Forensic MAGAZINE

Forensic Science In The Information Age

By Mark W. Perlin, Ph.D., MD, Ph.D. Article Posted: April 10, 2012

Reliable computer interpretation can address the scientific need for thorough, objective, and informative analysis of DNA evidence.

Forensic Identification

Ancient societies used distinguishing marks to identify people and their property. During the T'ang Dynasty, Chinese officials certified documents with handprints.[1] Babylonians in the pre-Islamic Sassanid Empire established ownership of lost items through identifying marks.[2] In the 19th century, more precise measurements of physical features[3] and fingerprints[4,5] enabled more accurate human identification and statistical association.[6]

The 20th century witnessed a flowering of diverse forensic modalities (hair, fiber, glass, ballistics, etc.) that could connect crime scene evidence with a suspect.[7] Investigative databases that could solve cold cases were established in the earliest decades for fingerprints in London and New York and evolved into DNA databases by the close of the century.[8] Today, the success of DNA identification,[9] coupled with outsized "CSI effect" expectations,[10] have instilled in the modern world a sense of forensic infallibility, with abiding faith in the power of DNA.

Great Expectations

There are many consumers of forensic information. Police investigate crimes, prosecutors present evidence that is weighed by judges and juries, while the public funds the forensic enterprise in order to secure better protection from crime. These societal consumers of crime lab information all assume that: a) the forensic evaluation process is thorough and objective, and b) the full measure of identification information has been accurately extracted from the available evidence. These are reasonable expectations. An incomplete or biased approach that discarded information might be ineffective (not find the right person), incorrect (implicate the wrong person), or unusable (inadmissible in court).

However, current forensic science practice does not always meet these expectations. The National Academy of Sciences (NAS) 2009 report on "strengthening forensic science" identified potential flaws or examiner bias in some methodologies.[11] Indeed, comparative bullet lead analysis was found to lack a sound scientific basis and is no longer used.[12] Since DNA evidence interpretation had not consistently accounted for natural data variation, the federal Scientific Working Group on DNA Analysis Methods (SWGDAM) issued new 2010 guidelines.[13] Recent studies report examination bias in DNA mixture interpretation[14] and a million-fold information loss in the human review of mixture data.[15,16]

A thorough and objective evaluation of evidence that yields all the data's identification information may be desirable, or even necessary, but is it feasible? Isn't a "match" between evidence and suspect inherently limited to just comparing the data features of these two items? And isn't this "match" unavoidably biased, since one of the two items is the suspect? If the items either fully "match" or they don't, how does one measure the amount of information?

Before the advent of computers and information theory, these might have been challenging questions. But in the modern age, the questions have workable, generally accepted answers. We next examine how information science can resolve these fundamental issues in forensic science.

Information Science

Computers and information theory were forged on the scientific battlefields of World War II. The earliest electronic computers used probabilistic search to solve the equations that helped build the atomic bomb.[17] Information theory and likelihood ratios (LR) cracked the German Enigma code, providing daily military intelligence on custom-built computers.[18] Scientific computing crunched raw data into the information that won the war, and ushered in our modern world.

Most people like certainty. During cross-examination, an expert witness wants solid fact to provide a shield of certainty against an onslaught of critical questioning. But science reasons from uncertain data, not blind faith, and so such certainty is illusory.

Uncertainty rules in nature. DNA differences provide abundant natural variation in biological populations, with no two (uncloned) individuals alike. Scientific data exhibit laboratory variation, with random fluctuations occurring between repeated experiments and seen within each measurement. In a random world, though, these same data are needed to reduce uncertainty and increase our rational belief in one hypothesis over another.

Uncertainty arises when there is more than one possible explanation. Our scientific belief in a particular explanation is its probability, a number between 0 and 1. An explanation's information is related to the reciprocal of its probability.

A rare event has a small probability, so its reciprocal gives a large information number, reflecting the considerable surprise experienced when the event happens. The occurrence of a common (high probability) event does not surprise us at all, and so its low information (the reciprocal of a high probability is a small number) indicates little or no surprise. Quantifying uncertainty through probability[19,20] provides the information needed to strengthen forensic science.

