LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

LLNL-TR-572992

Validation Strategies for Microbial Forensic Analysis

S. P. Velsko

August 13, 2012

Disclaimer

This document was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor Lawrence Livermore National Security, LLC, nor any of their employees makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States government or Lawrence Livermore National Security, LLC. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or Lawrence Livermore National Security, LLC, and shall not be used for advertising or product endorsement purposes.

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Validation Strategies for Microbial Forensic Analysis of Biological Agents: Beyond Sample Matching

Stephan P. Velsko
Lawrence Livermore National Laboratory
Livermore, California
April 10, 2007

Abstract

The validation of the *probative value* of microbial forensic techniques is a critical aspect of research and development that requires careful planning and a sound conceptual framework. This paper outlines a particular approach to the validation of certain types of forensic methods that naturally generates statistical measures of the relevance and weight of the scientific evidence derived from them. The suggested approach is based on the *likelihood ratio* interpretation of Federal Rules of Evidence 401 and 402 and allows measurement evidence to be to be presented in a format that resembles the forensic "gold standard" - human DNA typing. Examples of specific genetic and chemical and physical analysis methods are used to illustrate how this general strategy can be applied. This approach also provides a natural interpretation of the notion of "preliminary validation" that has been proposed in the literature.

1. Introduction: Technical and Interpretation Error

The term "microbial forensics" encompasses a set of methods that can be used to help investigators identify the provenance of a microbial agent that may have been used in a terror or criminal incident. Microbial forensic methods include genetic characterization of the microbe, chemical and physical analysis of the agent, and other techniques. Recent publications have outlined many of the basic tenets and requirements of this field, and several government agencies support R&D programs aimed at advancing this nascent area of forensic science[1-6]. This capability is recognized as a pillar of the U.S. counter-terrorism and counter proliferation posture[7]. However, it is important to recognize that Microbial Forensics is being born in an era when forensic science *in general* is facing unprecedented skepticism and new, sometimes daunting, challenges in the courtroom. Therefore, it is critical to build a scientifically sound and transparent body of methods and knowledge that can withstand scrutiny at both the national and international levels.

In the last two decades, the reliability and relevance of expert testimony in forensic science has come under increasing scrutiny[8-10]. A recent paper has noted that forensic science testing errors and misleading testimony by expert witnesses are significant factors associated with wrongful conviction in DNA exoneration cases[11]. The 2004 NRC study of compositional analysis of bullet lead as a forensic technique is an example of a serious challenge to a previously accepted area of forensic science[12-15]. Similarly, the scientific basis for inferences about the individualizing power of hair, fibers, bitemarks, ballistics, and even certain fingerprint evidence has been questioned[8-10,16]. A substantial reason for the more skeptical attitude towards classical forensic science lies in the widespread legal acceptance and utility of human DNA forensics, which has lead to re-examination of a number of past criminal cases and the subsequent recognition that inferences based on the original non-DNA scientific evidence presented in those cases must have been flawed. In addition, *Daubert* and related precedents[17-19], as well as amendments to the Federal Rules of Evidence[20] have lead to more critical examination of the grounds for admissibility.

Although the effects of these changes have not yet been felt fully in criminal cases due to a tendency for judicial conservatism in rejecting government proffered scientific evidence, we can expect that attorneys and judges will become increasingly more sophisticated in their understanding of the grounds for challenging the admissibility and weight of scientific testimony[20]. Under U.S. code, the prosecution of biological terrorism falls squarely in the Federal domain[21], making the "general acceptance" (Frye) test for admissibility that is still used in some state courts of little relevance. A bio-terrorism trial is likely to have a high public profile, so scientific evidence will get full scrutiny by defense lawyers in both admissibility and trial phases of the case, making rigorous validation procedures that have broad acceptance within the scientific community a *de-facto* requirement. It will be increasingly difficult for an expert witness to proffer testimony as to the significance a particular piece of scientific evidence simply based on "experience." As was learned in the 1998 al-Shifa incident, the scientific evidence used to justify a national security policy decision[22] may (eventually) have to satisfy similar scrutiny in an international forum[23]. Thus, it is important that the development of a national capability in microbial forensics be underpinned with a sound scientific foundation from its start, to prevent this nascent field from being stillborn.

Regardless of these considerations, it is clearly important that scientific evidence proffered in a forensic setting be presented accurately with regard to weight and interpretation. The term "validation" is often used to refer to the process of accumulating data that speaks to the precision and accuracy of a method and the interpretation of its results. A number of forensic science areas have codified guidance for validation, and it is often cited as an essential component of the microbial forensics discipline itself[4-6]. However, it is important to understand, but not often appreciated, that there are two essentially different categories of error that must be dealt with in any validation process. In this paper, we will refer to these as *technical* and *interpretation* error.

Technical errors arise when a technique is not performed properly; the equipment is faulty, mis-calibrated, or not properly cleaned; or the reagents are impure or degraded, or there were opportunities for the original sample to become contaminated. In practice, the probability of this type of error is reduced by standard laboratory quality assurance and control measures such as rigorous and documented adherence to detailed SOPs, the use of appropriate blanks and spikes, inter-laboratory comparisons, and proficiency testing. Laboratory accreditation under ISO or ASCLD standards is another important element of technical validation regardless of the method or technique.

However, technical error is often <u>not</u> the critical issue in courtroom challenges to scientific evidence. For example, as Imwinkelried[24] has pointed out:

> "Even if an instrument yields exquisitely precise measurements, the witness's inferences from the measurements may be badly flawed. As Justice Blackmun stressed in *Daubert*, it is the expert's ultimate inference which 'must be derived by the scientific method … [and] supported by appropriate validation…'"
>
> -

Similarly, in its analysis of elemental pattern matching of bullet lead[12] the National Research Council concluded:

> "The committee found the analytical technique used is suitable and reliable for use in court, as long as FBI examiners apply it uniformly as recommended. […] However, for legal proceedings, the probative value of these findings and how the probative value is conveyed to a jury remains a critical issue."

Understanding and reducing errors in the interpretation of scientific evidence proffered in forensic settings often represents the most challenging aspect of method validation, and clearly demands special attention in light of the potential for intense scrutiny cited above. Surprisingly, textbooks and other foundational literature seldom recognize interpretive validation as a separable and formalized part of forensic science in general, and it has not yet been discussed widely in the context of microbial forensics. Therefore it is the intention of this paper to describe an approach to interpretive validation that can be

applied to a wide variety of methods relevant to the microbial forensic discipline. An earlier study applied this framework narrowly to the problem of sample matching[25].

The concept of interpretive validation is closely connected to the most basic grounds for admissibility of evidence in court. Thus, we begin the discussion by reviewing those aspects of the federal rules of evidence that form this basis. We explicitly adopt the interpretation of the relevance criteria in terms of a probability concept, namely the *likelihood ratio*[26]. Subsequently we outline a general approach for implementing a validation process that leads to measures from which a likelihood ratio can be derived in a transparent way. Following this, the approach is applied to three microbial forensic analysis examples: using genetic sequence data to support infection source determination, analysis of residual signatures of production, and identification of the origin of microbial growth media based on isotopic signatures. In conclusion we discuss the idea of "validation-on-the-fly", which has been raised previously in the context of microbial forensic method development[5].

2. Relevance and admissibility of scientific evidence

The admissibility of scientific evidence depends on reliability and *relevance.* For the purposes of this paper, it is particularly important to note that the Federal Rules of Evidence[27] define relevance in a manner that explicitly invokes a probability concept:

> Rule 401: *"Relevant evidence" means evidence having any tendency to make the existence of a fact that is of consequence to the determination of the action* <u>*more probable*</u> *or* <u>*less probable*</u> *than it would be without the evidence.*
>
> Rule 402: *Evidence that is not relevant is not admissible.*

In addition to this definition, several other rules speak to the reliability of scientific evidence and implicitly raise the issue of validation:

> Rule 702: [Expert testimony is admissible if]
> (1)The testimony is based on sufficient facts or data,
> (2)The testimony is a product of reliable principles or methods,
> (3)The witness has applied the principles and methods reliably to the facts of the case.
>
> Rule 901: [There must be foundational evidence] showing that [a scientific] process or system produces an accurate result.

NRC studies of forensic science issues have consistently advocated a *likelihood ratio* interpretation of relevance[12,28]. In this explicitly probabilistic framework, the amount of support that a piece of scientific evidence lends to a hypothesis in question is quantifiable in terms of the probabilities that the evidence would be observed if the hypothesis were true or false[29,30]. Let H be the hypothesis in question, and E be the evidence. If we

denote the odds that the hypothesis is true in the absence of the evidence by $O_0(H)$, then the odds that H is true *given* the evidence is:

$$O(H|E) = \frac{P(E|H)}{P(E|\bar{H})} O_0(H) \tag{1}$$

where P(E|H) is the probability that the evidence would obtain if H were true, and $P(E|\bar{H})$ is the probability that E would obtain if H were false. The interpretation of equation (1) is that the evidence E strengthens or weakens the odds of H being true via the likelihood ratio LR, defined as:

$$LR(E) = \frac{P(E|H)}{P(E|\bar{H})} \tag{2}$$

The likelihood ratio LR is often referred to as the weight of the evidence, or the probative value of the evidence. The important connection between the rules of evidence and the likelihood concept is that when LR(E) > 1 or LR(E) < 1, then E is <u>*relevant*</u>, whereas evidence for which LR(E) ≈ 1 is not relevant or is only weakly relevant. Excellent discussions of the use of this framework in forensics are given in several textbooks[26,31,32].

As a matter of terminology, when the likelihood ratio is larger than 1, the evidence is said to *support* the hypothesis. As a matter of convention, when LR(E) > 10 the support is *strong*, while LR(E) values of order 1 are said to *weakly* support the hypothesis. When LR(E) < 1 the evidence *does not support* the hypothesis in question. It is very important to note that LR(E) > 1 does <u>not</u> mean that the evidence makes the hypothesis in question more likely than not. This only occurs if $LR(E) \bullet O_0(H) > 1$. For many hypotheses it is difficult to estimate the prior odds $O_0(H)$, and it is not possible to make this statement on the basis of the scientific evidence being presented.

In the next section we will suggest experimental designs for validating the interpretation of microbial forensic evidence that are explicitly constructed to permit the estimation of a likelihood ratio. A key element of this approach involves defining the relevant ”population” that is associated with the type of sample being analyzed, and determining how to acquire a “representative” sampling of this population. This approach allows one to present evidence based on measurements in a statistical format that resembles the forensic “gold standard” - human DNA comparisons (although not always with the extremely large likelihood ratios encountered in human DNA forensics!) Thus, a scientist may testify that his measurement of a certain value of some metric for a sample provides a particular level of support to the hypothesis in question, rather than stating that his values are “consistent with” the hypothesis, or worse, that his results make it “likely that the hypothesis is true.” While there is debate about the best format in which to

present the statistical weight of DNA evidence in the courtroom[33,34], the general acceptance of likelihood ratio arguments in this instance strongly argue for adoption of analogous methods for other types of evidence. In many respects, the most important aspect of this approach is the change it represents in the language used to present forensic science evidence.

We note that this point of view has not yet been widely implemented for other types of forensic analytical methods applied to trace evidence and, in fact, some authors are explicitly skeptical about the general practicality of any attempt to attach explicit statistical weight to the results of such methods[35,36]. For example, Houck[35] has (correctly) stated, "Any statistical interpretation of a trace evidence finding would have to be based on a thorough knowledge of the population under study." However, he goes on to argue that for a particular branch of trace evidence analysis – fiber analysis – the population is so complex and dynamic that it is not practically possible to determine probabilities. While it may be the case that it is difficult or costly to sample large and complex populations, nonetheless there is no other way to achieve a measure of the weight that can be assigned to this kind of evidence. Moreover, there is movement towards this approach to interpretive validation in several other fields where scientific findings are used to support critical decisions, most notably in the validation of medical diagnostics[37-40]. In this paper we freely take advantage of results and ideas that have been developed in the medical diagnostics arena.

The reader should note that the treatment below makes extensive use of the statistical construct known as the "receiver-operator characteristic", or ROC curve, and some elementary Bayesian statistical formulae. No tutorial is provided in this report, but those not acquainted with the basic principles of ROC curve construction and interpretation may wish to consult reference 40.

3. A general scheme for characterizing and validating microbial forensic methods

Many of the analytical methods that have been used, or are under development, for microbial forensics fall into one of three major classes. The first class of methods is concerned with comparing two samples (either a questioned sample and a reference sample, or two questioned samples) to establish that they are related in some way. For instance, the genetic sequences of two viral isolates may be compared to decide if they could be derived from a common source or if one could have been transmitted from one person to another. Another example is the comparison of elemental profiles or isotopic fingerprints to establish that two agent samples originate from the same batch of material, or were made by the same process. In a previous paper we outlined a framework for characterizing and validating sample matching using elemental composition data[25].

A second large class of analytical procedures aims to establish that a certain material (e.g. agar, detergent, or a solvent) was used in the manufacture of the agent, or that a certain process condition was used in the manufacture of the agent (e.g. growth on a solid medium, or lyophilization.)

The third class of techniques is concerned with parameter estimation. Examples of this might be methods to establish that the agent was made within a certain timeframe relative to the time it was released, or to establish that a sample was washed a certain number of times during processing.

Hypothesis testing

For the first two classes, the purpose of the assay or analysis is to find support for a particular hypothesis, such as "the HIV virus that infected person A came directly from source B", "these two samples originated from the same batch of agent" or "this sample was made by growth on agar plates". Each hypothesis specific to a questioned sample is an example of a general relationship that may exist between two samples or between a sample and the set of possible production methods. For instance, pairs of viral isolates that are related by direct transmission from one person to another, or pairs of samples drawn from the same batch of material, or samples that were grown on agar. We will refer to the hypothesis describing this general relationship as H, and its opposite (isolates that are not related by direct transmission, were not drawn from the same batch or were not grown on agar) as $\bar{H}$. A general paradigm for evaluating and validating an assay that is formulated as a hypothesis test has five steps:

Define the sample "population", This is the set of all samples relevant to the hypothesis, e.g. "all HIV viral isolates in the U.S." , or "samples made by all the growth and preparation methods that could be used to generate the bioagent in question and all of the potential sources for starting materials and additives". Clearly the population relevant to testing a hypothesis H must contain both H and $\bar{H}$ samples. Note that the population can be real (HIV isolates) or virtual (production methods.)

