Adventures in Forensic DNA:

Cold Hits, Familial Searches, and Mixtures

Sandy Zabell

Departments of Mathematics and Statistics

Northwestern University

JSM 2012, July 30, 2012

The approach of this talk

Some simple conceptual issues

. . . sometimes illustrated by examples from actual cases

. . . and an occasional reality check

Footnote for the bar: slide language is intentionally kept simple

Forensic DNA 101

PCR: *Polyermase chain reaction*

- "molecular xeroxing"
- very sensitive

  (potential for contamination and trace DNA)

STR: *Short tanden repeat*

Example: ATCC ATCC ATCC . . . ATCC (n times)

DNA Profile: typically 9–15 pairs of repeat numbers

Forensic DNA 101 concluded

Most common scenario: two sources of DNA:

"Known" (victim or suspect)

"Unknown" (evidence)

9–13 locus concordance usually very strong evidence of identity.

("one in a gizillion")

"Cold Hits"

"Probable cause" vs. cold hit scenario

Common intuition: Evidence more compelling in first case.

NRC committees, distinguished statisticians have differed on this!

NRC 2: if p is match probability, but searched database of size n,

$$\text{use } 1-(1-p)^n \approx np \text{ instead of } p.$$

Resolution of the (apparent) paradox

Use Bayes's theorem:

$$\frac{P(H_1 \mid E)}{P(H_0 \mid E)} = \frac{P(E \mid H_1)}{P(E \mid H_0)} \cdot \frac{P(H_1)}{P(H_0)};$$

- : E: DNA evidence
- : H_0: target not source
- : H_1: target is source

The likelihood ratio is (largely) unchanged, but prior odds differ.

For the non-Bayesians . . .

Suppose

- $p = 1/1,000,000$ (match probability)
- $n = 100,000$ (size of database)
- $N = 10,000,000$ (size of potential suspect pool)

$np = 1/10$ . . . but expect about 10 profiles in pool.

Explaining these issues to trier of fact can be complicated.

Familial Searches

Search a database for "near misses"

Rationale: relatives are much more likely to have matching profiles

Example: IN v. Flowers

1. Steven Myers *et al.*, *Forensic Science International: Genetics*, 5 (2011), pp. 493–500.

2. David H. Kaye, "The genealogy detectives: a constitutional analysis of "familial searching", to appear.

Mixtures

Two or more sources of DNA are present

For example, might see n alleles $A_1, \ldots, A_n$

There are $n + \binom{n}{2} = \frac{n(n+1)}{2}$ consistent genotypes

The CPI (*Combined Probability of Inclusion*): uses

$$(p_{A_1} + \cdots + p_{A_n})^2$$

The likelihood ratio

Recommended by NRC2, but less commonly used.

If all alleles are present and accounted for, a simple formula exists,

. . . thanks to

Weir, B.S., et al. (1997). Interpreting DNA mixtures. *Journal of Forensic Sciences* 42, pp. 213–222.

If . . .

In fact many complications exist:

- The number of contributors may be unknown
- The amounts of DNA may differ

. . . and most feared of all . . .

- **Allelic dropout**

Technical issues

Alleles are scored using *peaks* on an *electropherogram*

The fine print:

- "stochastic thresholds"
- "analytical thresholds"
- "peak height ratios"

Lab protocols leave the analyst great leeway about scoring alleles.

Enormous activity in this area recently

Resource: NIST, "STRbase"

In particular, see "Information on DNA Mixture Interpretation" (http://www.cstl.nist.gov/strbase/mixture.htm)

John Butler:

> *[M]any labs are doing or attempting more complex mixtures often without appropriate underlying validation support or consideration of complicating factors.*

The single most important consideration: one should:

> *Make decisions on the evidentiary sample and document them prior to looking at the known(s) for comparison purposes.*

[Again Butler, but my emphasis]

Many forensic scientists resist or do not understand this basic scientific principle.

This problem is not restricted to forensic DNA.

Nevertheless . . .

The use of DNA typing (justly) remains the gold standard of current forensic identification.

Questions?

Thank you!