

Transcript of Dr. Mark Perlin's talk on "Three Match Statistics, One Verdict" delivered on 25 February 2010 in Seattle, WA at the 62nd Annual Scientific meeting of the American Academy of Forensic Sciences. Copyright 2003-2010 Cybergenetics, all rights reserved.

Dr. Perlin: I will next be speaking about a case in which all this mixture theory played out. The title of the talk is "Three Match Statistics, One Verdict." This was a case that Dr. Robin Cotton and I both testified in. Again, this work was done by Cybergenetics for a prosecutor using the TrueAllele[®] system, which is a commercial product.

(next slide)

The case is Commonwealth of Pennsylvania vs. Kevin Foley. In April 2006, Blairsville dentist John Yelenic was brutally slashed to death, and bled out in his home. A year and a half later, Pennsylvania state trooper Kevin Foley was charged with the homicide. Kevin Foley is shown wearing a University of Pittsburgh t-shirt. In February of 2008, there was a hearing. There are 13 million people in the state of Pennsylvania, and so the defense questioned why a 13,000 CPI inclusion match score was all that helpful.

(next slide)

The DNA was the key evidence in the case. The DNA sample came from under the victim's fingernails. He had scratched somebody, so there were two contributors to this

DNA mixture. The computer tells us it's a 93% victim and 6.7% unknown. A lot of DNA was present, yielding one nanogram (ng). The STR analysis was done by the FBI's Quantico lab with ProfilerPlus and Cofiler, and we know the victim's contributor genotype because we have the victim. The TrueAllele computer interpreted the DNA mixture using a quantitative addition method to infer the unknown contributor genotype. Afterwards, as with all objective methods, the inferred genotype was compared with the suspect genotype.

(next slide)

Interestingly, there were three different match statistics here. The inclusion method that the FBI offered was 13,000. Robin Cotton's obligate allele subtraction method gave 23 million. Quantitative computer interpretation using statistical methods gave 189 billion. So the questions before the court were: "Why are there different match results?", "How do these mixture interpretation methods differ?", and "What should we present in court?" That is what I will endeavor to explain in the remainder of the talk.

(next slide)

What are the critical differences between these different interpretation methods? They involve how the data is used. An *inclusion* method does not use the victim profile, nor does it use quantitative data. That is, "every pair has equal share." A *subtraction* or obligate allele method does use the victim profile, and again gives every pair an equal

share. The *addition* method uses all of the information that has been presented from the crime scene, including quantitative data and the victim profile. We can see the victim's profile in the mixture data. As you'll see, his data peaks dominate every STR profile, since the DNA evidence was scraped from his fingernails. Moreover, we have the concept that "a better fits more likely it" to use the quantitative data.

(next slide)

There was a Frye hearing that established the general acceptance of the underlying principles in the relevant scientific community. The court determined that the relevant community comprised the statisticians, mathematicians, and various scientists who develop, test, discuss, publish and present DNA interpretation methods. We showed articles about quantitative STR peak information, and how that has been standard for 20 years. Genotype probability distributions have been around since Mendel, which goes back a hundred years. Computer interpretation of STR data has been around for a 20 years, so we showed more articles on that. Statistical modeling and computation with modern probability models and search has really come into its own in the last 20 years, and underlies much of the ongoing scientific inquiry in physics, social sciences, economics, gene chips and so on. This history was easy to describe. The modern likelihood ratio (LR) literature goes back to Turing and Good's work in the 1950s. We provided admissibility decisions showing that some jurisdictions find that mixture interpretations are always admissible, but are subject to cross-examination. There are

a number of computer systems that do mixture deconvolution, and of course we described TrueAllele.

(next slide)

We also described the results of some studies that we had done on NIST data for an NIJ grant. The paper came out last year, published in PLoS ONE. The paper discusses many things, but I'll describe what the data here mean. We had a set of two contributor DNA mixtures formed from known samples in different dilutions. Therefore, for each synthesized mixture, we know exactly how much DNA is present from the culprit (i.e., the total DNA times the mixture weight) whose genotype we are inferring. The x-axis of DNA quantity is on a logarithmic scale from 10 picogram (pg) to 1000 pg.