Define the signature, i.e. the set of molecular, chemical, or physical characteristics that provide the basis for decision (H or $\bar{H}$).

Define an objective metric for decision, i.e. a scalar quantity defined in terms of the signature that is used to decide H or $\bar{H}$. The objectivity of the metric is not strictly necessary, but if subjective criteria for decision are used, then the validation procedure strictly applies only to the operator making the subjective decision, not the method in general.

Characterize the population, e.g. collect the viral genetic sequence information, or the chemical or physical signature data on a set of bioagent samples generated by a *representative sampling* of the population. Sampling must include both H and $\bar{H}$ exemplars. In the case of a virtual population like the set of all production methods, the act of drawing representative samples is actually choosing recipes and making samples according to those recipes.

Evaluate the receiver - operating characteristic (ROC) This involves determining the distributions of the values of the scalar metric for the two subpopulations corresponding to H and $\bar{H}$, and plotting the parametric relationship between the true and false positive

rates determined by the value of the scalar metric. The slope of the ROC curve at any point is the likelihood ratio for H at the corresponding value of the metric.

In general, the definition of the relevant "population", and the requirement of sampling from that population in a "representative" or unbiased way is a critical part of the experimental design. Some specific examples of the considerations that apply to specific types of analyses will be provided in the next section. The ROC curve is a standard method for displaying the statistical predictive power of any hypothesis testing method that depends on the magnitude of a measured quantity, and is described in many books and publications[37-47]. It can be used in a transparent way to generate the likelihood ratio that associates any particular value of the metric observed in a questioned sample with the probability that H is true. The application of ROC curves in other areas of forensic science has been discussed in a number of papers[47-61], and the reader is encouraged to review this literature if unfamiliar with the method.

In the context of this scheme "validation" becomes an iterative process where the relevant population can be independently re-sampled and tested, and the ROC from this experiment is compared to the previous one. If the results are significantly different, then at least one of the samplings was *not* representative of the true population, and hence that at least one of the curves is not a valid way to determine the weight that the measurement lends to the truth of H. The independence of the re-sampling and measurement process can be assured in various ways, including adherence to randomized selection protocols, use of multiple independent laboratories, blind testing, etc. The ROC formulation allows the data sets from both experiments to be combined to form a new ROC, and the process can be repeated until additional re-sampling does not generate significantly different results.

Parameter estimation (calibration)

The evaluation and validation of a parameter estimation procedure involves a slightly more complicated, but similar algorithm. In this case we are trying to collect a set of data that will allow us to determine the likelihood that a certain parameter of a questioned sample lies within a certain range, based on a measurement (or measurements) of some other property. The steps in the validation procedure are:

Define the signature, i.e. the set of chemical, molecular or physical characteristics that provide the basis for estimating the parameter range.

Define an objective metric for estimating the parameter, i.e. a scalar quantity (defined by the signature) that is correlated with the value of the parameter to be estimated.

Define the sample "population", i.e. some complete, well-defined set of samples, or of possible sample origins and histories that might affect the "calibration curve" that relates the metric value to the value of the parameter in question. (As in the case of hypothesis testing, the population could be real or virtual.) The value of the parameter of each member of the population used for validation must be known.

Characterize the population, i.e. choose bioagent samples in an unbiased way from the complete population of interest (each of which has a known value of the parameter) and determine the value of the metric for each one.

Evaluate the calibration curve, i.e. construct the scatter plot that shows the dependence of the metric on the value of the parameter.

Normally, an analyst would fit the calibration curve to some appropriate function, which would then be used to map values of the metric determined for a questioned sample onto predicted values of the parameter for that sample. The analyst's testimony might then consist of an estimate of the parameter and some measurement of its uncertainty. While this approach is reasonable in many circumstances, the connection between the uncertainty of the parameter and how strongly the measurement "evidence" supports the hypothesis that the parameter lies in that range (i.e. within the stated uncertainty) is not transparent. Therefore, we suggest a slightly different approach to parameter estimation, whereby the calibration curve can be used to directly generate likelihood ratios associated with the hypothesis that a questioned sample has a parameter (e.g. age, number of washings, etc.) that falls into a certain range, given the measured value of the metric. This method of constructing likelihood ratios from calibration data is outlined in Appendix 1.

As in the case of the ROC curve, validation of a calibration curve consists of independently re-sampling the "population" and comparing that new calibration curve with the earlier one. Again, the two calibration sets can be combined to form a calibration curve that is more representative of the complete population than either separate curve, and the process may be iterated to generate a more and more accurate representation of the total population.

Some general considerations

Understanding the relevant "population", choosing a sampling strategy, and determining an appropriate size for the sample set, are key questions that arise in executing the processes outlined above. Because these elements differ markedly among different types of analyses it is easier to illustrate them with particular examples, as will be done in the next section. However, three important general observations can be made: First, as a prelude to any validation exercise it is necessary to consider all the possible factors that could affect the relationship between the measured value of the metric and the hypothesis or parameter in question but can't be controlled, or might not be known about a questioned sample. For example:

- The exact method of growth and production of an agent
- The exact source of materials used in the production process
- The temperature and humidity conditions under which an agent might be stored prior to dissemination

- The immune system condition (e.g. correlated with general health, age, treatment history) of a patient that a reference isolate was drawn from

- The exact passage history experienced by an isolate compared to some reference isolate

For a method to be applicable to a questioned sample for which these factors are not known, the set of samples used for validation must reflect a random selection from a population in which these factors are allowed to vary over their naturally occurring ranges. Thus the estimation of the likelihood ratio takes into account the probability of the various conditions that might have existed, but cannot be known with certainty about the questioned sample. On the other hand, if these factors happen to be known in a particular case, then, the ROC or calibration curve that is used could be one that is determined for the particular sub-population in which this factor was controlled.

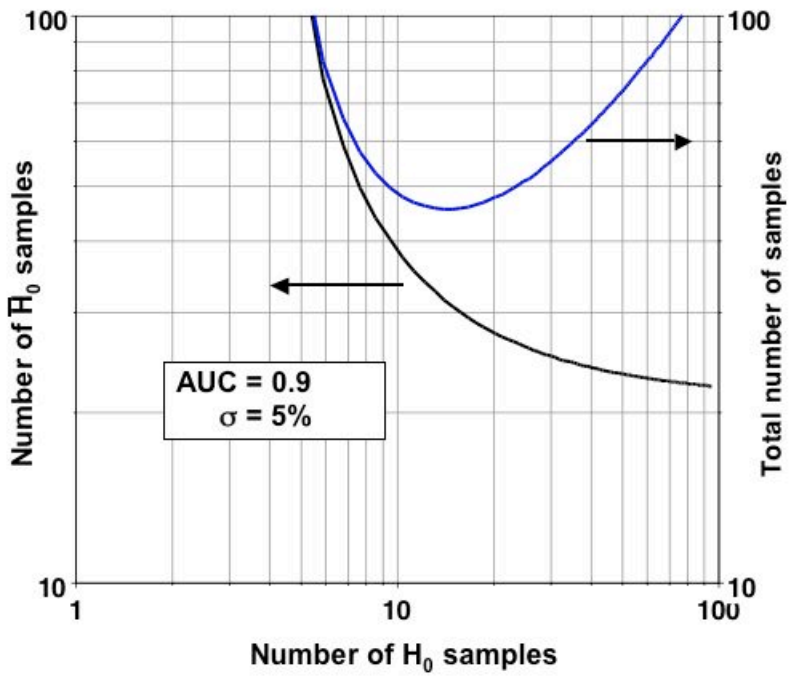

Figure 1. Plot of the number of H and $\bar{H}$ samples required to achieve an AUC precision of 5% when AUC ≈ 0.9.

Second, while validation usually requires a significant number of samples, this number need not be unreasonably large to obtain accurate results. Standard formulae for estimating sample size effects on the precision of ROC curves have been published[42]. One common way to characterize sample size effects is to look at the precision with which the area under a ROC curve (usually referred to as the AUC) depends on the number of H and $\bar{H}$ samples. Figure 1 shows the number of samples required to achieve a precision of 5% for a ROC curve with an AUC of 0.9 (which would correspond to a test with a relatively high diagnostic capability.) Note that the total number of samples required exhibits a minimum. To achieve the specified precision around 45 samples are

needed, with about 15 H_0 and 30 $\overline{H}_0$ samples. Thus, reasonable precision is possible with a fairly modest number of samples.

The third consideration touches on the difference between real and virtual "populations". For the forensic analysis of manufactured materials such as fibers and drugs, samples can be drawn from a real population. Biological agents are clearly not manufactured continuously in quantity, so sampling the "population" of manufacturing processes necessarily involves simulating the diversity in manufacturing methods by using different recipes and laboratories. This explicitly connects the validation process with intelligence about terrorist and state program knowledge and practice. Biological agent manufacturing information that a terrorist might use can come from many sources. This includes material derived from a wide variety of open sources such as recipes provided by underground cookbooks and internet sites, relevant knowledge from the open scientific literature, and inadvertent leaks of sensitive (but often inaccurate) information that are published in the news media. In the rare instance U.S. intelligence efforts may uncover information about the technical knowledge possessed by particular terrorist groups or state BW programs. Thus, even with the virtual population of manufacturing processes, validation will require periodic updating, and always leaves open the question of whether there may be unknown sub-populations that have not been sampled.

4. Applying the strategy to forensic assays – examples from the literature

This section discusses validation strategies for several analysis methods that could have utility in microbial forensic investigations. Because the field is relatively new, many methods are still in the exploratory stage of development, and the examples we have chosen are taken from the open literature such as it exists. At the risk of causing confusion, we have adopted the convention of denoting particular hypotheses by a capital letter with the subscript 0. This uniformly allows the symbol for a hypothesis to be distinguished from other variables whose most natural notation uses the same capital letter. For example, S_0 is the hypothesis that a victim was infected from the source S. Traditionally the subscript 0 has denoted the "null hypothesis" in statistical literature, but has no such interpretation here.

Viral transmission – HIV and HCV

One of the oldest methods of microbial forensics is the use of comparative genetic sequence data to establish a possible source of infection. Within the last two decades, a number of forensic/epidemiological investigations involving Human Immunodeficiency Virus (HIV) or Hepatitis C Virus (HCV) have been discussed in the literature[62-76]. Inferences about the likelihood of viral transmission events essentially rely on measures of genetic similarity between isolates taken from a victim and a putative source. Because isolates of RNA viruses from infected humans typically consist of a broad distribution of genetic sequences, there is never an exact "match", strictly speaking, between the sequence data obtained from any two samples (even when they originate from the same person, if the isolates are obtained at different times.) In this sense, testing the hypothesis of transmission by means of genetic similarity is very similar to using elemental profiles for sample matching, as discussed in reference 25.

Suppose we define T_0 to be the hypothesis that two isolates are related by direct transmission between a source and an infected person. $\overline{T}_0$ is the hypothesis that they are *not* related by direct transmission - hence either there are additional hosts in the transmission chain between the putative source and the infected victim, or both the victim and putative source were infected from some other (possibly common) source. Note that T_0 does not necessarily specify a direction of transmission, although the context provided in particular cases often provides the presumptive direction. The relationship between the direct transmission hypothesis T_0 and the attribution of infection to a particular source is explored in Appendix 2, but is not essential to our discussion.

With respect to T_0, the probative weight of the observation that two isolates are separated by a certain genetic distance Δ is expressed by the likelihood ratio $LR_t(\Delta)$:

$$O(T_0|\Delta) = LR_t(\Delta) \bullet O(T_0) \tag{3}$$

where

$$LR_t(\Delta) = P(\Delta|T_0)/P(\Delta|\ \overline{T}_0) \tag{4}$$

The genetic distance metric Δ that maps the genetic differences between two sequences onto a scalar quantity can be defined in various ways[77-81], and is functionally analogous to the metric defined by the difference between elemental profiles discussed in reference 25. It is important to note that genetic distance metrics are used in several common phylogenetic tree construction methods, and may be derived from others where the scalar metric is related to branch length. Thus, in a broad sense, phylogenetic expressions of relatedness that are often used in forensic source attribution are based on fundamental principles similar to (3) and (4).

The procedure for generating a ROC curve to estimate $LR_t(\Delta)$ would involve sampling sequences that were derived from a collection of isolates which contained some known transmission-related pairs. Databases for both HIV and HCV often contain annotation that includes known epidemiological relationships that can be used to select T_0 and $\overline{T}_0$ sequences[82,83]. An approximate representation of such a ROC curve is shown in Figure 2, which was constructed from measures of genetic distance (expressed simply as % of sites exhibiting substitutions) between transmission related and non-related pairs of HIV isolates reported in the literature[62,84,85]. However, this data set is far from ideal because in many cases only averages and/or ranges were reported, rather than precise values for individual clones. In addition, several different gene regions were used to determine the distance values. However, Figure 2 serves to illustrate several features of the ROC representation of this type of data.

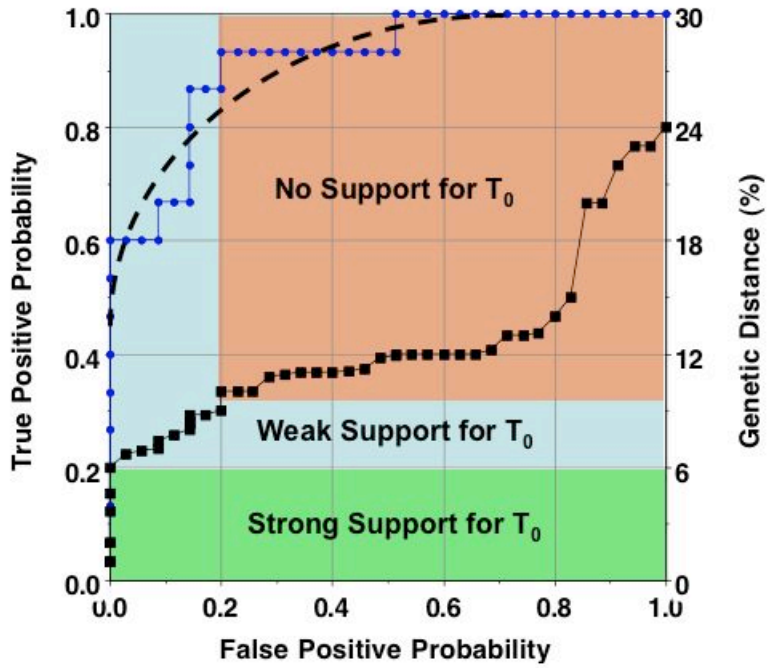

Figure 2. A notional ROC curve derived from available literature data on genetic distances between transmission and non-transmission related samples. Blue dots: ROC data; Dashed curve: smooth interpolation of the ROC data; Black squares: Genetic distance values (%).

