The problem of Y-STR rare haplotype

February 17, 2014

The likelihood ratio

When evidence is found on a crime scene, the expert is asked to provide a Likelihood ratio (LR), defined as

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

where $E =$ evidence, while H_p, and H_d represent the prosecution's and the defence's hypotheses.

E=crime stain's Y-STR haplotype is x AND suspect's Y-STR haplotype is x.

Note: x represents a vector of 7 (or more) repeat numbers.

H_p= the stain belongs to the suspect.

H_d= the stain belongs to someone else (the match happened by chance).

The numerator of the LR is $P(E|H_p)$=1, what about the denominator?

How to estimate $P(E|H_d)$?

Once we have established the population of possible perpetrators (e.g., Caucasian, the whole population, etc.), this probability is equal to the frequency of the profile x in the population.

We can only "estimate" $P(E|H_d)$, since its true value is unknown, unless we knew the haplotypes of the whole population.

We only have a database, with limited individuals.

The counting method

Normally this estimate is obtained by the "counting method", which consists in estimating the frequency of the haplotype in the population by the frequency of the haplotype in the database (www.yhrd.org).

$$f_x = \frac{N_x}{N}$$

So, one possibility is to estimate $P(E|H_d)$ by f_x.

The counting method

Advantages:

- Its simplicity, which is an advantage for explaining the method in court.
- It is the MLE for multinomial samples (it only represents factual information given in the database)
- unbiased

Drawbacks:

What happens for never observed haplotype?

$f_x = 0$

The counting method

Actually, more commonly used is the observed frequency in the extended database (by adding the suspect's copy x to the database).

$$f_x = \frac{N_x + 1}{N + 1}$$

It is conservative for rare haplotypes: frequencies cannot be smaller than $\frac{1}{N}$, so for rare haplotypes the frequency will be overestimated.
Note: "conservative" means that the frequency given by the method in the end should favor the suspect. If f_x is overestimated, than more people with same haplotype , weaker evidence.
Remark: The defence's might prefer to use a different estimate, namely:

$$f_x = \frac{N_x + 2}{N + 2}$$

because from its point of view the two different stains with the same haplotype are observed.

Problem

If the haplotype x, found at the crime scene and matching a suspect, does not appear in the database, according to the counting method we should give to this haplotype a frequency of $\frac{1}{N+1}$ (or $\frac{2}{N+2}$)
Sometimes, also 95% confidence intervals can be provided.
Example: Amanda Knox case: a previously unseen Y-STR haplotype was found on the crime scene, which matches Sollecito's haplotype.

The Good-Turing estimator

Assuming that X distinct species have been observed, numbered $x = 1, ..., X$
$\underline{R}$ is the frequency vector, $\underline{R} = (R_x)_{x=1,...X}$
Each R_x gives the number of individuals that have been observed for species x.
The frequency of frequencies vector, $(N_r)_{r=0,1,...}$ shows how many times the frequency r occurs in the vector $\underline{R}$;
$N_r = |\{x|R_x = r\}|$

The first step in the calculation is to find an estimate of the total probability of unseen species. This estimate is $p_0 = \frac{N_1}{N}$.
The next step is to find an estimate of probability for species which were seen r times: $p_r = \frac{(r+1)N_{r+1}}{N}$.

E: a never observed haplotype x has been found on the crime scene. The suspect y-STR profile matches x.

Prosecution point of view: a single new observation yields a previously unknown profile.
Defence's point of view: the next two observations are the same as one another and previously unknown

$$P(E|H_p) = p_0 = \frac{N_1}{N}$$
$$P(E|H_d) = p_1 \times \frac{1}{N+1} = \frac{2N_2}{N} \times \frac{1}{N+1}$$

$$LR = \frac{N_1(N+1)}{2N_2}$$

Other approaches

- Discrete Laplace (Andersen et al. 2013): This is a parametric method, which uses a mixture of independent Laplace distribution.
- Haplotype Surveying method (Roewer et al. 2000): a Bayesian approach which derives the parameter of the Beta prior distribution for the observed haplotype's frequency from its genetic distance to the other haplotypes in the database.

  This is motivated by the idea that haplotypes which are similar (near in term of distance between the repeated numbers) have similar frequencies. The posterior distribution is then constructed from it, via the likelihood function of the observed database frequencies.

References

Good, I.J. (1953)

The population frequencies of species and the estimation of population parameters

Biometrika 40, 237– 264.

Andersen, M.M., Eriksen, P. S., Morling, N. (2013)

The discrete Laplace exponential family and estimation of Y-STR haplotype frequencies

Journal of Theoretical Biology 329, 39–51.

Roewer, L and Kayser, M and de Knijff, P and Anslinger, K and Betz, A and Caglia, A and Corach, D and Furedi, S and Henke, L and Hidding, M and Kargel, HJ and Lessig, R and Nagy, M and Pascali, VL and Parson, W and Rolf, B and Schmitt, C and Szibor, R and Teifel-Greding, J and Krawczak, M (2000)

A new method for the evaluation of matches in non-recombining genomes: application to Y-chromosomal short tandem repeat (STR) haplotypes in European males

Forensic Science International 114, 31–43.