Each mixture sample can be interpreted using different methods. In the published study, we examined the two unknown inclusion (CPI) and one unknown subtraction (CLR) human review methods, as well as TrueAllele probability modeling with one and two unknowns. This plot shows the TrueAllele interpretation results on quantitative data, assuming the victim profile and looking for one unknown contributor genotype. The y-axis shows the amount of likelihood ratio information that was inferred, with the LR information gain plotted on a logarithmic scale from 1 to sextillion.

We can form a scatter plot of the 40 mixture-dilution experiments, and their x-y (quantity, information) pairs. We observe a straight-line fit to the data as we move

towards the left. At the midpoint of the x-axis is the 100 pg level beyond which human review dares not tread. However, the regression line shows that down to about 15 pg of DNA, the computer still infers a usable LR result of over a million to one.

Importantly, we can use these validation data as a calibration curve to predict what would happen in an individual case. We had 1000 pg of DNA, so with a 6.7% mixture we find where 67 pg ($1000 \text{ pg} \times 6.7\%$) lies on the x-axis. We move up from the x-axis to find the point of intersection with the calibration line. If we had 15 loci, we'd expect about a quadrillion as our LR match strength. Since 12 loci were used (80% of the total), the LR shrinks to about a trillion. Our observed LR match number in this homicide case is in the hundreds of billions. Thus the defense contention that all match statistics on the DNA evidence should give the same low LR information as CPI was not scientifically founded. We could predict from our calibration data roughly what the match statistic would be.

(next slide)

When I testified in the case, I explained to the jury why these methods were expected by science. Essentially, if you use more data going in, you will get more information out. TrueAllele, as far as we can make its models, uses all of the available DNA data, stutter, relative amplification, quantitative peaks, etc. The analogy that seemed to be somewhat persuasive was that of a microscope. We can take a microscope slide, trying to diagnose a pneumonia, and look at it in different ways. If you look at it with the naked

eye, you'll see the slide (and little else). If you're looking at it instead with a magnifying glass, maybe you'll go on to the next level of resolution (as with using the victim) and will see a bit more. But if you use the appropriate microscope, you can diagnose the bacterium, and know how to treat the disease. So all the computer is doing here is providing a better microscope on the same data.

(next slide)

Let's look at some data. All of these pictures are snapshots taken from the TrueAllele user interface. For the mixture weight, we can see its probability distribution. Here is a histogram centered around 6.7%, having a standard deviation of about 1%. We show some tables that confirm that.

(next slide)

This is the inferred genotype. What you see at each locus is a probability distribution of the 100 or so possible allele pairs. Which genotype values were most probable? Some have a probability of about 1. That is because they are four allele cases, and it is then easy (if you have the victim profile) to get that answer. Other loci are less informative, and so look a little more diffuse. As we'll soon see, genotype is much more informative overall, compared with inclusion methods. The genotype has been inferred solely from the data, and no suspect has yet been considered. We can now match this evidence genotype against a population of suspects (say, 10 million convicted offenders) or

compare it against one suspect in a case. The evidence genotype is objectively inferred, without ever knowing the suspect's genotype.

(next slide)

This is the inferred LR match information, shown for every locus. Black means that the log(LR) information is positive. The LR is shown on a log scale, with x-axis values of 1, 10 and 100. The log(LR) is positive in every case, and (because it's logarithmic) we add up the information to get a total LR of about 11, or 100 billion (i.e., a 1 followed by 11 zeroes).

(next slide)

Let's zoom in to one locus, and see what the data look like. Here is the STR locus D8. Towering over all the peaks are the victim's profile peaks. The victim comprises 93% of the mixture, so we see clearly see tall peaks at his alleles 12 and 14. That allele pair [12, 14] is his genotype. Now down at the bottom with low rfu values, we see some other low-level peak events. These peaks might be stutter, but the one at 15 is probably not stutter. It is probably a real allele event, and an obligate method would infer that allele 15 has to be in the genotype solution.

(next slide)

We next see the TrueAllele VUIer Explain interface. It is helpful when computers can explain their thinking to us visually. The Explain window lets us do "what-if" analyses. We see (in gray) the genotype of the victim [12 14] as 93% of the total DNA. The little blue rectangles represent the second unknown contributor alleles. The Explain windows lets you move the alleles around to wherever you want. Based on the genotype value that you choose, the computer will generate a proposed peak pattern from using statistical modeling that accounts for stutter, relative amplification, degraded DNA and so on. When the hypothesis of [12 15] is put in, we get a pattern that very closely (shown in gray) matches the quantitative peak data (shown in green). Another likely pattern could be from a possible [15 15] homozygote genotype.

