Assessment of statistical methods used in library-based approaches to microbial source tracking

Kerry J. Ritter, Ethan A. Carruthers[1], C. Andrew Carson[2], R. D. Ellender[3], Valerie J. Harwood[4], Kyle S. Kingsley[5], Cindy H. Nakatsu[6], Michael J. Sadowsky[1], Brian L. Shear[4], Brian R. West[5], John E. Whitlock[7], Bruce A. Wiggins[8] and Jayson D. Wilbur[10]

ABSTRACT - Several commonly used statistical methods for fingerprint identification in microbial source tracking (MST) were examined to assess the effectiveness of pattern-matching algorithms to correctly identify sources. Although numerous statistical methods have been employed for source identification, no widespread consensus exists as to which is most appropriate. A large-scale comparison of several MST methods, using identical fecal sources, presented a unique opportunity to assess the utility of several popular statistical methods. These included discriminant analysis, nearest neighbor analysis, maximum similarity and average similarity, along with several measures of distance or similarity. Threshold criteria for excluding uncertain or poorly matched isolates from final analysis were also examined for their ability to reduce false positives and increase prediction success. Six independent libraries used in the study were constructed from indicator bacteria isolated from fecal materials of humans, seagulls, cows and dogs. Three of these libraries were constructed using the rep-PCR technique and three relied on antibiotic resistance analysis (ARA). Five of the libraries were constructed using *Escherichia coli* and one using *Enterococcus* spp. (ARA). Overall, the outcome of this study suggests a high degree of variability across statistical methods. Despite large differences in correct classification rates among the statistical methods, no single statistical approach emerged as superior. Thresholds failed to consistently increase rates of correct classification and improvement was often associated with substantial effective sample size reduction. Recommendations are provided to aid in selecting appropriate analyses for these types of data.

INTRODUCTION

Many microbial source tracking (MST) methods rely on libraries of indicator organisms cultivated from known sources of fecal contamination to identify unknown sources (see Simpson et al, 2002 and Scott et al., 2002 for a recent review of these methods). These library-based methods involve the assembly of a variety of "fingerprints" from indicator organisms for several known animal sources (e.g., cow, human, and seagull). These fingerprints are stored as libraries that are used to compare with fingerprints from these same indicators isolated from water presumed contaminated with fecal material. In this way, the source of the unknown indicator bacterium can be identified, or at least predicted, based on similarity to members of the known-source libraries.

Library-based MST methods may be based on either genotypic or phenotypic "fingerprints" of fecal indicator organisms, frequently *E. coli* or *Enterococcus* spp. Antibiotic resistance analysis (ARA) is a phenotypic MST method that uses profiles of resistance to various antibiotics at different

[1] *Department of Soil, Water, & Climate, University of Minnesota, St. Paul, MN*
[2] *Department of Veterinary Pathobiology, University of Missouri, Columbia, MO*
[3] *University of Southern Mississippi, Hattiesburg, MS*
[4] *Department of Biology, University of South Florida, Tampa, FL*
[5] *Applied-Maths, Sint-Martens-Latem, Belgium*
[6] *Department of Agronomy, Purdue University, West Lafayette, IN*
[7] *Division of Math and Science, Hillsborough Community College, Tampa, FL*
[8] *Department of Biology, James Madison University, Harrisonburg, VA*
[9] *Department of Mathematical Sciences, Worcester Polytechnic Institute, Worcester, MA*

concentrations (Wiggins, 1996; Hagedorn et al., 1999; Harwood et al., 2000; Whitlock et al., 2002; Wiggins et al., 2003). The underlying assumption of ARA is that differential exposure of humans and animals to a variety of antibiotics will elicit specific resistance patterns for associated flora of host intestines. Subsequently, the antibiotic resistance patterns of indicators from unknown sources can be compared to a library of ARA profiles of indicators from known sources.

