Transcript of Dr. Mark Perlin's presentation on "DNA Identification as an Information Science" delivered on 20 May 2011 in New York, NY at the DCJS DNA Subcommittee.

(Next slide)

*Dr. Perlin:* Why are we interested in DNA? For evidence, DNA helps us solve crimes, match evidence to suspects, and provide a weight of evidence that there is some connection within a crime. For investigative purposes, with a DNA database of crimes and criminals, DNA can reach out across the ether into a collection of possible candidate suspects, and thereby make an identification, which can be used to match evidence to a convicted offender. For crime prevention, America has been working toward a vision that the British began about 10 years ago. This vision was nicely described in Ray Wickenheiser's 2004 article, "The Business Case for DNA", that if all evidence were processed (property crime, sexual assault, and so on), and all criminals (felonies, misdemeanors, etc.) were promptly put on DNA databases, then DNA databases could interrupt criminal careers and prevent further crime and victimization. That is our overall goal – not just apprehending and convicting criminals, but also preventing further victimization.

(Next slide)

The DNA evidence problem is that even though one might like to say something very definite in court (like "this is absolutely a match between these two items of evidence,"

 

or “an evidence and a suspect”), most DNA evidence actually is uncertain, with the result that genotypes can be uncertain. We see this with DNA mixtures, where there are two or more contributors. For example, the electropherogram data here shows four peaks; many methods of interpretation, like combined probability of inclusion (CPI), would produce a list of 10 possible allele pairs reflecting that uncertainty. DNA can be degraded or damaged. We can have low template DNA amounts, for example, here the data shows a peak height around 50 rfu. Different people might state different results. Clearly there is uncertainty in interpreting quantitative data.

(Next slide)

About 10 or 15 years ago, an interim approach (as John Buckleton wrote) of applying thresholds to the quantitative data was introduced. Thresholds have the advantages of simplicity, easy explanation, and being easy to apply. However, they do come with some drawbacks. This is what a threshold does. We pick a level, say 50 or a few 100 rfu (whatever our laboratory determines) that will be uniformly applied to data of that class. The threshold is applied, and every peak (there are some second-order decisions, but for the most part every peak) that is above the threshold is considered to be an allele and given equal treatment. Those peaks which fall below threshold are considered to not be alleles. The quantitative data are thus reduced to a series of all-or-none decisions as to what is in and what is out. The quantitative data are then not used, ignoring the information that some peak heights might be much greater than others, reflecting more DNA mass quantity.

 

Here are some well-known threshold issues. Peak heights have variation (blue), whereas with thresholds (red) an absolute value may not reflect the fact that there is vertical variation in the vicinity of a threshold. There is an issue of variance scaling. The probability models that have been published for the last 20 years, both theoretical and empirical, show that with peak variation, the variance scales linearly with the peak height. It is not a fixed value. The variation is not constant. Data have probability distributions, as we observe whenever we do replicate amplifications – the peak heights do not ever come out exactly the same because they cannot. The data are drawn from a probability distribution, and that can lead to a high false negative rate when using higher thresholds. We have done some studies on this showing that the number of missed alleles per locus can exceed 100% with imbalanced mixture weights.

Probability methods that make more use of the quantitative data (as I will show in a picture in just a few minutes) can concentrate the probability more around the allele pairs that the data supports. Not using the quantitative data, but instead treating it qualitatively, disperses the probability to other allele pairs that may be far less likely (given the quantitative data). The identification information that can be preserved using probability methods on quantitative data is discarded. These sorts of issues were nicely discussed in the ISFG article by Gill et. al. in 2006. By not preserving evidence, DNA identifications may fail instead of succeed, worsening the ability of science to protect society and reduce victimization.

 

That summarizes the interim threshold approach. The last 10 years have seen progress along the fronts of probability modeling of quantitative data, as well as measuring that information with likelihood ratios. I will give a quick historical background about what probability and information means throughout science and in DNA.

(Next slide)

We begin with probabilistic science, with a classic example from the 20th century. Newtonian physics, invented about 350 years ago, has done a fantastic job of building buildings, bombs, and all sorts of mechanistic things. Newtonian physics is one of the smashing successes of modern science. However, as experimental detection of small particles got better, and electrons were discovered, the theory did not work at that sub-atomic level the way it might have. In response to these issues, Niels Bohr proposed in 1913 an interim solution, which was that electrons live in planet-like orbitals. His interim theory helped a lot, explaining some of the early quantum effects. But it took Schrödinger, Heisenberg, and others a good deal more theory to introduce a probabilistic model, which (12 years later) in 1925, was the true solution (as far as we know), that electrons come with probabilities. They do not live in discrete orbitals. There is no threshold of where an electron might be in a particular energy level. They are just a diffuse distribution. That probability theory had some difficulty getting accepted.