Identification Science

A modern identification science must be thorough, objective, and informative. These properties, which we now consider in turn, make scientific identification less susceptible to legal challenge.

A *thorough* evidence examination must use all the data. A "model" helps explain observed data. When data count up how much of a quantity is present, a quantitative model is needed to express those counts with numbers. Random variation is accounted for by the model's statistical component. A computer model uses probability to mathematically assess how well a proposed hypothesis actually explains the data.

Thoroughness also entails considering all possibilities. Deductive reasoning proceeds forward from a hypothesis to explain observed data. Inductive reasoning instead starts from the data, and infers causal hypotheses. This inverse reasoning (from data, back to hypothesis) is done with Bayes' theorem,[21] which assesses all possible hypotheses and calculates the probability of each one. Most real world inductive problems (including forensic inference) are complicated and need a computer to solve their complex equations.

An *objective* assessment requires that the computer never see a suspect or defendant when evaluating evidence. That is, when the computer infers the probability of each hypothesis, its deliberation must be done without any knowledge of what the "answer" should be. Rather, the computer should reach its unbiased conclusions solely from the available evidence.

The hypotheses considered by the computer will ultimately be compared with known references. With

ballistics evidence, for example, the striations and marks of a crime scene bullet can be explained as arising from the barrel of a particular firearm. The make and model of a firearm is a possible hypothesis for the bullet evidence. Thorough inference considers all feasible firearm models, assigning each model a probability that accounts for its prevalence and how well it explains the bullet mark measurements. Performed objectively, only the bullet data is used.

When comparing evidence to a known suspect, we want to quantify the amount of information in a match. How much more does one hypothesis explain the evidence than alternative hypotheses? Equivalently,[22] how much more probable is a match between evidence and known than mere coincidence? This balance of probabilities is the match statistic (or "LR") that weighs the evidence supporting the hypothesis, relative to all other alternatives. In accordance with the Federal Rules of Evidence (FRE) Rule 403, LR match information is able to assess the probative value of a hypothesis and factor away prior prejudices.

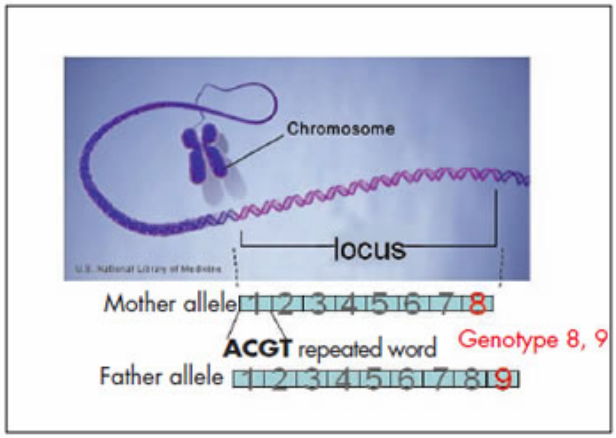

Figure 1: DNA genotype. A genetic locus has two DNA sentences, one from each parent. An STR allele is the number of repeated words. A genotype at a locus is a pair of alleles, as in the 8, 9 shown in the diagram. Since many alleles allow for a great many allele pairs, a person's genotype is relatively unique.

Genotype Inference

The genotype is the central concept of DNA identity. At a genetic locus (some location on a chromosome), a person has two copies of DNA, one inherited from each parent (Figure 1). The DNA content is called an "allele". A person's genotype is their allele pair at a locus.

A genetic locus having very many alleles can help identify people. For example, 20 alleles lead to over 200 allele pairs for a genotype. Testing a dozen such loci leads to quintillions of possible genotypes, more than enough to statistically distinguish billions of people.

Human identification uses short tandem repeat (STR) alleles, DNA sentences of repeated short words. The number of repeated words determines the allele. An allele's DNA molecule can be amplified and sized on a DNA sequencer, producing an allele peak. An STR allele is specified by peak position, since molecule length is proportional to the number of repeats. The amount of allele DNA is proportional to peak height.

