Transcript of Dr. Mark Perlin's talk on "Sherlock Holmes and the DNA Likelihood Ratio" delivered on 25 February 2011 in Chicago, IL at the American Academy of Forensic Sciences 63rd annual meeting.

Moderator: Dr. Mark Perlin's research centers on computer interpretation of DNA evidence. His company, Cybergenetics, develops the courtroom-tested TrueAllele® system for inferring probabilistic genotypes and matching them. Today, he will be talking about "Sherlock Holmes and the DNA Likelihood Ratio".

(Next Slide)

Dr. Perlin: How many people here have read Hound of the Baskervilles at some point in their life? As can be keenly remembered, in chapter 4 there was a letter that was received with little glued-on, cut-out typed words that said "as you value your life or your reason keep away from the moor." As Sir Henry Baskerville's group is trying to figure out what is going on, Sherlock Holmes picks up a newspaper, begins reading from the financial section, and announces that the words were cut from the front page, pointing to the sentence.

(Next Slide)

Astounding. The party was totally baffled. Dr. Mortimer, a friend of Sir Henry Baskerville, asked if this was just guesswork. Sherlock Holmes replied that it was a balance of probabilities. Holmes then looked up at the written address shown here and claimed that the address was almost certainly written in a hotel. Dr. Mortimer protested, "How in the world can you say that?" That is the topic of this talk. The answer is a balance of probability, as Sherlock Holmes had asserted.

(Next Slide)

What is our belief before observing data? On the left, we have a hypothesis shown as a light blue rectangle, labeled "H" for hypothesis. The alternative hypothesis (darker blue) is labeled "A" for alternative. Between the two hypotheses, we cover all possibilities (exhaustive), and there is no overlap (disjoint). The odds of this hypothesis is the probability of H divided by the probability of A. In this talk, probability will just be the area of the shape we see. So let us use the word "area" instead of probability. The hypothesis odds is then the area of the left side divided by the area of the right side.

(Next Slide)

Now we observe data. We want to know what our belief is after having observed the data. Once we have data, our world is now confined to the blue data ellipse. The region outside the blue ellipse is not part of the data we have observed, and we must ignore it because those possibilities are not part of our evidence. Thus, the odds of the hypothesis given the data is now the area of the left part of the ellipse (inside the light rectangle) divided by the area of the right part of the ellipse. Those odds form a bigger number than before we saw the data because there is relatively more ellipse area on the left than there is on the right.

(Next Slide)

That change in odds is a likelihood ratio, or the information gained by observing data. So, what is Sherlock Holmes thinking? Well, he knows that the likelihood ratio is the odds of our hypothesis after we have seen the data (the numerator) divided by the odds of the hypothesis before we have seen the data (the denominator). We can easily calculate that information gain ratio since we just saw what those expressions were. The numerator is all in the blue ellipse data region: the area of the left part of the ellipse divided by the area of right part of the ellipse. The denominator is what we believed before seeing data, which is the area of the left rectangle (the hypothesis) divided by the area of the right rectangle (the alternative). That is straightforward. There is no likelihood here yet. But if we look at the right side of the equation (on the bottom right of the slide), we have one thing divided by another divided by something else divided by something else. If we simply rearrange the order of the division operations …

(Next Slide)

… some arithmetic magic happens. Now, let us look at this picture in a different way. Instead of looking left to right, let us look from top to bottom. We see that the new numerator is the area of the blue ellipse on the left, "H & D", divided by the left hypothesis rectangle, H. That ratio is a conditional probability, or one area contained inside another. This numerator ratio is the probability of observing the data, D, assuming a particular hypothesis, H. This particular type of probability is called a likelihood, where the data is fixed, but we change the hypothesis assumption. In the denominator, we have a different conditional probability: the probability of the data, D, now conditioned on the alternative hypothesis, A. If we look at the picture on the right, we see that little area of the ellipse "A & D", the alternative and the data, is embedded inside the alternative rectangle, A. What fraction is the right ellipse region out of the total area of the rectangle on the right? Visually, this is what a likelihood looks like. The little algebra division trick lets us view the odds ratio as a ratio of likelihoods. It turns out that likelihoods are very easy to calculate and have some useful information properties. We will now see what Sherlock Holmes computed, and how he used this information.

