UCRL-TR-236634

LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

A validation framework for microbial forensic methods based on statistical pattern recognition

S. P. Velsko

November 16, 2007

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.

A validation framework for microbial forensic methods based on statistical pattern recognition

Stephan P. Velsko
Lawrence Livermore National Laboratory
November 12, 2007

Summary

This report discusses a general approach to validating microbial forensic methods that attempt to simultaneously distinguish among many hypotheses concerning the manufacture of a questioned biological agent sample. It focuses on the concrete example of determining growth medium from chemical or molecular properties of a bacterial agent to illustrate the concepts involved.

1. Introduction

The manner in which the *interpretation* of microbial forensic data is validated in the laboratory and reported in the courtroom is a crucial consideration in the research and development process. In two previous studies[1,2], it was suggested that validating the interpretation of microbial forensic analysis methods could be best accomplished within a framework that utilized ROC curves and estimates of likelihood ratios (ROC/LR). This framework was applied to the validation of sample matching, single hypothesis testing, and calibration in the context of current problems in microbial forensics. A key feature of the ROC/LR approach is that no decision criterion is assumed. Instead, the outcome of a measurement on a questioned sample is converted into an estimate of the likelihood ratio supporting the hypothesis being tested. This mode of expert testimony is consistent with the legal definition of relevance, and hence meets the standard for admissibility embodied in the Federal Rules of Evidence 401 and 402.

Following the anthrax incident of 2001, several published research papers have implied that it might be possible to infer the growth medium that was used to grow biological agent organisms from the chemical or molecular analysis of the agent. Mass spectrometric methods that determine elemental[3], isotopic[4], or biomolecular[5] composition of an agent were suggested for this application. A key feature in each type of analytical scheme is the use of multivariable data sets (i.e. two or more independent measured quantities per sample) to produce a "map" that segregates the data from each medium type into easily discriminated groups of points. This pattern recognition paradigm applies to many other kinds of analyses in which a multivariate data set is used to simultaneously discriminate among multiple hypotheses about the manufacture of a biological agent.

The problem of validating forensic methods that correspond to multiple hypothesis testing can also be treated within the ROC/LR framework. For example, when the analysis is intended to show that a particular growth medium was used, each particular growth medium corresponds to a separate hypothesis. By analogy to the case of the single hypothesis test[2], we are not interested in constructing classifiers, i.e. decision criteria for assigning a data point from a questioned sample to a particular cluster of points associated with a particular medium. Instead, we only seek to estimate the likelihood ratio that relates the prior and posterior odds that the questioned sample belongs to that cluster. Similarly, we avoid the estimation of posterior probabilities for a classification, since it is seldom possible to estimate prior probabilities and the costs of false positive and false negative determinations. (There may be other evidence that bears on prior probability, like a finding of an invoice for the purchase of a certain medium in a suspect's possession, but it is still hard to assign a numerical value to the probability.)

This report discusses the application of the ROC/LR approach to validating microbial forensic methods that attempt to simultaneously distinguish among many hypotheses concerning the manufacture of a questioned biological agent sample. It focuses on the concrete example of determining growth medium from chemical or molecular properties of a bacterial agent to illustrate the concepts involved, but the approach applies generally to any assay that involves multiple hypothesis testing and multivariable signatures. The remainder of this report is divided into 3 sections. Section 2 contains a basic description of how the ROC/LR method applies to assays that are based on statistical pattern recognition, or multiple hypothesis tests. The statistical pattern recognition language used in this section is primarily drawn from reference 6, and may differ from more contemporary usage. It is assumed that the reader is familiar with basic concepts from statistical pattern recognition, or has access to a basic text in this area. In section 3 we address some of the technical issues associated with deriving likelihood ratios from multivariate data. This includes the use of feature extraction techniques, choosing an appropriate metric, and how to do unbiased sampling of the relevant population. Much of this discussion is similar to that in earlier reports[1,2]. A particular metric, based on the k-nearest neighbor approach to pattern recognition is developed in an Appendix. Section 4 summarizes the steps required to establish a validated assay within the ROC/LR framework, and discusses some practical issues that arise in the case of the growth medium identification problem, such as how large a sample set is required for a defensible assay.

2. The ROC/LR method for statistical pattern recognition-type assays

In the pattern recognition approach, it is assumed that the signature associated with, say, a particular growth medium is contained in a set of measurements of individual characteristics of the agent sample such as elemental concentrations[3], isotope ratios[4], or the amplitude of peaks corresponding to particular mass fragment sizes[5]. This set of (usually quantitative) characteristics is termed the *pattern vector* in the literature on pattern recognition[6]. If the pattern vector of a sample consists of n separate measurements (e.g. the concentration of n different elements), then the (conceptual) n-dimensional space that pattern vectors reside in is called the *pattern space*. The pattern vectors obtained from samples that were made with a common medium, but differ in other aspects of the growth and processing are not expected to be identical, but rather to be distributed "near" each other within some readily discernable volume of the pattern space. Even pattern vectors determined on replicate samples will not be identical because of measurement uncertainty, and will cause the volume associated with each medium to have diffuse boundaries.

