Transcript of Dr. Mark Perlin's presentation on "Computer Interpretation of Uncertain DNA Evidence" delivered on 20 June 2011 in Arlington,VA at the National Institute of Justice.

(Next slide)

Dr. Perlin: Any process of DNA interpretation begins with biological evidence that comes from the field (left). The laboratory transforms those specimens into data, which we see as STR signals (middle), and then an objective interpretation infers an evidence genotype (right). When there is uncertainty in the data, as we see in a mixture, there are more than one or two allele peaks. Then, as in all science, we represent our possibilities as lists with probabilities. Notice that no reference is ever made to any suspect. That isolation guarantees objectivity, whether interpretation is being done by a person or by a machine. When (and only when) we are all done, a comparison can be made to a suspect, or to a million suspects if there is a database. This is the general paradigm that machines and people follow for interpreting with uncertainty in the evidence.

(Next slide)

Cybergenetics was awarded an NIJ grant 10 years ago, with the goals to provide quantitative mixture interpretation, handle stutter and other artifacts, and be very user friendly. The screenshots we will be seeing in the case I will show all come

from a Cybergenetics computer program called TrueAllele® Casework.

(Next slide)

The study design centered on 40 mixture items. There were five mixture weights: 90%, 70%, 50%, 30%, and 10%. There were four dilutions: 1, 1/2, 1/4, and 1/8 of a nanogram. With 2 different pairs of individuals, these combinations comprised comprising these 40 mixture items. The data were prepared by Margaret Kline at NIST.

(Next slide)

In computing mixture interpretation methods, we observed that different information can be inferred from the same data (Promega ISHI paper, 2005). What do we mean by information? There is a match statistic that is usually presented in court DNA evidence – all DNA match statistics are likelihood ratios, including inclusion statistics. The beauty of that fact is that the logarithm of the likelihood ratio (the number of zeros, for example one million has six zeroes (10^6), which gives an information value of 6) is a standard measure of information throughout science and law. So people and machines were obtaining different amounts of information. On all these charts, the y-axis of the information chart is going to be the number of zeroes in the match statistic. On this logarithmic information scale, the exponent 6 indicates one million, while the value 12 is for a

trillion.

(Next slide)

We observed this disparity in more detail in our validation study of a computer system for mixture interpretation (JFS, 2011). Here are 8 different cases, again with the y-axis on a logarithmic information scale. The blue and green bars show what the computer interpreted for each of those cases, with scores of a trillion, quadrillion, and so on. And to the right (orange) is what the inclusion inferred. The computer and the people assumed two unknown contributors. We found, here and in other studies, that the computer preserves far more information than human review. The computer went from 7 zeros average information up to 13 zeroes, with an information difference of about 6 zeroes, or a million-fold improvement.

(Next slide)

In this scatter plot from the data in our study, each point is one of the 40 NIST samples. The x-axis is DNA quantity on a logarithmic scale (10 pg, 100 pg, and 1 ng). The y-axis again is match information on a logarithmic scale. This log-log plot of quantity versus information lets us measure, assess, or predict the amount of information (the match statistic) from DNA interpretation, based on the quantity of DNA in the unknown genotype. In the constant region (right half), the

computer, between 100 pg and 1 ng, essentially preserved all the information. The linear region (left side) between 10 pg and 100 pg shows a straight-line on a log-log plot. This linear regression line shows exactly (with some random variation) how information decreases from the amount at 100 pg down about a million match statistic at 15 pg.

(Next slide)

Comparison with a human review of the same data (red) shows that about half of the samples of the mixtures were “inconclusive,” and had no human statistic result at all. There was a shifting of the million-level crossing point from 15 pg (computer) to about 150 pg (human) as human review with thresholds lost sensitivity. There is a clear information gap between human and computer DNA evidence interpretation.

(Next slide)

This information gap was important in the first case where computers were used (that I know of) for automatic statistical interpretation of DNA mixtures. I testified two years ago in Pennsylvania *Commonwealth v. Foley*. Fascinating background story; more information can be found in our newsletters or in books about the case. The DNA laboratory extracting a mixture with a 6.7% unknown fraction from the victim’s fingernails. The original inclusion method from a prominent

national laboratory gave a match score of 13,000 using thresholds. An independent expert, using the victim reference, came up with a score of 23 million. The computer's result was 189 billion, as described in the newsletter. So, there was about a 6 or 7 order of magnitude gap, the major difference being how much of the what data was used. The computer used all the quantitative information and the victim. Probability modeling preserved the information, whereas peak thresholds discarded it.