For example, the ROC data displayed in Figure 2 exhibits the typical "stair-step" form due to the sparse number of data points. Therefore, an interpolating curve has been added to aid discussion of the qualitative relationship between genetic distance and the likelihood ratio $LR_t(\Delta)$ defined in equation (1). For values of $\Delta < 6\%$ the slope of the ROC curve (which gives the value of LR) is clearly much greater than 1. Hence, a Δ in this range provides strong support for the hypothesis of transmission. In the region $6\% < \Delta < 10\%$, the slope of the ROC curve is greater than 1, but not by a large factor. Hence a Δ value in this range provides weak support for T_0. Values of Δ greater than 10% provide no support for T_0. It is of some interest to note the clarity of this *representation* of the data in contrast to the way that genetic distance data was used to support the transmission hypotheses in an early forensic case that generated some controversy in the literature [62,86].

A better approximation to this ROC curve would clearly be constructed by accessing actual individual clonal sequence data from a large population of HIV isolates that contains both known direct-transmission pairs and non-direct transmission-related sequences. Clearly the same segment of the genome should be used to create the ROC plot since different regions evolve at different rates. Ideally the entire genome could be used, but the primary effect of using a smaller portion would simply be to re-scale the % axis (assuming % difference is used as the genetic distance metric) in accordance with the evolutionary rate of that segment. It would be interesting to compare ROC plots created by using different genomic segments to see if there are differences based on insensitivity to selection pressure or other factors that can influence the relationship between the

distance metric and the hypothesis of direct transmission (see below). The effect of using metrics defined in different ways should also be explored.

Strictly speaking, the relevant "population" against which such an assay must be validated consists of all HIV sequences that presently exist within the human host population, plus those in isolates that have been archived in various laboratories around the world but not yet sequenced, and all archival sequence data contained in databases. An important consideration for sampling from this population is the accuracy with which the subpopulation of transmission related pairs has been identified. There are several additional considerations that could affect the relationship between T_0 and Δ:

- The length of time between the putative transmission event and the acquisition of clinical samples
- Whether one or both pairs of T_0 related hosts was treated with antiviral therapy
- Differences in the general health or immune system condition among T_0 related hosts.

Clearly any sampling of isolates or archived sequence data must be examined to ensure that significant biases related to these factors are not obviously present. While there is no reason to suppose that the geographical location of the subpopulation might itself introduce bias, a geographically diverse validation set might help meliorate unknown biases introduced by the above factors. Conversely, in certain cases it may be that particular sub-populations (e.g., isolates from persons treated with antiviral drugs) may be the most relevant, and generating the ROC curve using only these sequences produces the most accurate representation of the likelihoods.

Note that the ROC curve approach provides an alternative to the use of phylogenetic arguments in the forensic setting[75]. It should be noted that while phylogenetic trees have been declared admissible[87], they are only allowed as evidence that the genetic sequences are "closely related". Typically the trees used in court cases are constructed by taking a convenient sampling of HIV sequences that are geographically local to the crime[88]. It is clear that the prosecution argument, even if not explicitly stated, is that such a tree demonstrates that the victim's sequence is closer to that of the putative source than to other possible sources, and hence increases the likelihood that the victim and putative source isolates are related by transmission. The difficulty with this argument is that phylogenetic "closeness" as expressed by tree topology is not *per se* strongly supportive of the hypothesis of transmission. It is only the genetic distance measure associated with the branching that is relevant to this question. This caveat is reinforced by one of the case studies discussed in an excellent review of HIV forensics by Learn and Mullins, where phylogenetic construction led to an incorrect inference of transmission that was not supported by the calculated genetic distance value[89].

Determination of growth method: heme analysis

The determination of manufacturing method is an important general problem in microbial forensic analysis[90]. The "population" of possible growth and processing methods is

represented by a unit process matrix, shown schematically in Figure 3. Here each end-to-end process is broken down into unit process steps such as growth, separation of the microbe from the growth medium, washing, drying, milling, and combining with additives. For preparation of a toxin such as ricin, a similar matrix can be constructed with columns defined by the unit process steps appropriate to the particular toxin. For each unit process, there are a number of options, including the "null" option in which that particular unit process is not carried out. (Note that "null" is not an option for the growth step because we are assuming that at least a small amount of bulk agent is made prior to dispersal.) Any particular end-to-end production process, as represented by the shaded cells in Figure 3, draws from these possible unit process options. It should be noted that each unit process step may actually represent a rather complex combination of subunits, and that different laboratories might implement a particular unit process in a slightly different way, or use materials from different sources, adding another layer of potential variation to any end-to-end process. Even when the same nominal process is repeated in a given laboratory, some variation among the chemical and physical properties of the each batch of agent might be expected.

An analysis that purports to show that a particular unit process step was used in the agent preparation process is clearly a hypothesis test, and the approach to validation outlined in section 3 can be applied. The "population" that one samples to generate a ROC curve is (conceptually) all of the processes represented by the unit process matrix. At first glance this might seem like a daunting, if not impossible task, given the large number of potential methods for growing and processing spores. However, methods of production are not a-priori equally probable. Because many of the end-to-end processes are not commonly used, the process of random sampling from this matrix can be weighted by knowledge of preferences that exist among practitioners. This knowledge can be drawn from the existing literature as well as from consultation with experts. A detailed example of this weighted sampling process was provided in reference 25.

Growth	Separation	Washing	Drying	Milling	Additives
	∅	∅	∅	∅	∅
G_1	S_1	W_1	D_1	M_1	A_1
G_2	S_2	W_2	D_2	M_2	A_2
G_3	S_3	W_3	D_3	M_3	A_3
⋮	⋮	⋮	⋮	⋮	⋮

Figure 3. The unit process matrix for biological agent preparation. ∅ represents the case where that unit process is not carried out. Violet shaded cells represent a particular choice of unit processes that make up an end-to-end production method.

To account for variations in the execution of particular process steps, it is preferable to have samples made by different laboratories working independently as well as samples

drawn from different batches of material made by the same process. Thus, a set of samples that provides the best representation of "process space" will include laboratory, process and batch variations. Partial factorial sampling designs can be used to reduce the number of samples to reasonable values. It also may be possible to use expert judgment to delimit the set of samples if it is clear that certain changes in medium composition or post-growth processing ought to be irrelevant. (For example, changes in separation method, e.g. centrifugation vs. filtration, may not affect the isotopic content of the agent.) However, since it is always possible that some new scientific findings may later change this assessment, or new process variations may come to light, it is necessary to review the validation panel periodically.

One example of an analysis that has been proposed to determine a growth method is a recently published paper on the determination of heme content in Bacillus samples[91]. The intention of heme analysis is to find evidence that a bacterial agent (such as anthrax) used in a terrorist or criminal act was grown on blood agar plates, which might be indicative of the resources or training that were available to the perpetrator. This approach is based on the observation that harvesting *Bacillus* from agar plates usually entrains some residual solid medium within the agent. Propagation on blood agar plates is not a very common method for growing bulk samples of *Bacillus anthracis*, but is normally used for diagnostic purposes. A very crude estimate of the *a priori* probability that a perpetrator might choose to produce a sample this way can be obtained from a survey of the open literature where 7 out of 78 papers (9%) cite this method[90]. No papers cite growth in liquid media containing blood. On this basis the odds of a sample having been grown this way, in the absence of any other information, might be estimated as around 1 in 10. Thus, analysis for residual heme can be used to provide support to the hypothesis that the samples were grown on a blood-containing agar medium. The data presented in reference 91 provide a case for the plausible utility of this analysis, but no extensive validation has yet been attempted. (One unfortunate result was that irradiation of the sample to inactivate the micro-organism, which is necessary before it can be analyzed in a laboratory that is not able to handle live pathogens, appears to degrade the heme signature. However, other means of deactivation are potentially applicable.)

In the proposed analytical method, heme is extracted from bulk spore samples and the extracted material is analyzed by Matrix Assisted Laser Desorption/Ionization (MALDI) mass spectrometry[91]. The metric that is correlated with the presence of heme is the peak height or area of a mass spectral peak observed at the mass value of heme. A fitted calibration curve based on samples spiked with heme was used to relate this metric to a heme concentration. The authors estimate a "limit of detection" (LOD) that corresponds to approximately 0.12 nanograms of heme per milligram of spores, and a "limit of quantification" approximately twice this value. Table 1 of this reference summarizes the results for 7 un-irradiated samples that were grown using different methods, 3 on blood agar plates and 4 on other media. All three blood agar grown samples gave positive detections, presumably meaning signals greater than or equal to the "limit of detection", while the 4 non-agar samples gave negative detections. For the samples with positive detections, estimates for heme content in ng/mg were also provided.

Strictly speaking, in the absence of values of the metric for both the positive and negative samples, it is not possible to construct a ROC curve. However, presuming that none of the negative detection samples had values above the "limit of detection", and since the estimates for the lowest concentration in a positive sample was at least 2 times larger than the LOD, this data implies a perfect ROC curve. However, with only 7 points, this almost certainly is the result of under-sampling. Note that the confidence intervals for the true and false detection probabilities calculated for this data set are quite large[92], as shown in Table 1.

There are several reasons to believe that the distribution of apparent heme concentrations among samples produced by a more extensive sampling from the population of potential growth and processing methods would be wider than is implied by this small sample set, providing both false positive and false negative detections distributed around the nominal "limit of detection." Clearly different harvesting techniques might change the quantity of agar-derived heme present in the spore preparation. Different washing protocols might reduce heme content. Since heme containing enzymes are a component of many bacteria, including *B. anthracis*, there is probably some natural background level in spore preparations, and this might be detectable in certain samples[93-97]. Finally, complex medium components may contain molecules that have the same mass spectral signature as heme, which could raise the false positive detection rate. A summary of possible sources of false positive and negative detections, and the characteristics of a proper set of samples for validation is provided in Table 2.

Table 1. Calculated 95% confidence intervals for conditional detection probabilities for heme analysis data from reference 91.

	B_0 (Sample grown on blood agar)	$\bar{B}_0$ (Sample not grown on blood agar)	
Positive test result ($C_h \geq$ LOD)	$P(+\|B_0) \geq 0.47$ (3/3 samples)	$P(+\|\bar{B}_0) \leq 0.45$ (0/4 samples)	$N(+) = 3$
Negative test result ($C_h <$ LOD)	$P(-\|B_0) < 0.53$ (0/3 samples	$P(-\|\bar{B}_0) \geq 0.55$ (4/4 samples)	$N(-) = 4$
	$N(B_0) = 3$	$N(B_0) = 4$	$N_T = 7$

Table 2 raises two additional issues that are relevant to the concept of validation, but are not included directly in the framework we have outlined in this report. Note first that, strictly speaking, the analysis under discussion does not directly answer the question "Was the agent grown on blood agar medium?" The interpretation that a large value of the metric (i.e. an apparently large concentration of heme in an agent sample) supports the hypothesis of blood agar growth could be false, for example, if heme were deliberately added to the agent after it had been grown in, liquid culture. (The perpetrator

might do this to mis-direct investigators.) To rigorously validate such an interpretation, it would be necessary to have some independent method of analysis that could exclude the possibility of post-growth addition of some material that contains heme. Likewise, the mass of heme is not unique, so that if a different molecule with the same mass were present (either naturally or as the result of deliberate addition) it could lead to a false inference of support. Clearly, *by itself* the analysis under discussion can only support the hypothesis of agar growth under the assumption that heme is what is being detected, and that there is no mis-direction. By the same token, a low value of the metric could be due to the perpetrator having used, e.g., an unusually rigorous washing process designed to remove heme, so that the inference of weak support for blood agar growth is false. (According to reference 91 there is some indication that washing is not very effective in removing heme due to its hydrophobic nature. However, certain washing protocols are more likely to remove heme than others.) In effect, the validation process assumes that such unusual processes are rare, and thus have little effect on the likelihood ratio. An independent analytical method that can generate support for the hypothesis that this rare step was used is required to test this assumption. A more extensive discussion of the issue of washing is presented in Appendix 3.

Table 2. Potential errors and validation elements for the heme analysis example.

Type of error	Nature of potential error	Validation element
False negative (Non-detection of heme when agent was grown on blood agar)	Post-growth harvesting method does not entrain agar	Independent variation in harvesting technique among laboratories
	Post-growth processing removes heme	Inclusion of relevant processes in the validation panel
False positive (Detection of heme but sample not grown on blood agar)	Heme added post-growth	**Method to distinguish post-growth addition**
	Different molecule present, misidentified as heme	**Second, independent method/metric**
	Heme present in other growth medium	Wide variation in growth media used in validation panel
	Heme generated from bacterial metabolism or some process step	Wide variation in processes used in validation panel

Microbial forensic utility of isotopic signatures

Isotopic measurements have been used in several areas of forensics, including determination of the geographical origin of cocaine and heroin, and the authentication of other imported agricultural products[98]. The success of using isotopic ratios to identify geographical origin of agricultural products has encouraged the investigation of these methods for microbial forensics[99-103]. Of course, a typical bacterial culture incorporates medium components from many sources, so the direct determination of origin, if possible, would be considerably more complicated than for agricultural products like

cocaine or heroin. Nonetheless, at least three types of forensic utility related to microbial forensics have been suggested in the literature:

- Sample matching[25]
- Analysis of the provenance of medium components[101-103]
- Use of Hydrogen and Oxygen isotopes for geolocation[99,100]

In the case of sample matching, the isotopic "fingerprints" of two samples are compared to determine if the samples are likely to have originated from the same batch of material or were grown using the same growth medium components. The framework for validating this kind of analysis is similar, *mutatis mutandis*, to the one previously developed for elemental "fingerprints", so the reader is referred to that discussion for insight into this application. However, the remaining two applications are discussed in this section.