Let G_j be the hypothesis that the jth growth medium was used to produce an agent sample, and let $\{\mathbf{g}_j\}$ be the set of pattern vectors that belong to reference samples made using that medium. Let the pattern vector belonging to a questioned sample be $\mathbf{g}_q$. An analysis that purports to show that a particular growth medium was used in the agent preparation process is clearly a hypothesis test, and can be described using the standard likelihood ratio equation:

$$O(G_j|\mathbf{g}_q) = [P(\mathbf{g}_q|G_j)/P(\mathbf{g}_q|\overline{G}_j)] \bullet O(G_j)$$

While the standard approach to pattern recognition seeks to decide if $\mathbf{g}_q \in \{\mathbf{g}_j\}$, or estimate a posterior probability that $\mathbf{g}_q \in \{\mathbf{g}_j\}$, the ROC/LR method seeks to quantify the *probative value* of the observed position of $\mathbf{g}_q$ relative to other points with respect to the proposition that $\mathbf{g}_q \in \{\mathbf{g}_j\}$. This is simply determined by evaluating the quantity in brackets, namely the likelihood ratio.

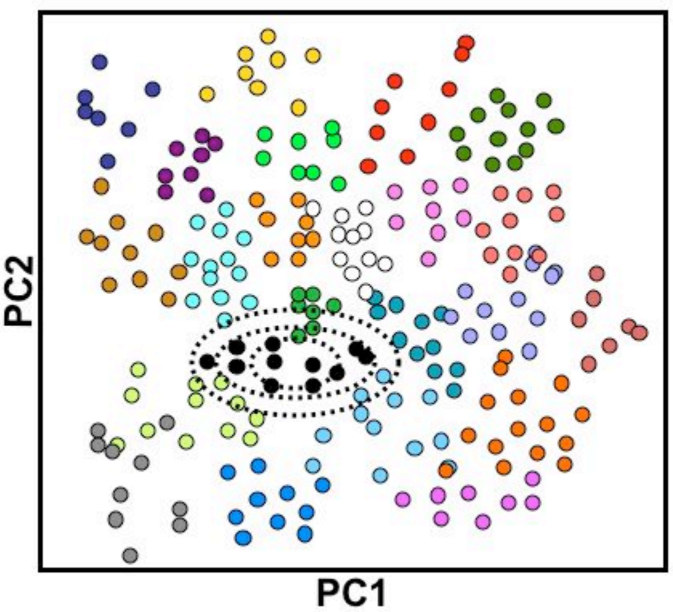


Figure 1. A 2-dimensional plot of simulated multivariate analytical data that is used to classify a biological agent according to the medium formulation used to grow it. Each color corresponds to a different growth medium. The nested ellipses represent three regions with different values of a distance metric for classifying a questioned data point.

In practice {$\mathbf{g}_j$} is determined by making measurements on a collection of reference samples that have been produced by various methods all utilizing the same growth medium, j. The pattern space {**G**} is then the union of all the subspaces {$\mathbf{g}_j$} that correspond to different growth medium choices. While the medium used to produce the reference samples that determine each subspace {$\mathbf{g}_j$} must be known, it is not absolutely necessary to know any other aspects of how the reference samples were produced.

Figure 1 is a simulated representation of a pattern space defined by two dimensions. These dimensions could be measurements themselves or could have been derived from some feature extraction procedure that reduces the dimensionality of the original pattern vectors, such as principle component analysis. Each medium choice j is represented by a different color in Fig. 1, and each cluster of identically colored points represents a different subspace {$\mathbf{g}_j$}.

As a concrete example, assume that a centroid can be calculated for those points contained in each cluster. (In section 3 and appendix A we will show that this is not necessary, but it is convenient for explaining how ROC curves can be derived from clustered data.) Various metrics can be devised to describe how close a point is to the centroid, and used to parametrically determine the ROC curve. Assume that we have chosen a metric, **d**. A ROC curve is constructed for each cluster by considering the number and types of data points contained within a specified value of **d** from the centroid. Referring to figure 1, the cluster of interest {$\mathbf{g}_k$} is represented by the black points, and the dashed ellipses represent different values of **d**. Clearly, as the value of the metric $\mathbf{d}$ gets larger, the probability that a point that is not a member of $\{\mathbf{g}_k\}$ has a distance metric value that is less than $\mathbf{d}$ increases (false positives increase.) Conversely, as $\mathbf{d}$ shrinks, the probability of a true member of $\{\mathbf{g}_k\}$ falling outside the ellipse determined by $\mathbf{d}$ increases (false negatives increase.) For each value of $\mathbf{d}$, a contingency table like Table 1 can be constructed.