Rep-PCR is a genotypic method that uses the polymerase chain reaction and primers based on conserved extragenic repetitive sequences to amplify specific portions of the microbial genome (Versalovic, et al., 1991, Versalovic, et al., 1994). Following electrophoresis and staining, a banding pattern or fingerprint is revealed that can be used for strain identification. The underlying assumption of this technology is that organisms having indistinguishable banding patterns can be regarded as being identical or nearly identical, and those having similar banding patterns are regarded as genetically related. As a result, hosts for the bacteria may be identified by comparing presence/absence of bands with those from known source fingerprints (i.e. band-matching). Rep-PCR DNA fingerprinting of *E. coli* has been previously used for MST (Carson et al. 2003; Dombek et al. 2000).

Several statistical methods, including discriminant analysis (DA) (Wiggins 1996, Carson et al., 2001, Harwood, et al. 2000, Whitlock et al., 2002), and maximum or average similarity (MS and AS) (Dombek et al. 2000, Carson et al. 2003) have been used to classify sources. These statistical approaches differ with respect to distributional assumptions, measures of distance or similarity, and strategies for prediction. Many approaches, such as DA and AS, take into account the central tendency and variability of each source group as a whole. Other methods, such as nearest neighbor (NN) and MS predict source membership based on similarity to an individual isolate within each source. Consequently, rates of correct classification may differ depending on the method used.

To reduce error inherent in these statistical methods, some have suggested that uncertain or poorly matched isolates should be removed from classification . This is especially relevant when false positives are a concern, such as when response to presence of specific sources of fecal contamination results in costly management action. Some have argued that imposed thresholds decrease noise and may eliminate false positives resulting from statistical or measurement error (Wiggins, personal communication, 2003). Others have suggested threshold values should be based on an average rate of misclassification of sources estimated from known library isolates (Whitlock, et al. 2002) or effects of inter-gel variability on similarity between identical control isolates (Wheeler et al., 2002, Sadowsky, personal communication, 2003). Others advocate exact matching, removing any isolates whose fingerprints were not represented in the library (Samadpour, personal communication, 2003).

While statistical methods for classification and threshold criteria to reduce error have been used extensively with these library-based methods, there is no widespread consensus as to when each of these methods or criteria is appropriate. Further, little attention has been given to the consequences associated with applying these various statistical approaches to microbial data. In this study we consider several statistical approaches for identifying source membership using these (binary) data resulting from rep-PCR and ARA fingerprints. Statistical methods were selected based on their popularity and availability in standard software packages. Following a brief review of each method, we assess the ability of each statistical technique to successfully identify sources of fecal contamination from blind test samples. Further, we investigate the use of threshold criteria to reduce rates of false positive classifications within several of the statistical methods.

Methods

Six libraries of indicator bacteria were used to assess the utility of various statistical approaches. Three of the libraries (A1, A2, and A3) were constructed using antibiotic resistance analysis and three were constructed using rep-PCR (R1, R2, and R3). In this study, only BOXA1R primer was used to generate fingerprints from *E. coli* isolates. Neither the three ARA libraries nor the three rep-PCR libraries were congruent in any way other than having a roughly a similar library size, aliquots of the same starting fecal materials and the same sources of blind samples. Library samples consisted of swab fecal samples of human, dog, cow, and seagull. Blind test samples consisted of pure (100%) sources of human, cow and seagull, composited from the same fecal samples used to create the library samples. None of the test samples contained pure dog sources. Around 60 isolates per known source sample were used to create libraries and approximately 50 isolates per unknown source sample were used for testing the rates of correct classification by the method. All libraries were constructed using *E. coli* except one (AR2) that used *Enterococcus* spp. For a more detailed description of the study design see Griffith and Weisberg in this issue.

