel
Meyer
Andersen

Introduction

Estimating frequencies

Dimension reduction
Existing methods
New methods

Frequency surveying
Ancestral awareness
Classification models
Kernel smoothing

Comparing models
Calculating evidence
Y-STR: Haplotype Frequency Estimation and Evidence Calculation
by Mikkel Meyer Andersen

Introduction

- Estimating frequencies
- Dimension reduction
- Existing methods
- New methods

Frequency surveying
- Ancestral awareness
- Classification models
- Kernel smoothing

Comparing models
- Calculating evidence

Idea: use second moment

\[ W_i = \frac{1}{N - N_i} \sum_{i \neq j} \frac{N_j}{d_{ij}} = \frac{1}{N - N_i} \sum_{i \neq j} \frac{N_j}{\sum_{k=1}^{r} d_{ijk}} \]
uses first moment of the allele differences on each locus of \( r \) loci

- Maybe also use the second moment to introduce

\[ Z_i = \frac{1}{N - N_i} \sum_{i \neq j} \left( \frac{N_j}{\sum_{k=1}^{r} (d_{ijk} - \frac{d_{ij}}{r})^2} \right) \]

- Fit \( \mu_i \)'s and \( \sigma_i^2 \)'s by multiple regression using a grid of \( W_i \) and \( Z_i \) values

- 15 groups only correspond to a \( 4 \times 4 \)-grid, which is way too coarse – requiring 15 groups of \( W_i \)'s and \( Z_i \)'s, the grid would have size \( 15 \cdot 15 = 225 \): requires a lot of observations!

- Too few observations in \textit{berlin}, \textit{dane}, and \textit{somali}, but it would be interesting to see how it would perform compared to just using the \( W_i \)'s
Generalised frequency surveying

- Assume a priori that $f \sim \text{Dirichlet} (\alpha)$ where $f = (f_1, f_2, \ldots, f_K)$ is the vector of frequencies for all possible haplotypes
- Use the likelihood $N|f \sim \text{Multinomial} (N_+, f)$
- The posterior becomes
  $f|N \sim \text{Dirichlet} (\alpha_1 + N_1, \ldots, \alpha_K + N_K) = \text{Dirichlet} (\alpha_1 + N_1, \ldots, \alpha_n + N_n, \alpha_{n+1}, \ldots, \alpha_K)$ where $f_1, f_2, \ldots, f_n$ are the frequencies for the observed haplotypes and $f_{n+1}, f_{n+2}, \ldots, f_K$ are for the unobserved haplotypes
Marginal distribution

- Let \( \alpha_+ = \sum_{i=1}^{K} \alpha_i \)

- The marginal posterior distribution for the \( i \)'th haplotype is
  \[
  f_i | N_i \sim \text{Beta} \left( \alpha_i + N_i, \sum_{j=1}^{K} (\alpha_j + N_j) - (\alpha_i + N_i) \right) = \text{Beta} \left( \alpha_i + N_i, \alpha_+ - \alpha_i + N_+ - N_i \right)
  \]

- \( \mathbb{E} [f_i | N_i] = \frac{\alpha_i + N_i}{\sum_{j=1}^{K} (\alpha_j + N_j) - (\alpha_i + N_i) + (\alpha_i + N_i)} = \frac{\alpha_i + N_i}{\alpha_+ + N_+} \)

- \( \sum_{i=1}^{K} \mathbb{E} [f_i | N_i] = (\alpha_+ + N_+)^{-1} \sum_{i=1}^{K} (\alpha_i + N_i) = 1 \)