Table 1. Contingency table for a given value of $\mathbf{d}$.

	Belongs to $\{\mathbf{g}_k\}$	Does not belong to $\{\mathbf{g}_k\}$	In/out totals
Falls inside $\mathbf{d}$	$N(\mathbf{d}, \{\mathbf{g}_k\})$	$N(\mathbf{d}, \{\bar{\mathbf{g}}_k\})$	$M(\mathbf{d})$
Falls outside $\mathbf{d}$	$N(\bar{\mathbf{d}}, \{\mathbf{g}_k\})$	$N(\bar{\mathbf{d}}, \{\bar{\mathbf{g}}_k\})$	$M(\bar{\mathbf{d}})$
Category totals	$N(\{\mathbf{g}_k\})$	$N(\{\bar{\mathbf{g}}_k\})$	N_{total}

In Table 1, $N(\mathbf{d}, \{\mathbf{g}_k\})$ (for example) is the number of points belonging to class $\{\mathbf{g}_k\}$ that fall inside the region defined by $\mathbf{d}$, while $N(\{\mathbf{g}_k\})$ is the total number of points in the cluster $\{\mathbf{g}_k\}$. The set $\{\bar{\mathbf{g}}_k\}$ represents all points that are not members of $\{\mathbf{g}_k\}$. For each value of $\mathbf{d}$, the fraction of true positives is given by:

$$F(TP) = N(\mathbf{d}, \{\mathbf{g}_k\}) / N(\{\mathbf{g}_k\}),$$

and the fraction of false positives is given by:

$$F(FP) = N(\mathbf{d}, \{\bar{\mathbf{g}}_k\}) / N(\{\bar{\mathbf{g}}_k\})$$

The ROC curve is determined by plotting F(TP) vs. F(FP) for each value of $\mathbf{d}$. For any questioned point $\mathbf{g}_q$, the value of the likelihood ratio is determined from the slope of the ROC curve at $\mathbf{d}_q$, which is $\mathbf{g}_q$'s distance from the centroid.

Note that the likelihood of $\mathbf{g}_q$ belonging to any other cluster $\{\mathbf{g}_j\}$ can be computed in a similar manner. Obviously, if the pattern vector of a questioned sample falls near the centroid of the cluster of vectors belonging to a particular class (e.g. a particular growth medium) it will have a higher likelihood ratio for that hypothesis than if it fell near the border of the cluster. Clearly it is possible that if $\mathbf{g}_\mathbf{q}$ lies in the boundary region among several clusters then it may have strong levels of support (i.e. likelihood ratios >1) for belonging to several $\{\mathbf{g}_\mathbf{j}\}$. This ambiguity in the data can be reported naturally in the ROC/LR framework without detracting from the (possibly important) fact that $\mathbf{g}_\mathbf{q}$'s position does provide a particular level of support for a case-relevent hypothesis.

One disadvantage of using this purely empirical technique for constructing the ROC curve is that if the clusters are well separated, as was the case with the data in reference 3, then the empirical ROC curve is "perfect" and the likelihood ratio strictly speaking, infinite over a certain range of the metric, and zero over an adjacent range. In this case, further statistical analysis is necessary to establish lower bounds on the likelihood ratio in the range where it is nominally infinite, and these lower bounds are reported.

In general, for the ROC/LR approach to be defensible, several conditions must be met:

(a) The hypotheses G_j and the associated subspaces $\{\mathbf{g}_j\}$ must span the entire space of possible distinct medium formulations. However, the loss in accuracy associated with the exclusion of less probable media types may be negligible in practice.

(b) The hypotheses must be mutually exclusive. Thus, it would not be possible to use major nutrient components such as peptone, tryptone, and yeast extract as the basis for multiple hypothesis testing because some medium formulations mix these components.

(c) The production methods used to generate each $\mathbf{g}_j$ must be drawn in an unbiased way from the space of possible production methods.

(d) The subspaces $\{\mathbf{g}_j\}$ should have diffuse overlapping boundaries characteristic of randomness, rather than complex, intertwined, but sharply defined boundaries. Strictly speaking it is difficult to prove that the latter characteristic is absent when analyzing a finite number of reference samples. However, unexpected complexity in subspace shapes may become evident even when finite sample sizes are used to create a map like that in Fig.1. This may make it difficult to choose a simple metric for constructing the ROC curve.