When the two alleles in a person's genotype are the same size, the data show one tall peak. With two different alleles, the data show two peaks. Inferring a genotype from reference data is easy, since there is only one allele pair explanation for the peak data. Starting from the hundreds of possible allele pairs in a population, reference data quashes that uncertainty down to a single answer, generating considerable identification information.

A DNA mixture contains two or more individuals. This combination of allele pairs from different people

leads to evidence data more complex than just one or two peaks. Therefore, there is usually more than one genotype explanation, since different allele pair combinations can account for the peak data. These multiple explanatory possibilities create genotype uncertainty. DNA interpretation must thoroughly and objectively examine the mixture data, assigning accurate probabilities to genotype allele pair hypotheses.

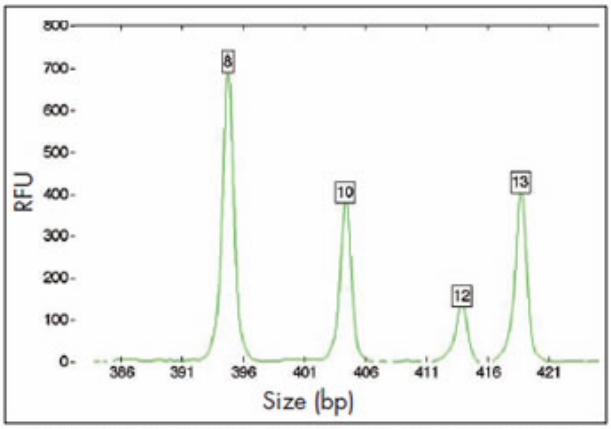

Figure 2: Mixture data. A mixture sample has genotype allele pairs from two or more contributors. Shown is a DNA signal (green curve) having four peaks at the Penta D locus. The horizontal x-axis indicates an allele's DNA molecule length, while the vertical y-axis measures a peak height that reflects the allele's DNA quantity.

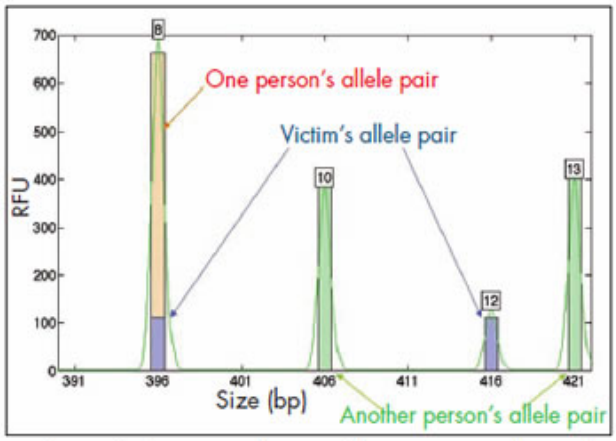

Figure 3: Genotype inference. The computer explores all possible genotype combinations, trying to explain the observed data peak pattern. Better explanations lead to a higher genotype probability. Shown is an explanation that combines different amounts of allele pairs (colored bars) from three contributors.

Case Example

Two years ago in Stafford County, a Virginia woman awoke to find a man she knew on top of her. She screamed, he fled. DNA from her underpants showed a mixture containing multiple contributors (Figure 2). Comparison was made with the DNA of a sergeant from the nearby Quantico Marine base, but human review following the 2010 SWGDAM guidelines was inconclusive. The DNA evidence was therefore reinterpreted using a validated probabilistic genotyping system[15] in order to determine a match statistic.

Starting from the Promega PowerPlex® 16 amplification signals stored in an Applied Biosystems 3100® sequencer file format, the DNA data was uploaded to a Cybergenetics TrueAllele® Casework interpretation computer system. The computer processed the evidence item's data under different

scenarios: two versus three contributors, with or without assuming the victim was a contributor, and so on. To ensure objectivity, the computer was given no knowledge of the sergeant's genotype, and so the evidence examination was unbiased.