(Next slide)

I used this NIJ validation data during the Foley admissibility hearing prior to trial. The defense's contention was that if one match score was 13,000 and another one was 189 billion, then surely DNA mixture interpretation is unreliable or invalid. The validation data (PLoS ONE, 2009) showed that a DNA quantity of 67 pg (6.7% times 1000 pg), predicted a match score around 10^{15} with all 15 loci. With the 12 loci that were used, that predicts a match score of about 10^{12}, or 1 trillion. The actual computer match score was 189 billion, the judge was persuaded, I testified, and the man was convicted largely on DNA of first-degree murder.

(Next slide)

I testified in a Pennsylvania trial earlier this month that illustrates how easy it is to

testify about computer interpretation. There was DNA mixture evidence on a sweatshirt that was discarded after a homicide. The front of the sweatshirt contained the victim's blood, and in the neck area of the sweatshirt was a mixture that was 75% of the victim and 25% of some other person.

(Next slide)

These pictures come from the TrueAllele Visual User Interface for easy review (VUIer™) computer program. Shown is the original STR data, where the computer has worked out the extent of DNA degradation. The degradation extent is used to compensate for high molecular weight marker signal loss during mixture interpretation.

(Next slide)

On the bottom left of this slide is an icon of a jury. Every slide with that icon was shown and explained to the jury. I showed the jury the quantitative peak height data. We see two very tall peaks in the major component from the victim. There is one other peak here, presumably an allele from another person, evident in the quantitative data.

(Next slide)

I then explained to the jury how the computer thinks. What the computer does is try out every possibility. There are thousands of parameters, including PCR stutter and peak uncertainty, how much DNA is present and the genotype allele pairs. In gray is the victim's allele pair, and in blue is a hypothesis as to what the unknown allele pair might be. The allele pair possibilities are moved everywhere, tried in different heights and different sizes. Those genotype explanations of the data that generate patterns (including stutter and peak imbalance and so on) that best explain the data are the ones that are assigned a higher probability.

That is the basic idea of how the computer thinks. The machine tries out hundreds of thousands of times most every possible data explanation using statistical simulation. That extensive search determines the probability of all the variables: stutter, the data uncertainty, and (in particular) the genotypes.

(Next slide)

When the computer has finished, it ends up with an objective determination of genotype probability. In this case there were two unknown contributors to the mixture, so it had worked out the victim probability (dark blue) as well as the other contributor's genotype probability (light blue). Notice that at this locus there is probability mass concentrated at the allele pair [13,14]. The computer has looked at the data, and this genotype is highly probable.

Why is the computer solution objective? There is still no suspect considered, there is still no comparison made. The computer had no idea what answer we were looking for. It just looked at the data and, *ab initio,* determined the genotypes, up to uncertainty (i.e. probability, of whatever the data are trying to express).

(Next slide)

We now compare the inferred evidence genotype ([13,13]; [13,14]; [13,15]; etc.) with the suspect's [13,14] genotype. This comparison is done after the computer has written its answer to the database. It never knew a suspect genotype.

Now we can ask, "How much more does the suspect match the evidence that a random person?" One answer is to divide the evidence probability afterwards (blue) by the population allele pair probability (in brown) at the suspect's genotype (Promega ISHI, 2010). In words, at the suspect's genotype, how much more concentrated is the evidence genotype than a coincidental genotype.

Before we saw a suspect, before we saw the evidence data, before we saw anything in this case, we believed that there were 100 possible allele pairs (brown). After we saw the data, the computer can now tell us the ratio (visually, or mathematically) of the blue evidence probability to the brown coincidence probability. It is about 90% to 15%, or 6. This is a visually interactive way to

describe the likelihood ratio.

(Next slide)

The prosecutor asked me to explain to the jury why thresholds do not work very well, and the reason is that thresholds discard data (Gill, 2006). They create all-or-none peak events by slicing through the quantitative data and saying "Anything over the threshold must be there. Anything under maybe is not there," which is very confusing to statisticians. Thresholds remove quantitative information.

Also, inclusion methods like CPI do not pay any attention to the victim, so the gray peaks from the victim are not given any special consideration.

(Next slide)

The result is that instead of concentrating the probability on where the data is telling us the genotype is supposed to be (remember we had a very tall concentration at [13,14]), it diffuses the probability. From the three threshold alleles, we have six possible allele pairs, and diffuse that probability across many impossible events. That is how CPI works.

All genotype inference methods with uncertainty, including CPI, produce

probabilities. Using less of the data produces less informative probabilities.