3. Technical issues in applying the ROC/LR method to multivariate data

Reducing data dimensionality

Pattern recognition approaches to data analysis often involve feature selection or feature extraction steps to reduce the dimensionality of the feature space. Feature selection is the deliberate selection of a smaller set of components out of the full feature vector, presumably based on the observation that these are the most important components for classification. Feature extraction is the mathematical transformation of the full feature vector into a vector of lower dimensionality. Principal components analysis (PCA, also called the Karhunen-Loeve Expansion) is a typical feature extraction technique that was used in references 3 and 5, for example. Although, PCA is commonly employed to help visualize the discriminating power of multi-dimensional data, and for reducing computational labor in pattern classification, it is not an essential part of the analysis. However, this and other feature extraction techniques can be important for transforming complexly shaped subspaces into ones for which simple metrics can be defined, thus fulfilling element (d) of the list of conditions given above.

Choosing a comparison metric

The most natural comparison metrics are those based on distance from a centroid. Standard distance measures such as the Euclidian or Mahalanobis distance[6] can be used as the metric **d**. This can be computed directly from the multivariate data, e.g. the complete elemental concentration vector, or in a reduced-dimensionality space such as that produced by PCA. However, not all metrics need to be based on distance from a centroid.

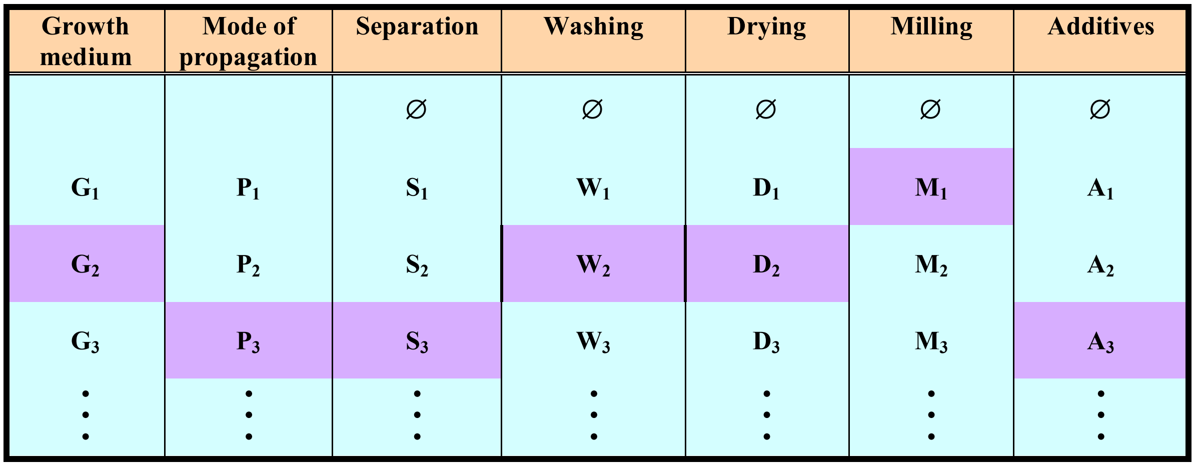
A ROC curve can also be formed by using a metric derived from a generalization of the "majority rules" or "k nearest neighbors" (kNN) approach to classification[6]. This approach is described in detail in Appendix A. In the kNN scheme we calculate the k nearest neighbors of a point, declare the point to be in $\{\mathbf{g_0}\}$ if a specified fraction of its neighbors is in $\{\mathbf{g_0}\}$, and in $\{\overline{\mathbf{g}}_\mathbf{0}\}$ otherwise. The fractional value represents a continuous metric that can be related to the (unknown) prior odds ratio of choosing a point from $\{\mathbf{g_0}\}$ vs $\{\overline{\mathbf{g}}_\mathbf{0}\}$. Both distance and neighbor-counting metrics have the property that a point that maps close to the interior of a cluster will have a higher likelihood ratio for belonging to that cluster, while a point that maps near the edge of the cluster or beyond will have a lower likelihood ratio for belonging to that cluster. The kNN based metric may have an advantage over simpler distance type metrics in the case of more complexly shaped pattern subspaces since it only depends on the local density of points from $\{\mathbf{g_0}\}$ and $\{\overline{\mathbf{g}}_\mathbf{0}\}$ within the distance from the questioned point to its kth nearest neighbor.