A thorough genotype explanation should account for the peak height data pattern observed at a locus. The computer can suggest an allele pair assignment for each contributing genotype. The software then expands the height of a contributor's allele pair in proportion to how much DNA came from that contributor. The computer adds up these weighted allele pair components to construct an allele peak pattern whose hills and valleys might explain the DNA data (Figure 3). Since "a better fit's more likely it", genotype values that form more explanatory patterns receive higher probability.

The computer thoroughly considered hundreds of thousands of genotype explanations. Those explanations that better explained the data conferred higher probability to their genotype allele pairs. Genotype possibilities that could not explain the data received very low probability. Intermediate genotype explanations received intermediate probability. When done deliberating, the computer wrote into a database its inferred genotypes, providing a probability distribution for each locus and contributor (Figure 4).

The computer's objectively inferred mixture evidence genotype was then compared with the suspect's genotype. This comparison was done relative to a black population genotype (representing all conceivable alternative people) in order to form a match statistic. The LR match statistic reports on how much the evidence changed our belief from a random match to a specific suspect match. There was a four-fold gain in probability at the Penta D locus, as seen visually (Figure 5). Multiplying together the match scores at 15 genetic loci, and accounting for human relatedness, a match between the victim's underpants and the sergeant was 284 million times more probable than coincidence.

The computer's DNA mixture interpretation and reported match statistic were introduced as evidence at the court martial. The direct and cross-examination of the expert witness took under an hour. The defendant was found guilty on all charges and sentenced to three years in the brig and a dishonorable discharge from the Marine Corps.

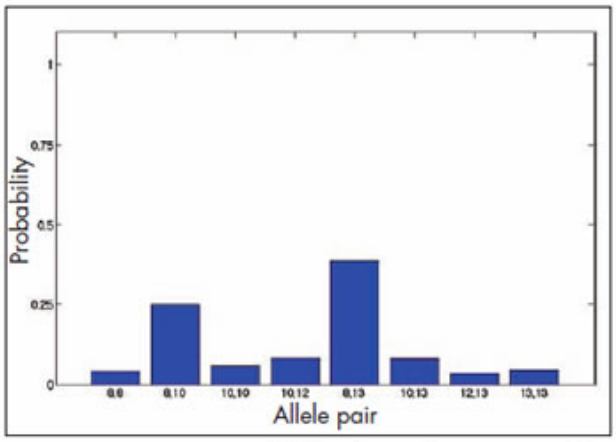

Figure 4: Genotype probability. The evidence genotype for one of the unknown contributor genotypes, objectively inferred without any knowledge of the suspect. Shown are probability bars for eight allele pairs that account for about 99% of the probability. Virtually no probability is assigned to the hundreds of other allele pair possibilities at this locus, since they cannot adequately explain the data. The data imposes a constraint that reduces the number of possible genotype values; fewer possibilities increases genotype information.

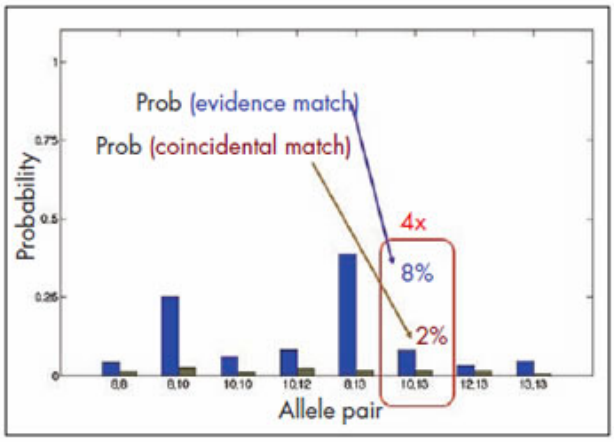

Figure 5: Match statistic. DNA match information tells us how much more the suspect matches the evidence than a random person. With computer inference done, we now note that the suspect's genotype at locus Penta D is the allele pair 10,13. The LR focuses our attention on just this one genotype value (red box), taking the ratio of the evidence genotype probability (8%) to the population genotype probability (2%). This ratio gives us a match statistic of 4, expressing the information gain from the Penta D locus data.

