# Deconvolution of MPS STR DNA Mixtures

*Søren B. Vilsen*

*2017-11-03*

## Contents



## 1 Introduction

The central equation to statisticians in forensic genetics is the posterior odds of two competing hypotheses:

$$\frac{\mathbb{P}\left(\mathcal{H}_d|\mathcal{E}\right)}{\mathbb{P}\left(\mathcal{H}_p|\mathcal{E}\right)} = \frac{\mathbb{P}\left(\mathcal{E}|\mathcal{H}_d\right)}{\mathbb{P}\left(\mathcal{E}|\mathcal{H}_p\right)} \frac{\mathbb{P}\left(\mathcal{H}_d\right)}{\mathbb{P}\left(\mathcal{H}_p\right)} \tag{1}$$

where $\mathcal{E}$ is the evidence (quantitative and qualitative), $\mathcal{H}_d$, and $\mathcal{H}_p$ are the hypotheses of the defence and prosecution, respectively, $\mathbb{P}\left(\mathcal{E}|\mathcal{H}_d\right)/\mathbb{P}\left(\mathcal{E}|\mathcal{H}_p\right)$ is the likelihood ratio of the evidence under the two hypotheses, and $\mathbb{P}\left(\mathcal{H}_d\right)/\mathbb{P}\left(\mathcal{H}_p\right)$ are the prior odds of the hypotheses. We cannot and should not infer the prior odds, they should be determined by the court, which leaves the likelihood ratio:

$$LR(\mathcal{H}_d, \mathcal{H}_p) = \frac{\mathbb{P}\left(\mathcal{E}|\mathcal{H}_d\right)}{\mathbb{P}\left(\mathcal{E}|\mathcal{H}_p\right)}. \tag{2}$$

Assuming that the evidence $\mathcal{E}$ consists of quantitative information $\mathcal{Q}$ (the coverage in MPS or peak height/area in CE) about genetic information $g$, then we can factorise $\mathbb{P}\left(\mathcal{E}|\mathcal{H}_i\right)$ as:

$$\mathbb{P}\left(\mathcal{E}|\mathcal{H}_i\right) = \mathbb{P}\left(\mathcal{Q}, g|\mathcal{H}_i\right) = \mathbb{P}\left(\mathcal{Q}|g\right)\mathbb{P}\left(g|\mathcal{H}_i\right). \tag{3}$$

Thus, the probability of the evidence, given a hypothesis, as two parts:

(1) **The probability of the quantitative information given allelic information**, determinded by the likelihood under a model.

(2) **The probability of the allelic information under the hypothesis in question**, assuming either HWE or $\theta$ correction [**?**].

Let $\mathcal{U}_p$ and $\mathcal{U}_d$ be the set of unknown profiles under the prosecutors and defence hypotheses, respectively, then we can generally write the $LR$ as:

$$LR(\mathcal{H}_d, \mathcal{H}_p) = \frac{\sum\limits_{\boldsymbol{g}_{U_d} \in \mathcal{U}_d} L\left(\boldsymbol{\mathcal{Q}}|\boldsymbol{g}_{K_d}, \boldsymbol{g}_{U_d}\right)\mathbb{P}\left(\boldsymbol{g}_{U_d}|\boldsymbol{g}_K\right)}{\sum\limits_{\boldsymbol{g}_{U_p} \in \mathcal{U}_p} L\left(\boldsymbol{\mathcal{Q}}|\boldsymbol{g}_{K_p}, \boldsymbol{g}_{U_p}\right)\mathbb{P}\left(\boldsymbol{g}_{U_p}|\boldsymbol{g}_K\right)}, \tag{4}$$

where $\boldsymbol{g}_K = \{\boldsymbol{g}_{K_d}, \boldsymbol{g}_{K_p}\}$.

The sums in the numerator and denominator of the general $LR$, seen in Eq. (4), are intractable. Even the sum of the numerator in Eq. (**??**) can be computationally intensive to evaluate. In order to get around this problem we will approximate the LR, using both simple step approximations and more complicated sampling methods, see Section **??**. We will sample from a distribution centred at the most probable combination of profiles. Therefore, we will find the most probable combination of profiles under the model introduced in Section **??**, given the number of known and unknown contributors. [**?**] finds this combination of profiles using a greedy algorithm on a database of possible profile combinations. We will find the optimal combination of profiles using an Evolutionary Algorithm described in Section **??**.

In the remainder of these notes, we will assume the following (unless otherwise stated):

(1) We have sequenced a sample using an $\mathcal{M}$ marker multiplex panel and denote an arbitrary locus by $m$.

(2) We observe $\mathcal{A}_m$ alleles at locus $m$ and denote an allele (using MPS these are represented as sequenced DNA strings) by $a \in \{a_{m1}, ..., a_{m\mathcal{A}_m}\}$. For ease of notation we will sometimes denote this as $n_m$.

(3) The DNA sample is a mixture of $\mathcal{C}$ DNA profiles, i.e. that the sample contains exactly $\mathcal{C}$ contributors.

(4) The data is correct, i.e. contains no drop-outs/ins or other artefacts.