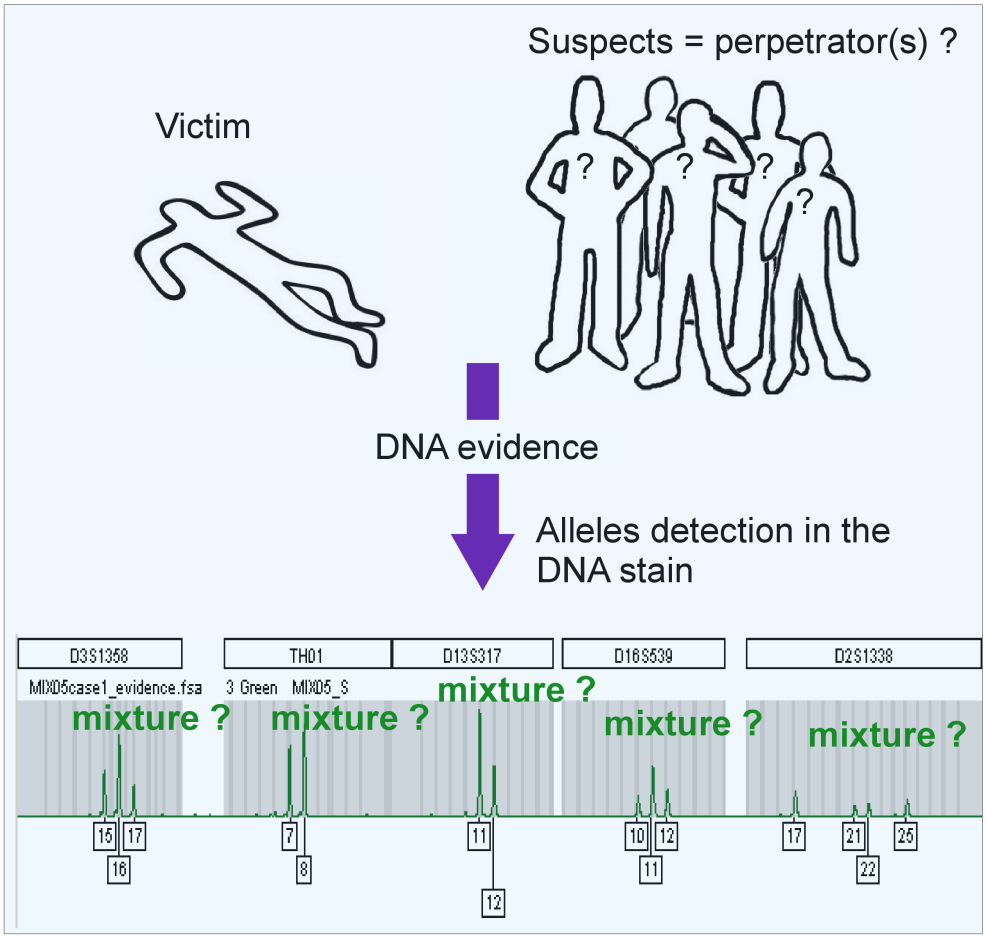
forensim: a freeware initiative for statistical methods' evaluation in forensic genetics

Hinda Haned

Laboratory of Biometry and Evolutionary Biology, University of Lyon, France

November 25th 2009

Forensic DNA mixtures : A challenging task



Interpretation issues

- Is it a mixture ?
- How many people involved ?
- Weight of the stain as an evidence ?

Available methods

- Several methods dedicated to mixtures interpretation are available :

LR in case of population substructure	Curran *et al* 1999
Number of contributors	Egeland *et al* 2003
Unknown related contributors	Fung and Hu 2003
Genotyping errors	Thompson *et al* 2003

$\Rightarrow$ **Lack of evaluation of methods' efficiency and robustness**

How to evaluate these methods ?

On simulated DNA stains where the circumstances of the hypothetical crime are known by the experimenter.

The experimenter would evaluate method's efficiency :

1. While varying accurate parameters :
   - type of markers analyzed
   - number of markers analyzed
   - number of contributors to the DNA evidence
2. In critical situations :
   - population subdivision (co-ancestry)
   - partial profiles
   - relatedness between contributors to the DNA stain
   - allele dropout

How to evaluate these methods ?

- **Laboratory simulated DNA stains :**
  - Some scenarios are hard to test in laboratory (ex. population substructure)
  - Cost issues : new experiments are to be conducted for each tested scenario

- Computer simulated DNA stains :
  - Complex scenarios can be simulated
  - No cost issues

Currently, there is no free software providing simulation tools specific to forensic genetics.