- Incorporate prior knowledge can be done by specifying the prior parameter \( \alpha_i \) for all possible haplotypes, but with this approach \( \alpha_+ \) might be problematic to calculate for large \( K \).
Y-STR: Haplotype Frequency Estimation and Evidence Calculation

by Mikkel Meyer Andersen

Introduction

Estimating frequencies

Dimension reduction

Existing methods

New methods

Frequency surveying

Ancestral awareness

Classification models

Kernel smoothing

Comparing models

Calculating evidence

Approximating $\alpha_+$

Histogram of Wi for Berlin

$\text{mean}(W) = 0.186$

Gamma: $s = 25.926, r = 139.146$

Beta: $s_1 = 21.296, s_2 = 93.016$
Y-STR: Haplotype Frequency Estimation and Evidence Calculation

by Mikkel Meyer Andersen

Introduction

Estimating frequencies

Dimension reduction

Existing methods

New methods

Frequency surveying

Ancestral awareness

Classification models

Kernel smoothing

Comparing models

Calculating evidence

Histogram of Wi for dane

\[
\text{mean}(W) = 0.153
\]

Gamma: \( s = 15.291, r = 100.14 \)

Beta: \( s_1 = 13.002, s_2 = 72.151 \)
Approximating $\alpha_+$

Y-STR: Haplotype Frequency Estimation and Evidence Calculation

by Mikkel Meyer Andersen

Introduction

Estimating frequencies

Dimension reduction

Existing methods

New methods

Frequency surveying

Ancestral awareness

Classification models

Kernel smoothing

Comparing models

Calculating evidence

Histogram of $W_i$ for somali

$\text{mean}(W) = 0.242$

$\text{gamma: s = 3.112, r = 12.84}$

$\text{beta: s1 = 2.44, s2 = 7.636}$
As earlier stated, $0 \leq W_i \leq 1$ so the Beta distribution is the right choice theoretically.

Assume that $\alpha_i = h(W_i)$ for some function $h$ such that $\alpha_+ = \sum_{i=1}^K h(W_i)$.

Denote by $f_\beta$ the density of a fitted Beta-distribution, then

$$
\alpha_+ \approx K \int_0^1 f_\beta(W) h(W) dW
$$
Approximating \( \alpha_+ \)

- To get equality between the first prior parameter in surveying and the generalised surveying, let
  \[
  \alpha_i = u_i = \frac{\mu_i^2 (1 - \mu_i)}{\sigma_i^2}
  \]

- Because \( \mu_i = \mu(W_i) = \exp(aW_i + b) \) can result in \( \mu_i \geq 1 \) then \( 1 - \mu_i \leq 0 \) such that \( \alpha_i \leq 0 \) which is now allowed

- For \textit{berlin}, \( \mu_i \geq 1 \) for \( W_i \geq 0.377 \) so that all contributions to the integral in the \( \alpha_+ \) approximation is negative for \( W_i \geq 0.377 \)

- \textit{berlin}: \( \alpha_+ = 50205.04 \) and the uncovered probability mass is estimated to \( 0.986 \)

- \textit{dane}: \( -b/a = 0.271 \) and \( \alpha_+ = -7897176 \)

- \textit{somali}: \( -b/a = 0.648 \) and \( \alpha_+ = -1535169 \)

- \( \alpha_i = u_i \) and the exponential regression is unusable, but the distribution of the \( W_i \)'s might be helpful for other choices
Problem

- Maybe getting too much attention because it is the only published method for estimating haplotype frequencies
- At the 7th International Y Chromosome User Workshop in Berlin, 2010, Michael Krawczak (one of the authors of the original articles) gave a talk where the associated slides included the statement "[frequency surveying has] never [been] thoroughly studied and validated"
Y-STR: Haplotype Frequency Estimation and Evidence Calculation

by Mikkel Meyer Andersen

Introduction

Estimating frequencies
Dimension reduction
Existing methods
New methods
Frequency surveying
Ancestral awareness
Classification models
Kernel smoothing

Comparing models
Calculating evidence

Ancestral awareness
The idea: approximate when identified a common (partial) ancestor

The basic idea is to find \( I = \{i_1, i_2, \ldots, i_q\} \subseteq \{1, 2, \ldots, r\} \) such that

\[
P \left( \bigcap_{j \notin I} L_j = a_j \bigg| \bigcap_{i \in I} L_i = a_i \right) \approx \prod_{j \notin I} P \left( L_j = a_j \bigg| \bigcap_{i \in I} L_i = a_i \right)
\]

is a good approximation.
Example

Assume that we have loci \( L_1, L_2, L_3, L_4 \) and \( I = \{1, 2\} \)

