Estimating the number of contributors to forensic DNA mixtures: Does maximum likelihood perform better than maximum allele count ?

Hinda Haned*, Dominique Pontier*, Jean R. Lobry *, Laurent Pène*, Anne B. Dufour*

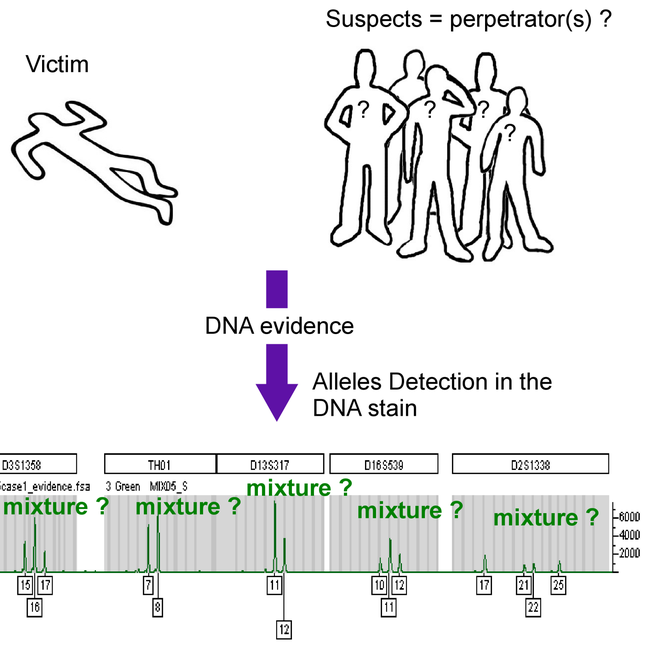
* Université de Lyon, France ⋆ Institut National de Police Scientifique, Laboratoire de Lyon.

Abstract

We introduce a new likelihood-based estimator of the number of contributors to forensic DNA mixtures and compare it to maximum allele count in two situations relevant to forensic casework: population subdivision and partial profiles. Our simulation investigations revealed that our method is the most efficient and the most robust for complex mixtures resolution.

Background

DNA stains composed of a mixture of several people's DNA are commonly encountered in forensic studies.



Determining the number of contributors

Most forensic laboratories use the **maximum allele count method** consisting on setting the number of contributors to the **minimum required to explain the observed profiles**.
Several methods were proposed to bound the number of contributors to a mixture, however, these methods do not make use of the available data, except for the **maximum observed number of alleles overall loci** [1].
Previous studies showed that maximum allele count gives biased estimates which are likely to increase in case of population subdivision [2, 3].

Our contribution

We propose a maximum-likelihood estimator that makes explicit use of allele frequencies [4]. Uncertainty about subpopulation allele frequencies in case of population subdivision is taken into account in the estimator through Wright's Fst coefficient, often called the coancestry coefficient θ in forensic casework. Our methodology is freely available in the package *forensim* for the R statistical software.

Open discussion

Our main results:

- Maximum likelihood performs better than maximum allele count for complex mixtures of more than two individuals.
- Maximum likelihood is more robust to loci loss than maximum allele count for four- and five-person mixtures.
- In case of subdivision, maximizing the likelihood is the most efficient method.

Our perspectives:

Our estimator takes into account qualitative information consisting of the allele types present in the stain, further work could thus include the modification of the likelihood estimator to allow the use of allele peak heights or areas information.

References

[1] Egeland T, Dalen I, Mostad PF. Estimating the number of contributors to a DNA profile. Int J Legal Med 2003;117:271-5.

[2] Paoletti DR, Doom TE, Krane, CM, Raymer ML, Krane DE. Empirical analysis of the STR profiles resulting from conceptual mixtures. J Forensic Sci 2005;50(6):1361-6.

[3] Buckleton JS, Curran JM, Gill P. Towards understanding the effect of uncertainty in the number of contributors to DNA stains. Forensic Sci Int: Genetics 2007;1:20-8.

[4] Haned H, Pontier D, Lobry JR, Pène L, Dufour AB. Estimating the number of contributors to forensic DNA mixtures: Does maximizing the likelihood performs better than the maximum allele count ? Submitted, 2009.

[5] Curran JM, Triggs CM, Buckleton J, Weir BS. Interpreting DNA Mixtures in Structured Populations. Int J Forensic Sci 1999;44(5):987-95.

A maximum likelihood method for mixtures resolution

The likelihood function. The likelihood of having x individuals giving the alleles observed at a locus A, in the case of all individuals belonging to the same subpopulation, is given by the general formula:

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

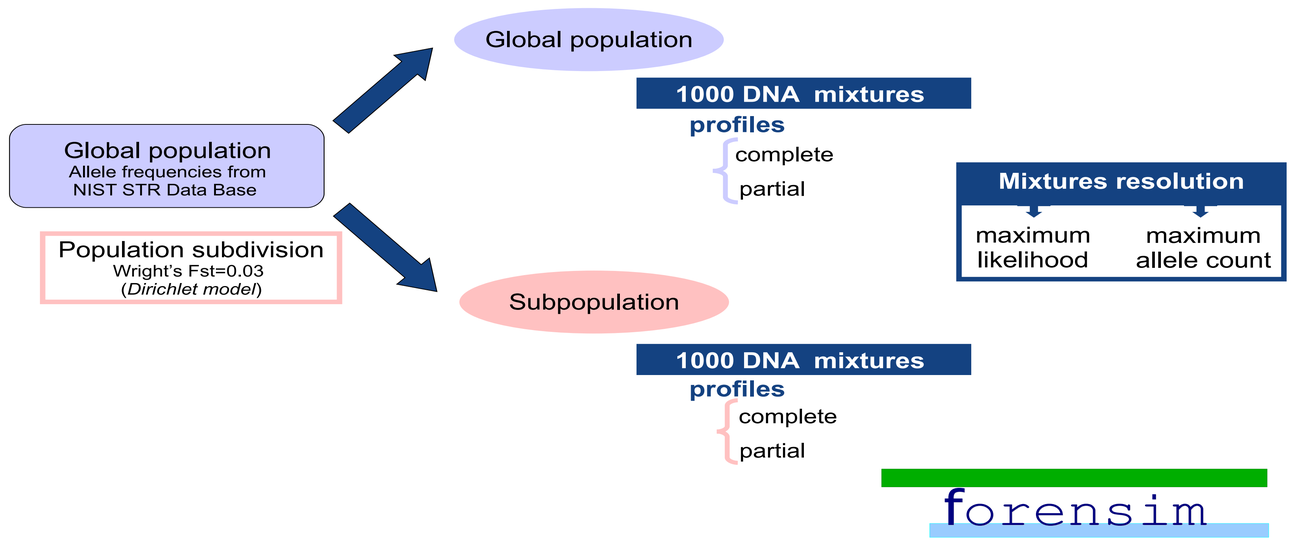
We use the same notations as Curran *et al* [5]:
x: The unknown number of contributors to the mixture
c: The distinct number of alleles observed in the DNA stain
p_i: The frequency of allele A_i
r: The number of unconstrained alleles, $r = 2x - c$
r_i: The unknown number of copies of allele A_i among the r unconstrained alleles of the stain
u_i: The unknown number of copies of allele A_i in the stain , with $\sum_{i=1}^{c} u_i = 2x$ and $u_i = r_i + 1$
θ: Wright's Fst coefficient

The likelihood estimator. Maximum likelihood estimation of x, when a single marker A is considered, verifies:

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

Maximum likelihood *vs* Maximum allele count

Procedure



Simulations were computed using functions from the *forensim* package for the free R statistical software. *forensim* can be downloaded at the following address: `http://forensim.r-forge.r-project.org/`.

Results

Estimators' efficiency: complete profiles case

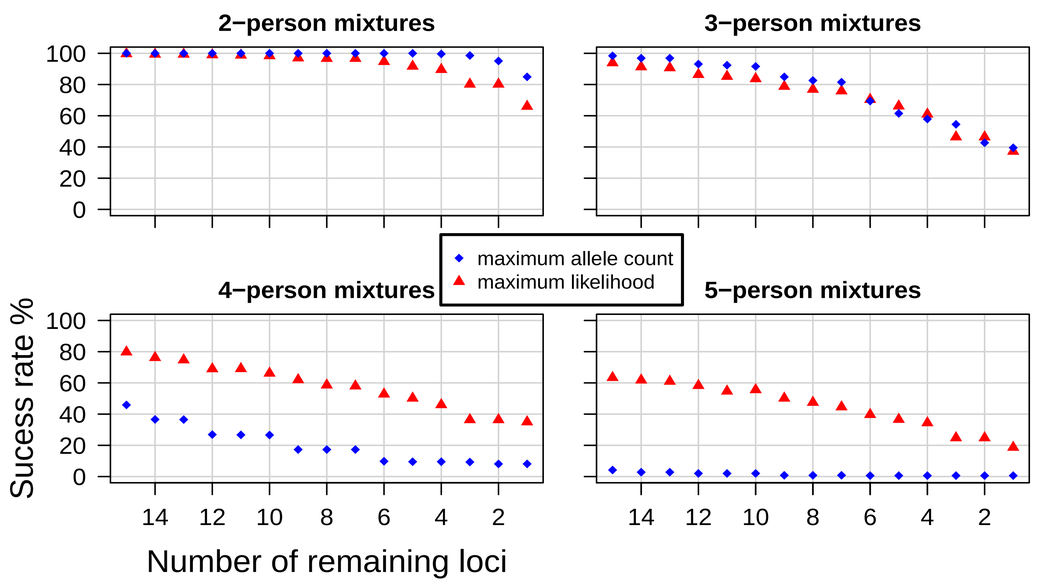
The percentages of correctly identified mixtures of 2 to 5 individuals were investigated in case of **homogeneous populations**: all mixtures' contributors belong to the same population ($\theta = 0$), and in case of **subdivided population**: all contributors belong to the same subpopulation, with a coancestry coefficient of $\theta = 0.03$.

Homogeneous population		
x	Maximum Allele count	Maximum likelihood
2	100%	100%
3	99%	94%
4	45%	79%
5	5%	67%

Subdivided population		
x	Maximum Allele count	Maximum likelihood
2	100%	99%
3	94%	91%
4	21%	76%
5	0.7%	60%

Estimators' efficiency: partial profiles case

The percentages of correctly identified mixtures of 2 to 5 individuals belonging to the same homogeneous population, in case of **partial profiles**. Loci markers were successively removed according to their alleles' expected median length. This corresponds to what happens in the case of a degraded DNA sample: longer DNA fragments drop out first. Similar results were obtained in case of population subdivision (results not shown).



Both estimators' precision decreases with the number of available loci. But, maximum likelihood was more robust to partial profiles for complex mixtures of more than three individuals.