Transcript of Dr. Mark Perlin's talk on "Investigative DNA Databases that Preserve Identification Information" delivered on 23 February 2012 in Atlanta, GA at the American Academy of Forensic Sciences 64th annual meeting.

Dr. Perlin: Today I will be talking about investigative DNA databases and how they can be used to preserve identification information. This study was done using a TrueAllele® Casework investigative database.

(Next Slide)

DNA uncertainty occurs in much evidence, such as mixtures, low DNA templates, kinship, and stochastic effects, but these data still contain considerable information. The slide shows a mixture of two people (three allele peaks). Our goal is to preserve the information that the mixture contains, and then use it to make identifications.

(Next Slide)

The key idea is the probabilistic genotype. We know there is uncertainty, but that genotype uncertainty can be expressed through probability. There are dozens of possible alleles at locus CSF, hence a hundred or so possible allele pairs in the population for CSF. The evidence at this locus indicates that only some of these allele pairs are feasible. At the five allele pairs shown, more probability is

concentrated at two of the allele pairs, [10,12] and [12,12], than at the others. This concentration of probability, reducing the possibilities and increasing probabilities where the data indicate, is how information is expressed.

Probability and probabilistic genotypes have a long history. Probability goes back at least 200 years to Laplace. Gregor Mendel used probabilistic genotypes with his Punnett squares and pea plant experiments 150 years ago. More recently, SWGDAM and ANSI/NIST standards have accepted probabilistic genotypes as representations. The cited paper describes a validation of the TrueAllele® Casework system, and how probabilistic genotypes can be determined using a computer.

(Next Slide)

In order to measure the identification information that has been preserved in the genotype, likelihood ratios are used. What a likelihood ratio (or match statistic) shows us is the gain in genotype probability from the background distribution, shown in brown, to what has been inferred after seeing the data, shown in blue. In this particular example, that ratio of probabilities of an evidence match to a coincidental match, blue to brown, is about three. So that number is the DNA match statistic at this locus following computer review.

Notice that the computer did not know who the reference individual would be, nor

what genotype we would compare with. The computer interpretation process is completely objective. It has no concept of what the answer should be. Afterwards, we slide a cursor (red) over the allele pair of the reference we want to compare with; no other genotype value is germane. We use logarithms, which are a number's powers of 10 to measure information. For example, 10 has a log of 1, 100 has a log of 2, and so on. The log counts up the number of zeros after the leading digit.

(Next Slide)

Once the computer has inferred the probabilities, we obtain an evidence genotype. This locus row shows five probability bars for CSF. There are 15 different STR loci shown. Many of these locus rows have just one bar, indicating a probability of one. We can then make a comparison between the probabilistic evidence genotype and a reference genotype, and compute a likelihood ratio match statistic. In this case, the LR is 10^{17}, which is about a quintillion.

The concept of an investigative DNA database is to (1) infer probabilistic genotypes using computers that preserve all the information that is present in the data, and (2) measure that information quantitatively using a match statistic. The computer performs both of these steps (infer and quantify) automatically.

(Next Slide)

Why would we want to have an information preserving DNA database? It is good for investigations that can solve cold cases. We can match evidence to suspect, connect serial crimes with evidence to evidence matches, find missing people with evidence to kinship comparisons, or conduct familial searches from kinship to references. Matching remains to missing can identify disaster victims, as well. The items on the left are the probabilistic evidence and the items on the right can be reference samples or probabilistic genotypes. Our work on the World Trade Center is described in the cited article, and was mentioned today in another talk.

(Next Slide)

Our study looked at 40 mixtures. These were generated originally for an NIJ study spanning a wide range of mixture proportions from 10% to 90%, and a range of quantity of DNA from 1,000 pg down to 125 pg for two different pairs of individuals. This data was developed by Margaret Kline at NIST 10 years ago. The higher end of where there is more DNA is possibly more representative of sexual assaults and homicides. Where we see less DNA is more representative of touch applications, property crimes, terrorism, and so on.

(Next Slide)

There are two main assessment metrics, sensitivity and specificity. Sensitivity is

the ability to detect matches. In this experiment, we examine the computer-inferred genotypes of these two-contributor mixtures. First, we assume that the victim is known, as seen in a sexual assault or homicide, where the computer infers one probabilistic genotype, trying to preserve all the information that is present in the data.

Once we have this set of genotypes, we compare them against the reference genotypes. As our metric, we measure the identification information and see how much has been preserved. Again, the information measurement is the logarithm of the match statistic. 10 would mean 10^{10} or 10 billion; 20 is 10^{20}.

We see a histogram that gives a sense of the probability distribution measuring sensitivity. All 40 of the profiles genotypes appeared with a likelihood ratio greater than one, and the average information preserved was about 18, or a quintillion (10^{18}). The histogram shows that the detection capability of a system like this is excellent.

(Next Slide)

But what about the specificity? Are spurious matches made? To assess these questions, we generated 1,000 random genotypes. We put them on the reference side of the database, and then compared them with the 40 inferred mixture evidence genotypes. In red, we see a probability distribution histogram

comprised of 40,000 random matches to the wrong person.

With the random matches shown, the information scale extends to the left. Here, on the logarithmic x-axis, a -10 means one over 10 billion (10^{-10}), and so on. The red bars show the extent of evidence against the match, as opposed to evidence for a match, in blue.

The observed match statistic behavior shows specificity, the ability to reject matches that are not real matches when making a genotype comparison. The system nicely separates out the real matches with a high likelihood ratio (blue) from spurious ones (red). The false positive overlap is less than one in 10,000 genotypes, less than .01%, showing a very good separation.