Seven statistical analyses were performed on each of the six libraries. These included DA (pooled and non-pooled estimates of covariance), NN (Mahalanobis and Euclidean distances), MS (Jaccard), AS (Jaccard), and ID Bootstrap Maximum Similarity (Jaccard). For three of the statistical methods (DA, MS, and ID Bootstrap), threshold criteria for excluding isolates from source identification were applied and resulting changes in percent correct classification (% CC) were determined for each blind test sample. Threshold criteria for DA and ID Bootstrap were based on estimates of posterior probabilities, or the probability of correct classification. Threshold criteria for MS were based on a quality quotient. These criteria, along with the corresponding statistical methods, are described in more detail in the following section

Both DA and NN analyses were performed using PROC DISCRIM in SAS (Cary, NC) while MS, AS, and ID Bootstrap were performed using the identification function in BioNumerics (Applied Maths, Sint-Martens-Latem, Belgium). These software packages were chosen due to their versatility, usefulness, and popularity among the microbial source tracking community. BioNumerics is a software package predominantly chosen by those using rep-PCR, while choice of SAS is more common among those using ARA. All rep-PCR patterns were processed in BioNumerics prior to statistical analyses.

The predictive abilities of the various statistical methods for each of the library methods were assessed by calculating the percent of correctly classified isolates (%CC) from each of the three blind test samples (human, cow, and seagull). Trade-offs for applying the different threshold criteria for removing uncertain isolates were examined by calculating both the change in %CC and the proportion of the sample (i.e. percentage of isolates) eliminated from final analyses. Ties, or those isolates that were predicted by the statistical algorithm to belong to more than one source, were excluded from final analyses.

Statistical Analyses

Discriminant analysis (DA) is a commonly used technique for classifying unknown samples into predefined groups and is especially popular among microbiologists using ARA. In DA, a classification rule is developed from a calibration data set where group membership among the samples is known (the library). In SAS's PROC DISCRIM the rule is based on estimated posterior probabilities, or the probability that an isolate belongs to a specific group. The isolate is classified into the source group yielding the highest estimated posterior probability among all source categories. Given a normal distribution and assuming equal covariances among groups, classification using posterior probabilities is equivalent to placing the observation into its closest group. By default, distances to groups are defined by Mahalanobis distance which takes into account both distance of each observation to the mean and the variability within the group. The "pool= yes" option in PROC DISCRIM estimates a single covariance structure for all groups while the "pool= no" option allows for estimation of covariances separately for each group (for more detailed description see SAS/STAT User's Guide, Vol. 1. Chapter 20).

In SAS's PROC DISCRIM a "threshold" option is available for excluding observations from classification in discriminant analyses based on a minimum threshold for posterior probabilities. If an isolate's posterior probability for its predicted group falls below the threshold value, the isolate is classified into an "other" category and excluded from classification into one of the known groups.

Maximum Similarity (MS) is another commonly used statistical algorithm that is particularly popular among those using rep-PCR. In MS, observations are classified into the group to which its closest or most similar known member belongs. BioNumerics offers several alternative measures of similarity, including Jaccard, Dice, and simple matching. We chose to use the Jaccard similarity coefficient for our analyses because Jaccard targets only those bands that are present in at least one of the pairs being compared, ignoring potentially large numbers of missing bands that may dilute or mask differences. The Jaccard similarty coefficient is given by,

$$\frac{N_{AB}}{N_A + N_B - N_{AB}}$$

where N_{AB} is the number of shared bands, N_A is the total number of bands in pattern A, and N_B is the total number of bands in pattern B.