Then

\[
P(L_1, L_2, L_3, L_4) = P(L_1) P(L_2|L_1) P(L_3, L_4|L_1, L_2) \quad (1)
\]

\[
\approx P(L_1) P(L_2|L_1) \prod_{j=3}^{4} P(L_j|L_1, L_2) \quad (2)
\]
How to chose \( I \)

- \( I \) is called an ancestral set, because it can be interpreted as a set of alleles that is common with one’s ancestors.
- \( I \) can be found using a greedy approach adding the \( j \) to \( I \) that maximises \( P(L_j = a_j \mid \cap_{i \in I} L_i = a_i) \)
- Stop adding elements to \( I \), e.g. when only a percentage of the observations is left to use for calculating the marginal probabilities conditional on \( I \).
The approach is simple, but like it is not a statistical model
Introduction

Estimating frequencies
Dimension reduction
Existing methods
New methods
Frequency surveying
Ancestral awareness
Classification models
Kernel smoothing

Comparing models
Calculating evidence
Let $L_1, L_2, \ldots, L_r$ be the $r$ different loci available in the haplotype. Then fit

$$L_{i_1} \sim \sum_{k \notin \{i_1\}} L_k$$  \hspace{1cm} (3)

$$L_{i_2} \sim \sum_{k \notin \{i_1, i_2\}} L_k$$  \hspace{1cm} (4)

$$\vdots$$

$$L_{i_{r-2}} \sim \sum_{k \notin \{i_1, i_2, \ldots, i_{r-2}\}} L_k$$  \hspace{1cm} (6)

$$L_{i_{r-1}} \sim \sum_{k \notin \{i_1, i_2, \ldots, i_{r-1}\}} L_k = L_{i_r}$$  \hspace{1cm} (7)

and use the empirical distribution for $L_{i_r}$. 

The idea: alternately perceive different loci as a response variable
Properties

- A class of statistical models
- The classifications can be done with some of the several available classifications methods such as classification trees, ordered logistic regression, or support vector machines
- Selection of \( i_j \) should be done using standard model selection criteria depending on the classification model used
- Does not incorporate prior knowledge
Kernel smoothing and model based clustering

Y-STR:
Haplotype
Frequency
Estimation
and Evidence
Calculation

by Mikkel
Meyer
Andersen

Introduction
Estimating
frequencies
Dimension
reduction
Existing
methods
New methods
Frequency
surveying
Ancestral
awareness
Classification
models
Kernel
smoothing

Comparing
models
Calculating
evidence
The idea: create a density around each haplotype

- Put a scaled density/mass (called a kernel) around each haplotype with mass equal to its relative frequency \( \frac{N_i}{N} \)
- In this way unobserved haplotypes get probability mass from the (near) neighbours
Choice of kernel

- A straightforward approach is the Gaussian kernel
  \[ K(z|x_i, \lambda) = \left(2\pi\lambda^2\right)^{-r/2} \det(\Sigma)^{-1/2} \exp \left( -\frac{1}{2\lambda^2} (x_i - z)\Sigma^{-1}(x_i - z)^\top \right) \]
  where \( \lambda \) is called a smoothing parameter that has to be chosen.

