Transcript of Dr. Mark Perlin's talk on "Explaining the likelihood ratio in DNA mixture interpretation" delivered on 14 October 2010 in San Antonio, TX at the 21st International Symposium on Human Identification.

Dr. Perlin: Good morning. Today I will be talking about explaining the likelihood ratio (LR) idea and DNA mixture interpretation. The goal is to demystify things a little bit.

(Next Slide)

What is the likelihood ratio? Well, the likelihood ratio is many things in most fields across biological, physical, and social science. The LR is a standard measure of information. It is a single number that summarizes the data support for a hypothesis. It is a way of accounting for all of the evidence in favor or against a particular hypothesis or proposition. The LR is also the match statistic that is used in all DNA reporting, as our poster, #85, shows downstairs. It is also in many ways forensic science's credibility in court, since the LR has good legal and scientific standing. In some sense though, it tells us how the data updates our belief in a hypothesis.

(Next Slide)

A quick show of hands here might indicate that the likelihood ratio is not all that

Copyright 2003-2010 Cybergenetics

popular in the US just yet for reporting DNA. There are several reasons. One is that most forensic analysts outside of DNA do not yet have an LR available for their disciplines. Moreover, within DNA, analysts sometimes find the LR hard to explain. However, since all DNA match statistics, including inclusion, are likelihood ratios, the stronger likelihood ratios can preserve the DNA match information, while the weak ones discard match information. So, it can be very helpful to know what the differences are if our goal is to preserve information. Without a likelihood ratio, one can misreport DNA that is highly informative as inconclusive. We want to find a better way to explain the likelihood ratio so that everybody in the US can become comfortable with LRs and start using them.

(Next Slide)

Here is a quick and largely British history of the likelihood ratio. It was in the 18th century that the Reverend Thomas Bayes first came up with the idea of updating one's beliefs or probability based on data. He showed how to use evidence to revise our belief in a hypothesis. In the 1940s, Alan Turing, the father of computer science, used likelihood ratios for the Enigma code breaking project. Jack Good, a statistician who worked with Alan Turing, ushered the LR into mainstream scientific thought with his classic book, Probability and the Weighing of Evidence. Chapter 6 of that book is a fantastic read that was written primarily for scientists, not statisticians. In the 1970s, Dennis Lindley, a Bayesian statistician in England, introduced likelihood ratios into forensic science in a

Copyright 2003-2010 Cybergenetics

rigorous way, starting with glass evidence. His Understanding Uncertainty is an amazing book written for lawyers, judges, and nonscientists. Many non-specialists have found it to be a fantastic introduction to probability. Over the last two decades, John Buckleton, Ian Evett, Bruce Weir, and others have brought the likelihood ratio into the interpretation of DNA and mixtures.

(Next Slide)

I am going to quickly describe four different forms of the likelihood ratio. Each one has a different mathematical justification, but we will ignore the math for this talk. I will not show any math proofs here, but rather leave them for an Appendix. These four different forms provide four equivalent ways of stating the same likelihood ratio.

Here is the original form that came out sixty years ago in Jack Good's book, The Weighing of Evidence. This is the *hypothesis* form of the LR. It focuses on the identification hypothesis, which for DNA is that "the suspect contributed to the evidence." We start off with our prior belief about that hypothesis, which is based on random people in a population *before* we examine any data in a case. *After* we have seen the data, we update our hypothesis. Now, we look at the odds of the identification hypothesis given that we have seen the data relative to what we knew before. In other words, what was the information gain based on the data? Let us state this in English. Suppose that the factor was a billion. This is the

Copyright 2003-2010 Cybergenetics

likelihood ratio number that I will use throughout this talk's examples. We could say "the evidence increased our belief that the suspect contributed to the DNA by a factor of a billion." This sentence sounds like English.