BioNumerics also offers a quality factor (QF) option that qualifies the relative uncertainty of correct classification for each isolate and may be used for eliminating potential false positives. The QF is the ratio of the average distance between the unknown and the library source members and the average internal distance where:

$$QF = D_{UN}/D_{LU}$$

$$\text{where } D_{UN} = \frac{\sum_{i=1}^{n}(1 - s_i)}{n}$$

$$\text{and } D_{LU} = \frac{2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(1 - s_{i,j})}{n(n-1)}$$

n is the number of entries in the library unit and *s* is the similarity. If QF ≤ 1, it can be inferred that the unknown fits into the library source as well as or better than other members of that source group; whereas if QF >> 1, a poor fit is indicated. Qualitative scores or "grades", each representing a numerical range of the QF for each isolate, may be exported to a data file along with similarity and scores and predicted source classifications. The numerical range of these grades are as follows: 0-0.5=A, 0.51-1.0=B, 1.01-1.5=C, 1.51-2=D, and >2=E. Recently a script file has been made available to BioNumeric users that allow numerical values to be exported as well, however, the program was not available at the time of this study.

Average Similarity (AS), a common alternative to MS, is available in BioNumerics, and is also very popular among microbiologists using rep-PCR. Rather than classifying unknown isolates into source groups based on proximity to a single known isolate, AS classifies an unknown isolate into the source category yielding the largest average similarity to all library isolates within that source.

Nearest Neighbor (NN) is a nonparametric alternative to DA and is available in SAS using PROC DISCRIM. Source assignment is based on nonparametric estimates of posterior probabilities based on k-nearest neighbors. With the k=1 option, as was specified in this study, NN assigns source membership based on proximity to closest known individual. Therefore, NN is analogous to MS, only NN defines proximity using Mahalanobis or Euclidean distances, rather than Jaccard, Dice, or simple matching similarity measures.

ID Bootstrap is a script file designed to improve correct classification rates and has only recently been made available to BioNumerics users (see http://www.applied-maths.com/bn/bn.htm). The script applies a bootstrap algorithm to a matrix of similarity values in order to estimate the probability of obtaining an observed similarity score relative to chance. A similarity matrix is first calculated for all known samples, each of which is assigned to a particular group, and then each unknown is compared to each group of known samples, providing an average (or maximum) similarity for each unknown to each group. Each unknown is tentatively identified as belonging to the group with which it has the highest average (or maximum) similarity. The distribution of similarities between group non-members and group members is then approximated by resampling, with replacement, group members and non-members. Each bootstrap iteration involves the random selection of 30 or more group members and a single non-member. The proportion of bootstrap iterations in which the unknown is more similar to the re-sampled group than the known non-member approximates the probability that the unknown belongs to the group. In this study, the script was applied to MS and 1000 iterations were specified. As with DA, thresholds for ID Bootstrap were based on estimated probabilities of correct identification.

Results

The ability of each MST method to correctly predict source membership relied heavily on which statistical analysis was used (Table 1). Depending on the choice of statistical method, %CC could be quite high; at other times, %CC fell well below chance. In half of the cases where blind test samples were classified, %CC changed by more than 40 percentage points as a result of varying the statistical approach. In one case %CC increased from 0% to 90% by switching the statistical analysis from average similarity to maximum similarity. Even simply changing the measure of distance often resulted in substantial changes in the proportion of correctly identified isolates. For example, in NN, changing from Euclidean to Mahalanobis distance increased the percentage of correctly identified human isolates for researcher A1 from 29% to 75%.

Table 1. Percent Correct Classification for Seven Statistical Methods.

Source	StatMethod	RepPCR			AR		
		R1	R2	R3	A1	A2	A3
Human	DA^{P}	50	66	27	2	100	9
	DA^{NP}	35	20	12	0	100	25
	NN^{M}	43	**80**	33	2	100	**33**
	NN^{E}	57	74	**51**	0	100	13
	Max. Sim	65	74	22	**3**	100	13
	Ave. Sim	**74**	66	12	0	100	17
	ID Boot Max	65	74	22	**3**	100	13
Cow	DA^{P}	30	16	**71**	25	51	51
	DA^{NP}	39	**58**	39	70	59	33
	NN^{M}	17	20	27	**75**	27	38
	NN^{E}	31	16	24	29	41	35
	Max. Sim	53	18	47	29	**90**	37
	Ave. Sim.	**59**	0	53	4	0	**60**
	ID Boot Max	53	18	47	29	**90**	39
Gull	DA^{P}	53	8	**34**	41	93	57
	DA^{NP}	21	**42**	23	44	93	25
	NN^{M}	47	0	28	33	93	47
	NN^{E}	62	8	38	33	98	52
	Max. Sim	**67**	5	29	36	98	**65**
	Ave. Sim.	33	0	0	**67**	**100**	57
	ID Boot Max	**67**	5	29	35	98	**65**