Forensic Applications

Last year the author filed case reports with computer match statistics in over 75 criminal cases. The offenses include sexual assault, homicide, weapons or drug possession, bank robbery, and home invasion. Expert testimony was given in state, federal, military, and foreign courts. Most cases involved police or prosecutors needing a match statistic for arrest or court on DNA evidence where human review was inconclusive or unpersuasive. The DNA was usually crucial evidence, with the defendants typically convicted of their crimes. Several cases involved the defense assessing actual or post conviction innocence.

The courts have accepted computer DNA evidence interpretation. The Pennsylvania appellate Superior Court upheld the 2009 Commonwealth v. Kevin Foley homicide conviction and computer admissibility in a published precedential ruling.[23] In the Real IRA Massereene Barracks attack, which killed two unarmed British soldiers, the Northern Ireland court admitted computer methodology into evidence and used the DNA match statistic in its ruling.[24] These legal precedents are based on extensive scientific validations of the probabilistic genotype method[15,16] and regulatory approval.[25] Computer interpretation is not novel, just a useful implementation based on established mathematics and science.

Highly informative computer-inferred evidence genotypes can help investigators find criminals and missing people. However, some government DNA databases institutionalize the low information yield of human interpretation methods. One crime lab study showed that out of fifty DNA mixtures, human review did not produce a match statistic on half the evidence items, whereas the computer succeeded every time. Other crime lab studies show that even when human review of mixture data does yield a match statistic, the computer's numbers are (on average) about a million times greater. More informative genotypes translate into a more powerful investigative database.

Investigative DNA databases of information-rich genotypes were used for identifying victim remains in the World Trade Center disaster.[26] The statistical computer inferred probabilistic genotypes from victim remains evidence, and from missing person kinship and personal effects data; the investigative database then compared these genotypes to form matches having LR statistics.[27] Probabilistic genotypes are written into the ANSI/NIST forensic data exchange standards (sect. 18.020-18.021),[28] and are specifically allowed by SWGDAM interpretation guidelines (par. 3.2.2).[13]

Path Forward

An information age demands information. We expect a Google search to return thorough, objective, and informative results, using the best available probability computer model methods. Our human minds ask questions, and we rely on the computer to calculate the best answers. Whether cracking a code, diagnosing disease, piloting a plane, or working on Wall Street, our lives and livelihoods depend on computer thought. Apprehending criminals through forensic intelligence is no exception—we want the most informative computers working 24/7 to provide protection.

DNA laboratories are now bringing computers on board to extend their forensic examiners' analytic capability. A scientist can organize evidence and frame forensic questions; robots and computers can then automate the mechanics. A forensic scientist can incorporate informative DNA match statistics from complex mixture calculations into their case reports, and provide testimony in court. Experts excel at human activities, while computers are better calculators. Even before their crime labs deploy computer interpretation, police investigators and trial attorneys can rely on the private sector to deliver computer processing, case reports, and expert witness services.

The NAS report identified ways to strengthen forensic science. In addition to sound scientific data, the criminal justice system relies on thorough, objective, and informative interpretation of such data. DNA has paved the way once more and shown how reliable computer interpretation can address these scientific needs, complementing human cognition. Forensic science is now embracing the information age, extending the human mind with objective and informative computer solutions.