- A frequency estimate for any given haplotype \( z \) is
  \[ g(z) = \frac{1}{N_+} \sum_{i=1}^n N_i K(z|x_i, \lambda) \]

- To incorporate prior knowledge, \( K(z|x_i, N_i, \lambda) = \left(2\pi\frac{\lambda^2}{N_i}\right)^{-r/2} \det(\Sigma)^{-1/2} \exp \left( -\frac{1}{2\frac{\lambda^2}{N_i}} (x_i - z)\Sigma^{-1}(x_i - z)^\top \right) \)
  could be used.
Problem

- The model can be inaccurate if the kernel has small variance, because then the actual mass when evaluated over the discrete grid can differ greatly from the relative frequencies
- Discrete kernels could be tried instead, e.g. the multinomial
Model based clustering

- Estimating a frequency for a haplotype using kernel smoothing require evaluating as many densities as the number of haplotypes in the database.
- Model based clustering can be used to perform clustering first to minimise the required number of density evaluations.
- Same problem as with kernel smoothing if the variances are too small.
Comparing models

- Model verification is as always crucial
- One important feature of a model is to be able to efficiently obtain further samples of haplotypes according to their probability under a model
Different ways of comparing models

- Estimated unobserved probability mass
- Marginal deviances (for a model’s single and pairwise compared to observed)
- Several more should be definitely considered
Unobserved probability mass: point estimate

- $K_0$: the number of singletons
- $N$: the number of observations

In 1968, Robbins showed that

$$V = \frac{K_0}{N + 1}$$

is an unbiased estimate of the unobserved probability mass.
Unobserved probability mass: limiting consistent variance estimate

- \( K_1 \): the number of doubletons
- In 1986, Bickel and Yahav showed that under some regularity conditions,

\[
\hat{\sigma}^2 = \frac{K_0}{N^2} - \frac{(K_0 - 2K_1)^2}{N^3}
\]

is limiting consistent estimate of the variance of the unobserved probability mass

- Both \( \hat{\sigma} \) and the variance estimate can be verified by simulation
Unobserved probability mass: simulation study

Y-STR:
Haplotype
Frequency
Estimation
and Evidence
Calculation
by Mikkel
Meyer
Andersen

Introduction
Estimating
frequencies
Comparing
models
Unobserved
probability
mass
Marginal
deviances
Calculating
evidence
Further work
Questions

Confidence interval coverage for true population size $Q = 10000$

Based on 10000 simulations and 10–90 probabilities. Line at 0.95.

- Symmetric standard confidence interval
- Logit–transformed confidence interval
Unobserved probability mass

- The estimate seems like a good and simply way of perform model verification, but it cannot stand alone as we shall soon see
- It can also be used to fit model parameters, e.g. the smoothing parameter in the kernel smoothing model
### Unobserved probability mass: approximate confidence intervals

<table>
<thead>
<tr>
<th></th>
<th>berlin</th>
<th>dane</th>
<th>somali</th>
</tr>
</thead>
<tbody>
<tr>
<td>( V )</td>
<td>0.364</td>
<td>0.602</td>
<td>0.277</td>
</tr>
<tr>
<td>( \hat{V} )</td>
<td>0.061</td>
<td>0.078</td>
<td>0.120</td>
</tr>
<tr>
<td>95% conf. int.</td>
<td>[0.321; 0.408]</td>
<td>[0.510; 0.694]</td>
<td>[0.212; 0.342]</td>
</tr>
<tr>
<td>( S: \beta_1/\beta_4 \neq 0 )</td>
<td>NA</td>
<td>0.643</td>
<td>NA</td>
</tr>
<tr>
<td>( S: \beta_1/\beta_4 = 0 )</td>
<td>0.479</td>
<td>0.703</td>
<td>0.335</td>
</tr>
<tr>
<td>rpart</td>
<td>0.478</td>
<td>0.71</td>
<td>0.42</td>
</tr>
<tr>
<td>svm</td>
<td>0.526</td>
<td>0.792</td>
<td>0.43</td>
</tr>
<tr>
<td>polr</td>
<td>0.639</td>
<td>0.886</td>
<td>NA</td>
</tr>
<tr>
<td>Ancestor: 10%</td>
<td>0.39</td>
<td>0.589</td>
<td>0.182</td>
</tr>
<tr>
<td>Ancestor: 15%</td>
<td>0.454</td>
<td>0.668</td>
<td>0.22</td>
</tr>
<tr>
<td>Ancestor: 20%</td>
<td>0.466</td>
<td>0.713</td>
<td>0.246</td>
</tr>
</tbody>
</table>
Marginal deviances