(Next Slide)

Now, this next formulation is our problem. Those of us looking at this LR expression are breaking into cold sweats knowing that there are some who are very happy with it but many others who are not. This is an LR form that supposes that there is an *alternative* hypothesis that someone else contributed to the evidence. We then do what statisticians and computers love to do – we contrast the two hypotheses using easy (for them) mathematics. The likelihood ratio is thus the probability of the data given the identification hypothesis divided by the probability of the data given the alternative. In English, we might say something like this: "the probability of observing the evidence assuming that the suspect contributed to the DNA is a billion times greater than the probability of observing the evidence assuming that someone else was the contributor." Those are a lot of words for an American. The British are really comfortable with sentences that long, but in the United States, we cannot expect a jury with members who may be innumerate or math phobic and are used to short sound bites to process that sentence. Therefore, we do not generally use LRs in this country.

(Next Slide)

Copyright 2003-2010 Cybergenetics

Let us examine two more forms of the likelihood ratio that are mathematically equivalent. First, if we know what a genotype is (I will describe this visually later on), then we can state the genotype information gain at the suspect's genotype. Now, this LR *genotype* form is starting to look a bit more understandable. Since there is no conditional probability in this form, there is no way that we can transpose a conditional probability, which might lead to a testifying problem in court. The LR genotype form simply compares the probability of the evidence genotype relative to a coincidental genotype. In other words, before we saw the case data, there was a random population genotype based on the product rule. After we have seen the data, that genotype was updated by Bayes' theorem. We might read the math in English as, "at the suspect's genotype, the evidence genotype is a billion times more probable than a coincidental genotype."

(Next Slide)

Well, that is great if everyone knows what a genotype is and can perhaps show pictures. However, maybe our jury does not know what a genotype is yet, and we do not have a picture to show them. So, here is the simplest LR form that is written in plain English. It is mathematically equivalent to the original LR. This LR is the *match* form, and it addresses the question, "how much more does the suspect match the evidence than some random person?" What is the match information gain, not in the hypothesis or in the genotype, but in the DNA match?

Copyright 2003-2010 Cybergenetics

People have an intuitive idea of what it means to match. This LR compares the probability of an evidence match to a coincidental one. In English, this LR form would read as "a match between the suspect and the evidence is a billion times more probable than a coincidental match." That is plain English. I think that is even more comprehensible than our random match probability (RMP) statement for single source DNA. I am going to use this match form of the LR and illustrate it with some examples throughout the remainder of this talk.

(Next Slide)

Mixture interpretation is common in forensic DNA practice. Mixture interpretation is interesting because more than one allele pair often appears on a genotype list at each locus for a contributor. The likelihood ratio compares an evidence match relative to coincidence, but different interpretation methods yield different DNA information. Let us review three examples. Random match probability is done when we do not have a mixture or when there is a clear major contributor. We write the LR here as *one* over the probability of a coincidental match, as computed from the product rule. RMP describes the chance of seeing a random match in the population.

The "inclusion" mixture interpretation method, though, as we will see in a minute, diffuses genotype probability over many allele pair possibilities. The result is a *small* matching genotype probability relative to a coincidental match.

Copyright 2003-2010 Cybergenetics

Quantitative interpretation methods can preserve more DNA identification information by inferring a *large* matching genotype probability relative to a coincidental match. The LR is therefore greater.

Note that in all three mixture interpretation methods, the denominator of coincidental match is the same. The differences occur in the numerator with the probability of an inferred genotype based on the evidence. The data indicates how much weight to assign each allele pair, and more informative methods use more of the data to place greater probability on the correct solution.

(Next Slide)

Let us look at quantitative mixture interpretation. Here is quantitative STR mixture data at a genetic locus. The x-axis is DNA fragment length in base pairs, while the y-axis shows relative fluorescent units (*rfu*). We see two taller alleles [28,30] that might come from a major contributor and two shorter alleles [29,32.2] perhaps from a minor contributor. The question we ask here is, "what are the underlying genotypes? How can we infer the major and minor contributor genotypes at this locus?”

(Next Slide)

Copyright 2003-2010 Cybergenetics

Most DNA analysts in the US use *qualitative* thresholds. The threshold level of 50 *rfu,* shown here, is lower than a stochastic threshold, which would raise this threshold three times higher to 150 *rfu* and make the STR data disappear entirely. When we apply a threshold, we take the quantitative data and slice away the information. All of the quantitative data disappears. In this case, we are hopefully left with four all-or-none allele events that might not even be alleles at all. Forming all possible allele pairs would produce ten candidate pairings of those four "allele" events. This genotype listing diffuses the probability across ten allele pairs and reduces the likelihood ratio.