Bold = method with highest percent correct classification
DA^{P} = Discriminant Analyses, pooled covariance
DA^{NP} = Discriminant Analyses, non pooled covariance
NN^{M} = Nearest Neighbor, Mahalanobis distance
NN^{E} = Nearest Neighbor, Euclidean distance
R1, R2, and R3 = rep-PCR
A1, A2, and A3 = ARA

The consistently high %CC for human and seagull across the various statistical methods for researcher A2 was due primarily to the fact that variability of fingerprints was low within each of these unknown samples. In fact, all isolates analyzed in the human blind sample for this researcher had the exact same antibiotic fingerprint (so that only 100% or 0% were possible for %CC) and only a handful of fingerprints were found among the blind seagull isolates.

Despite the large differences in correct classification rates among the statistical methods, no single statistical approach emerged as superior across any of the MST methods or sources. Improvements in correct classification rates for one source as a result of applying an alternative statistical analysis were often followed by a decrease in correct classification rates for another source. For example, for researcher R2 switching statistical analysis from DA (non-pooled) to NN (Mahalanobis) resulted in a 60-percentage-point increase in %CC for human, while decreasing %CC for cow and seagull by approximately 40 percentage points.

In the majority of the cases, the "right" choice of statistical method improved substantially the researcher's ability to identify unknown sources of fecal contamination. However, in some cases, none of the statistical methods were particularly satisfying. In many cases, %CC fell below 60%, regardless of which statistical method was applied. In one case (A1), the maximum %CC for human was less than 4% regardless of the statistical approach (less than expected by chance alone).

Attempting to reduce false positives by applying various threshold criteria to exclude uncertain isolates from classifications did not always result in improved %CC (see Tables 2-4). Using thresholds based on estimated DA posterior probabilities between 80% and 95% resulted in changes in %CC for individual samples anywhere from -16 to +33 percentage points depending on the library and source (Table 2). Thresholds based on QF scores

Table 2. Effect of Imposing threshold value for posterior probabilities (Discriminant Analyses, Pooled Covariance).

Researcher	%Thresh	Human		Cow		Gull	
		%CC	%Remain	%CC	%Remain	%CC	%Remain
R1	None	**50**	100	**30**	100	**53**	100
	80	48	78	**42**	70	**53**	68
	90	48	78	**45**	57	**53**	57
	95	**69**	54	**46**	52	29	32
R2	None	**66**	100	**16**	100	**8**	100
	80	**73**	88	0	80	6	68
	90	**81**	62	0	74	7	56
	95	**81**	62	0	74	**10**	42
R3	None	**27**	100	**71**	100	**34**	100
	80	26	78	**78**	73	**36**	77
	90	**28**	59	**84**	61	31	68
	95	26	55	**85**	51	24	62
A1	None	**2**	100	**25**	100	**41**	100
	80	0	98	**25**	98	**49**	80
	90	0	98	**26**	95	**50**	78
	95	0	85	**32**	73	**51**	76
A2	None	**100**	100	**51**	100	**93**	100
	80	**100**	100	**67**	29	**93**	100
	90	**-**	0	**60**	24	**93**	61
	95	**-**	0	**56**	22	**93**	61
A3	None	**9**	100	**51**	100	**57**	100
	80	0	49	**54**	44	**85**	49
	90	0	36	**62**	38	**88**	32
	95	0	26	**61**	33	**90**	19

Bold = same or increased as a result of threshold
Grey = decreased as a result of threshold

Table 3. Effect of Imposing Cutoff based on Quality Factor (Maximum Similarity QF).

