Transcript of Dr. Mark Perlin's talk on "Casework Validation of Genetic Calculator Mixture Interpretation" delivered on 25 February 2010 in Seattle, WA at the 62nd Annual Scientific meeting of the American Academy of Forensic Sciences.

Dr. Perlin: Today I would like to first talk about casework validation of genetic calculator mixture interpretation. This work was done in collaboration with Dr. Barry Duceman at the New York State Police. For financial disclosure, I work at Cybergenetics, which produces DNA interpretation technology including the TrueAllele® program, which I will be mentioning because that is what we validated.

(Next Slide)

There are many ways to look at a likelihood ratio (LR). Before I begin, I want to state a few things. Human interpretation of DNA and computer interpretation of DNA can be viewed as being essentially identical processes with maybe one difference. I am going to be describing it that way and pointing out that difference as a data analysis rule. Moreover, the likelihood ratio can be written down in at least 20 different ways. Instead of using the form that Adele wrote down in the last talk that is used by many people, I am going to use Jack Good's original LR definition from 1950, which is *information gain*. The forms are mathematically equivalent, but I find this one conceptually easier. The information gain, or likelihood ratio, is what we learn about an identification hypothesis by examining

data. Before we look at the data, we have initial odds on our hypothesis or event, such as somebody is guilty or contributed to the evidence. Then, we look at our DNA data and arrive at a posterior state afterwards. Now, we have new information. We ask, what are the odds of that hypothesis *after* we have seen the data? That ratio of the hypothesis odds after and before is the information gain or likelihood ratio. In statistics and other sciences, we take the logarithm, which is the order of magnitude or powers of 10, because we can then add together likelihood ratio information.

(Next Slide)

Where do mixtures come from? Well, suppose we have a first contributor (in blue), and there is a lot of him. Then, in orange, we have a second contributor. These colored bars represent their alleles at some locus. There is less of one than the other. We add a lot of A to a little bit of B, amplify them, and get a data pattern. That is mixture data with the genotypes of contributors A and B. The data are independent of our theory. They are just what we observe.

(Next Slide)

Let me go through mixture interpretation, for both quantitative and qualitative review (computer or human). There are only two steps. The first one is that we *infer* a genotype only from the data without ever looking at a suspect to produce

this genotype. The second step is to make a comparison against one, ten, or, if we are using CODIS, a million potential suspects and compute a strength of match for a likelihood ratio.

Let us look at quantitative mixture interpretation. Every step is the same as what we all do for a human review except for the data analysis rule. Step one: infer a genotype. We consider every possible allele pair. We are required to do that in order to make a valid statistical inference. We then compare the hypothesized pattern that we form from that pair with the DNA data. Here is an example. Suppose that we have a lot of the first contributor (blue), a little bit of the second contributor (orange), and we add them together to predict the shape of a peak height pattern. We compare this predicted pattern with the data following the rule "a better fit's more likely it." This is used to replace a lot of mathematics in this talk. It is not the best rhyme, but it fits in four beats and is good enough to stand in for equations. Assuming a major contributor and a minor contributor, we see their corresponding alleles – two big ones and two little ones. These genotype values have a high likelihood because "a better fit's more likely it."

Suppose we try assuming that there are only three alleles. Maybe the orange shares one with the blue. Then, we try to fit that hypothesized pattern to the quantitative data peaks. That pattern does not fit the quantitative data as well, and so its genotype has a lower likelihood. Depending on the problem, it could have a low likelihood or almost zero likelihood.

The result of trying out every genotype possibility, comparing it against the data, looking at its likelihood, and then using statistical reasoning (like Bayes' Theorem) is that for every one of the 100 possible allele pairs there is a probability. If the data are very definite, then there will be one answer with a probability of one. If there is more uncertainty, then we will typically get some allele pairs with larger probability and some with smaller probability. There is no statement as to what is right. We do not know the suspect genotype. We do not know what right means. We only know what the data can tell us.

What the data can tell us is a *genotype*. A genotype is a probability distribution over allele pairs with a few high probability winners and a lot of low probability losers. We will see in a minute how human review shares these principles.

(Next Slide)

In step two, we match our inferred genotype against some other genotype, and we look at the information gain. At the suspect's genotype allele pair we ask, "what is the information gain at that locus?" Again, before we looked at the data, what was the prior probability of that allele pair? The population distribution is what we would believe about a genotype if we had studied genetics. We then look at the data and get an after (or posterior) probability of that allele pair based on the data. That ratio is one of the many forms of writing down a likelihood ratio.

Notice that this is a guaranteed objective procedure because the computer is programmed to have no choice but to complete step one without looking at a suspect and only afterwards do step two.

(Next Slide)

Let us take a look at our cases. Jamie Belrose will be speaking two talks from now about the larger scale validation study from which these data were drawn. Here we look at where we had a match score. We are going to look at two-person mixtures. First, assume that we do not have a victim reference. Therefore, in these eight cases, the task is to solve for the two unknown genotypes. The TrueAllele computer infers them, stops, and then makes a comparison to determine the strength of match. The computer's average match score on these eight cases was 10^{13} (i.e. 13 log units) or about 10 trillion.