(Next Slide)

Let us return to the text of our story. We will go through and identify the color-coded clauses we need from the author's paragraph. As we examine the data, please keep in mind that the "pens" used in those times were fountain pens, not the less troublesome ballpoints that we use today. In blue, Sherlock says that the "pen has spluttered twice in a single word." We record that fact under "Data". In brown, he also says that the pen has "run dry three times in a short address." That is more evidence for us to record under "Data". In orange, the detective says that a "private pen or ink bottle is seldom allowed to be in such a state", which forms our alternative hypothesis: the letter was not written in a hotel. But in blue, with "hotel ink and the hotel pen, where it is rare to get anything else," we form our main hypothesis that the letter was written in a hotel. We have just copied out what he said and restated it into the logical forms of hypothesis, alternative, and data. Mr. Holmes also talks about how to combine evidence, which is very important in Victorian (or modern) forensics, noting that the "combination of the two" must be quite rare. There is also an action that he is considering taking: spending money out of his own pocket to hire a boy to go around and check all of the waste paper baskets in Charing Cross hotels. He asks himself: is the likelihood ratio high enough to warrant this expenditure? In the real world, we typically use probability to help us decide on some action that we might want to take. That is, science in service to making decisions.

(Next Slide)

Now let us again look at the pen data and their likelihoods. The pen data is below with two splutters having red arrows beside them. The hypothesis is that the letter was "written in a hotel"; the alternative is that it was "privately written". The data is that the "pen spluttered". We see this event happen twice in the one word "Northumberland" in the address of this letter. We now form a likelihood – the conditional probability of the fixed data given a variable hypothesis. Here, the likelihood is the probability of the pen splutter data given the possible hypothesis that it was written in a hotel. What likelihood does Sherlock Holmes give? He says that "it is rare to get anything else", which could be as much as 90%. Since some forensic scientists like to be "conservative", let us use a smaller value of 50%. Now, we look at the denominator likelihood. What is the probability of the pen’s splutter data assuming the alternative hypothesis that it is a private pen (i.e., was not written in a hotel). Mr. Holmes says that it is "seldom allowed to be in such a state". We will say "seldom" means perhaps 10%.

(Next Slide)

Combining these likelihood numbers into a ratio of 50% divided by 10%, we find a likelihood ratio of 5. If we had used a higher numerator value of 90% and divided that by 10%, then we would have gotten a likelihood ratio of 9. In this range of likelihoods, we are going to calculate ratios around 4, 5, 9, or 10, no matter what specific values we put in. We have the pen likelihood ratio. What about the other data?

(Next Slide)

We also have the ink data and roughly know their likelihoods. In the slide, we see that in the address data on the bottom, there are three arrows in places where the script lettering is broken because the ink ran dry. Assume the hypothesis that the letter was written in a hotel. We want the likelihood, or the probability of the observed data that the ink ran dry, under this hotel hypothesis. Holmes says that it is "rare to get anything else", which is at least 50%. Next, what is the conditional probability of that fixed data (the ink ran dry) assuming the alternative hypothesis that the address was written using a private pen? Holmes said that private pens are "seldom allowed to be in such a state", so the likelihood may be around 10%.

(Next Slide)

The likelihood ratio with this data is a ratio of a likelihood of the hotel hypothesis divided by the likelihood of the private hypothesis. The likelihood of 50% over the likelihood 10%, as we see on the bottom of the slide, is 5. So, now we have our likelihood ratios for both data events – pen splutter and the ink running dry.

(Next Slide)

To get the most information out of our data, we have to combine evidence. We never hear a prosecutor summing up with just one piece of evidence; usually there are a lot more events are going on. Nor do scientists typically report on just one piece of data. They form averages or other combinations. A detective combines independent events by multiplying their individual likelihood ratios together to form a joint likelihood ratio. We estimate that for a single splutter data event (shown in blue), the likelihood ratio, LRS, of the hotel hypothesis relative to its private alternative is 5. Similarly, for a running dry data observation (shown in brown), the likelihood ratio, LRR, support for the hotel hypothesis relative to its alternative is about 5. At the bottom of the slide, we see that the likelihood ratio in favor of the hotel hypothesis is a combination of all of the evidence. In blue, we combine the splutter data likelihood ratios – it spluttered twice, so we get 5 x 5. In brown, the pen ran dry three times in the word Northumberland, and so we multiply its likelihood ratio by itself three times. The result is 5 times itself 5 times, which is over 3,000 (or 3,125, to be precise). If we had believed that each event provided a stronger likelihood ratio, perhaps 10 if we were less conservative, the joint likelihood ratio would have been 100,000 (10 times itself 5 times). If we had thought that each observation's likelihood ratio had updated our belief in the hypothesis by a smaller amount, say by 4, then the joint value would have been over 1,000 (4 times itself 5 times is 1,024). Regardless, Sherlock Holmes decided that the likelihood ratio showed evidential support for the hypothesis that the letter was written in a local hotel. So, based on this sufficiently large likelihood ratio, he hired a boy to check all the Charing Cross hotel waste paper baskets for newspaper cuttings.