(Next slide)

The effect of diffusing away from the correct genotype answer is that when we make a comparison we have not put as much probability where the true answer is. The probability is instead spread out to where the answer cannot be. The result at this locus is that instead of getting 6 times the probability ratio we get 2 times the ratio. The likelihood ratio has dropped from 6 down to 2.

(Next slide)

The computer produced a match statistic of 9.5 trillion. Human review in this case was only 42,000. Cybergenetics was called in because the DNA was the main evidence in this particular homicide case.

We have done math research to make the statement of likelihood ratios become understandable English. So this match statement looks like plain English, and it is. “A match between the suspect and the evidence is 9.5 trillion times more probable than coincidence.” The jury understood that, and the press liked it enough to quote it the next day – it certainly seemed like straightforward English. It was not like those formidable likelihood ratio statements we sometimes read about in foreign papers.

(Next slide)

Using the NIJ mixture data, we conducted a study (Cybergenetics newsletter, 2011) of error rates. To determine the error rates, we looked at different mixture ratios (50/50, 30/70 and 10/90) at different thresholds. We also looked at different amounts of DNA (1ng, 0.5 ng, 0.25ng, 0.125 ng). In the bar chart, we can see one error (missed allele) per locus, reaching a 100% false negative rate. What is fascinating about thresholds is that their error rates, in terms of false negatives for matches we cannot detect, can be well over 100%. This large error rate is unusual in science, but that is what the NIJ data show.

(Next slide)

How does this loss of information translate into casework? This is a study that we recently completed with Joe Galdi from Suffolk County in New York. Shown are 52 items, and matches that were made by computer or by person. The y-axis again shows the information, the number of zeros in the match statistic. There were 25 matches from human review and 27 that were called ‘inconclusive.’ The lab went back and looked through the cases – there may be two or three that were non-probative, but for the most part the uninformative results were truly inconclusive. That is, an analyst tried to form a match answer, but was prevented from doing so by threshold protocols.

(Next slide)

Here is the match information that the computer found (blue). The computer doubled the yield from 25 evidence items to roughly 50 items. Where a person got an answer, the computer generally got a better answer. On the same taxpayer-funded DNA evidence, the computer doubled the yield for public safety.

(Next slide)

This “inconclusive” study has implications for triaging under the new 2012 SWGDAM guidelines. When interpreting DNA evidence, if we use human review with stochastic thresholds, and we get an answer, fantastic! But, the other half of the time, when we do not get an answer on a mixture, then we can do more with the “inconclusives.” We can run it through a computer program, infer a probabilistic genotype in accordance with paragraph 3.2.2 of SWGDAM, and get an answer.

We have a new project we are starting with a state. We are building up a library of match information, preprocessing over a hundred cases where SWGDAM human rules might be eliminating informative DNA evidence from trial. By running the same data through the computer, it only took a few days by machine to reprocess forty of these mixture cases (about 100 items). We are building a

library of DNA cases and match information. As these cases go to court, and the prosecutors need results, in a few days we can send out a final report with the match statistics. There is no delay as the cases go to trial. We preprocess case evidence into information, which lets us rapidly reprocess to confirm and issue a report.

Forensic DNA databases (like CODIS) are formed of inclusion method allele lists on the evidence side, losing investigative power. When we build up evidence libraries, and use them as databases, we preserve far more information, as we had done with the World Trade Center DNA a few years ago.

(Next slide)

In comparing computer versus human DNA evidence interpretation, the theme of this session, there is no competition for two reasons. The first is that the information yield by computer on mixtures is on average one million times greater. Computer interpretation is reproducible, with validation studies quantifying that reproducibility. Computers resolve mixtures with one, two, three, or four unknown contributors. They can quantify degraded DNA and solve unsolved cases. The computer automates complex case interpretation, working for us as a useful tool. The interpretation is easy and fast because it is being done by a calculator, not by a person using paper or monitors. The likelihood ratios are easily explained, so the preservation of informative evidence by

machine can be brought into court, and the crime impact can be maximized.

But, another way of looking at it is that there is no competition because it makes no sense for people to compete with calculators on mathematical problems. People use their tools, they do not fight against them and say “I can compute hyperbolic sines faster and better.” What are they talking about? That does not make any sense. If a person has a calculator, that person works with the machine. People do what they do well: thinking forensically, generating phenomenal data, testifying in court. Calculators do what they do well: calculating on what is basically a thousand dimensional problem. It is a miracle people can do anything with these mixture problems at all.

(Next slide)

For more information, at our website there is an entire noncommercial section of courses, newsletters, presentations, and publications. Most everything we write is up there. Thank you very much.