THE CONTRIBUTOR PROBLEM IN DNA FORENSICS

Mark W. Perlin[1], **Joseph B. Kadane**[2]
[1]*Cybergenetics, Pittsburgh, PA*
[2]*Department of Statistics, Carnegie Mellon University, Pittsburgh, PA*

There are diverse DNA interpretation problems in forensics. Analysis situations include single source profiles, mixed stains, unknown suspect rape kits, degraded DNA samples, and inconsistent data. Reporting issues with uncertain data require identifying the correct suspect from a database search, and providing the court with highly informative statistics. We are developing and validating a fully automated DNA casework interpretation system that treats all these issues as different versions of a common "contributor problem."

The contributor problem starts with genotype data on samples. Some samples may be references that correspond to known contributors. These data are the only input. The forensic analyst (human or computer) then has the task of determining the number of contributors present in the data, the contributor genotypes, and the relative weights of contributor DNA present in each sample. That is, the general contributor problem is to factor the (locus vs. sample) data into (locus vs. contributor) genotypes and (contributor vs. sample) mixing weights. Since real world data contains error, these results are best described as probability distributions.

As an illustrative example, consider the (simplified) unknown suspect rape kit data comprising STR amplifications of two samples, a clean single source victim reference (V) and a mixed sperm fraction from a differential extraction (S). The analyst might believe that there are two contributors, C1 and C2.

- First consider the mixing weights. Say contributor C1 corresponds to the victim sample, then sample V contains 100% of C1 and 0% of C2. Contributor C2 is the unknown suspect. The analyst infers from the probability distributions that sample S contains about 75% of C1 (the victim), and 25% of unknown suspect C2.

- Next consider the genotype results. C1's genotype is unambiguously that of the victim. However, the genotype data from contributor C2 produce small peaks that generate genotype uncertainty. This uncertainty in C2's genotype is described by a set of full-length profiles at a fixed confidence level (say, 1,000 profiles at 99% probability).

These results can be used to find the unknown suspect by matching each of the 1,000 full-length profiles against a convicted offender database, or against a likely suspect. If a match is found, the match probability is determined by combining the genotype uncertainty (roughly 1 in 1,000) together with the random match probability. The resulting match statistics are entirely rigorous, and can provide a million-fold more specificity than current likelihood methods.

We have built a statistical computer system that mathematically solves the DNA contributor problem. Starting from quantitative peak data, the system adjusts for PCR artifacts (stutter, preferential amplification) and data uncertainty, and automatically considers all feasible genotypes for every unknown contributor. It can solve unknown suspect problems in several minutes, providing a complete statistical and visual description of its results (genotypes, weights). We are currently validating the system for forensic use.