Certain bacterial growth medium components are, in effect, agricultural products, so inferences about provenance based on their isotopic composition are plausible. This clearly applies to cases where the growth medium components themselves are available for analysis. In an extensive study, Kreuzer-Martin, et. al. showed that the isotopic composition of growth medium components exhibit strong correlations with their origin[102]. This extensive collection of results provides an excellent opportunity to illustrate the interpretation of such data in the ROC curve/likelihood format. Particularly good examples are the $^{13}C/^{12}C$ and $^{15}N/^{14}N$ signatures of peptone samples that originate from meat proteins or casein. The following analysis utilizes the data that is displayed in Figure 2 of reference 102, kindly provided by the author.

Let C_0 is the hypothesis that the peptone medium is derived from casein, while $\overline{C}_0$ is the hypothesis that it is not. Assuming that peptones are only derived from casein or meat, and no other sources are possible, $\overline{C}_0$ is equivalent to M_0, the hypothesis that the peptone is derived from meat. Figures 4 and 5 show the ROC curves derived from the $\delta^{13}C$ and $\delta^{15}N$ values obtained by Kreuzer-Martin on a variety of peptone samples. Clearly the measurement of a $\delta^{13}C$ value below -22‰ is strong support for a casein origin of the peptone, while values greater than -18‰ argue for M_0.

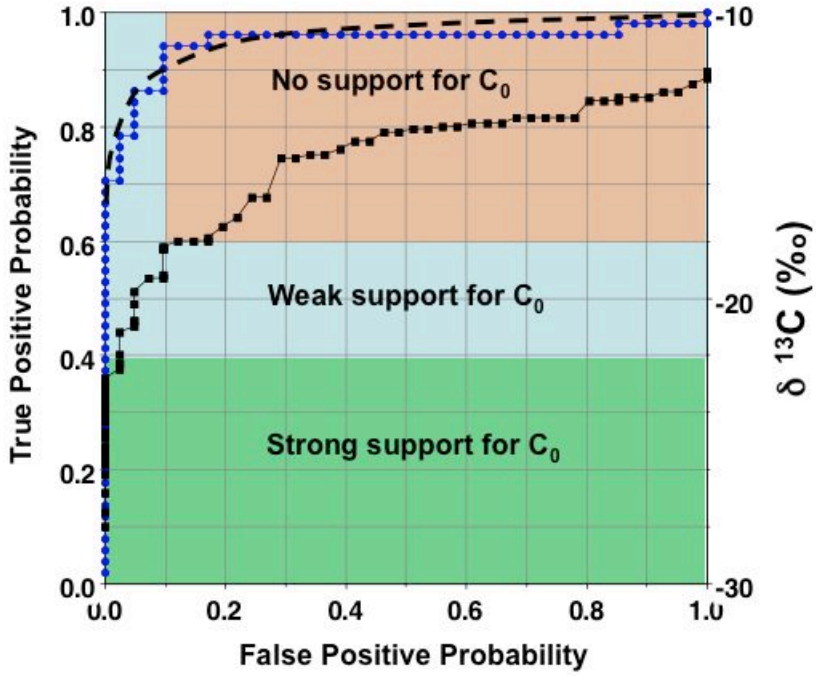

Figure 4. ROC curve generated from $\delta^{13}C$ data from peptones of casein and non-casein origin (data from reference 102.) Blue dots: ROC curve; Black squares: $\delta^{13}C$ values. The dashed curve is an interpolation of the ROC data.

The ROC curve for the $\delta^{15}N$ data has much weaker inferential value since the ROC is much closer to the slope = 1 line that indicates no-better-than-chance discriminating power. Nonetheless, finding that the $\delta^{15}N$ is less than 6 adds to the strength of the hypothesis that the medium is derived from casein, while values greater than 6 support a non-casein origin. The likelihood ratio derived from combining the data in Figures 4 and 5 (slopes of the ROC curves) is simply the product of the likelihood ratio from each.

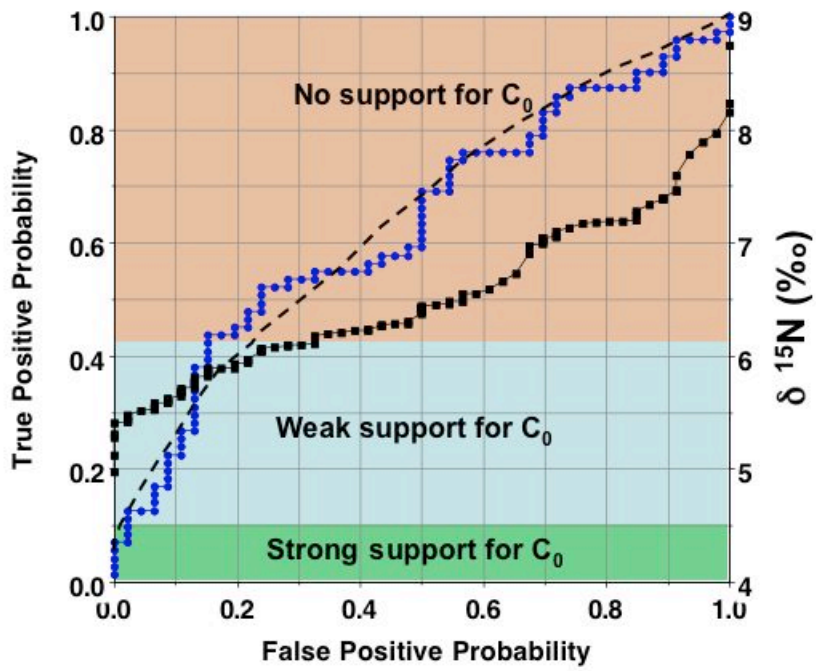

Figure 5. ROC curve generated from $\delta^{15}N$ data from peptones of casein and non-casein origin (data from reference 102.) Blue dots: ROC curve; Black squares: $\delta^{15}N$ values. The dashed curve is an interpolation of the ROC data.

A potentially powerful application that has been suggested in the literature is the use of hydrogen and oxygen isotopes to narrow the geographic origin of a bioagent sample[99,100]. This suggestion is based, in part, on the observation that the hydrogen and oxygen isotope ratios in *Bacillus* are correlated with those of the water they are grown in. To the degree that water used to reconstitute dry culture medium components and wash harvested agent is likely to be local in origin, and that hydrogen and oxygen isotope ratios in water have strong geographic correlations, this approach is clearly plausible. However, extensive validation of the approach has not yet been undertaken.

From the discussion in references 99 and 100 we can infer a basic algorithm for using the hydrogen and/or oxygen isotopes to bound the geographical origin of the water used to manufacture an agent.

(1) From a sample of the questioned agent, measure the $^{2}H/^{1}H$ and $^{18}O/^{16}O$ ratios and determine $\delta^{2}H$ and $\delta^{18}O$.

(2) Construct a set of calibration data that relate $\delta^{2}H$ and $\delta^{18}O$ in samples of the agent to $\delta^{2}H$ and $\delta^{18}O$ in the water used to grow and process it.

(3) Use the calibration data set to estimate the range of possible values of $\delta^{2}H$ and $\delta^{18}O$ for the water used to grow and process the agent.

(4) Finally, from the ranges of possible values of δ^2H and $\delta^{18}O$ of the water, identify the geographic contour band(s) that contain water sources that lie within these ranges.

The second and third steps in this procedure, which involve the concept of calibration described in Appendix 1, require careful consideration. In most cases, the precise method by which the agent was produced will not be known. Sources of hydrogen and oxygen other than water can influence the isotopic ratios in microbes, as illustrated schematically in Figure 6. While these additional O and H sources may have small contributions, they constitute a major source of uncertainty in the final values of δ^2H and $\delta^{18}O$ of the agent.

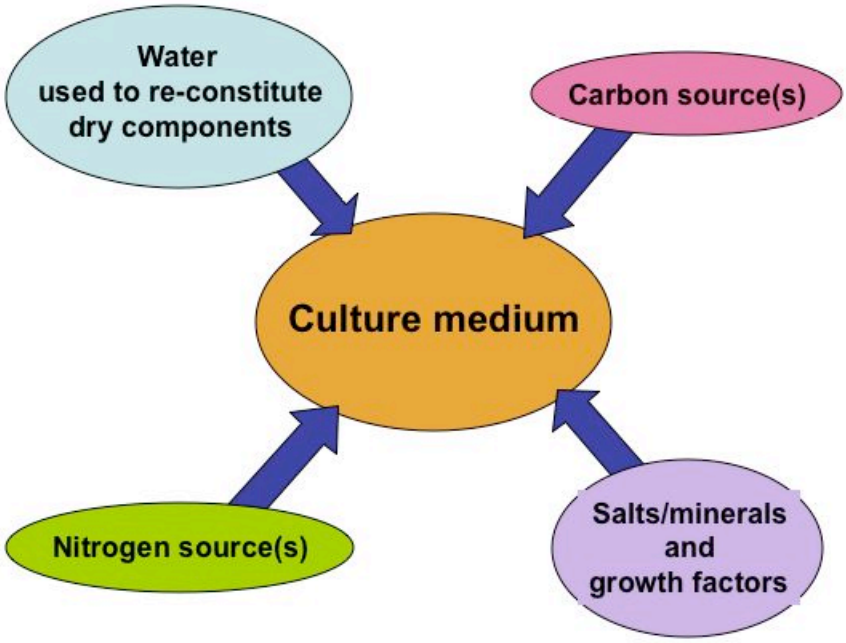

Figure 6. Different components of a typical bacterial growth medium that might contribute hydrogen and oxygen to the bacterial composition.

Additionally, there are potential fractionation effects that depend on details of the culture, including the aeration rate and temperature, and post processing steps such as drying. Thus, to make the calibration set that is used to relate δ of the microbes to δ of the water, it is necessary to sample many possible variants in growth and production process in which the water is spiked with ^{18}O and ^2H. The spiking quantities should, of course reflect δ values in the ranges that are extant among real water supplies. While such a data set does not yet exist, it is possible to simulate a partial set for illustrative purposes by using data provided in reference 99. In this study spores were grown in five different media and four isotopically distinct waters (but were all identically processed,) and a set of best-fit straight lines relating the δ^2H of the spores to the δ^2H value of the water were obtained. The points in the simulated calibration data shown in Figure 7, were generated using a two-step process. First, we generated 100 values of δ^2H_{water} randomly chosen from the range -100‰ to +300‰ (which encompasses the natural range of δ^2H for terrestrial water.) For each value of δ^2H_{water} we randomly chose one of the slope intercept pairs from among the 5 choices corresponding to the 5 media, and calculated a

corresponding value of δ^2H_{spore}. To simulate experimental uncertainty in the data points we added a small noise term sampled from a Gaussian distribution with a standard deviation of 5‰.

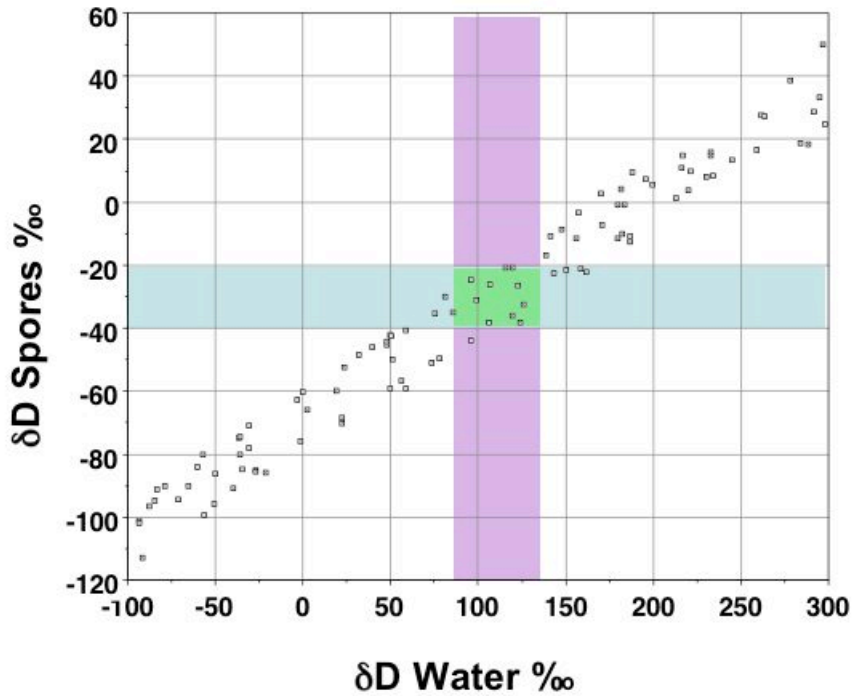

Figure 7. Simulated calibration data for relating the δ^2H of a spore sample to the δ^2H of the water it was grown in. The blue band is the range of measured δ^2H_{spores}, the violet band indicates a range of δ^2H_{water} values.

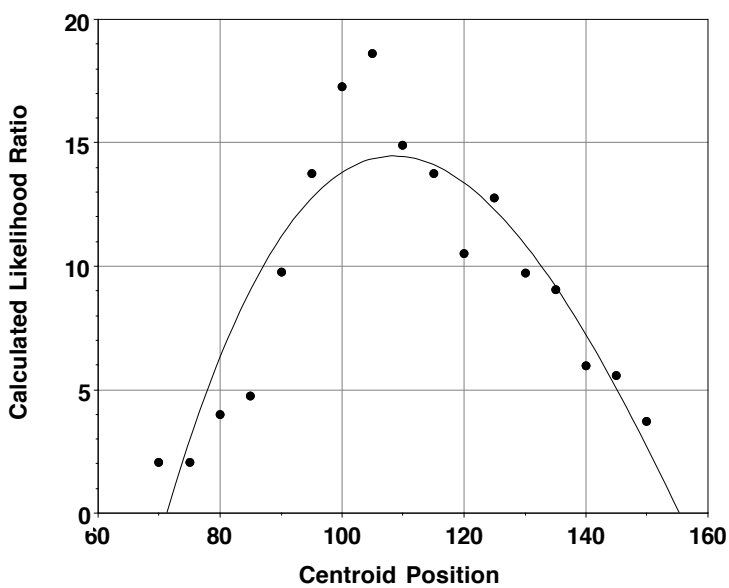

Figure 8. Variation of LR as the centroid of the band predicting the range δ^2H_{water} is moved across the calibration set. The width of the band was fixed at 60‰.