It should be mentioned that there is a third approach which avoids the use of metrics and ROC curves entirely, and uses the empirically determined subspace populations $\{\mathbf{g}_\mathrm{j}\}$ and $\{\overline{\mathbf{g}}_\mathrm{j}\}$ to estimate the conditional probability densities $\mathrm{P}(\mathbf{g}_\mathrm{q}|\ \mathbf{g}_\mathrm{q} \in \{\mathbf{g}_\mathrm{j}\})$ and $\mathrm{P}(\mathbf{g}_\mathrm{q}|\mathbf{g}_\mathrm{q} \in \{\overline{\mathbf{g}}_\mathrm{k}\}$ and calculate the likelihood ratio directly. One standard method for doing this involves the Parzen density estimation method, described in many texts on pattern recognition[6]. Another method for determining the conditional probability densities is given in reference 7. It is not clear if these methods provide very accurate results if the pattern subspaces have complex shapes, and they appear to be considerably more complicated to implement than distance or k-NN metrics.

Sampling the test population

A key condition on the accuracy of the ROC/LR framework is that the set of reference samples be drawn in an unbiased way from the complete population of biological agent production methods. The "population" of possible growth and processing methods for a bacterial agent is virtual[1,2], and can be represented by a unit process matrix, shown schematically in Figure 2, where 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 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. Each growth method is characterized by a choice of growth medium and a mode of propagation, such as agar plate, shake flask, or fermentor. The complete set of growth medium formulations corresponds to the set of hypotheses that we are trying to test

Referring to the discussion in section 2, the "population" {G} that one samples to generate a ROC curve is (conceptually) all of the processes represented by the unit process matrix, *conditional on choosing a particular growth medium.* 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.

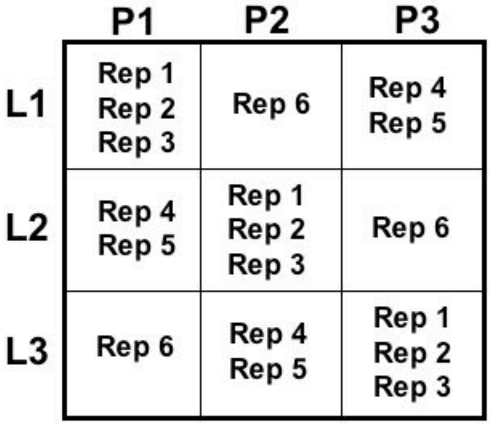


Growth medium	Mode of propagation	Separation	Washing	Drying	Milling	Additives
		∅	∅	∅	∅	∅
G_1	**P_1**	**S_1**	**W_1**	**D_1**	**M_1**	**A_1**
G_2	**P_2**	**S_2**	**W_2**	**D_2**	**M_2**	**A_2**
G_3	**P_3**	**S_3**	**W_3**	**D_3**	**M_3**	**A_3**
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Figure 2. 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.

Although the number of potential methods for growing and processing agents may be large, all 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 Ref. 1.

To account for variations in the execution of particular process steps, it is preferable to have samples made by different laboratories working independently and 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. An example of a partial factorial design that was introduced in reference 1 is shown in Figure 3. 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.



	P1	P2	P3
L1	Rep 1 Rep 2 Rep 3	Rep 6	Rep 4 Rep 5
L2	Rep 4 Rep 5	Rep 1 Rep 2 Rep 3	Rep 6
L3	Rep 6	Rep 4 Rep 5	Rep 1 Rep 2 Rep 3

Figure 3. A partial factorial design for producing reference samples grown in a particular medium. L1-L3 represent different laboratories, while P1-P3 represent different processes, drawn from the unit process matrix.

4. Practical implementation of the ROC/LR method for medium identification

In light of the above discussion, there are a number of steps that must taken to produce a ROC curve that would permit the estimation of a likelihood ratio that supports the hypothesis that a certain medium type was used to grow a questioned sample:

- Identify the complete set of possible growth media for the agent of interest.
- For each medium, make reference materials that sample the space of production methods using an unbiased selection method such as that outlined in reference 1.
- Make multi-parameter measurements on the set of reference samples using the analytical method of choice (or several analytical methods, if desired.)
- Choose a feature selection method such as PCA, if desired.
- Choose a metric for deciding which class a questioned sample belongs to. Two easy-to-implement choices are distance and kNN-based metrics.
- Construct a ROC curve from the contingency table results derived by calculating the metric from the multi-parameter data obtained for each member of each reference sample subspace corresponding to a different medium.
- Validate the ROC curve by drawing an independent set of reference samples from the same space of media and production methods and repeating the measurements and analysis. Compare the resulting new ROC curve with the original to see if there are significant differences. The new and old data can be combined to produce a composit ROC curve that has improved sampling of the complete production method space.