(Next Slide)

Quantitative mixture interpretation does not use thresholds or "allele" events. Rather, a computer system proposes the possible combinations of allele pairs, mixture weights, stutter, peak uncertainty, degraded DNA, etc. to generate patterns that are compared with the experimental data. In the figure, we see the same STR data (green) now with a quantitative superimposed pattern (gray) that fits well. When this quantitative pattern (e.g., assuming the major and minor genotypes and some stutter and relative amplification values and so on) explains the data very well, as in the figure, the likelihood function returns a high value. A high likelihood confers a higher probability to the underlying parameters, such as the genotype allele pair values.

Copyright 2003-2010 Cybergenetics

When we propose genotypes allele pairs or other parameters that do not explain the data well, we obtain patterns that have nothing to do with the data. These ill-fitting patterns produce a low likelihood, which give very low or no probability. Intermediate patterns, which sort of fit the data but not all that well, yield likelihoods in between.

This complete search across all parameter values that follows the laws of probability is how computers infer genotypes and other parameters. The result is a probability distribution over all possible allele pairs with probability weights that are not equal. A quantitative DNA interpretation method considers all possibilities and describes experimental uncertainty through scientific probability. A valid statistical inference is not permitted to change the observed quantitative data (e.g., no thresholds allowed), but rather it tries to explain all of it mathematically.

(Next Slide)

We will use Cybergenetics TrueAllele® Casework system in the talk examples. TrueAllele is a quantitative computer interpretation method. It does statistical search using a rich probability model with thousands of variables. The main one that we care about is the genotype. Only a genotype or probability distribution goes into the likelihood ratio. TrueAllele preserves all of the identification information in the DNA evidence. It first objectively infers genotypes without ever seeing a suspect. Only afterwards does it make any comparison with a suspect,

Copyright 2003-2010 Cybergenetics

ten suspects, or an entire country's database of possible suspects.

TrueAllele can use any number of mixture contributors, four, five, etc. The system has models for PCR stutter, allele imbalance, degraded DNA, etc. – all of the experimental components that are familiar. Most importantly, TrueAllele calculates the uncertainty of every peak. This is critical because by knowing the uncertainty around each peak, it then knows to what extent how well the genotype patterns are comparing with the data.

I gave a live demo here at Promega seven years ago that solved a two person mixture on a laptop finding the genotype in 30 seconds. Much of Cybergenetics research for five years after that presentation has been on how to calculate the uncertainty at every STR peak. Think of data uncertainty as a bell curve around each peak that describes exactly how confident we are in its height. That peak uncertainty modeling is what lets us confidently proceed with reliable genotype inference. These standard data uncertainty methods first appeared in computational statistics about twenty years ago. If a lazier computer did not bother working out the data uncertainty, then it might as well just use thresholds. It would only be guessing about data confidence and inviting a major court challenge.

The TrueAllele system was created over ten years ago. It has been in version 25 now for two years. TrueAllele has been used on over 100,000 evidence samples.

Copyright 2003-2010 Cybergenetics

The technology is offered as a product, a service, or as a combined product and service.

(Next Slide)

This landmark case, Commonwealth v. Foley, was the first time that rigorous statistical computer interpretation of DNA mixtures was admitted into evidence after a pretrial hearing and used in court. The original inclusion statistic from a national laboratory was an LR of 13,000. An independent expert's obligate allele method gave an LR of 23 million. The TrueAllele computer reported an LR of 189 billion.

How would we state this result using straightforward match LR language? First, we would state our assumptions, such as there are two contributors to the DNA mixture including the known victim. Then, in English, we would say: "A match between Mr. Foley and the fingernails is 189 billion times more probable than a coincidental match to an unrelated Caucasian." That is the likelihood ratio stated in a match form that ordinary people can understand.