References

1. Laufer B. History of the Finger-Print System. Washington: Smithsonian Institution, Government Printing Office, 1913.
2. ha-Nasi J. Bava Metzia. Babylonian Talmud. Mesopotamia: Sura Academy; 500.
3. Bertillon A. Signaletic Instructions: including the theory and practice of anthropometrical identification. Chicago: The Werner Company, 1896.
4. Herschel WJ. Skin furrows of the hand. Nature. 1880 Nov 25;23:76.
5. Faulds H. On the skin-furrows of the hand. Nature. 1880 Oct 28;22:605.
6. Galton F. Finger Prints. London: Macmillan and Co, 1892.
7. Saferstein R. Criminalistics: An Introduction to Forensic Science. 7th ed. Upper Saddle River, NJ: Prentice Hall, 2000.
8. Gill P, Werrett D. Interpretation of DNA profiles using a computerised database. Electrophoresis. 1990;11:444-8.
9. Butler JM. Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers. Second ed. New York: Academic Press, 2005.
10. Toobin J. The CSI Effect. The New Yorker. 2007 7 May 2007.
11. National Research Council. Strengthening Forensic Science in the United States: A Path Forward. Washington, DC: National Academies Press, 2009.
12. Tobin WA, Thompson WC. Evaluating and challenging forensic identification evidence. The Champion. 2006 July;30(6):12-21.
13. SWGDAM. Interpretation guidelines for autosomal STR typing by forensic DNA testing laboratories. 2010.
14. Dror IE, Hampikian G. Subjectivity and bias in forensic DNA mixture interpretation. Science & Justice. 2011;(in press).
15. Perlin MW, Legler MM, Spencer CE, Smith JL, Allan WP, Belrose JL, Duceman BW. Validating TrueAllele® DNA mixture interpretation. Journal of Forensic Sciences. 2011;56(6):1430-47.
16. Perlin MW, Sinelnikov A. An information gap in DNA evidence interpretation. PLoS ONE. 2009;4(12):e8327.
17. Harlow FH, Metropolis N. Computing and computers - weapons simulation leads to the computer era. Los Alamos Science. 1983;7(Winter/Spring):132-41.
18. Singh S. The Code Book: The Science of Secrecy from Ancient Egypt to Quantum Cryptography. New York: Doubleday, 1999.

19. Kadane JB. Principles of Uncertainty Boca Raton, FL: Chapman & Hall, 2011.
20. Lindley DV. Understanding Uncertainty. Hoboken, NJ: John Wiley & Sons, 2006.
21. McGrayne SB. The Theory That Would Not Die: How Bayes' Rule Cracked the Enigma Code, Hunted Down Russian Submarines, and Emerged Triumphant from Two Centuries of Controversy. New Haven: Yale University Press, 2011.
22. Perlin MW. Explaining the likelihood ratio in DNA mixture interpretation. Promega's Twenty First International Symposium on Human Identification, 2010; San Antonio, TX. 2010.
23. Commonwealth of Pennsylvania v. Kevin James Foley. Superior Court of Pennsylvania; 2011.
24. The Queen v Colin Duffy and Brian Shivers. Crown Court in Northern Ireland; 2011.
25. Approval for the use of TrueAllele® technology for forensic casework. New York State Commission on Forensic Science; 2011.
26. Perlin MW. Mass casualty identification through DNA analysis: overview, problems and pitfalls. In: Okoye MI, Wecht CH, editors. Forensic Investigation and Management of Mass Disasters. Tucson, AZ: Lawyers & Judges Publishing Co; 2007;23-30.
27. Perlin MW. Identifying human remains using TrueAllele® technology. In: Okoye MI, Wecht CH, editors. Forensic Investigation and Management of Mass Disasters. Tucson, AZ: Lawyers & Judges Publishing Co; 2007;31-8.
28. Carey S. Data format for the interchange of fingerprint, facial & other biometric information. In: Wing B, editor. Gaithersburg, MD: American National Standards Institute (ANSI) and National Institute for Standards and Technology (NIST); 2011.

Dr. Mark Perlin *is Chief Scientific and Executive Officer for Cybergenetics. He has twenty years experience developing computer methods for information-rich interpretation of DNA evidence and providing TrueAllele products and services to the criminal justice community. Cybergenetics, 160 North Craig Street, Suite 210, Pittsburgh, PA 15213; (412) 683-3004; perlin@cybgen.com; www.cybgen.com.*