Good answers to bad questions about DNA match statistics

Mark W. Perlin, PhD, M.D., PhD

Direct examination is over, and the DNA evidence has been presented. You've addressed chain-of-custody, crime lab processing, data analysis, and mixture interpretation. The match statistic is a trillion. You're almost done.

And then opposing trial counsel asks about simple arithmetic. The questions don't seem relevant to your DNA match statistic. But the jury wakes up, and is now paying close attention. They wonder, is the defendant's DNA really in the evidence?

This DNA mixture scenario has played out in courtrooms for decades. The match statistic captures uncertainty, so an adversary tries to evoke doubt. His math is wrong, his arguments irrelevant [1]. But twisting probability can confuse the jury, and affect the outcome.

Uncertain Genotypes

Forensic scientists test chromosome locations ("loci") that have genetic variation. One locus has a dozen different genetic variants ("alleles"). A person's genetic type ("genotype") at a locus is a pair of these alleles, one inherited from their mother, and one from their father.

A locus has about a hundred possible genotype values. Scientists test a dozen or so genetic loci, giving trillions upon trillions of possible population genotypes.

A reference sample taken from a person produces a definite genotype. At a genetic locus, a person has one genotype value. A multi-locus genotype lists allele pairs at all loci.

Mixtures arise when two or more people contribute DNA to the same evidence item. A crime laboratory generates mixture data that superimposes DNA signals from these contributors. Different genotype combinations can explain the same mixture data.

With multiple genotype explanations, probability enters the equation. The simplest mixture statistic, combined probability of inclusion (CPI), assigns probability to included allele pairs [2]. More sophisticated methods, like Cybergenetics TrueAllele® technology, use DNA peak heights to compute Bayesian probability at every genotype value [3].

Match Statistics

Scientists compare evidence DNA with a person through their genotypes. A mixture genotype represents allele pair uncertainty using probability. For a given reference, the match statistic is the genotype probability ratio of evidence to coincidence.

In the match ratio, the top number (numerator) is the evidence genotype probability at a reference. The bottom number (denominator) is the chance of coincidence, based on the reference genotype's prevalence in the population.

Every valid DNA match statistic is a likelihood ratio (LR) of evidence to coincidence probability. The random match probability (RMP) reciprocal is an LR. The weak CPI mixture statistic reciprocal is an LR [2]. A drop out match statistic is an LR. More sophisticated match methods that preserve quantitative information are LRs [3].

All these match statistics have evidence probability in the numerator, and use coincidence probability in the denominator. With reciprocal RMP, the definite numerator probability is 1. With mixtures, the uncertain numerator is usually less than 1.

A valid match statistic is a vertical ratio – numerator to denominator – evaluated at a reference genotype. Absolute probability doesn't matter, just the probability ratio.

Misleading Arithmetic

A clever adversary may take the jury's eye off the match ratio. He will distract them with other numbers. He will mislead them with superficial arithmetic that looks convincing, but lacks scientific validity.

These diversionary tactics shift attention away from the defendant's genotype and the vertical match ratio. An adversary can rearrange, add and multiply numbers to raise irrelevant issues that sound like serious science. Here is how it's done.

Ploy 1. Highest probability

In the "highest probability" ploy, your adversary ignores the coincidence denominator, and only looks at numerator evidence probabilities. He then shifts attention from the defendant to someone else's genotype.

At one locus, suppose the data puts 10% of evidence probability on the defendant's genotype, and there is a 5% chance of coincidental match. Then the locus match ratio is 2 (i.e., 10% divided by 5%). This ratio is larger than 1, pointing to the defendant.

Your adversary shows just the numerator probabilities. He agrees that 10% of the evidence probability is on the defendant. But then he points to the highest probability, say 20%, at another genotype. The DNA evidence now seems to point away from the defendant – it's probably someone else!

But that's not how probability works. All numerator genotype possibilities will have some evidence probability – high or low – based on the DNA data. The adversary focuses solely on numerator probabilities, and ignores the equally important denominator.

Absolute evidence probability is not enough. What matters is the ratio of evidence to coincidence probability. This vertical ratio can be inclusionary (greater than one) at some genotype values, and exclusionary (lower than one) at others. The defendant's ratio, evaluated at her genotype, is the relevant match statistic.

The data changed the defendant's locus probability from 5% to 10%, a ratio of 2, supporting her DNA's presence in the evidence. The other genotype values are not relevant for this defendant.