(Next Slide)

Here is a recent case where I testified in a pretrial hearing over the summer in Oxford – the Queen of England versus an arsonist. The biological evidence was

Copyright 2003-2010 Cybergenetics

a low template mixture of three DNA contributors. The PCR amplifications were done in triplicate with post-PCR enhancement. Accounting for the post-PCR enhancement, the pre-enhancement peak heights would all be under 50 *rfu*. As can be seen, the three patterns are highly dissimilar because there is a lot of stochastic PCR variation. After no match score was found by human review, the British did what the British do when no man nor machine in England can solve a DNA mixture problem: they called Cybergenetics. So, we applied TrueAllele to the data examining all three amplifications in a joint likelihood function. The computer spent most of its time working out the peak uncertainty and modeling the variance distribution by trying out all possibilities.

(Next Slide)

This is perhaps the single most important picture to take home from this talk. We are looking at the same locus, vWA, from the quantitative data. The population distribution of the allele pairs based on the product rule (2pq, p2) is shown in brown. The small amount of probability at each allele pair is what we believe *before* observing the data. *After* looking at all of the quantitative data, the computer updates its genotype belief, changing its probability distribution from the population to whatever the data has indicated, as shown in blue. There was a probability gain at some allele pairs and a loss at others. The computer did not know the suspect genotype, so its inference was entirely objective. We now have the objectively inferred genotype, or allele pair probability distribution of the

Copyright 2003-2010 Cybergenetics

unknown contributor at the vWA locus. We take the suspect's allele pair [14,18] at this locus, slide a window over it (shown in red), and look only at this particular allele pair. That is what the likelihood ratio tells us to do – focus solely on the suspect's genotype. We see that the posterior genotype probability (blue) is six times higher than the prior population probability (brown). I showed this picture in a TrueAllele visual interface to the prosecutor. He then insisted on explaining the LR himself to his fellow prosecutors and police using visualizations like this at all of the other loci. He was quite successful, and I did not have to intervene. He was quite pleased with himself. He did not know the underlying science or math, but this visual form of the LR was completely obvious to him and everyone else.

The picture visualizes the genotype probability LR approach. For the match LR statement, we give our assumptions of a co-ancestry with a theta value of 1% and the presence of three contributors. In understandable English, we can now state the LR as "a match between Mr. Broughton and the fuse is 3 million times more probable than a coincidental match to an unrelated Caucasian." This LR is given plain English, and it is mathematically correct.

(Next Slide)

Scientists gather, and of course all they ever do is explain the likelihood ratio to each other, as well as other physical phenomena, such as the physics of

Copyright 2003-2010 Cybergenetics

helicopters, based on the data that they see.

(Next Slide)

LR methods vary, as Dr. John Butler from the National Institute of Standards and Technology (NIST) has shown. His classic slide from five years ago shows LR results from independent done by over 50 laboratories on a single two-contributor mixture sample. There was a range from an inclusion LR of 31 thousand (10^4) to more quantitative human interpretation of 213 trillion (10^{14}). These match scores represent over 10 orders of magnitude of very different likelihood ratios produced from identical DNA data.

(Next Slide)

Here is a study that we published last year in PLoS ONE. Dr. Margaret Kline of NIST prepared the DNA samples. The data were a series of mixture combinations (90:10, 70:30, 50:50, 30:70, and 10:90) of known genotypes. There were two different pairs of individuals, serially diluted at 1 nanogram (ng), 1/2 ng, 1/4 ng, and 1/8 ng for a total of 40 prepared mixtures.

Let us first focus on the blue scatter plot. For each point, the x-axis shows the amount of unknown culprit DNA on a logarithmic scale: 10 picograms (pg), 100 pg up to 1000 pg. We know this by multiplying the total DNA amount times the

Copyright 2003-2010 Cybergenetics

mixture weight. The y-axis, also on a log scale, shows the likelihood ratio: thousand, million, billion, trillion, quadrillion, and so on. The blue scatter plot shows the TrueAllele computer interpretation using a quantitative method where for the two-person mixture, we assume the victim and solve for the unknown genotype (probability distribution). The paper compared four different mixture interpretation methods. As we move from 1000 pg leftwards, we see that down to about 100 pg that all of the DNA match information is preserved. Then, from 100 pg leftwards down to 10 pg, there is a predictable linear decrease in LR match information. At about a million-to-one likelihood ratio level (the jury “convincing" level), the regression line crosses at 15 pg, which is a measure of TrueAllele's genotyping sensitivity.