How to evaluate these methods ?

- Laboratory simulated DNA stains :
  - Some scenarios are hard to test in laboratory (ex. population substructure)
  - Cost issues : new experiments are to be conducted for each tested scenario

- **Computer simulated DNA stains :**
  - Complex scenarios can be simulated
  - No cost issues

Currently, there is no free software providing simulation tools specific to forensic genetics.

The *forensim* package

Main features

- *forensim* is a package for the  statistical software
- *forensim* is freely available
- Relies on object oriented programming
- Sources freely available on 

- Compiles and runs on a wide variety of UNIX platforms, Windows and MacOS

forensim's structure

Simulation tools

Simulation of data commonly encountered in forensic casework

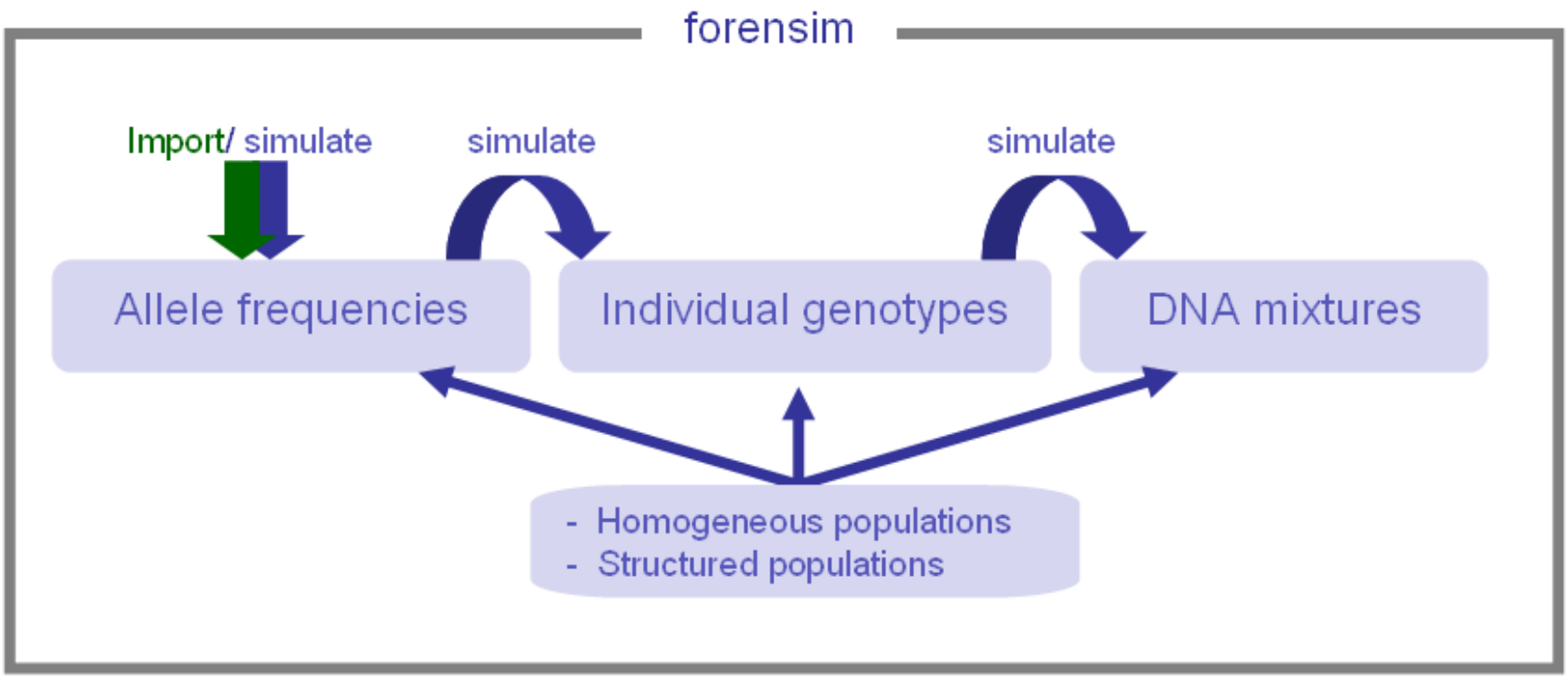
Statistical tools

Main statistical methods for forensic DNA evidence interpretation

Simulation tools

Object oriented programming

Structured code, can easily be modified/ enriched ⇒ allows a wide variety of scenarios