(Next slide)

What about cases where we do not have a victim profile, where there are two unknown contributors? We used the same mixture study, except now we have two inferred genotypes for each of the 40 evidence samples. So, there are 80 probabilistic genotypes that were inferred, and all 80 again appear with a positive likelihood ratio. This time, the average match statistic sensitivity is somewhat reduced to 10^{15}, or only a quadrillion.

We again tested the specificity against the thousand random profiles. The red

bars on the left represent about 80,000 matches to random (i.e., incorrect) people. Again, we see that there is tremendous evidence that there is no match, and that such a person would not be detected in an investigative database. The crossover to a LR>1 is again less than .01%, or less than one in 10,000. We observe a very good separation for detecting real matches, as well as for rejecting spurious matches.

(Next slide)

Most human mixture review does not strive to preserve all the identification information. Rather, it is done to identify a match in a way that a person can. So the human method is different than the computer's, and is not information preserving. When the data are examined, instead of using all of the quantitative data, thresholds are applied which convert peaks into all or none events that (hopefully) correspond to alleles. These alleles are considered to be either present or absent, and that decision forms a list.

In the case of our CSF example, that list would be the alleles 10, 11, and 12. This list of three alleles gives six allele pairs, [10,10], [10,11], [10,12], [11,11], [11,12] and [12,12]. Note that even if we use inclusion methods and allele lists, we still obtain a probabilistic genotype. It just has not been inferred by a computer, and so does not give us the sharpest probability distribution. Instead, the human-inferred genotype diffuses the probability over many possibilities.

When we measure the inferred information from thresholds, focusing on the reference sample allele pair, we no longer see a large probability ratio. The ratio is now much closer to one, and a likelihood ratio of one corresponds to a log of zero, or no information. That is how thresholds lose identification information.

(Next slide)

Once people have generated an evidence allele list, they can put the allele list onto an investigative allele list database, such as CODIS. On the left, the slide shows one such CODIS uploadable allele list. For this study, we applied different thresholds, and had people score the data (in triplicate) for thresholds of 50, 100, 150, and 200 rfu. The mixture allele list shown used a threshold of 100, and was uploaded to the evidence part of an allele list database. This is not what our probabilistic computers infer, but it is how many crime labs currently work. We can also represent references as allele lists, as shown to the right.

A comparison then gets made, and at every locus (say with what CODIS calls “moderate stringency”), an allele set comparison can ask, “Are the alleles in the reference a subset of the mixture’s alleles?” The allele database returns a number – how many of those loci were hit. In this example, 7 of 13 hit. This mixture allele example is the one where the probabilistic genotype computer gave a match score of 10^{18}.

(Next slide)

We can measure the sensitivity and specificity of these allele list databases. The sensitivity measure is the fraction of uploadable allele lists. Based on the threshold used, this fraction can vary from 50% to 65%. On the low threshold end, there are too many alleles. With the higher thresholds, there are too few alleles. Regardless, the fraction is between 1/2 and 2/3. Thus, about a third to a half of the evidence is thrown out.

(Next slide)

We can also measure allele list data specificity. We uploaded the uploadable manually scored genotypes at different thresholds (showing the results for 150 rfu since that is common after SWGDAM 2010). When comparing against a thousand random reference genotypes, we counted up the number of alleles that were hit. In this particular mixture set (it could be different in another lab), there were a lot of low level mixtures, representative of today's DNA evidence.

From the locus distribution, we formed a tail probability distribution. This is the probability of a spurious hit against a random person, generating a false hit that would give a false lead. For example, suppose we set a criterion that when seven or more loci hit, we examine the allele list database matches. Adding up the blue

bar probabilities for the 7,8, …, 13 loci forms the error rate tail distribution. At seven loci, the false positive rate is about 35% on this data. At 8 loci, it is about a quarter of the random references, and so on. The error fraction decreases as we set the locus number threshold higher.

(Next slide)

We looked at a CODIS-like allele list database in order to compare an information preserving probabilistic genotype with current practice. Probabilistic genotype database sensitivity, the ability to detect matches, has a match score (on average) of a quadrillion, with a false negative rate on this data set of less than .01% of false hits. Allele lists (under moderate stringency) on this data show that upload failed about a third of the time, with false hits in the 5% to 25% range, depending on where we set the locus criteria.

Since a 1% spurious rate hits 10,000 profiles out of a million, less informative methods can generate a lot more work than more informative probabilistic genotype methods, whose goal is to preserve and measure all the information. This additional, unnecessary work is done by the lab in reviewing spurious hits, and by the police when they follow up on false leads.

(Next slide)

In the information age, information translates into saved work. We do not need to send police around to follow up on many false leads. That reduced work translates into saved time and money, better evidence that can be brought into court, and greater protection of the public from crime.

Fifteen years ago, in the pre-Internet era, we used to drive our cars to go shopping. We would drive from store to store, maybe looking for some particular item. Maybe a store had it in stock; maybe they did not. So, we would then go on to the next store, and (using a lot of time and energy and work and gasoline) we would visit a lot of places. Maybe we would go back to the first store because the price was lower there; but someone else had bought it. Some of you are old enough to remember this from the Stone Age of the late 20th century. But the world has changed.

We are now in the 21st century. We use the Internet. We can check out on the Internet what stores have. In fact, we can use Amazon, and not even leave our house. The information does the work for us, and the goods arrive on our doorstep. That is really the point of using computers to infer genotypes and quantify the match information. Fully informative investigative DNA databases preserve information in the ways that we now expect in the year 2012. Thank you.