REWRITE

2.3 How Forensic Algorithms Work

Probabilistic genotyping software (PGS)

A genotype samples human DNA at a dozen or so chromosome locations. At each location (or "locus"), the genotype is a pair of numbers listing the genetic variants (or "alleles") inherited from parents. Since there are about a hundred possible allele pair values at each locus, there are a trillion-trillion possible multi-locus genotypes, far more than actual people.

Forensic science uses short tandem repeat (STR) loci, where alleles have different DNA lengths. Amplifying the STR loci using polymerase chain reaction (PCR) produces genotyping data, where an allele's peak height corresponds to DNA amount. The PCR amplifier introduces random peak height variation, so that for one DNA sample, no two data experiments are alike.

Definite genotype systems (DGS) data interpretation tries to remove PCR variation. With small DNA amounts, or mixtures of two or more people in a sample, DGS often fails to produce useful identification information. Manual DGS approaches discard data, forcing all-or-none results. Except for the simplest STR data, DGS is subjective, inaccurate, and uninformative.

Probabilistic genotyping software (PGS) uses all the STR data, and objectively considers all genotype possibilities. PGS exploits PCR variation to derive accurate genotypes. A probabilistic genotype is a statistic that summarizes DNA data, assigning probability to possible genotype outcomes. When examining evidence, the computer doesn't know the genotype "answer."

PGS compares genotypes, not data. A likelihood function compares expected data patterns formed from different genotype combinations (and other variables) with observed STR data. Better genotype explanations have higher likelihood. Comparing an evidence genotype with a reference genotype, relative to coincidence, yields a "likelihood ratio" (LR) match statistic.

The LR is a standard measure of information, quantifying the change in probability from coincidence to evidence. Large numbers (e.g., a million) statistically support an STR feature match between evidence and suspect. Small numbers (e.g., a millionth) support the absence of a match. The greater the LR magnitude, the stronger the support.

Error rates are easily calculated from (probabilistic) genotypes. A genotype has two associated LR probability distributions, one for coincidence and one for evidence. These distributions allow determination of false positive and false negative error rates for an LR result. Pooling groups of genotype LR distributions together enables comprehensive PGS validation.

PGS has had a profound impact on forensic DNA in criminal justice. Reanalyzing previously "inconclusive" DGS results has freed innocent men from prison, helped acquit defendants, and helped convict violent criminals. PGS computer results can be replicated by involved parties, eliminating the black box of the biased human mind from forensic DNA evidence.