(Next Slide)

Let us turn now to DNA mixtures. DNA mixture evidence is interpreted exactly the same way as the envelope's address. We begin with quantitative STR (short tandem repeat) data, and following the laws of probability and valid inference, we are not allowed to touch that data – no thresholds, no modification, nothing. Holmes would not have changed anything, and neither can we. The STR peak heights are proportional to the DNA amount. The likelihood function is how we explain the data under each alternative hypothesis, which in this case is the different genotype possibilities. The joint likelihood within a locus lets us scientifically combine information in a valid way, as it has been done since the Victorian era, by combining the independent data that we have for each locus about the genotypes. Once we have inferred our genotypes, we can then compute a likelihood ratio of the support for a match to a particular suspect relative to the alternative that it is someone else in the population. Once we have finished with the likelihood ratio at one locus, we can then find a joint likelihood ratio (at least for autosomal DNA) by multiplying the independent locus likelihood ratios to derive a joint match statistic.

(Next Slide)

We now look at a DNA example from a case where I testified in a pretrial hearing last summer, the Queen versus Mel Broughton, an alleged (now convicted) arsonist. On the right, we see some data at the vWA locus. The data showed a low template mixture, and there were presumed to be three contributors to the DNA item. The Cellmark lab (Abingdon, UK) performed triplicate PCR (polymerase chain reaction) amplifications. They then did post-PCR enhancement. So, those tall peaks of 200 *rfu* (relative fluorescence units) that we see here were originally more like 10 or 20 *rfu* in the pre-enhanced data. With the enhancement, we can see some dissimilarity between the three STR peak patterns at the same locus from the same item. Human data review did not produce a DNA match score. However, the computer could infer a result. First, TrueAllele computation produced a likelihood function by examining every possible combination of the three contributor genotypes (i.e., all possible triples of allele pairs) along with PCR stutter, relative amplification, peak uncertainty, etc. The computer model explored thousands of variables, trying to explain the observed data with predicted patterns. Because the data are derived from independent experiment repetitions, a joint likelihood function yields far more information and sharpens the inferred genotype's probability distribution. Thus, the likelihood ratio computed for locus vWA was 6. And, as we saw from Hound of the Baskervilles example, if we multiply 10 independent loci with numbers like 6, we can end up with a large valid joint likelihood ratio across all the loci. The joint LR in this case was over three and a half million. We used the same joint inference principles as Holmes but applied them instead to DNA evidence.

(Next Slide)

Sherlock Holmes used likelihood inference, and the Victorian scientist found the likelihood ratio to be a very reliable way of solving crimes. Since the Hound the

Baskervilles was written in 1904, likelihood ratios have had a long-standing general acceptance in the forensic community. The likelihood principle applies to DNA evidence. With any scientific inference, likelihoods must apply exploring the probability of the data under different hypotheses. Using quantitative likelihood functions enables scientific combination of DNA evidence. A joint likelihood function examines all of the data simultaneously, which is different than a consensus approach. With quantitative data and valid statistics, the math that lets us multiply the loci together also lets us simultaneously look at multiple amplifications or multiple items. Clearly, Sherlock Holmes would not choose to only look at each piece of data in isolation. With five items available, he would have to examine them jointly since he was a really good scientist. If you would like to read more about likelihood ratios, please see my paper from this fall's Promega meeting, which is online at both the conference and Cybergenetics websites. The narrated movie presentation of that talk and this talk are also on our website. If you have an interesting DNA case where you think a solid quantitative likelihood ratio analysis with probabilistic genotypes might be informative, please send me an email. I would be happy to take a look at it with TrueAllele and show how we can make the transition from the Victorian era to the 21st century. Thank you.