The red scatter plot shows the LRs for an inclusion mixture interpretation method. As expected, below 150 pg inclusion no longer reaches a "convincing" LR of a million-to- one; the DNA identification information is then gone. That relative paucity of information is why, with human review methods, labs tend to not interpret evidence much below 100 pg. Computer search with rich probability models does not have such human review limitations and can reliably achieve 10x the sensitivity, reaching down to 15 pg of DNA.

(Next Slide)

This is a study that will appear next year in the Journal of Forensic Sciences. It

Copyright 2003-2010 Cybergenetics

was done collaboratively with Dr. Barry Duceman of the New York State Police. The paper shows how quantitative computer interpretation preserves likelihood ratio information, whereas qualitative human review discards identification information.

Here are eight adjudicated cases without a known victim genotype. The y-axis again shows the log likelihood ratio: 5, 10, 15 (a quadrillion to one), etc. The TrueAllele computer LR values are shown in blue for each case. Also shown are the human review inclusion LR scores (orange). These scores were the LR values reported in the case folder from the calculated CODIS combined probability of inclusion (CPI) match statistics. Comparison was made using the same population databases without co-ancestry theta correction. The LR comparisons for another eight cases, each having a known victim genotype, comparing TrueAllele with the combined likelihood ratio (CLR) method, are not shown here. They are reported in the paper. We see that, on average, the computer (blue) LRs of about 10 trillion (10^{13}) preserve identification information, but human inclusion review of the same case mixture data (orange) averages only 10 million (10^{7}). Relative to quantitative TrueAllele interpretation, using thresholds typically discards a million-fold (six powers of ten) worth of DNA identification information.

All match statistics are likelihood ratios and can be explained within the same scientific framework. Therefore, the relative efficacy of mixture interpretation

Copyright 2003-2010 Cybergenetics

methods can be compared in studies, such as these, using the LR logarithm as a universal information measure.

(Next Slide)

More dramatically, Dr. Duceman and I then looked at what happened when we did not assume that human review produced any match score We just looked at all 85 mixture items, as listed on the x-axis. The y-axis again shows DNA match information measured by the log likelihood ratio. The items are sorted by decreasing match information.

The TrueAllele computer inferred match information for each item is shown as the large blue background. In the different foreground colors, RMP (gray), CLR (green), and CPI (orange), we see the human LR result for each case. Human review lost about two thirds of the information where they did get an answer. However, most importantly, fewer than 30% of the items were even assigned an LR match score. That 30% of human reviewed cases having any match score at all is an amazing statistic if we think about it. From a productivity standpoint, suppose that a lab wants to get some LR match statistic in a case with mixture items. They must keep processing items until they are lucky enough to get a match score, sort of like DNA Russian roulette. With a 70% LR failure rate, they would have to process, on average, 3.5 samples to get some LR, doing 3.5 times the work (effort, expense, time, etc.). That LR number is typically far less than

Copyright 2003-2010 Cybergenetics

what they would have obtained had they used an informative quantitative TrueAllele calculator in the first place.

(Next Slide)

The likelihood ratio also applies to investigative DNA databases. This is how quantitative LR matching works in TrueAllele or any other highly informative DNA investigative database.

The allele database approach discards information. Its information handling is just like CPI. In fact, it simply lists "included" alleles to give a very weak representation of the genotype probability distribution. Some identification exists there, but it is not very informative. Like inclusion, an allele-based DNA database makes very poor use of the data. The CODIS-like approach can only store and match information-poor genotypes. But a probabilistic genotype database preserves more of the DNA evidence information. It can store and match those genotypes as probability distributions. When a new convicted offender or evidence genotype comes in, a likelihood ratio is then computed. This information-rich DNA database approach exploits the sensitivity and specificity that likelihood ratios are known for throughout science.