From a practical point of view, implementing this scheme for the growth medium identification problem raises several issues. Foremost is the large number of reference samples that would have to be produced and measured in order to have a reasonable sampling of {G}. The smallest partial factorial design (a 2 x 2 version of Fig, 3) for sampling production methods generates 6 samples for each medium. For *Bacillus anthracis*, at least 25 media have been reported in the open and classified literature, and at least 12 of these are used commonly[1]. If only one analytical replicate of each sample were analyzed, this minimal experiment would still require 72 samples. Expansion of the number of media, labs, processes, and analytical replicate measurements easily drive the number of samples into the hundreds.

On the other hand, microbes other than *B. anthracis* may involve far fewer choices of growth medium either because of fastidiousness, or simply because of microbiological tradition. Moreover, exploratory work could be staged to establish the effectiveness of the method on a smaller subset of media before generating a full set of data. For example, such a study could begin by choosing only a few media that are felt to be harder to discriminate against, and seeing how the method performs against them.

It is easy to imagine that after producing and analyzing the large reference set containing all the media types it is found that there are certain subsets that can be discriminated well, while others are difficult to distinguish. It might then be possible to combine subsets into distinguishable classes of media, providing less, but still useful discrimination.

References

1. Velsko, S., "Bioagent Sample Matching using Elemental Composition Data: An Approach to Validation", Lawrence Livermore National Laboratory Report UCRL-TR-220803, April, 2006.

2. Velsko, S., "Validation Strategies for Microbial Forensic Analysis of Biological Agents: Beyond Sample Matching", Lawrence Livermore National Laboratory Report UCRL-TR-229944, April 2007.

3. Cliff, J.B., et. al., "Differentiation of Spores of Bacillus subtilis Grown in Different Media by Elemental Characterization Using Time-of-Flight Secondary Ion Mass Spectrometry", App. Environ. Microbiol 2005; **71**: 6524-6530.

4a.Kreuzer-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 Sci., 2004; **49**:954-960; b. Kreuzer-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 Sci., 2004; **49**:961-967.

5. Madonna, A.J., Voorhees, K.J., Hadfield, T.L., Hilyard, E.J., “Investigation of Cell Culture Media Infected with Viruses by Pyrolysis Mass Spectrometry:Implications for Bioaerosol Detection”, J. Am. Soc. Mass Spectrom. 1999; **10**: 502-511.

6. *Classification, Estimation, and Pattern Recognition*, by T.Y. Young and T.W. Calvert (American Elsevier Publishing Company, Inc., New York, 1974.

7. Patrick, EA, and Fischer FP, “A Generalized k-Nearest Neighbor Rule”, Information and Control, 1970; **16**:128-152.

Appendix A – The k-nearest neighbor metric and ROC curve

It is possible to define a simple metric for constructing a ROC curve based on a generalization of the k-nearest neighbor (kNN) criterion for assigning a point in pattern space to a particular cluster. Consider two clusters of points defined by two parameters as shown in Figure A.1, corresponding to samples from subpopulations $\{H_0\}$ and $\{\overline{H}_0\}$ in which some hypothesis H_0 is true and false, respectively. For any point **x** = (parameter 1, parameter 2), the set of its k nearest neighbors can be defined in terms of a (usually) Euclidian distance measure. The set of k nearest neighbors will, in general, contain points from both $\{H_0\}$ and $\{\overline{H}_0\}$.

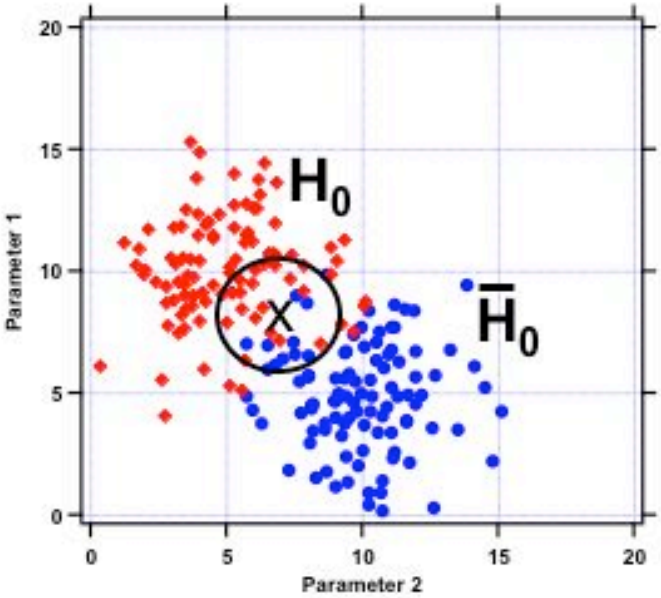


Figure A.1. Two clusters of points defined by membership in $\{H_0\}$ and $\{\overline{H}_0\}$. The X represents the coordinates of a questioned sample, and the circle is defined by a Euclidian distance metric between X and its kth nearest neighbor. Note that both H_0 and $\overline{H}_0$ points are contained within the kth nearest neighbor distance from X.