(Next slide)

 

Fortunately for quantum mechanics, the theory was validated by building the atom bomb. This is because the theory of probability and probabilistic electrons was able to account for, predict, and help engineers in splitting the uranium atom. It took perhaps half of the known physicists, mathematicians, and eventual computer scientists in the Western world at that time to do this. But the atomic bombs ended the war in the Pacific, validating quantum mechanics and demonstrating that probabilistic methods worked.

(Next slide)

Interestingly, forensic DNA computing employs the same methods that were used during World War II and developed by the likes of Enrico Fermi, John von Neumann, and other physicists, mathematicians, and eventual computer scientists. They built the first digital computers during the war, as primitive as they were, and used them to make calculations. The first calculations ever done on a computer, for this sort of sophisticated scientific effort, were Monte Carlo methods, random probabilistic searching to solve very hard physics equations. The classic paper Metropolis et. al. (1951) that describes how to use probabilistic methods to solve complex, apparently deterministic, problems, is one of the most referenced papers in all of science. These sorts of Markov chain Monte Carlo (MCMC) or statistical search methods are now ubiquitously used in every field of science. If we look at the *Journal of the American Statistical Association* (*JASA*), most of the articles just assume that we are doing probability modeling, and running probabilistic search methods.

 

These computers appeared in forensics about ten years ago. Groups worked on this in Europe and in Australasia, with an MCMC method for resolving two-person mixtures published by James Curran of New Zealand in 2008. Niels Morling's group in Denmark has been doing this same type of research. Every one of these papers describes how to do probability modeling of genotypes using all the quantitative data, and representing it as multivariate normal (or gamma, or some other) distribution. These probabilistic methods, typically statistical search, search through and solve the DNA mixture equations. Probabilistic genotyping has become fairly normative science for scientists whose goal it is to extract as much information from the evidence as possible.

(Next slide)

Once we have computed a probability result, how do we quantify the information? Claude Shannon brought this together in the 1940s with his mathematical theory of communication. Today, we now take the notion of bits and bytes as something ordinary. Bits were first described by Claude Shannon, in this sort of work, as a way of measuring information, the logarithm of a number of possibilities. If an information source can be transmitted (even with noise) it can be received, and then at the destination the message can be decoded. It makes no difference whether it is digital information, computer information, DNA information, continuous telephone signals (which is what he created this for), or images such as video transmission. There is a general theory that quantifies the information in a system. By taking more time and exploiting redundancy,

 

we can preserve virtually all the information that is present.

(Next slide)

This informative theory approach was also validated during World War II, done in a dramatic way by Alan Turing, who is often called the “Father of Computer Science.” The Nobel Prize equivalent, the Turing medal, is named after him. Turing used information theory to quantify information and crack the Enigma code. The Enigma machine was a secret message coding machine that the Germans had believed was unbreakable. Alan Turing used information methods to decipher that code, constructed as hundreds of giant computers. These “Bombe” machines decoded thousands of German war messages. Intercepting and properly decrypting them contributed to winning the war in Europe and defeating the Nazis.

(Next slide)

In DNA, one of the holdovers from information theory was the likelihood ratio (LR). The LR was first described by Alan Turing's group, and used in cracking the Enigma code. The likelihood ratio is a way of directly quantifying the gain in information. What was our probability distribution before examining data? Now, what is our probability distribution afterwards? That difference is the gain in information.

With LRs, as we see in courts and in DNA, the notion of uncertainty can be quantified,

 

testified to, and made quite precise. A number like a quadrillion to one is a match statistic, a likelihood ratio. Likelihood ratios were introduced in DNA from the very beginning by many scientists. There are hundreds of papers on the likelihood ratio, its use and quantifying information difference, kinship analyses, etc. We cannot look at a journal article now where people are quantifying genotyping information from phenotypic or genetic DNA identification without seeing plots of log likelihood ratios. Recent efforts help explain the likelihood ratio in ways that DNA analysts, particularly in the US, can be comfortable with in court. Courts in England, New Zealand, Australia, often require likelihood ratios. The effort now is on how we apply these more information-preserving methods. Here again, the Gill paper states that the LR ways of reporting are more informative (and are preferred to) inclusion methods. Even though inclusion is a type of likelihood ratio, it is less informative, since it uses less of the available data.