Ploy 2. Add up others

In the "add up others" ploy, your adversary adds together numerator probabilities, without considering the denominator. He then shifts attention away from the defendant to all the other genotype values.

As before, with 10% of evidence probability on the defendant's genotype, and a 5% chance of coincidental match, the locus match ratio is 2. This ratio exceeds one, so points to the defendant.

But your adversary shows just the numerator probabilities. He says that with only 10% of probability on the defendant, 90% (i.e., 100% – 10%) of the remaining probability points away from her. So now probability seems to side with the defendant – it's not her!

There are two flaws in this argument. First, a valid match statistic must balance numerator (chance of match) with denominator (chance of coincidence). If the ratio is greater than one for the defendant, collectively it must be less than one for everyone else.

For example, a 5% coincidence for the defendant means a 95% (i.e., 100% – 5%) coincidence for the other genotypes. So the adversary's 90% for all non-defendant genotypes, divided by 95% for all non-defendant coincidental matches, gives a statistic less than one. By his own reasoning, the match statistics point away from other people.

The second flaw is asking the wrong question. The defendant is on trial, not everyone else in the world. The locus match statistic for the defendant is the match probability (10%) divided by the coincidence probability (5%). This ratio is greater than one, supporting a match to the defendant.

These flaws highlight how irrelevant arithmetic can confuse a valid match statistic. The numbers don't relate to a relevant fact of consequence in the trial.

Ploy 3. Match probability

Your adversary forms a numerator-only "match probability" product to produce an irrelevant number that omits crucial denominator information.

Multiplying together small numbers gives a very small number. For example, multiplying numerators like 10% or 50%, across 20 loci, can give a product of a trillionth (one over a trillion). Each numerator value is a match probability, which can't exceed one, so the match product can become very small.

Your adversary shows this very small match number to the jury, declaring a match probability of a trillionth. The DNA evidence now seems to agree with him; it can't be the defendant – the match probability is too small!

But the adversary has intentionally ignored the denominator. With an inclusionary match statistic, at most loci, coincidence is less probable than match. Multiplying tiny coincidence denominators like 1% or 5%, across 20 loci, can give a very tiny product of a trillionth-trillionth.

A match statistic is a vertical probability ratio of evidence to coincidence. Dividing the very small trillionth (match probability) numerator, by the very tiny trillionth-trillionth (coincidence probability) denominator, gives a very large match ratio of a trillion.

When both numerator and denominator are considered, the correct large inclusionary match statistic ratio is restored.

Professional Responsibility

A lawyer has a duty of candor to the court (American Bar Association, Professional Conduct Rule 3.3). The attorney may not "knowingly make a false statement of fact or law." A forensic scientist should not "materially misrepresent data or scientific principles" (American Academy of Forensic Sciences, Code of Ethics and Conduct).

When confronted with flawed arithmetic, a lawyer can request a Rule 403 relevancy hearing to keep invalid or misleading DNA assertions away from the jury. Unfounded expert testimony may be blocked via a Rule 702 admissibility hearing to keep junk science out of the courtroom.

Re-examination of a knowledgeable expert can correct bad DNA match arithmetic. Cross-examination of a misleading witness may reveal lapses in ethics or expertise.

Conclusions

Accurate and objective DNA evidence can help a jury understand crime scene activity. Valid DNA match statistics quantify the strength of evidence, whether inclusionary or exclusionary. Match arithmetic that intentionally misleads has no role in criminal justice, and should be kept out the courtroom.

A twenty-minute YouTube talk (with case examples, courtroom testimony, color figures and evidence rules) expands on the ideas presented in this article [1]. ●

References

1) M.W. Perlin, "Distorting DNA evidence: methods of math distraction," American Academy of Forensic Sciences 70th Annual Meeting, Seattle, WA, 22-Feb-2018. On-line video presentation at: https://www.cybgen.com/information/presentations/2018/AAFS/Perlin-Distorting-DNA-evidence-methods-of-math-distraction/page.shtml

2) M.W. Perlin, "Inclusion probability is a likelihood ratio: implications for DNA mixtures," Promega's Twenty First International Symposium on Human Identification, San Antonio, TX, 14-Oct-2010. On-line poster presentation at: https://www.cybgen.com/information/presentations/2010/ISHI/Perlin_Inclusion_probability_is_a_likelihood_ratio_implications_for_DNA_mixtures/page.shtml

3) M.W. Perlin and B. Szabady, "Linear mixture analysis: a mathematical approach to resolving mixed DNA samples." Journal of Forensic Sciences, 46(6), pp. 1372-77, 2001.