Researcher	QF	Human		Cow		Gull	
		%CC	%Remain	**%CC**	%Remain	**%CC**	%Remain
R1	None	**65**	100	**53**	100	**67**	100
	B	16	15	46	55	**94**	37
R2	None	**74**	100	**18**	100	**5**	100
	C	**74**	100	**18**	100	**5**	97
	B	38	16	0	52	4	59
R3	None	**22**	100	**47**	100	**29**	100
	B	**40**	31	**56**	34	**27**	74
A1	None	**3**	100	**29**	100	**36**	100
	D	0	98	**29**	100	**36**	100
	C	0	98	**29**	100	**34**	98
	B	0	86	**73**	27	**77**	29
A2	None	**100**	100	**90**	100	**98**	100
	C	**100**	100	**88**	85	**98**	100
	B	**100**	100	**83**	30	**98**	100
A3	None	**13**	100	**37**	100	**65**	100
	D	**16**	81	**38**	98	**65**	100
	C	**21**	45	29	76	**67**	69
	B	17	19	0	7	**100**	38

Bold = same or increased as a result of threshold
Grey = decreased as a result of threshold

Table 4. Effect of Imposing threshold value for estimated relative likelihood (ID Bootstrap, Maximum Similarity).

Researcher	%Thresh	Human		Cow		Gull	
		%CC	%Remain	%CC	%Remain	%CC	%Remain
R1	None	**65**	100	**53**	100	**67**	100
	80	**86**	55	**61**	61	**81**	37
	90	**91**	28	**30**	20	**90**	23
	95	**100**	20	**25**	8	**100**	12
R2	None	**74**	100	**18**	100	**5**	100
	80	**40**	30	**17**	96	**7**	61
	90	**56**	18	**17**	94	**5**	43
	95	**100**	4	18	88	**8**	27
R3	None	**22**	100	**47**	100	**29**	100
	80	**43**	29	**46**	60	**25**	69
	90	**25**	16	**53**	32	**25**	46
	95	**20**	10	**43**	15	**20**	14
A1	None	**3**	100	**29**	100	**35**	100
	80	**0**	81	**0**	39	**22**	20
	90	**0**	81	**0**	9	**22**	20
	95	**0**	59	**-**	0	**0**	9
A2	None	**100**	100	**90**	100	**98**	100
	80	100	100	**92**	60	**98**	100
	90	100	100	**100**	20	**98**	100
	95	100	53	**100**	5	**98**	98
A3	None	**13**	100	**39**	100	**65**	100
	80	**-**	0	**-**	0	**-**	0
	90	**-**	0	**-**	0	**-**	0
	95	**-**	0	**-**	0	**-**	0

Black = same or increased as a result of threshold
Grey = decreased as a result of threshold
- = posterior probabilities for classification of isolates fell below threshold

between B and D (there were no A's), resulted in changes in %CC for individual samples between -49 to +44 percentage points (Table 3). For ID Bootstrap posterior probability thresholds between 80% and 95% resulted in changes in %CC for between -35 and +35 percentage points (Table 4). For one researcher (A3), estimated posterior probabilities from ID Bootstrap were below 80% for all isolates.

Where increases in %CC occurred as a result of applying a threshold for classification, large proportions of isolates were often eliminated from final analyses. Increases in %CC above 10% as a consequence of applying a posterior probability threshold in DA often resulted in the elimination of more than half of the sample. Similarly, for exclusion of isolates based on QF's, increases in %CC above 10% were typically accompanied by removal of 60-75% of the sample isolates. For IDBootstrap thresholds, %CC above 10% resulted in the removal of 60%-95% of the sample.