- Depending on the model, exact marginals can be difficult to obtain.
- If haplotypes can be sampled according to their probability under a model, then marginals can be approximated by simulating a huge number of haplotypes under that model.
- Using only the observed marginals should correspond to this, but – at least for small databases – this is not the case according to simulations studies performed with the classification models.
Deviance for pairwise marginals

- Let \( \{ u \}_{ij} \) be the two-way table with the observation counts, \( \{ \tilde{p} \}_{ij} \) the table of predicted probabilities under a model \( M_0 \), \( \{ \hat{p} \}_{ij} \) the relative probabilities such that \( \hat{p}_{ij} = \frac{u_{ij}}{u_{++}} \).

- For the pairwise marginal tables, the deviance is
  \[
  d = -2 \log \left( \frac{L(\{ \tilde{p} \}_{ij})}{L(\{ \hat{p} \}_{ij})} \right)
  \]
  where \( L(\{ p \}_{ij}) = \prod_{i,j} p_{ij}^{u_{ij}} \) is proportional to the likelihood (the constant \( \frac{u_{++}!}{\prod_{i,j} u_{ij}} \) is cancelled out in the fraction).

- Then \( d = -2 \sum_{i,j} u_{ij} \log \left( \frac{\tilde{p}_{ij}}{\hat{p}_{ij}} \right) \sim \chi^2_{\nu} \)
Comparing models

- A deviance is calculated for each pair of loci
- To compare models these deviances can summed and used for relative comparisons (the sum is not $\chi^2$ distributed)
- The deviance is calculated similar for single marginals
### Deviance sums for single marginals

<table>
<thead>
<tr>
<th></th>
<th>berlin</th>
<th>dane</th>
<th>somali</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpart</td>
<td>1.236</td>
<td>1.449</td>
<td>0.268</td>
</tr>
<tr>
<td>svm</td>
<td>13434.632</td>
<td>2363.861</td>
<td>5664.264</td>
</tr>
<tr>
<td>polr</td>
<td>3.114</td>
<td>4.623</td>
<td>NA</td>
</tr>
</tbody>
</table>

Sum of deviances for observed single marginals vs. simulated single marginals for the classification method specified in each row.
Deviance sums for pairwise marginals

<table>
<thead>
<tr>
<th></th>
<th>berlin</th>
<th>dane</th>
<th>somali</th>
</tr>
</thead>
<tbody>
<tr>
<td>rpart</td>
<td>Inf</td>
<td>768.444</td>
<td>772.936</td>
</tr>
<tr>
<td>svm</td>
<td>Inf</td>
<td>24971.264</td>
<td>53797.852</td>
</tr>
<tr>
<td>polr</td>
<td>1761.153</td>
<td>Inf</td>
<td>NA</td>
</tr>
</tbody>
</table>

Sum of deviances for observed pairwise marginals vs. simulated pairwise marginals for the classification method specified in each row.
Calculating evidence

Y-STR:
Haplotype
Frequency
Estimation
and Evidence
Calculation

by Mikkel
Meyer
Andersen

Introduction
Estimating
frequencies
Comparing
models
Calculating
evidence
Two
contributors
n contributors
Further work
Questions
The purpose is to get an unbiased opinion in a trial

Often formulated as the hypothesis

\[ H_p : \text{The suspect left the crime stain together with } n \text{ additional contributors.} \]

\[ H_d : \text{Some other person left the crime stain together with } n \text{ additional contributors.} \]

Then the likelihood ratio given by \( LR = \frac{P(E|H_p)}{P(E|H_d)} \) is calculated.

In a courtroom it can then be stated that the evidence \( E \) is \( LR \) times more likely to have arisen under \( H_p \) than under \( H_d \) (formulation is from "Interpreting DNA Mixtures" by Weir et al., 1997).
The actual evaluation of $LR$ can be a computation heavy task: the number of combinations giving rise to the same trace grows with the number of contributors.
Assume

\(H_p\) : The suspect left the crime stain together with one additional contributors.