(Next slide)

TrueAllele® is simply a system that infers probabilistic genotypes, and then matches them by quantifying the information with likelihood ratios. That is it. That is, in essence, all it does. From the user's perspective, data comes in, they ask questions, and answers come back out. TrueAllele has a large database server, and a number of interpretation processes that run in parallel to solve each problem by doing statistical search of a probability model. That probability model includes genotype, the peak heights, the weighting of the different contributors, PCR stutter, relative amplification, and (perhaps most importantly) the variation of the uncertainty around every peak in the data. The

 

data is never changed. There are no thresholds used. The observed quantitative data are the data. Leaving data unchanged is standard in modern probability calculation. But the uncertainty around the data can be calculated from the data itself within the model, and that is why thresholds are not needed. TrueAlllele® customizes the uncertainty to every peak element in the data, as opposed to using one blanket solution.

The system is parallel for a very good reason: effective statistical sampling takes time. A typical mixture might run about eight hours to sample 100,000 times from the probability distributions. Therefore, one processor can solve three items in one day. New York's computer, for example, has 16 processors, so they can solve 48 items per day, on average, for 15,000 to 20,000 items per year. Reference samples are fast, but evidence items, such as mixtures, take more time. The computer always runs in the background 24/7. Should the caseload increase and the lab need to move up to 35,000 or 50,000 items, the system is expandable. Computer modules can connect into the same database, adding more interpretation processors.

(Next slide)

The system infers genotypes up to probability and then quantifies the information with the likelihood ratio. This is mixture data from an interesting case, with three replicate amplifications of some very low template DNA with much uncertainty. As described in my Promega talk from 2010, here at the vWA locus (blue), we see the posterior genotype probability distribution. The computations were done objectively, without any

 

knowledge of a suspect or a database of suspects. Mathematically, the computer just does not know. It can only produce a posterior genotype probability distribution (blue) after it has seen the data.

How do we quantify the identification information from the genotype? There are ten different ways to mathematically write down a likelihood ratio. They are all equivalent, but one that is particularly intuitive for lawyers and juries is to describe the LR as the gain of information. This was pretty much Turing's view of the likelihood ratio.

Here (in brown) is part of the population genotype probability distribution across all allele pairs. Out of hundreds of possible allele pairs, we now note that the suspect has [14,18] as his allele pair. We did not know that fact until this instant. So now we take our slider (red), move to that point, and ask, “What was the change in probability? What is the ratio of probabilities?” The ratio of the posterior (blue) to the prior (brown), the after to the before, is six. That ratio contributes a factor of six to the likelihood ratio from the locus. Typically, we would take the logarithm of that ratio, and add the logarithms up at all ten to fifteen loci to get the exponent.

(Next slide)

To validate the reliability of any genotyping mechanism, human or computer, we first infer a genotype, that is, a probability distribution of allele pairs. Then the logarithm of the likelihood ratio is a standard measure of information used throughout science. It has

 

nothing to do with DNA in particular. The LR is used in statistics, natural language processing, and many other fields of science to measure how much information was obtained. As in our case, the LR translates all of the details (the complexities of the alleles, probability distributions of the genotype, what it is being compared to, the reference population, etc) into one number. The logarithm of that likelihood ratio number is a standard measure of information that can be used to validate reliability. Here, for example, is a slide that I showed the subcommittee a year ago with New York State TrueAllele validation results. There is a whole section on this information-based validation approach in the *JFS* paper that is appearing in November.

We validate a DNA interpretation method by measuring its information efficacy and reproducibility. Here we see eight different cases, and the information obtained from inferred genotypes via match against the known (or presumed) contributor. These cases are mixtures containing two unknown contributors, without a known victim. The y-axis here is $10^5$, $10^{10}$, $10^{15}$, and so on, showing LR information on a logarithmic scale. For each case, the computer ran a statistical search in duplicate. The average efficacy (the mean log(LR)) is $10^{13}$. As reported in the *JFS* paper, this was about a one million-fold improvement over the $10^7$ of human review. We visually see reproducibility as match numbers that are close. That can be translated mathematically into a within-item standard deviation that gives the information variation within each item. The resulting number measures interpretation reproducibility, which in this study was about a tenth of a log unit.

 