To illustrate the use of this data, suppose that a questioned sample of agent was analyzed and found to have a δ^2H value falling into the 95% confidence interval -20 to -40 ‰. This defines the blue band in Figure 7. The violet band represents a range of possible values for δ^2H of the water used in the manufacture of the agent. The centroid of this band was fixed by noting that the likelihood ratio was maximized near this position, around 110‰, as shown in Figure 8. For the calculation of the maximum likelihood position, the width of the band was fixed at 60‰. For the band centered at 100‰ the dependence of the likelihood ratio on the width is shown in Figure 9.

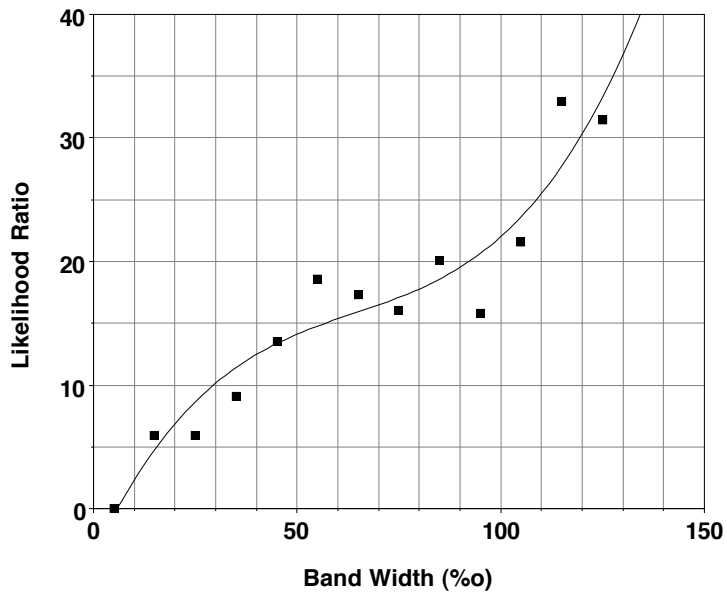

Figure 9. Dependence of the likelihood ratio on the range of ‰. For this calculation, the band centroid was placed close to the maximum likelihood position.

Note from Figure 9 that the calibration data does not strongly support the hypothesis that the range of possible δ^2H_{water} values is smaller than 30‰, but provides very strong support for a range smaller than 140 ‰. Support for a range of 100‰ is roughly twice a strong as for a range of 50‰. If this example were a real case, the values for the centroid and range of δ^2H_{water} could then be used to delineate the geographical areas that might have been the source of the water used to manufacture the agent.

Clearly a much larger choice of production processes would have to be sampled to provide adequate validation for this method. Considering the wide variety of growth media and washing processes that comprise the matrix of unit processes shown in Figure 3, it is likely that the calibration data set generated by sampling actual processes would be more disperse than that exhibited in Figure 7. Conversely, if information were available to narrow the range of growth and production processes that could have been used to generate a questioned sample, then only that sub-set of calibration data would be sufficient to perform the analysis.

5. Concluding Remarks

The likelihood ratio paradigm allows us to present evidence in a format that resembles forensic human DNA typing. Table 3 summarizes this analogy for two broad classes of microbial forensics evidence. In this table, the "populations" are sometimes real (humans, isolates) and sometimes virtual (passage histories, manufacturing methods) and the act of sampling from the relevant population will differ accordingly. Regardless of the particular format chosen to present the results, the process of validation begins by questioning the fundamental assumptions of the method under scrutiny, and then seeking out the sub-populations of isolates or production conditions that will test the limitations of these assumptions. The validation process establishes the range of conditions for which the interpretation of the analytical results is correct – and more importantly, identifies those (hopefully rare) conditions when the standard interpretation is not valid. The full range of relevant conditions must be represented in the panels of reference samples used in the various tests.

Expert review can be used to increase our confidence that the relevant range of conditions have been taken into account, or to delimit the range of conditions that must be considered. Expert panels help to answer questions such as

- Has the "population" that we are sampling from been defined correctly and completely?

- Are there certain weights that should be used in the sampling process, based on prior probabilities that certain subpopulations are more likely to be encountered than others?

- Are there exceptional processes or conditions that might represent rare but confounding outliers?

Thus the utilization of experts becomes a natural and often essential element of method validation, consistent with the Daubert requirement of scientific peer review. In addition, the validation process must be by definition dynamic, subject to periodic review to assess the impact of knew scientific knowledge, changes in the type of agent considered, and the evolution of production technologies for bacterial agents.

In this context the term "preliminary validation" has been introduced to describe validation over a restricted range of conditions that may be necessary for evaluating a new technique in the midst of an on-going investigation[5]. In terms of the framework advocated in this paper, there are two natural interpretations of "preliminary validation." The first is that a relatively small number of samples are used, with concomitant increases in the uncertainty with which the probative value of the assay can be stated. The second interpretation is that it utilizes only sub-populations of samples that are judged to be of the highest relevance to the case at hand. In this version, "preliminary validation" is not intrinsically less rigorous than ordinary validation, as long as the inferential power of the method is not extended to cases that clearly involve other sub-populations. In fact, "preliminary validation" may require more short-term resources and

planning than are usually associated with forensic technique development, because in practice, validation is often a slow community-driven process in which work is done piecemeal by many individual investigators over many years. Thus, a more systematic and planned approach to the transition from exploratory signature discovery to validation may sometimes be necessary to make "preliminary validation" a reality.

Table 3. Analogies between human DNA and microbial forensic evidence.

	Human DNA Evidence	Microbial Genetic Evidence	Chemical and Physical Analysis Evidence
Hypothesis	Source of DNA S_0 = A particular person is the source	Relationship between source and attack isolates S_0 = A particular isolate is the source H_0 = A particular passage history separates the source and attack strain	Bioagent manufacturing method M_0 = A particular method was used to produce the agent
Population	Human population	All possible source isolates & passage histories	All possible production methods and sources of materials
Metric	Sequence match	Sequence differences $\mathbf{\Delta}$	Measurement data $\mathbf{D}$
Probative Value	$LR_{match} = P(match\|S_0)/P(match\|\bar{S}_0)$	$LR_{seq} = P(\mathbf{\Delta}\|H_0,S_0)/P(\mathbf{\Delta}\|\bar{H}_0,\bar{S}_0)$	$LR_{data} = P(\mathbf{D}\|M_0)/P(\mathbf{D}\|\bar{M}_0)$

Finally, we might note that there are other forms of microbial forensic evidence that may or may not fit neatly into the validation framework outlined above. One of these is morphology comparisons based on image data obtained using electron (EM), ion (IM), or atomic force (AFM) microscopy. Here one is trying to decide if certain structural features of agent particles are indicative of a particular production method or can be used to match two samples[90]. Modern EM, IM, and AFM instrumentation produce digitally manipulated maps of elemental composition and topography, often using sophisticated statistical computations. There are certain analogies between interpreting such images and diagnostic imaging in medicine, such as X-ray, CT, and MRI imaging. While digital imagery itself is generally thought to meet the admissibility requirements of scientific testimony[104], and medical imaging has long been formulated in terms of ROC curves and likelihood ratios[44,45], challenges to admissibility can and do arise[105]. Thus, the validation of EM and IM microbial forensic methods deserve careful study in this regard.

It is generally recognized that scientific evidence must meet two basic criteria for admissibility, *reliability* and *relevance*. While this paper emphasizes the latter criterion, we do not intend to diminish the importance of the former. Both of these criteria determine critical aspects of the way that analytical methods must be validated to make them useful forensic tools. To establish the reliability of scientific data in judicial proceedings it is necessary to demonstrate that adequate quality assurance and control steps were followed during the testing procedure. A considerable literature exists on QA-QC, including excellent documents generated by several FBI scientific working groups. Three relevant documents are the Trace Evidence Recovery Guidelines[106] formulated by

the Scientific Working Group on Materials Analysis (SWGMAT), the Quality Assurance Guidelines for Laboratories Performing Forensic Analysis of Chemical Terrorism published by the Scientific Working group on Forensic Analysis of Chemical Terrorism (SWGFACT)[107,108], and the Quality Assurance Guidelines for Laboratories Performing Microbial Forensic Work[109], published by the Scientific Working Group on Microbial Genetics and Forensics (SWGMGF). In addition to standards for single laboratory validation, inter-laboratory proficiency testing and round-robin exercises are a critical part of this process. In a robust national program, continuous testing through laboratory and field exercises would be required to improve the system and provide level-of-readiness assurance.

Nonetheless, the *logical relevance* criterion is increasingly recognized to be the more salient stumbling block in forensic science. In this paper we have shown that, when logical relevance is interpreted in terms of a likelihood ratio associated with a test or analytical result, a common approach to validation is possible for a variety of forensic evaluations, and leads naturally to a representation of the data in a form that does not overstate its probative value. In addition to providing the most straightforward and objective way to describe the result of an analysis, this methodology also makes it easy to compare two methods designed for the same purpose or to combine the results of two independent analyses using orthogonal methods. In the absence of technical errors, characterizing a method in this way ought to ensure its acceptability under the Daubert standard.

References

1. B. Budowle, "Defining a new forensic discipline: microbial forensics", (2001) available at www.promega.com/geneticidproc/ussymp13proc/contents/budowle.pdf

2. P. Keim, "Microbial Forensics: A Scientific Assessment", American Academy of Microbiology, (2003)

3. R.S. Murch, "Forensic Perspective on Bioterrorism" in *Firepower in the Lab: Automation in the Fight Against Infectious Diseases and Bioterrorism*, S.P. Lane, J.T. Beugeldijk, and C.K.N. Patel, eds. (National Academies Press, Washington D.C., 2001); R.S. Murch, "Microbial Forensics: Building a National Capacity to Investigate Bioterrorism", Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science Vol. 1, 117-122, (2003).

4. Budowle, B, Schutzer SE, Einseln A, Kelly, LC, Walsh AC, Smith JA, Marrone, BL, Robertson J, Campos J, "Public Health: Building microbial forensics as a response to bioterrorism", Science 301, 1852, (2003).

5. B. Budowle, et. al., "Toward a system of microbial forensics: from sample collection to interpretation of evidence", Applied and Environmental Microbiology 71, 2209-2213, (2005).

6. B. Budowle, et. al., "Microbial Forensics: the next forensic challenge", International Journal of Legal Medicine, 119, pp.317-330, (2005)

7. B. Budowle, et. al., "Genetic analysis and attribution of microbial forensics evidence", Crit. Rev. Microbiol. **31**, pp.233-254, (2005).

8. Homeland Security Presidential Directive HSPD-10, April 2004.

9. D. Kennedy and R.A. Merrill, "Assessing Forensic Science", Issues in Science and Technology, Fall 2003, pp.33-34; D.M. Risinger and M.J. Saks, "A House with No Foundation", Issues in Science and Technology, Fall 2003, pp.35-39; A.P.A. Broeders, "Of fingerprints, scent dogs, cot deaths, and cognitive contamination - a brief look at the present state of play in the forensic arena", Forensic Science International, (2005) Available on-line at www.sciencedirect.com; W.G. Schulz, "Judging Science", Chemistry and Engineering News, February 27, 2006, pp.36-39.

10. P.C. Giannelli, "Forensic Science", Journal of Law, Medicine and Ethics, Fall 2005, pp.535-544.

11. M.J. Saks and J.J. Koehler, "The Coming Paradigm Shift in Forensic Identification Science", Science **309**, pp.892-895, (2005).

12. National Research Council, Committee on Scientific Assessment of Bullet Lead Elemental Composition Comparison, Forensic Analysis: Weighing Bullet Lead Evidence, (National Academies Press, Washington DC, 2004).

13. M.O. Finkelstein and B. Levin, "Compositional Analysis of Bullet Lead as Forensic Evidence", Journal of Law and Policy **13**(1), pp.119-142, (2005).

14. "FBI Laboratory Announces Discontinuation of Bullet Lead Examinations", Press Release, Federal Bureau of Investigation, Washington, DC, September 1, 2005.

15. D.H. Kaye, "The Current State of Bullet-Lead Evidence", Jurimetrics **46**, 99-114, (2006).

16. S.L. Zabell, "Fingerprint Evidence", Journal of Law and Policy **13**(1), pp.143-179, (2005).

17. Daubert v. Merrell Dow Pharmaceuticals, 509 U.S. 579 (1993).

18. General Electric v. Joiner, 522 U.S. 136 (1997).

19. Kumho Tire v. Charmichael, 526 U.S. 137 (1999).

20. M.J. Saks and D.L. Faigman, "Expert Evidence After Daubert", Ann. Rev. Law Soc. Sci. **1**, 105-130, (2005); M.A. Berger, "The Supreme Court's Trilogy on the Admissibility of Expert Testimony", in the *Reference Manual on Scientific Evidence* 2nd edition, Federal Judicial Center (2000.)

21. U.S. Code Title 18, Part1, Chapter 10.

22. U.S. Code Title 22, Chapter 65, §5604.

23. M. Barletta, "Chemical Weapons in the Sudan: Allegations and Evidence", The Nonproliferation Review, Fall 1998, pp. 115-136.

24. Imwinkelried, E.J., *The Methods of Attacking Scientific Evidence*, 4th Ed. (LexisNexis, 2004).

25. S.P. Velsko, "Bioagent sample matching using elemental composition data: an approach to validation", UCRL-TR-220803 Lawrence Livermore National Laboratory, April 21, 2006.

26. B. Robertson and G.A. Vignaux, *Interpreting Evidence: Evaluating Forensic Science in the Courtroom*, (J. Wiley and Sons, Ltd, 1995.)

27. Federal Rules of Evidence, December 31, 2004.

28. National Research Council, Committee on DNA Forensic Science, "The Evaluation of Forensic DNA Evidence", (National Academy Press, Washington, D.C. 1996)

29. R. Lempert, "Modeling Relevence", Michigan Law Review Vol. 75, pp.1021-1057, (1977)

30. D.H. Kaye and J. J. Koehler, "The Misquantification of Probative Value", Law and Human Behavior 27, 645-659, (2003).

31. Lucy, D. *Introduction to Statistics for Forensic Scientists*, (John Wiley & Sons Ltd., West Sussex, UK, 2005)

32. C. Aitken and F. Taroni, Statistics and the evaluation of Evidence for Forensic Scientists, 2nd ed., (John Wiley & Sons Ltd., 2004).