\(H_d\) : Some other person left the crime stain together with one additional contributors.
Two contributors: notation

Let \( a = (a_1, a_2, \ldots, a_n) \) and \( b = (b_1, b_2, \ldots, b_n) \) and define

\[
T = a \oplus b = (\{a_1, b_1\}, \{a_2, b_2\}, \ldots, \{a_1, b_n\}) \\
T \ominus a = (\{b_1\}, \{b_2\}, \ldots, \{b_n\})
\] (8) (9)

where the sets are multisets.
Two contributors

Let $T$ be the trace, $h_s$ the suspect’s haplotype, and $h_1$ the additional contributor’s haplotype. Then $T = h_s \oplus h_1$ and $h_1 = T \ominus h_s$. Say that $(h_1, h_2)$ is consistent with the trace $T$ if $h_1 \oplus h_2 = T$, which is denoted $(h_1, h_2) \equiv T$. This makes

$$LR = \frac{P(E|H_p)}{P(E|H_d)} \quad \text{(10)}$$

$$= \frac{P(h_s, T \ominus h_s)}{\sum_{(h_1, h_2) \equiv T} P(h_s, h_1, h_2)} \quad \text{(11)}$$

$$= \frac{P(h_s) P(T \ominus h_s)}{P(h_s) \sum_{(h_1, h_2) \equiv T} P(h_1) P(h_2)} \quad \text{(12)}$$

$$= \frac{P(T \ominus h_s)}{\sum_{(h_1, h_2) \equiv T} P(h_1) P(h_2)} \quad \text{(13)}$$

by assuming that haplotypes are independent.
Let \( T = (T_1, T_2, \ldots, T_r) \), \( T_i \) is a set of alleles such that \(|T_i| \in \{1, 2\}\)

Let \( \mathcal{H}_T = T_1 \times T_2 \times \cdots \times T_r \)

In the non-trivial case a \( j \in \{1, 2, \ldots, r\} \) exists such that \( T_j = \{a_1, a_2\} \) with \( a_1 \neq a_2 \)

Let \( T'_j = \{a_1\} \) (such that one of the alleles is removed) and

\[
\mathcal{H}'_T = T_1 \times \cdots \times T_{j-1} \times T'_j \times T_{j+1} \times \cdots \times T_r
\]
Now the denominator of LR can be written as
\[ 2 \sum_{h_1 \in \mathcal{H}'_T} P(h_1) P(T \ominus h_1) \]

If \( k \) denotes the number of loci in the trace with only one allele, and we assume that we have the non-trivial case with \( 0 \leq k < r \), we have that \(|\mathcal{H}_T| = \prod_{i=1}^{r} |T_i| = 2^{r-k}\) such that \(|\mathcal{H}'_T| = \frac{|\mathcal{H}_T|}{2} = 2^{r-k-1} \leq 2^{r-1}\)

This means that for \( r \) loci, a maximum of \( 2 \cdot 2^{r-1} = 2^r \) haplotype frequencies have to be calculated, e.g. \( 2^{10} = 1024 \)

If a trace has two contributors with no known suspects, the two most likely contributors can be chosen to be
\[ \arg \max_{h_1 \in \mathcal{H}'_T} P(h_1) P(T \ominus h_1) \]
Y-STR: Haplotype Frequency Estimation and Evidence Calculation by Mikkel Meyer Andersen

Introduction

Estimating frequencies

Comparing models

Calculating evidence

Two contributors

$n$ contributors: formulation of the $LR$ for one locus

- $LR$ defined generally in "Forensic interpretation of Y-chromosomal DNA mixtures" by Wolf et al., 2005

- At a given locus, let $E_t$, $E_s$, and $E_k$ be the set of alleles from the trace, the suspect, and the known contributors, respectively

- Assume $n$ unknown contributors and let $A_n$ denote the set of alleles carried by the these $n$ unknown contributors