(Next Slide)

What is qualitative human review going to do? Well, it is the same idea. We are first going to infer a genotype and then match it. Again, we have to consider all allele pairs and make all data comparisons. The data comparison rule is changed for simplicity. Instead of the computer's "a better fit's more likely it" using the quantitative peaks, a person follows "every pair gets equal share." Any included allele pair is considered to have the same likelihood whether it exactly matches

the pattern or it does not match the pattern. The orange alleles shown are included either way, so the result is equal likelihoods. In this case, there are four alleles, which gives 10 allele pairs all weighted equally.

The human inference mechanism is valid and uses the same principles, but regardless of the true answer, human review does not place bets on what the data says is most likely. Rather, their review places genotype bets equally. That is the inclusion method. So, now, the match uses lower genotype probabilities than we had before. We thus expect a lower probability ratio and lower information gain. Indeed, we see a lower likelihood ratio.

(Next Slide)

Human review with CPI is shown on the same eight cases (in orange). Instead of an average of 10^{13}, we see an average of 10^{7}, or about 10 million. The average per-item computer information improvement (in red) was about 10^{6}, or about a million to one. This is the same data with an identical approach to interpreting it. The procedure may not be what exactly what everyone does, but the principles are identical. The only interpretation difference is in the computer's use of quantitative peak heights for a more informative likelihood function. Both interpretation methods were looking for two unknown genotypes.

(Next Slide)

Let us look at the amount of identification information relative to mixture weight. We see on a 50:50 mixture that there was a dip. Therefore, we only gain about a factor of a thousand (not maybe 10 billion or so). This is because we lose the deconvolving weight information to separate out genotypes with the 50:50 contributors.

(Next Slide)

How reproducible is this interpretation method? Well, with the computer we can run a statistical program twice. On duplicate runs, we can calculate a standard deviation, which was on the order of a few tenths of a log likelihood ratio unit. That is far less than the population sample variation that we observe with likelihood ratios. So the computer mixture interpretation method is quite reproducible.

(Next Slide)

We conducted a similar study on another eight case items. This time the victim was known in these two-person mixtures. Therefore, the task was different: infer one unknown contributor. It was done manually with combined likelihood ratio (CLR) using the victim profile. The victim's genotype supplies more information. For each case, each bar represents the log likelihood ratio. The average

computer inferred information was 10^{17}, which is 100 quadrillion.

(Next Slide)

Comparing with human review, we see that the computer gets more information. In every case, we see the CLR (orange) that New York State reported. We find an identification information improvement in every case. The overall per-item information gain was $10^{4.6}$, or about 50 thousand.

(Next Slide)

The information loss with 50:50 mixtures was not seen anymore because we have the extra information of the victim's genotype.

(Next Slide)

The computer mixture interpretation method was reproducible. This time it was reproducible on the order of a standard deviation measuring hundredths of a log likelihood ratio unit.

(Next Slide)

Here is a summary, in numbers and words, of the validation study. We used a

quantitative computer method (better fit's more likely it) relative to qualitative human review (every pair gets equal share) and looked at the improvement. With two unknowns (without using the victim) the human statistic is CPI, and the computer is solving for two genotypes. The computer averages 10^{13} (or about 10 trillion). However, without using quantitative data, a person averages only 10^{7} (or about 10 million). This information ratio gives the improvement, which was about a million.

When we have the victim genotype, we tend to get more match information because there is less uncertainty, and the task is easier – solve for only one unknown genotype. The match numbers are higher, about a hundred quadrillion. Human review improves also. The information difference between computer and human was on the order of 50 thousand. It is also interesting to compare, as we will see in the next talk, computer interpretation with one unknown having the victim relative to CPI human review. The 17 minus 7 log(LR) information difference was 10 orders of magnitude, or about 10 billion. That large information gain is what to expect when using these different mixture interpretation methods.

(Next Slide)

In conclusion, the information gain, or likelihood ratio, is a universal DNA metric that can be applied to any way of interpreting mixtures. Implicitly or explicitly, one infers a genotype as a probability distribution. We plug this genotype into a

likelihood ratio formula and obtain the match score. We looked at efficacy and saw that the computer always extracted useful information. We looked at the computer improvements and saw that with the victim known, inferring one unknown gave an average 50,000-fold information gain relative to human review. With a two unknown genotypes, the computer showed a million-fold information gain relative to human inclusion review. The computer's mixture interpretation method was reproducible within tenths or hundredths of a log(LR) unit. The computer procedure is objective. Some have proposed sequential unmasking, but a computer inherently does parallel unmasking. It does not know anything about the other data that we are looking at. It does not know about suspects when interpreting evidence. It only knows the data we are giving it. So, when the computer infers evidence genotypes, it can only look at the evidence data. When it infers the suspects, it only sees the suspect data. All of these genotypes are computed in parallel (at the same time) and then matched afterwards. TrueAllele DNA inference is inherently objective.

In the larger study, New York State reported one match statistic for roughly every three items, whereas the computer reported a usable statistic on every item. So, if the concern is about reagent costs or time, then using quantitative review might be getting more results faster. Our study results have applicability in *scientific studies* of different methods, *investigative* searching of suspect databases, and for *evidentiary* court reporting, as we will see next.

(Next Slide)

I would like to thank the staff at Cybergenetics, Jamie Belrose from NERFI, and the New York State Police examiners who identified these cases. There are no endorsements expressed or implied by any organization. We have written this paper up as a long but interesting manuscript where all of the math is in the Appendix, which provides many examples, descriptions, and results. If you would like a copy, please send me an e-mail at perlin@cybgen.com expressing your interest and I will send you a manuscript. Thank you.