An estimator of the probability of finding a point at coordinates **x**, given that it is drawn from $\{H_0\}$ is:

$$P(\mathbf{x}|H_0) = N(\mathbf{x}, H_0|k)/N_T(H_0), \qquad \text{(A.1a)}$$

where $N(\mathbf{x},H_0|k)$ is the number of points from $\{H_0\}$ that are contained in the set of k nearest neighbors to **x**, and $N_T(H_0)$ is the total number of points from $\{H_0\}$ in the cluster. Similarly,

$$P(\mathbf{x}|\overline{H}_0) = N(\mathbf{x},\overline{H}_0|k)/N_T(\overline{H}_0). \qquad \text{(A.1b)}$$

Given an questioned point $\mathbf{x_q}$, the Bayesian criterion for deciding that $\mathbf{x}_q \in \{H_0\}$is that

$$P(H_0|\mathbf{x_q})/P(\overline{H}_0|\mathbf{x_q}) > 1 \qquad (A.2)$$

Since

$$P(H_0|\mathbf{x_q})\bullet P(\mathbf{x_q}) = P(\mathbf{x_q}|H_0)P(H_0), \qquad (A.3a)$$

and

$$P(\overline{H}_0|\mathbf{x_q})\bullet P(\mathbf{x_q}) = P(\mathbf{x_q}|\overline{H}_0)P(\overline{H}_0), \qquad (A.3b)$$

the Bayes decision criterion can be re-written as:

$$P(\mathbf{x_q}|H_0)P(H_0) > P(\mathbf{x_q}|\overline{H}_0)P(\overline{H}_0) \qquad (A.4)$$

Using equations (A.1), and rearranging, we can re-write expression (A.4) as a generalized version of the k-nearest neighbor rule for assigning $\mathbf{x}_q$ to $\{H_0\}$:

$$N(\mathbf{x_q},H_0|k) > [(P(\overline{H}_0)/P(H_0)]\bullet[N_T(H_0)/N_T(\overline{H}_0)]\bullet N(\mathbf{x_q},\overline{H}_0|k) \qquad (A.5)$$

In the standard kNN rule, it is implicitly assumed that all of the points are drawn at random from the total space $\{H_0\} \cup \{\overline{H}_0\}$ so that $N_T(H_0) \propto P(H_0)$ and $N_T(\overline{H}_0) \propto P(\overline{H}_0)$, and thus the product of the two bracketed ratios in equation A.5 is 1. However, in the general expression (A.5) it is permitted to have independently chosen $N_T(H_0)$ samples from $\{H_0\}$, and $N_T(\overline{H}_0)$ samples from $\{\overline{H}_0\}$ to form the "training set", and to interpret $P(\overline{H}_0)/P(H_0)$ as the prior odds that $\mathbf{x}_q \in \{\overline{H}_0\}$. These prior odds weight the decision to assign $\mathbf{x_q}$ based on the number of near-neighbors in each class. For example, if nearly all of the k nearest neighbors of a point $\mathbf{x_q}$ were members of $\{H_0\}$, then $P(\overline{H}_0)/P(H_0)$ would have to be very large before we would consider assigning $\mathbf{x_q}$ to $\{\overline{H}_0\}$.

However, in general $P(\overline{H}_0)/P(H_0)$ is not known. The unknown prior odds ratio $P(\overline{H}_0)/P(H_0) = \rho$ can be taken as a metric which can be varied to change the decision threshold for assigning $\mathbf{x}_q \in \{H_0\}$. For any point in a cluster we can derive the statistic

$$\rho = [N_T(\overline{H}_0)/N_T(H_0)]\bullet[N(\mathbf{x},H_0|k)/N(\mathbf{x},\overline{H}_0|k)], \qquad (A.6)$$

which represents the largest value that $P(\overline{H}_0)/P(H_0)$ could have without causing us to change our decision about the assignment of $\mathbf{x}$ to $\{H_0\}$.

Since ρ is an odds ratio it can vary from 0 to ∞, and it is convenient to define a metric

$$p = \rho/1+\rho \qquad (A.7)$$

which varies from 0 to 1 in order to generate the ROC curve. Combining (A.6) and (A.7) leads to:

$$p = N_T(\bar{H}_0)\bullet N(\mathbf{x},H_0|k)/[\ N_T(\bar{H}_0)\bullet N(\mathbf{x},H_0|k) + N_T(H_0)\bullet N(\mathbf{x},\bar{H}_0|k)] \qquad (A.8)$$

Figure A.2 displays the histograms of p obtained for the two sets of points displayed in Figure A.1. The distributions of p values for $\{H_0\}$ and $\{\bar{H}_0\}$ are decidedly non-gaussian, due to the non-linear mapping (A.7).