- Let $P_n (V; W) = P (W \subseteq A_n \subseteq V)$

- Then

$$LR = \frac{P_n (E_t; E_t \setminus (E_s \cup E_k))}{P_{n+1} (E_t; E_t \setminus E_k)} = \frac{P (E_t \setminus (E_s \cup E_k) \subseteq A_n \subseteq E_t)}{P (E_t \setminus E_k \subseteq A_{n+1} \subseteq E_t)}$$

(14)
For $m$ loci we have:

$$LR = \frac{P_n \left( \bigcap_{i=1}^{m} \{ E_{t,i}; E_{t,i} \setminus (E_{s,i} \cup E_{k,i}) \} \right)}{P_{n+1} \left( \bigcap_{i=1}^{m} \{ E_{t,i}; E_{t,i} \setminus E_{k,i} \} \right)}$$
Let $E_{t,1} = \{1, 2\}$, $E_{t,2} = \{2\}$, $E_{s,1} = \{1\}$, $E_{s,2} = \{2\}$, $E_{k,1} = E_{k,2} = \emptyset$. Then

$$LR = \frac{P_1 \left( \bigcap_{i=1}^{2} \{E_{t,i}; E_{t,i} \setminus (E_{s,i} \cup E_{k,i})\} \right)}{P_2 \left( \bigcap_{i=1}^{2} \{E_{t,i}; E_{t,i} \setminus E_{k,i}\} \right)}$$

$$= \frac{P \left( \bigcap_{i=1}^{2} \{E_{t,i} \setminus (E_{s,i} \cup E_{k,i}) \subseteq A_i^1 \subseteq E_{t,i}\} \right)}{P \left( \bigcap_{i=1}^{2} \{E_{t,i} \setminus E_{k,i} \subseteq A_i^2 \subseteq E_{t,i}\} \right)}$$

$$= \frac{P \left( \{E_{t,1} \setminus E_{s,1} \subseteq A_1^1 \subseteq E_{t,1}\} \cap \{E_{t,2} \setminus E_{s,2} \subseteq A_1^2 \subseteq E_{t,2}\} \right)}{P \left( \{E_{t,1} \setminus E_{k,1} \subseteq A_2^1 \subseteq E_{t,1}\} \cap \{E_{t,2} \setminus E_{k,2} \subseteq A_2^2 \subseteq E_{t,2}\} \right)}$$

$$= \frac{P \left( \{\{2\} \subseteq A_1^1 \subseteq \{1, 2\} \} \cap \{A_2^1 \subseteq \{1, 2\}\} \right)}{P \left( \{\{1, 2\} \subseteq A_2^1 \subseteq \{1, 2\} \} \cap \{\{2\} \subseteq A_2^2 \subseteq \{2\} \} \right)}$$

$$= \frac{P \left( \{2\} \cap \{2\} \right)}{P \left( \{1, 2\} \cap \{2, 2\} \right)}$$

$$= \frac{P (h_1 = (2, 2))}{P (h_1 = (1, 2), h_2 = (2, 2)) + P (h_1 = (2, 2), h_2 = (1, 2))}$$
It is complicated

One key ingredient in calculating the LR is to be able to estimate frequencies for unobserved haplotypes

The next step is to be able to calculate the LR efficiently even for a large number of contributors
Further work
Further work

■ Estimating Y-STR haplotype frequencies
  ▶ Better incorporation of prior knowledge in a statistical model, e.g. graphical models with other test statistics (ordinal data and incorporating prior knowledge such as the single step mutation model)
  ▶ More and better ways to verify models
  ▶ Larger datasets (http://www.yhrd.org has gathered a lot of data, both publicly available in journals and directly from laboratories, but none is available for others, yet)

■ Y-STR Mixtures
  ▶ Efficient calculation of LR
  ▶ Use quantitative information (the amount of DNA material which can be seen in the EPG) instead of only the qualitative

■ Model the signal in the EPG (electropherogram)
Questions?