Statistical tools

Statistical methods usually used to report the weight of a DNA evidence are implemented :

Random man exclusion probability

- θ correction for allele dependencies Weir, In Buckelton *et al*, 2005

Likelihood ratios

- General formula for likelihood ratios Curran *et al*, 1999

Random match probabilities

- Accounts for :
  - relatedness
  - allele dependencies

Balding & Nichols, 1994

Simulation tools : Focus on DNA mixtures

Two kinds of information stored :

Usual information

- Alleles present in the stain
- Marker names
- Allele frequencies of the putative population

Simulation-related information

- Number of individuals involved
- Contributors' genotypes
- Contributors' populations

Simulating forensic DNA mixtures

Simulating a 3-person mixture, using the African American allele frequencies (Butler *et al*, 2003) :

Step1 : load the package

```
> library(forensim)

   ### forensim 1.1.2 is loaded ###
```

Simulating forensic DNA mixtures

Simulating a 3-person mixture, using the African American allele frequencies (Butler *et al*, 2003) :

Step1 : load the package

```
> library(forensim)

   ### forensim 1.1.2 is loaded ###
```

Step2 : generate the data

```
> data(strusa)
> geno <- simugeno(strusa, n = c(100, 0, 0))
> mix3 <- simumix(geno, ncontri = c(3, 0, 0))
```

Simulating forensic DNA mixtures

Mixture representation in forensim

```
> mix3

   # Simumix object: simulated mixture #

@which.loc: vector of  15  locus names
@ncontri: 3
@mix.prof:  3 x 15 data frame of the contributors genotypes
@mix.all: list of the alleles found in the mixture
@popinfo: populations of the contributors
```

Simulating forensic DNA mixtures

Mixture representation in forensim

```
> mix3

   # Simumix object: simulated mixture #

@which.loc: vector of  15  locus names
@ncontri: 3
@mix.prof:  3 x 15 data frame of the contributors genotypes
@mix.all: list of the alleles found in the mixture
@popinfo: populations of the contributors
```

Display stain profiles at locus FGA

```
> mix3$mix.all$FGA

[1] "20" "21" "24" "25"
```

Simulating forensic DNA mixtures

Mixture representation in forensim

```
> mix3

   # Simumix object: simulated mixture #

@which.loc: vector of  15  locus names
@ncontri: 3
@mix.prof:  3 x 15 data frame of the contributors genotypes
@mix.all: list of the alleles found in the mixture
@popinfo: populations of the contributors
```

Display stain profiles at locus FGA

```
> mix3$mix.all$FGA

[1] "20" "21" "24" "25"
```

Display contributors profiles at locus FGA

```
> mix3$mix.prof[, "FGA"]

  ind70   ind58    ind1
"21/24" "24/25" "21/20"
```

Reporting the weight of the evidence

What is the exclusion probability of the DNA evidence ?

```
>PE(mix3, freq = strusa, refpop = "Afri", theta = 0, byloc =FALSE)

       PE
0.999989
```

Reporting the weight of the evidence

What is the exclusion probability of the DNA evidence ?

```
>PE(mix3, freq = strusa, refpop = "Afri", theta = 0, byloc =FALSE)

       PE
0.999989
```

Help page

- mix : the DNA mixture
- freq : the allele frequencies to use
- refpop : the reference population, used only if freq contains allele frequencies for multiple populations
- theta : θ correction for allele dependencies
- byloc : logical indicating whether the PE is computed by/overall loci

Reporting the weight of the evidence

By locus exclusion probability

```
> PE(mix3, freq = strusa, refpop = "Afri", byloc = TRUE)

           PE_1
CSF1PO  0.6315
FGA     0.6320
TH01    0.4140
TPOX    0.2629
VWA     0.1739
D3S1358 0.2893
D5S818  0.2018
D7S820  0.2259
D8S1179 0.6082
D13S317 0.1739
D16S539 0.4404
D18S51  0.5828
D21S11  0.5426
D2S1338 0.6339
D19S433 0.8437
```

Determining the number of contributors to a DNA mixture

- In many situations, scarce data is available about the origin of the stain
  - No available suspect
  - Unknown contributors
  - Scarce non genetic evidence

An estimate of the number of contributors can help the investigators !

Determining the number of contributors to a DNA mixture

- A common laboratory practice : the number of contributors set to the minimum required to explain the profiles (maximum allele count)

An alternative approach :

- A maximum-likelihood estimator of the number of contributors to a forensic DNA mixture

Egeland *et al.* Estimating the number of contributors to a DNA profile. Int J Legal Med 2003 ;117(5) : 271-5.