33. D.A. Nance and S.B. Morris, "An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Large and Quantifiable Random Match Probability", Jurimetrics **42**, 403-453, (2002).

34. D.A. Nance and S.B. Morris, "Jury Understanding of DNA Evidence: An Empirical Assessment of Presentation Formats for Trace Evidence with a Relatively Small Random Match Probability", The Journal of Legal Studies **34**, 395-444, (2005).

35. M.M. Houck, "Statistics and trace evidence: The tyranny of numbers", Forensic Science Communications **1**(3), October 1999.

36. M.M. Houck, R.E. Bisbing, T.G. Watkins, and R.P. Harmon, "Locard Exchange: The Science of Hair Comparisons and the Admissibility of Hair Comparison Evidence: Frye and Daubert Considered", Modern Microscopy Online Journal, March 2004.

37. Zweig, M.H., "Evaluation of the Clinical Accuracy of Laboratory Tests", Arch. Pathol. Lab. Med., **112**, pp.383-386, (1988).

38. NCCLS. "Assessment of the Clinical Accuracy of Laboratory Tests Using Receiver Operating Characteristics (ROC) Plots; Approved Guideline. NCCLS Document GP10-A, December, 1995.

39. M.S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction* (Oxford Statistical Science Series 28; Oxford University Press, New York, 2003)

40. An extensive tutorial on ROC analysis that is oriented towards medical applications is provided at: http://www.anaesthetist.com/mnm/stats/roc/index.htm.

41. Zweig, M.H., "Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine", Clin. Chem. **39**, pp.561-577, (1993).

42. J. A. Hanley and B.J. McNeil, "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve", Radiology 143, 29-36, (1982).

43. Gryzbowski, M., "Statistical Methodology III. Receiver Operating Characteristic (ROC) Curves", Academic Emergency Medicine **4**, pp.818-826, (1997).

44. Zhou, K.H., et. al., "Smooth Non-parametric Receiver Operating Characteristic (ROC) Curves for Continuous Diagnostic Tests", Statistics in Medicine, **16**, pp.2143-2156, (1997); Zhou, K.H., et. al., "Statistical Validation Based on Parametric Receiver Operating Characteristic Analysis of Continuous Classification Data", Academic Radiology, 10, pp.1359-1368, (2003).

45. van Erkel, A.R. and Pattynama, "Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology", European Journal of Radiology **27**, pp.88-94, (1998).

46. M. Greiner et. al., "Principles and practical application of the receiver-operating characteristic analysis for diagnostic tests", Preventive Veterinary Medicine **45**, pp.23-41 (2000).

47. Kester, A.D.M, and Buntinx, F., "Meta-Analysis of ROC curves", Medical Decision Making, **20**, pp.430-439, (2000).

48. Spiehler, V.R., et. al., "Confirmation and Certainty in Toxicology Screening", Clin. Chem. **34**, pp.1535-1539, (1988)

49. Spiehler, V., "Enzyme immunoassay validation for qualitative detection of cocaine in sweat", Clin. Chem. **42**, pp. 34-38, (1996).

50. D. C. Fuller, "A Statistical Approach to the Prediction of Verifiable Heroin Use from Total Codeine and Total Morphine Concentrations in Urine", J. Forensic Sci., **42**, 685-689, (1997).

51. C.L. O'Neal and A. Poklis, "An Evaluation of the Role of ROC Plots in the Prediction of Heroin Use from Total Codeine and Total Morphine Concentrations in Urine", J. Analytical Toxicology, **22**, 487-492 (1998).

52. D.K. Wittaker, M.R. Brickley and L. Evans, "A comparison of the ability of experts and non-experts to differentiate between adult and child human bite marks using receiver operating characteristic (ROC) analysis", Forensic Science International, **92**, 11-20, (1998).

53. I.B. Collison, et. al., "Setting Cutoff Concentrations for Immunoassay Screening of Postmortem Blood", J. Forensic Sci., **43**, 390-394, (1998).

54. I.A. Pretty and D. Sweet, "Digital Bite Mark Overlays – An Analysis of Effectiveness", J. Forensic Sci., **46**, 1385-1391, (2001).

55. K.L. Arheart and I.A. Pretty, "Results of the 4th ABFO Bitemark Workshop – 1999", Forensic Science International **124**, 104-111, (2001).

56. C. Bennett and D.V. Canter, "Linking commercial burglaries by *modus operandi*: tests using regression and ROC analysis", Science and Justice **42**, 153-164 (2002).

57. B. Pfister and R. Beutler, "Estimating the Weight of Evidence in Forensic Speaker Verification", Proceedings of Eurospeach, 2003.

58. F. M. Wurst, et. al., "Concentration of Fatty Acid Ethyl Esters in Hair of Alcoholics: Comparison to Other Biological State Markers and Self Reported-Ethanol Intake", Alcohol and Alcoholism **39**, 33-38, (2004).

59. V.R. Spiehler, et. al. "Screening Postmortem Whole Blood for Oxycodone by ELISA Response Ratios", J. Forensic Sci., **49**, 621-626, (2004).

60. S.E. Fienberg and P.C. Stern, "In Search of the Magic Lasso: The Truth About the Polygraph", Statistical Science **20**, 249-260, (2005).

61. S. Martin-de las Heras, et. al., "Effectiveness of Comparison Overlays Generated with DentalPrint Software in Bite Mark Analysis", J. Forensic Sci, **52**, 151-156, (2007)

62. Ou,C.Y..; Ciesielski, C.A.,; Myers, G.; Bandea, C.I.,; Luo, C.C.; Korber, B.T.M.; Mullins, J.I.; Schochetman, G.,; Berkelman, R.L.; Economou, N.A.; Witte, J.J.; Furman, L.J; Satten, G.A.; Maclnnes, K.A.; Curran, J.W.; Jaffe, H.W.; Laboratory Investigation Group, Epidemiologic Investigation Group "Molecular Epidemiology of HIV Transmission in a Dental Practice". Science, 1992, Vol 256, pp. 1165-1171.

63. United States General Accounting Office "AIDS CDC's Investigation of HIV Transmissions by a Dentist" GAO/PEMD-92-31, 1992, pp. 2-50.

64. Rogers, A.S. PhD MPH; Froggatt III J.W. MD; Townsend, T MD; Gordon, T. ScD; Brown Leigh, A.J. PhD; Holmes, E.C. PhD; Zhang, L.Q. PhD, MSc; Moses III, H. MD "Investigation of Potential HIV Transmission to the Patients of an HIV-Infected Surgeon". JAMA,1993, Vol 269 No. 14, pp. 1795-1801.

65. Holmes, E.C.; Zhang, L.Q.; Simmonds, P.; Rogers, A.S.; Brown Leigh, A.J. "Molecular Investigation of Human Immunodeficiency Virus (HIV) Infection in a Patient of an HIV-Infected Surgeon". The Journal of Infectious Diseases, 1993, Vol 167, pp. 1411-14.

66. Albert, J.; Wahlberg, J. Leitner, T.; Escanilla, D.; Uhlen, M. “Analysis of a Rape Case by Direct Sequencing of the Human Immunodeficiency Virus type 1 pol and gag Genes”. Journal of Virology, 1994, Vol 68, No 9, pp. 5918-5924.

67. Court of Appeal of Louisiana, Third Circuit, State of Louisiana v. Richard J. Schmidt “ No. K97-249”. Westlaw, 1997, 669 So. 2d 448, 97-249, pp. 1-8.

68. Blanchard, A.; Ferris, Stephane; Chamaret, Sophie; Guetard, D.; Montagnier, L. “Molecular Evidence for Nosocomial Transmission of Human Immunodeficiency Virus from a Surgeon to One of His Patients”. Journal of Virology, 1998, pp.4537-4540.

69. Frenkel, L.M.; Mullins, J.L.; Learn, G.H.; Manns-Arcuino, L.; Herring, B.L.; Kalish, M.L.; Steketee, R.W.; Thea, D.M.; Nichols, J.E.; Liu, S.L; Harmache, A.; He, X.; Muthui, D.; Madan, A.; Hood, L.; Haase, A.T.; Zupancic, M.; Staskus, K.; Wolinsky, S.; Krogstad, P.; Zhao, J.; Chen, I.; Koup, R.; Ho, D.; Korber, B.; Apple,R.J.; Coombs, R.W.; Pahwa, S.; Roberts Jr., N.J. “Genetic Evaluation of Suspected Cases of Transient HIV-1 Infection of Infants”. Science, 1998, Vol 280, pp. 1073-1077.

70. Lot, F. MD.; Seguier, J.C. MD; Fegueux, S.MD; Astagneau, P.MD, PhD; Simon, P. MD; Aggoune, M.; van Amerongen, P. MD; Ruch, M. MD; Cheron, M. MD; Brucker, G. MD; Desenclos, J.C. MD; Drucker, J. MD, MSc, Probable Transmission of HIV from an Orthopedic Surgeon to a Patient in France”. Annals of Internal Medicine, 1999, Vol 130, pp. 1-6.

71. Birch, C.J.; McCaw, R.F.; Bulach, D.; Revill, P.A.; Carter, J.T.; Tomnay, J.; Hatch, B.; Middleton, T.V.; Chibo, D.; Catton, M. G.; Pankhurst, J.L.; Breschkin, A.M.; Locarnini, S.A.; Bowden, D.S. “ Molecular Analysis of Human Immunodeficiency Virus Strains Associated with a Case of Criminal Transmission of the Virus”. The Journal of Infectious Diseases, 2000, pp. 941-4.

72. “Court of Appeal of Louisiana Third Circuit, State of Louisiana v. Richard J. Schmidt No. 99-1412”, 2000, pp. 1-30.

73. Banaschak, S.; Werwein, M.’ Brinkmann, B; Hauber, L.” Human Immunodeficiency Virus Type 1 Infection after Sexual Abuse: Value of Nucleic Acid Sequence Analysis in Identifying the Offender”. Clinical Infectious Diseases, 2000, pp.1098-100.

74. Machuca, R.; Jorgensen, L.B.; Theilade, P.; Nielsen, C., “Molecular Investigation of Transmission of Human Immunodeficiency Virus Type 1 in a Criminal Case”. Clinical and Diagnostic Laboratory Immunology, 2001, Pp. 884-890.

75. Metzker, M.L.; Mindell, D.P.; Liu, X.M.; Ptak, R.; Gibbs, R.A.; Hillis, D.M.,”Molecular evidence of HIV-1 transmission in a criminal case”. PNAS, 2002, Vol 99, No 22, pp. 14292-14297.

76. Ceballos, A.; Andreani, G.; Gonzalez Ayala, S.E.; Romer, Y.; Rimoldi, I.; Agosti, M.R.; Peralta, L.M. "Epidemiological and Molecular Evidence of Two Events of Father-to-Child HIV Type 1 Horizontal Transmission". AIDS Research and Human Retroviruses, 2004, Vol 20, No. 8, pp. 789-793.

77. N. Kazarinova-Fukshansky and K. Hummel, "A new definition of genetic distance", Hum. Genet. **87**, 745-747, (1991).

78. E. Baake and A. von Haeseler, "Distance Measures in Terms of Substitution Processes", Thoretical Population Biology **55**, 166-175, (1999).

79. S.T. Kalinowski, "Evolutionary and statistical properties of three genetic distances", Molecular Ecology **11**, 1263-1273, (2002).

80. D.C. Hoyle and P.G. Higgs, "Factors Affecting the Errors in the Estimation of Evolutionary Distances Between Sequences", Mol. Biol. Evol. **20**, 1-9, (2003).

81. I. Elias and J. Lagergren, "Fast Computation of distance estimators", BMC Bioinformatics **8**:89, (2007).

82. C. Kuiken, B. Korber, and R.W. Shafer, "HIV Sequence Databases", Aids Rev. 5, 52-61, (2003).

83. C. Kuiken, K. Yusim, L. Boykin, and R. Richardson, "The Los Alamos hepatitis C sequence database", Bioinformatics 21, 379-384, (2005).

84. Burger, H.,Weiser, B.; Flaherty, K.; Gulla, J.; Nguyen, P.; Gibbs.R. "Evolution of human immunodeficiency virus type 1 nucleotide sequence diversity among close contacts". Proc. Natl. Acad. Sci. USA, 1991, Vol. 88, pp. 11236-11240.

85. Wolfs, T.F.W.; Zwart, G.; Bakker, M.; Goudsmit, J. "HIV-1 Genomic RNA Diversification following Sexual and Parenteral Virus Transmission". Journal of Virology, 1992, Vol 189, pp. 103-110.

86. Neiburger, E.J. DDS, "Fuzzy science: A look at CDC's analysis of the Acer case". CDC Review, 1996, pp. 22-27.

87. Budowie, B.; Harmon, R. "HIV Legal Precedent Useful for Microbial Forensics". Croat Med J, 2005, pp. 514-521.

88. Leitner, T.; Albert, J., "Reconstruction of HIV-1 transmission chains for forensic purposes". AIDS Review, 2000, pp. 241-251.

89. Learn, G.H.; Mullins, J.L. "The Microbial Forensic Use of HIV Sequences". HIV Forensics, pp. 22-37.

90. S.P. Velsko, "Physical and Analytical Chemical Analysis: A key component of Bioforensics", Presented at the AAAS Annual Meeting, February 2005, Washington D.C. available from Lawrence Livermore National Laboratory as UCRL-CONF-209735, (2005).

91. Whiteaker, JR, Fenselau, CC, Fetterolf, D, Steele, D and Wilson, D, "Quantitative Determination of Heme for Forensic Characterization of Bacillus Spores Using Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry", Anal. Chem. **76**, 2836-2841, (2004).

92. A number of programs for calculating exact confidence intervals for proportions are available on the Web. This calculation used a Bayesian estimator for the shortest confidence interval at www.causascientia.org/math_stat/ProportionCI.html.

93. F. M. Stone and C.B. Coulter, "Porphyrin Compounds Derived from Bacteria", J. General Physiology **15**, 629-639, (1932).