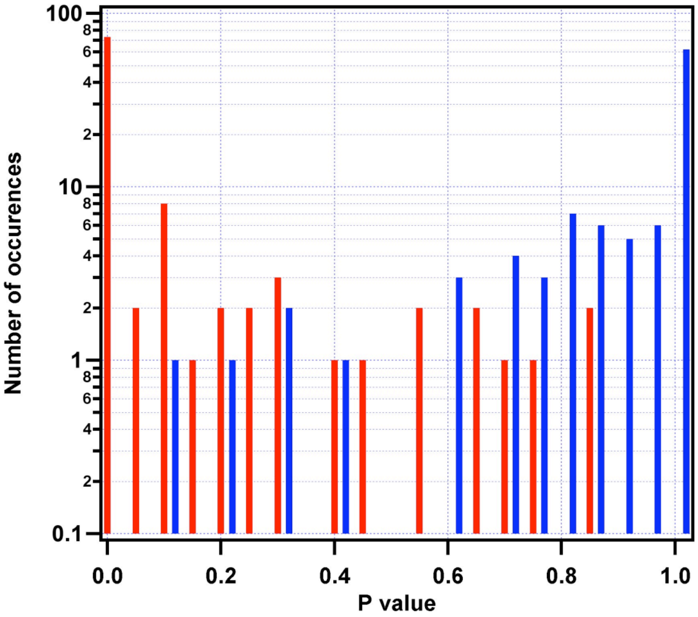


Figure A.2. Histograms of p values of the points shown in Fig. A.1. Red: $\{H_0\}$; Blue: $\{\bar{H}_0\}$. The p values were calculated for k = 30. Note: the red and blue lines have been displaced slightly along the p axis to improve the clarity of the figure.

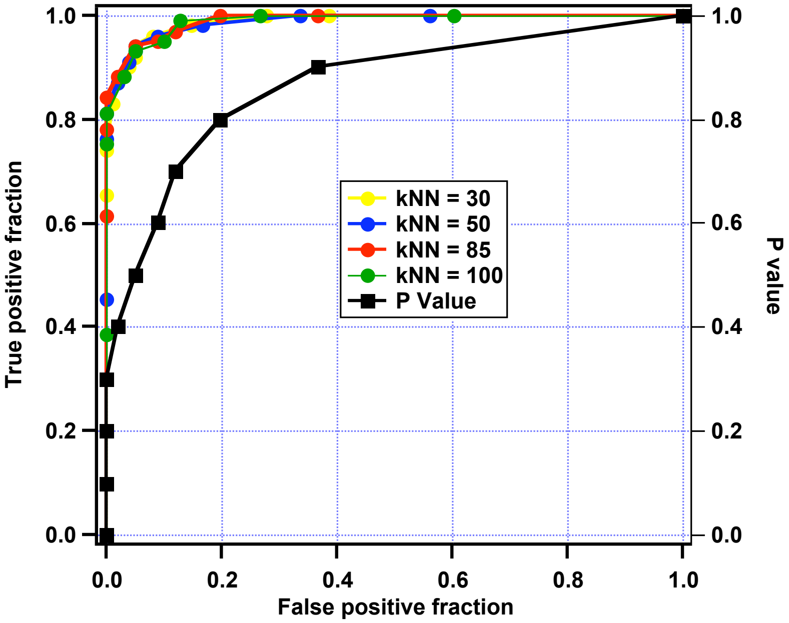


Figure A.2. ROC curves generated for various values of kNN for the data shown in Fig. A.1.

The p distribution for each cluster of points peaks at the extreme values ($p = 0$ for $\{H_0\}$ and $p = 1$ for $\{\overline{H}_0\}$), representing points close to the center of each cluster. Points from $\{H_0\}$ which lie close to the $\{\overline{H}_0\}$ cluster have large values of p and vice versa.

ROC curves generated for various values of k are displayed in Figure A.2, which shows that the ROC curve is relatively stable for different k values. Below k = 10 the stair-step nature of the ROC curve becomes pronounced. The bend in the ROC curve, where the slope = 1, corresponds to $p = ½$, and for p greater than this value support for $\mathbf{x_q}$ belonging to $\{H_0\}$ is weak. For $p < ½$ the slope of the ROC curve is greater than 1, and there is positive support for $\mathbf{x_q} \in \{H_0\}$. Note that if $p < ½$, then $\rho < 1$ which, by its definition implies that $P(H_0) > P(\overline{H}_0)$. Conversely if $p > ½$, then $\rho > 1$ and $P(\overline{H}_0) > P(H_0)$.