(next slide)

We can do a likelihood comparison in the Report interface. For the computer inferred genotype, we see (in dark blue) a quantitative likelihood bar chart at locus D8. Since "a better fit's more likely it," 90% of the probability ends up at [12 15] (which matches the suspect).

We also see the inclusion likelihood (light blue). We can upload any profile into the system, including what inclusion would have inferred. Inclusion dissipates its probability over six possible allele pairs constructed from the three alleles. The inclusion method follows the same inference procedure (which is why it is valid), but a different rule, a different philosophy of "every pair gets equal share." The inclusion inference process

reduces the suspect-matching genotype probability down from 90% to 17% – it does not infer a more definite belief based on the data, but instead tries to be fair to all candidates. Notice that three allele pairs out of inclusion's six possibilities don't include allele 15, which any obligate review method must do.

(next slide)

This a simple report looks that can be automated generated by the TrueAllele system. Although we did not consider coancestry at the trial, here I used a theta value of 0.01. The likelihood ratio of each allele is the genotype probability after seeing the data, divided by the probability before.

We can see how that LR works at locus D8. If you take a look at the row for D8 (fourth from the bottom), you'll see allele pair [12 15], which matches the suspect. Before we had looked at the data, the probability of the genotype value in the population was 3.6%. Afterwards, the inferred probability became about 90%. 90% divided by 3.6% is a likelihood ratio about 25. That's the LR shown here. Now, we take the log of the LR, which is 1.4. Add up the log(LR)'s for all the loci, and you compute the joint identification information from the independent locus experiments, which is a log of about 11 (or, 100 billion).

(next slide)

We presented this LR comparison bar chart at the trial. It shows that considering more data going in leads to more information coming out. Inclusion (purple) gives low-level LR values across all the loci. When Robin Cotton did an obligate allele analysis (orange), subtracting out the victim, the two tall orange bars (from four allele loci inferring a unique genotype) show large LR values.

TrueAllele (blue) extracted the same high identification information as Robin at those two (four allele) loci. However, it also found D8 to be highly informative, as well as D18 and some other loci. Combining the separate locus LRs, the joint LR increased three orders of magnitude from the 10 thousand of inclusion to a 10 million LR for subtraction. The computer's addition method also considered quantitative peak data, which added another four orders of magnitude, bringing the LR up to 10^{11} (or 100 billion).

(next slide)

What did we learn from the case? This was an objective review that never saw the suspect. The computer result was easy to testify about in court. One of the reasons it was easy was we always had genotypes available. We could point to the genotype, show the data, and make comparisons (e.g., "Robin did this", "we did that"). The concepts of information gain, and of how much data goes in and how much information comes out, were quite understandable to the judge and jury. The defense attorney kept saying there was no precedent. Well, now there is precedent of having computers that infer genotypes admitted into evidence and introduced in testimony.

A key point is that a probability inference method cannot create anything that isn't already there. The computer merely preserves the identification information that resides in the data. We know the victim's profile. We know the quantitative peaks. They are all available data that can be used. Perhaps the greatest lesson learned was that presenting multiple match statistics to the jury was fine. We didn't decide for them that "he did it." We didn't make some binary decision for them of inclusion or exclusion. The prosecution presented three different likelihood ratios (CPI, obligate allele, and quantitative), each based on certain assumptions that were easily explained to a jury.

(next slide)

Here is the verdict. After Kevin Foley testified in his defense, there was still one piece of physical evidence that contradicted his story that he wasn't there. As the prosecutor said, it was the DNA. The victim had scratched his assailant, and captured his DNA. We describe this case in a newsletter that talks about this case in more detail, if you want to see it. The verdict of the jury, having seen a likelihood ratio of a 100 billion, was the defendant was guilty of first-degree murder.

As a forensic science community, you have looked at the scientific studies. These studies compare the likelihood ratios of quantitative and non-quantitative mixture interpretation methods, showing that the computer preserves a million-fold more

identification information than human review. You have seen criminal case results, with similar findings.

I would hope that your ultimate verdict on moving toward more powerful probability methods will be decided on the weight of evidence – likelihood ratios and objective information. What are the most informative methods you can use to reliably process your evidence with efficacy and reproducibility? The verdict on how you should interpret your data in your own lab is up to you; I cannot decide that for you. Thank you very much.