The maximum likelihood approach

- Let A be a specific locus with alleles $A_1, ..., A_k$ with frequencies $p_1, ..., p_k$ in a given population.
- Crime scene profiles : A_1 and A_2 .

What is the likelihood of these profiles, if there were two contributors supplying these alleles ?

The maximum likelihood approach

7 genotype pairs are possible :

(A_1A_1, A_2A_2)	(A_2A_2, A_1A_1)	(A_1A_1, A_1A_2)
(A_2A_2, A_1A_2)	(A_1A_2, A_1A_1)	(A_1A_2, A_1A_2)
(A_1A_2, A_2A_2)		

Assuming the independence of alleles between and within individuals :

$$Pr(A_1A_1) = p_1^2 \text{ and } Pr(A_1A_2) = 2p_1p_2$$

$$Pr(A_1A_1, A_1A_2) = Pr(A_1A_1) \times Pr(A_1A_2)$$

Adding the genotype probabilities for all 7 genotype pairs

$$L_A(x = 2) = 4p_1^3p_2 + 6p_1^2p_2^2 + 4p_1p_2^3$$

The likelihood function

- Generalization :
  - Multiallelic loci
  - Allele dependencies due to population subdivision
- Automation :
  - Inspired from the general formula for likelihood ratios from Curran *et al.* (1999)

$$L_A(x) = \sum_{r_1=0}^{r} \sum_{r_2=0}^{r} \cdots \sum_{r_{c-1}}^{r-r_1-r_2-\ldots-r_{c-2}} \frac{(2x)!}{\prod_{i=1}^{c} u_i!} \times \frac{\prod_{i=1}^{c} \prod_{j=0}^{u_i-1} [(1-\theta)p_i + j\theta]}{\prod_{j=0}^{2x-1} [(1-\theta) + j\theta]}$$

Haned *et al* Estimating the number of contributors to forensic DNA mixtures : Does maximum likelihood perform better than maximum allele count ? *J. Forensic. Sci., accepted (2010)*

Maximum likelihood estimation

The maximum likelihood estimation of x, when a single marker A is considered, verifies :

$$\max_{j=1,2,3,\ldots} L_A(x = j)$$

When multiple loci are considered simultaneously :

$$\max_{j=1,2,3\ldots} \prod_A L_A(x = j)$$

Methods' evaluation

Does maximum likelihood perform better then maximum allele count ?

Implementation

Maximum allele count

```
>mincontri(mix3)

[1] 3
```

Maximum likelihood

```
>likestim(mix = mix3, freq = strusa, refpop = "Afri"', theta = 0)

max   maxval
3     2.6e-26
```

Implementation

Maximum allele count

```
>mincontri(mix3)

[1] 3
```

Maximum likelihood

```
>likestim(mix = mix3, freq = strusa, refpop = "Afri"', theta = 0)

max   maxval
3     2.6e-26
```

Help page

- mix : the DNA mixture
- freq : the allele frequencies to use
- refpop : the reference population, used only if freq contains allele frequencies for multiple populations
- theta : θ correction for allele dependencies

Methods' evaluation procedure

- 1000 DNA stains comprising x contributors, x=1,...,5.

```
> Mix2<-replicate(1000,simumix(geno, ncontri = c(2, 0, 0)))
```

- For each mixture : an error is scored if the value of x that maximizes the likelihood is different from the true number of contributors.

```
> res<-sapply(Mix2,likestim,strusa,"Afri")
```

Mixture simulated with African American allele frequencies

- DNA stains comprising 1 to 5 individuals belonging to the same population : African Americans

x	1	2	3	4	5
Max. Likelihood	1	1	0.94	0.79	0.67
Max. All. count.	1	1	0.99	0.45	0.05

Other situations can be investigated

More functionalities available via other packages :

- Basic statistical inference
- Bayesian inference
- Familial analysis
- Population genetics

How to get help

You are not familiar with  :

Do not worry ! A detailed tutorial with practical and reproducible examples is available online :
http://forensim.r-forge.r-project.org/

You are encountering problems using *forensim* :

- Post a message on *forensim* mailing list :
  forensim-help@lists.r-forge.r-project.org.
- Contact me :
  haned@biomserv.univ-lyon1.fr

Contributions are greatly encouraged !

forensim is evolving, and you can participate !

- Suggestions ?
- Particular needs ?
- Contributions to the package : data, methods... are welcome !

`http://forensim.r-forge.r-project.org/`