The TrueAllele system provides for this LR-based evidence versus suspect (e.g., convicted offender) genotype match. When Cybergenetics reanalyzed the World

Copyright 2003-2010 Cybergenetics

Trade Center data, we showed how the TrueAllele LR database could be used for disaster victim identification.

The same probabilistic genotyping and LR methods can be used with kinship to find missing people. In fact, TrueAllele can completely automate familial search for a laboratory with no human involvement or additional costs. Here, "automate" means not having people do anything except letting the computer completely solve the DNA matches for them in the background. Cybergenetics can set up customized LR-based DNA database matching for each state and each country in accordance with the laws and regulations of their jurisdiction.

(Next Slide)

The 2010 SWGDAM DNA mixture interpretation guidelines provide for reliable scientific computing in paragraph 3.2.2. The paragraph essentially says that a stochastic threshold is not necessary when using a validated probabilistic genotype method. That provision can help make the most of our DNA evidence data. Many labs already throw out half of their DNA mixture evidence because they cannot put a match number to them. However, using stochastic thresholds raises the qualitative peak cutoff, as many labs have recently observed. Higher thresholds discard more peak data, and so fewer evidence items can be reported with a match statistic. With low-level mixtures, such as property crimes, the information yield becomes even less. SWGDAM now lets labs use

Copyright 2003-2010 Cybergenetics

probability modeling to preserve DNA identification information. Moreover, we can measure the efficacy improvement because all match statistics are likelihood ratios.

(Next Slide)

What is the point of forensic science? Why do we do it in the first place? Certainly, the public and the police view is that we want to preserve all of the DNA identification information that is present in the data. We want to provide an accurate DNA match result. If the true match number is a trillion to one, we want to report a trillion to one and not a million to one. If the true number is a million to one, we do not want to say the data are inconclusive. The point is to serve the criminal justice system for law enforcement and the courts in an objective way to help protect the public from crime.

Forensic DNA science is not about making DNA labs and analysts feel comfortable with their methods. The purpose is to bestow the benefits of science on the public, much as any doctor or engineer would. Scientist professionals are concerned primarily about the safety of society and doing the most accurate job that they can.

The likelihood ratio is an essential tool for preserving and accurately presenting DNA match evidence. American forensic scientists want to communicate with

Copyright 2003-2010 Cybergenetics

their triers of fact in the English language, not with conditional probability, which risks transposed conditionals, or with arcane sentences that stroll on for 50 words. A solution is simply this: "A match between the suspect and the evidence is a billion times more probable than a coincidental match."

(Next Slide)

In conclusion, we can rearrange the likelihood ratio into many different mathematical forms. Many analysts would prefer to not talk about the "probability of data" ratio that works so wonderfully for computers and statisticians but can appear opaque to ordinary people. Fortunately, the likelihood ratio is easy to understand and easy to explain in court. I do it all of the time and show lawyers how to explain it to their friends. The approach described here is to state the LR in an appropriate form. The match form is particularly accessible. This LR always has the same denominator for coincidental match, just like in the RMP. In the numerator, we find the strength of match between the evidence and suspect. This match strength decreases with weaker interpretation methods (e.g., inclusion), and it stays high and is preserved with more informative likelihood ratio methods (e.g., TrueAllele). When we take the ratio of these two probabilities, the DNA identification information is preserved. That preservation of evidence is one scientist's view of the primary purpose of forensic science.

The handout slides for this talk can be downloaded from the Presentations page

Copyright 2003-2010 Cybergenetics

on our website (http://www.cybgen.com/information/presentations.shtml). Actually, on that page for every talk we present, we try to have the handout, along with a transcript of the talk and a narrated movie of the slides. This dissemination lets you see it again, tell your friends who were not at the meeting, or show it for continuing education when you get back to your laboratory.

If you are interested in reading our scientific papers, you can download the manuscripts from our website as well. If you want to learn more about quantitative TrueAllele interpretation and LR reporting, you can send Cybergenetics some interesting case data for TrueAllele processing (at no charge) and a follow up customized webinar. If you need further information, please email me (perlin@cybgen.com), and I would be happy to answer your questions. Thank you very much.

Copyright 2003-2010 Cybergenetics