94. N.J.Jacobs, J.M. Jacobs, and P. Brent, "Formation of Protoporphyrin from Coproporphyrinogen in Extracts of Various Bacteria", J. Bacteriology, **102**, 398-403, (1970).

95. N. Wolff, et. al. , "Histidine pK_a shifts and changes of tautomeric states induced by the binding of gallium-protoporphyrin IX in the hemophore $HasA_{SM}$", Protein Science **11**, 757-765, (2002).

96. S. Midha, et. al., "Cloning, expression, and characterization of recombinant nitric oxide synthase-like protein from *Bacillus Anthracis*", Biochemical and Biophysical Research Communications **336**, 346-356, (2005).

97. I. Salard, et. al., "Analogies and surprising differences between recombinant nitric oxide synthase-like proteins from *Staphylococcus aureus* and *Bacillus anthracis* in their interactions with L-arginine analogs and iron ligands", J. Inorganic Biochemistry **100**, 2024-2033, (2006).

98. E.M. Galimov, V.S. Sevastyanov, E.V. Kulbachevskaya, and A.A. Golyavin, "Isotope ratio Mass spectrometry: $\delta^{13}C$ and $\delta^{15}N$ analysis for tracing the origin of illicit drugs", Rapid Commun. Mass Spectrom. **19**, 1213-1216, (2005), and references therein.

99. Kreutzer-Martin H.W., et. al., "Microbe forensics: Oxygen and hydrogen stable isotope ratios in Bacillus subtilis cells and spores", PNAS **100**, pp.815-819, (2003).

100. Horita, J. and Vass, A.A., "Stable-isotope Fingerprints of Biological Agents as Forensic Tools", J. Forensic Sci. **48**, pp.122-126, (2003).

101. Kreutzer-Martin H.W., et. al. "Stable Isotope Ratios as a Tool in Microbial Forensics – Part 1. Microbial Isotopic Composition as a Function of Growth Medium", J. Forensic Science **49**, pp.954-960, (2004).

102. Kreutzer-Martin H.W., et. al. "Stable Isotope Ratios as a Tool in Microbial Forensics – Part 2. Isotopic Variation Among Different Growth Media as a Tool for Sourcing Origins of Bacterial Cells or Spores", J. Forensic Science **49**, pp.961-967, (2004).

103. Kreutzer-Martin H.W., et. al. "Stable Isotope Ratios as a Tool in Microbial Forensics – Part 3. Effect of Culturing on Agar-containing Growth Media", J. Forensic Science **50**, pp.1372-1379, (2005).

104. L. Jerrold, "The admissibility of digital images", American Journal of Orthodontics and Dentofacial Orthopedics, September 2006, pp.417-419.

105. L. Berlin, "Letter: Missed Mammographic Abnormalities, Malpractice, and Expert Witnesses: Does Majority Rule in the Courtroom?", Radiology 299, 288-289, (2003).

106. *Trace Evidence Recovery Guidelines*, Scientific Working Group on Materials Analysis (SWGMAT) Evidence Committee; Forensic Science Communications Vol. 1, (1999).

107. M.A. LeBeau, "Quality Assurance Guidelines for Laboratories Performing Forensic Analysis of Chemical Terrorism", Forensic Science Communications Vol.6, April 2004.

108. Scientific Working Group on Forensic Analysis of Chemical Terrorism (SWGFACT) "Validation Guidelines for Laboratories Performing Forensic Analysis of Chemical Terrorism", Forensic Science Communications, Vol.7, April 2005.

109. Scientific Working Group on Microbial Genetics and Forensics (SWGMGF) "Quality Assurance Guidelines for Laboratories Performing Microbial Forensic Work", June 20, 2003. *Science* Supporting Online Material, doi:10.1126/science.1090270.

Appendix 1. Likelihood ratios for parameter estimation.

This appendix discusses a simple method for constructing likelihood ratios from calibration data. Let $\mathbf{R_0}$ be the hypothesis that the parameter ρ associated with a questioned sample (for example, the age of the sample) falls into a certain range of values. The parameter of the sample is correlated with the value of some scalar metric (i.e. scalar quantity) μ that can be measured on the sample (for example, the ^{14}C concentration in the DNA of the agent particles.) If the measured range of values of μ in the questioned sample is $\mathbf{m_0}$, we want to know how much the observation of values falling into $\mathbf{m_0}$ increases the odds that the parameter falls within range $\mathbf{R_0}$. That is, we want to calculate the likelihood ratio $\mathbf{P(m_0|R_0)/P(m_0|\ \bar{R}_0)}$ where $\mathbf{P(m_0|R_0)}$ is the probability that a sample whose parameter ρ falls into range $\mathbf{R_0}$ will exhibit a metric value μ that falls into range $\mathbf{m_0}$, and $\mathbf{P(m_0|\ \bar{R}_0)}$ is the probability that a sample whose parameter falls <u>outside</u> $\mathbf{R_0}$ will exhibit a metric value that falls into the range $\mathbf{m_0}$.

Assume that we have a set of reference samples with known values of the parameter ρ, and these were drawn at random from among the complete population of possible types of such samples (examples of this are discussed in the main body of the paper.) For each reference sample, the value of the metric μ is measured. The calibration curve is a plot of all pairs of μ,ρ obtained from these samples, covering a total range $(\mathbf{m_T},\mathbf{R_T})$. $(\mathbf{m_T},\mathbf{R_T})$ forms the working range of the calibration set. Such a plot is illustrated in Figure A1.1.

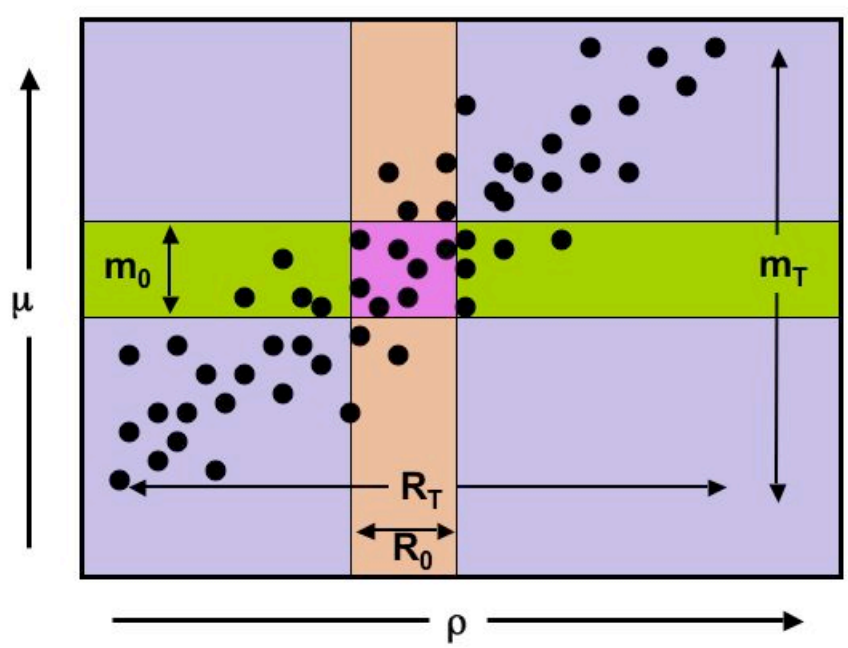

Figure A1.1. Plot of calibration data with color coded regions corresponding to the subsets of points in the contingency table A1. Purple: $\mathbf{(m_0,R_0)}$; Green: $\mathbf{(m_0,\ \bar{R}_0)}$; Orange: $\mathbf{(\ \bar{m}_0,\ R_0)}$; Blue: $\mathbf{(\ \bar{m}_0,\ \bar{R}_0)}$. The total working range of the data is $(\mathbf{m_T},\mathbf{R_T})$.

Suppose now that several replicate measurements are made on a questioned sample, and fall into a range of values $\mathbf{m_0}$. This range is represented by the green band in Figure A1. How much support does this provide to the hypothesis that the parameter we are trying to determine falls into the range $\mathbf{R_0}$? The orange band in Figure A1 denotes $\mathbf{R_0}$. These two bands divide the set of calibration points into four regions. The blue region captures points that fall outside of both $\mathbf{m_0}$ and $\mathbf{R_0}$, while the green region captures points that fall within $\mathbf{m_0}$, but not within $\mathbf{R_0}$, etc. Figure A1.2 is a contingency table denoting the correspondence between the colored regions and the number of points captured by each.

Figure A2. Contingency table corresponding to figure A1.

	R_0	$\bar{R}_0$	
m_0	$N(R_0;m_0)$	$N(\bar{R}_0;m_0)$	$N(m_0)$
$\bar{m}_0$	$N(R_0;\bar{m}_0)$	$N(\bar{R}_0;\bar{m}_0)$	$N(\bar{m}_0)$
	$N(R_0)$	$N(\bar{R}_0)$	N_T

$N(\mathbf{m_0})$ is the total number of points falling into the band defined by $\mathbf{m_0}$, while $N(\mathbf{R_0})$ is the total number of points falling into $\mathbf{R_0}$. Similarly $N(\bar{\mathbf{m}}_0)$ and $N(\bar{\mathbf{R}}_0)$ represent the number of points that fall into the complementary regions, respectively. Clearly:

$$N(\mathbf{m_0}) = N(\mathbf{R_0}, \mathbf{m_0}) + N(\bar{\mathbf{R}}_0, \mathbf{m_0}) \tag{A1.1}$$

$$N(\bar{\mathbf{m}}_0) = N(\mathbf{R_0}, \bar{\mathbf{m}}_0) + N(\bar{\mathbf{R}}_0, \bar{\mathbf{m}}_0) \tag{A1.2}$$

and

$$N(\mathbf{R_0}) = N(\mathbf{R_0}, \mathbf{m_0}) + N(\mathbf{R_0}, \bar{\mathbf{m}}_0) \tag{A1.3}$$

$$N(\bar{\mathbf{R}}_0) = N(\bar{\mathbf{R}}_0, \mathbf{m_0}) + N(\bar{\mathbf{R}}_0, \bar{\mathbf{m}}_0) \tag{A1.4}$$

so that

$$N(\mathbf{R_0}) + N(\bar{\mathbf{R}}_0) = N(\mathbf{m_0}) + N(\bar{\mathbf{m}}_0) = \mathbf{N_T}. \tag{A1.5}$$

Assuming that the points in the calibration set are representative of the population, and are distributed uniformly over the total working range $\mathbf{R_T}$, the joint probabilities of finding a point within a certain colored region are estimated by dividing each entry in the contingency table by $\mathbf{N_T}$, the total number of points in the calibration set. For example,

$$P(\mathbf{R_0}, \mathbf{m_0}) = N(\mathbf{R_0}, \mathbf{m_0})/\mathbf{N_T} \tag{A1.6}$$

and

$$P(\bar{\mathbf{R}}_0, \mathbf{m_0}) = N(\bar{\mathbf{R}}_0, \mathbf{m_0})/\mathbf{N_T}\ , \tag{A1.7}$$

while

$$P(\mathbf{m_0}) = N(\mathbf{m_0})/\mathbf{N_T} \tag{A1.8}$$

and

$$P(\mathbf{R_0}) = N(\mathbf{R_0})/\mathbf{N_T}. \tag{A1.9}$$

Finally, the conditional probabilities are formed in the usual way:

$$P(\mathbf{m_0}|\mathbf{R_0}) = P(\mathbf{R_0},\mathbf{m_0})/P(\mathbf{R_0}), \tag{A1.10}$$

$$P(\mathbf{m_0}|\ \bar{\mathbf{R}}_0) = P(\bar{\mathbf{R}}_0,\mathbf{m_0})/P(\bar{\mathbf{R}}_0), \text{ etc.} \tag{A1.11}$$

Thus, the likelihood ratio that relates the measurement of $\mathbf{m_0}$ to the range hypothesis $\mathbf{R_0}$ can be determined directly from the distribution of points on the calibration graph.

In figure A1.1, the centroid of the band $\mathbf{R_0}$ was deliberately placed so that the purple intersection area overlapped with the distribution of data points. One could imagine sliding the centroid of $\mathbf{R_0}$ band back and forth along the ρ axis, changing the count statistics in the contingency table A1.2. Clearly for a fixed $\mathbf{m_0}$ band, there is a maximum in the likelihood ratio when the centroid of $\mathbf{R_0}$ is near the position shown in the graph, because that position simultaneously maximizes the number of points that fall into ($\mathbf{m_0},\mathbf{R_0}$) while minimizing those falling into ($\mathbf{m_0}, \bar{\mathbf{R}}_0$). This corresponds to the fact that the measurement $\mathbf{m_0}$ lends the most support to the $\mathbf{R_0}$ whose centroid is given by that position.

Once the maximum likelihood centroid for $\mathbf{R_0}$ is determined, the width of $\mathbf{R_0}$ can be varied as well. Clearly increasing the width of $\mathbf{R_0}$ increases the likelihood ratio, but at the expense of making the uncertainty in the parameter value larger. A reasonable convention for choosing the width of $\mathbf{R_0}$ might be to fix it at the value that provides a given "false alarm probability" $P(\mathbf{m_0}|\ \bar{\mathbf{R}}_0)$, e.g. 5% or 1%. Note that since $N(\mathbf{m_0},\mathbf{R_0})$ and $N(\mathbf{m_0}, \bar{\mathbf{R}}_0)$ are *both* approximately proportional to the width of $\mathbf{m_0}$, the likelihood ratio does not depend on this value, as long as it is not too large. Thus, in practical calculations from actual sparse data sets, one can choose a width for $\mathbf{m_0}$ that is larger than

the actual uncertainty in the metric in order to increase the precision of the likelihood calculation.

Appendix 2. Source attribution in disease transmission

In this appendix we consider the relationship between the hypothesis T_0, that two isolates (actually two viral populations) are related by a direct transmission event, and the hypothesis S_0, that a certain source S (e.g. another infected person) is the origin of the infection. These two hypotheses are not strictly equivalent because although S_0 may be true, the disease may have infected the victim via a third (or n^{th}) party route. Obviously the prosecution in a criminal HIV transmission case would like to argue that evidence that lends weight to T_0 also lends weight to S_0. The reader should note that there is a close analogy between the analysis in this appendix and that given appendix 1 of reference A.1, which treats the case of sample matching.

To be precise, let

$S_0 \equiv$ the hypothesis that S is the source of the infecting isolate

$T_0 \equiv$ the hypothesis of direct transmission from S

We can define the joint hypotheses:

S_0T_0 = S is the source, and infection of the victim was by direct transmission

$S_0\ \overline{T}_0 \equiv$ S is the source, but infection was through other intermediate parties

$\overline{S}_0\ \overline{T}_0 \equiv$ S was not the source, and there was no direct transmission event

The fourth possible hypothesis, $\overline{S}_0T_0$, is automatically false because if S_0 is false then so is T_0 by definition.

Now consider two isolates, one from S and the other from the victim, that are suspected to be related by direct transmission. The odds that S is the source, given the measured genetic distance Δ between the isolates is given by the usual equation:

$$O(S_0|\Delta) = LR_s(\Delta)\bullet O(S_0) \qquad \text{(A2.1)}$$

where $LR_s(\Delta) = P(\Delta|S_0)/P(\Delta|\ \overline{S}_0)$. We are interested in the relationship between LR_s and LR_t, defined as $P(\Delta|T_0)/P(\Delta|\ \overline{T}_0)$.

We can expand the conditional probability expressions that appear in the likelihood ratio $LR_s(\Delta)$ in terms of S_0 and T_0 by using standard chain rule relationships:

$$P(\Delta|S_0) = P(T_0|S_0)\bullet P(\Delta|S_0T_0) + P(\ \overline{T}_0|S_0)\bullet P(\Delta|S_0\ \overline{T}_0) \qquad \text{(A2.2)}$$

$$P(\Delta|\ \bar{S}_0) = P(T_0|\ \bar{S}_0)\bullet P(\Delta|\ \bar{S}_0T_0) + P(\ \bar{T}_0|\ \bar{S}_0)\bullet P(\Delta|\ \bar{S}_0\ \bar{T}_0) \tag{A2.3}$$

Equation (A2.3) can be simplified by noting that by definition $P(T_0|\ \bar{S}_0) = 0$, thus

$$P(\ \bar{T}_0|\ \bar{S}_0) = 1 \tag{A2.4}$$

and

$$P(\Delta|\ \bar{S}_0) = P(\Delta|\ \bar{S}_0\ \bar{T}_0). \tag{A2.5}$$

Equation (A2.4) has the natural interpretation that if S is not the source of the infection, then there had to be some other transmission route as well.

Thus,

$$LR_s(\Delta) = P(T_0|S_0)\bullet\ [P(\Delta|S_0T_0)/P(\Delta|\ \bar{S}_0\ \bar{T}_0)]\ + P(\ \bar{T}_0|S_0)\bullet\ [P(\Delta|S_0\ \bar{T}_0)/P(\Delta|\ \bar{S}_0\ \bar{T}_0)] \tag{A2.6}$$

The conditional probability functions in equation (A2.6) are defined in Table A2.1. Clearly the conditional probabilities $P(T_0|S_0)$ and $P(\ \bar{T}_0|S_0)$ depend on other evidence that might bear on whether direct transmission between S and the victim is likely. Note that because both terms in equation (A2.6) are always ≥ 0,

$$LR_s(\Delta) \geq P(T_0|S_0)\bullet[P(\Delta|S_0T_0)/P(\Delta|\ \bar{S}_0\ \bar{T}_0)] \tag{A2.7}$$

Note also that

$$P(\Delta|T_0) = P(\Delta|S_0T_0) \tag{A2.8}$$

And since

$$P(\Delta|\ \bar{T}_0) = P(\Delta|S_0\ \bar{T}_0) + P(\Delta|\ \bar{S}_0\ \bar{T}_0) \tag{A2.9}$$

We have

$$P(\Delta|\ \bar{T}_0) \geq P(\Delta|\ \bar{S}_0\ \bar{T}_0) \tag{A2.10}$$

So that

$$P(\Delta|S_0T_0)/P(\Delta|\ \bar{S}_0\ \bar{T}_0) \geq P(\Delta|T_0)/P(\Delta|\ \bar{T}_0) \tag{A2.11}$$

Thus

$$LR_s(\Delta) \geq P(T_0|S_0)\bullet LR_t(\Delta) \tag{A2.12}$$

Expression (A2.12) shows how the strength of the evidence for T_0 bounds the strength of the evidence that S is the source. Suppose there is a case where there is a clear

opportunity for S to have infected the victim and there are no other plausible indirect routes for transmission, one could argue that $P(T_0|S_0) \approx 1$. Hence a value of Δ that supports the hypothesis of direct transmission also supports the hypothesis that S is the source. Conversely, if there is evidence for other possible indirect transmission routes (e.g. common sexual partners in an HIV transmission case) so that $P(T_0|S_0) < 1$, then $LR_t(\Delta)$ must be larger than $1/P(T_0|S_0)$ in order to support the hypothesis that S directly transmitted the virus to the victim.

Table A2.1. Definitions of the probability functions contributing to $LR_s(\Delta)$.

Function	Interpretation	Comment
$P(T_0\|S_0)$	Probability of direct transmission from S, *given* that S is the source	Depends on other evidence showing that S had the opportunity to transmit the disease to the victim
$P(\bar{T}_0\|S_0)$	Probability of an indirect transmission route from S, *given* that S is the source	In the case of HIV, for example, this would depend on whether there were mutual sexual partners of S and the victim
$P(\Delta\|S_0T_0)$	Probability of observing a genetic distance Δ between the isolate from S and the victim's isolate, *given* that S is the source *and* there was direct transmission.	Estimated by sampling pairs of isolates known to be related by direct transmission
$P(\Delta\|S_0\ \bar{T}_0)$	Probability of observing a genetic distance Δ between the isolate from S and the victim's isolate, *given* that S is the source, but the transmission is via intermediate parties	Estimated by sampling pairs of isolates known to be related by indirect transmission via one or more nodes
$P(\Delta\|\ \bar{S}_0\ \bar{T}_0)$	Probability of observing a genetic distance Δ between the isolate from S and the victim's isolate, *given* that S is *not* the source.	Estimated by sampling pairs of isolates that are not directly or indirectly linked by transmission

Equation (A2.12) is an important relationship from a practical point of view, because the process of validating $LR_t(\Delta)$ is considerably easier than validating $LR_s(\Delta)$ directly. To understand this, one must consider the collection of isolates that constitute the population from which the conditional probability distributions $P(\Delta|S_0T_0)$, $P(\Delta|S_0\ \bar{T}_0)$, and $P(\Delta|\ \bar{S}_0\ \bar{T}_0)$ are estimated. The set of archived viral isolates (or genetic sequence data) available for constructing these distributions can itself be considered (conceptually) a sampling of the complete set of viral subpopulations that are defined by the worldwide (and historical) transmission network that relates all persons ever infected by the virus. In Figure A2.1 we display a small portion of this network, whose nodes represent infected persons. Clearly not all such nodes have been sampled, which implies that in many cases we do not know the precise relationship between isolates in an archive.

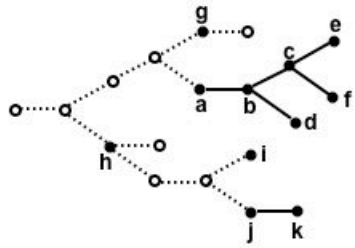

Figure A2.1. Part of an imaginary viral transmission network. Open circles are unknown nodes in the transmission chain (un-sampled isolates), while filled circles represent sampled isolates from known nodes. Dotted and solid lines represent unknown and known transmission events, respectively. Nodes a-f represent a known cluster of epidemiologically related isolates. Nodes j and k represent a known transmission related pair. Other relationships are assumed to be unknown.

If we define

$\{S_0T_0\} \equiv$ the set of all pairs of isolates related by direct transmission,

$\{S_0\ \overline{T}_0\} \equiv$ the set of all pairs of isolates related by indirect transmission, and

$\{\ \overline{S}_0\ \overline{T}_0\} \equiv$ the set of all pairs of isolates that are not related by direct or indirect transmission,

then the various pairwise comparisons between the labeled isolates in Figure A2.1 can be classified as shown in Table A2.2. Note that in some cases we know the relation between isolates. For example, in the large cluster of epidemiologically related isolates a-f in Figure A2.1, the relation of each pair is known. Similarly, an archive might also contain isolated pairs of isolates like j and k whose relationship is known. However, other isolates (e.g. g,h and i) may not have any known relation to any other, and this generally limits their utility in validating $LR_s(\Delta)$, but not $LR_t(\Delta)$.

Referring again to Figure A2.1, a real archive will have isolates like (h,j) where there is an indirect transmission relationship that is not known. Similarly the pair (i,j) belong to the set of pairs not linked by transmission, but this would not be known. Hence, when arbitrary samples are drawn from such an archive it is not clear how to categorize a pair as $S_0\ \overline{T}_0$ or $\overline{S}_0\ \overline{T}_0$, and the presumption is that they are members of the set of unrelated isolates, $\{\ \overline{S}_0\ \overline{T}_0\}$ may not be warranted. Note that 39 of 55 of the pairwise relationships in Figure A2.1 are unknown. Similarly, in a large archive of isolates (or a sequence database derived from it) a majority of the relationships among arbitrary pairs will not be known.

These considerations imply that only isolates that belong to a known epidemiological cluster with a known transmission history can be used to construct separate distribution functions $P(\Delta|S_0\ \overline{T}_0)$, and $P(\Delta|\ \overline{S}_0\ \overline{T}_0)$. On the other hand, since $\{\ \overline{T}_0\} = \{S_0\ \overline{T}_0\} \cup \{\ \overline{S}_0\ \overline{T}_0\}$, there is less difficulty in constructing $P(\Delta|\ \overline{T}_0)$. It should be noted that this

situation is very different from the case of sample matching[A1] where the "archive" is only virtual and one can choose the labs, methods and batches from which samples are drawn so that their relationships are always known.

Table A2.2. Classification of islolate pairs corresponding to figure A2.1.

Node	a	b	c	d	e	f	g	h	i	j	k
a		S_0T_0	$S_0\ \overline{T}_0$	$S_0\ \overline{T}_0$	$S_0\ \overline{T}_0$	$S_0\ \overline{T}_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$
b			S_0T_0	S_0T_0	$S_0\ \overline{T}_0$	$S_0\ \overline{T}_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$
c				$\overline{S}_0\ T_0$	S_0T_0	S_0T_0	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$
d					$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$
e						$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$
f							$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$
g								$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$
h									$S_0\ \overline{T}_0$	$S_0\ \overline{T}_0$	$S_0\ \overline{T}_0$
i										$\overline{S}_0\ T_0$	$\overline{S}_0\ T_0$
j											S_0T_0
k											

Key:

- Known S_0T_0
- Known $S_0\ \overline{T}_0$
- Known $\overline{S}_0\ \overline{T}_0$
- Unknown $S_0\ \overline{T}_0$
- Unknown $\overline{S}_0\ \overline{T}_0$

References

A.1 S. Velsko, "Bioagent Sample matching using elemental composition data: an approach to validation", Lawrence Livermore National Laboratory Technical Report UCRL-TR-220803, April 2006.

Appendix 3. Washing

A factor that complicates bulk analyses for residual growth medium components and similar material signatures is that repeated washing of the harvested agent during post-growth processing can remove such signatures, producing false negative test results. In example 4c it was assumed that common washing protocols would not very effective at removing heme due to its hydrophobic nature. However, for water-soluble materials common washing protocols are more likely to have significant removal rates, and uncontrollable variations in the rigor of the washing process as well as the number of times it is repeated will have a greater affect on the ROC curve of the assay. This raises the practical problem that assays for residual components of growth media are often limited in utility when only very small sample volumes are available for analysis. It also raises a conceptual problem with validating such assays by sampling the unit process

matrix because of the "open ended" nature of washing protocols, which in effect, can have an unlimited number of repetitions.

Consider an analysis for a particular analyte A that is a signature for growth on a particular medium M. Let A_0 be the hypothesis that A is present in the sample and $\bar{A}_0$ the hypothesis that it is absent. We will assume that the analysis for A is based on the measurement of some scalar quantity δ_A (e.g. a mass spectral peak height or area associated with a molecular fragment characteristic of A.) For the sake of argument, we will assume that A_0 is equivalent to M_0, the hypothesis that medium M was used.

Referring to the unit process matrix defined in section 4, the wash step effectively divides the population of processing methods into sub-populations that are defined by the number of times the washing step is repeated. We will designate the number of repetitions associated with each sub-population as N_w.

Given a value of N_w, the conditional probabilities $P(\delta_A|A_0N_w)$ and $P(\delta_A|\ \bar{A}_0N_w)$ can be determined from samples that were grown and processed by different methods drawn from the unit process matrix, but always washed the same number of times. A set of ROC curves, each for a different value of N_w could then be constructed. Clearly one would expect that the higher the value of N_w, the less likely it would be to detect large values of δ_A in a sample of a given size, because of systematic removal of A in the repetitively washed samples.

However, this information by itself would not generally be useful for analyzing a questioned sample unless it were known how many times the sample had been washed. It is more likely that the number of wash steps is an unknown. The composite probability $P(\delta_A|A_0)$, which averages over the number of wash steps, is difficult to construct from the individual probabilities $P(\delta_A|A_0N_w)$ because it requires an unknown weighting factor:

$$P(\delta_A|A_0) = \Sigma\ P(N_w|M_0)\bullet\ P(\delta_A|A_0N_w)$$

where the sum is over each value of N_w. Clearly the probability $P(N_w|M_0)$ that a certain number of washes would be used given that the sample was grown on medium M is difficult to determine with any precision (although it might be estimated in principle by extensive testing among independent laboratories.)

One solution to this problem is to find an independent metric δ_w that is correlated with the number of wash steps. From the calibration curve of δ_w vs N_w (see appendix 1,) it would be possible to estimate N_w, and hence choose the appropriate ROC curve for the agar assay. One example of a possible metric is the total organic carbon (TOC) content in the aqueous extraction medium used to extract A. A more general solution would be to find a metric that is *independent* of the number of washes. For example it is plausible that the ratio δ_A/TOC might be relatively independent of the number of washes because both δ_A and TOC are both proportional to the number of washes. It is also possible that some micro-structural feature of the spores or bacteria depends on whether M is used to grow it, and would not be modified by washing.