Consequences for both %CC and percent exclusion of sample isolates resulting from applying threshold criteria were not constant across sources. In fact, increases in %CC for one source were often accompanied by decreases in %CC for another source. For researcher R2 using rep-PCR, applying a threshold of 95% for estimated posterior probability for DA increased %CC for human from 66% to 81% and for gull from 8% to 10%, yet decreased %CC for cow from 16% to 0%. For this same researcher, applying a 95% threshold eliminated 38% of the human isolates, 58% of the seagull isolates, and 36% of the cow isolates. Similar trade-offs occurred with applying other threshold criteria as well.

Discussion

The results of this study demonstrate that the statistical method used to predict host source membership can significantly affect the ability of library-based methods to correctly identify sources of fecal contamination. While no one statistical method consistently performed better than another across all MST methods and sources, some statistical approaches were better suited than others for identifying certain patterns in the data and for assigning source membership based on those patterns. Choice of similarity or distance measure defined fingerprint distribution within each library and these distributions, in turn, affected the ability of the statistical algorithm to differentiate among sources. Clustering of sources, multimodality, overlap, and variability of fingerprints within sources had substantial effects on which statistical tool performed best.

The major challenges for each of the statistical methods were lack of library representativeness (no apparent match of unknown) and lack of significant host specificity (overlap between groups and lack of discriminatory characters). Although not addressed in this study, the issue of representativeness is one of the most important issues for library-based MST methods and has recently been addressed with respect to ARA of *Enterococcus* spp. (Wiggins et al., 2003). Representativeness is the link between sampling and the successful use of statistical methods. Even in this particular study, where fingerprints from the unknown isolates should be more similar to those of the library isolates because the same sources of fecal material were used in both the test and library samples, factors such as differential survival (Gordon 2002) or cultivability of indicators (Lleo et al. 2001, Bissonette et al., 1975, Boualam et al., 2002) may have affected the representativeness of the various libraries. While libraries were constructed directly from fecal material (primary habitat for indicators), indicators from the blind samples were isolated from water (secondary habitat). Gordon et al., 2002 suggested a significant change in population structure can take place between these two habitats for *E. coli*. In this study, many of the patterns observed in the blind test samples were distinct from those represented in the library.

Because most of the statistical strategies investigated in this study relied upon either central tendency or similar patterns within the same source, host specificity and similarity of patterns within a source are key to the success of these statistical methods. For the data sets observed in this study there was quite a bit of overlap between various sources (lack of host specificity). Such overlap has been documented by others. For example, McLellan et al. 2003 noted that rep-PCR banding patterns for seagulls overlapped with those obtained from sewage samples. Gordon et al. 2001 has suggested *E. coli* lack sufficient host specificity to be useful in MST methods

While various threshold criteria have been proposed for decreasing false positive errors and thus improving %CC, our investigation showed that such actions do not always produce favorable results. One explanation for the decrease in %CC as a result of applying a particular threshold criteria is the presence of “subtypes” of bacteria within a given host source group. In many of the libraries in this study, we saw clustering of fingerprints into multiple subgroups within the same source and these clusters were often interspersed among multiple clusters within other sources. These subgroups often belonged to the same individual within a given source. As a result test samples often contained isolates that were more similar to another source group or subgroup than the one to which they truly belonged. Applying a threshold value then discarded those isolates that were more poorly matched yet correctly classified while keeping those more similar to the true source yet incorrectly classified.

Recommendations

Given that choice of statistical method is important and that no statistical method emerged as superior to all others, how do we choose among the various statistical approaches available? In order to choose the approach likely to yield the highest %CC, we recommend a careful examination of the library, including visualizing the data, calculating estimates of predictive success, and assessing library representativeness. Repeating this process across the various statistical methods provides the researcher with a measure of potential success and a comparative basis from which to select the most appropriate statistical method. Finally, we recommend options for managing classification ties and fingerprint patterns resulting in all-zeros that may distort or bias %CC.

Visualizing the Data

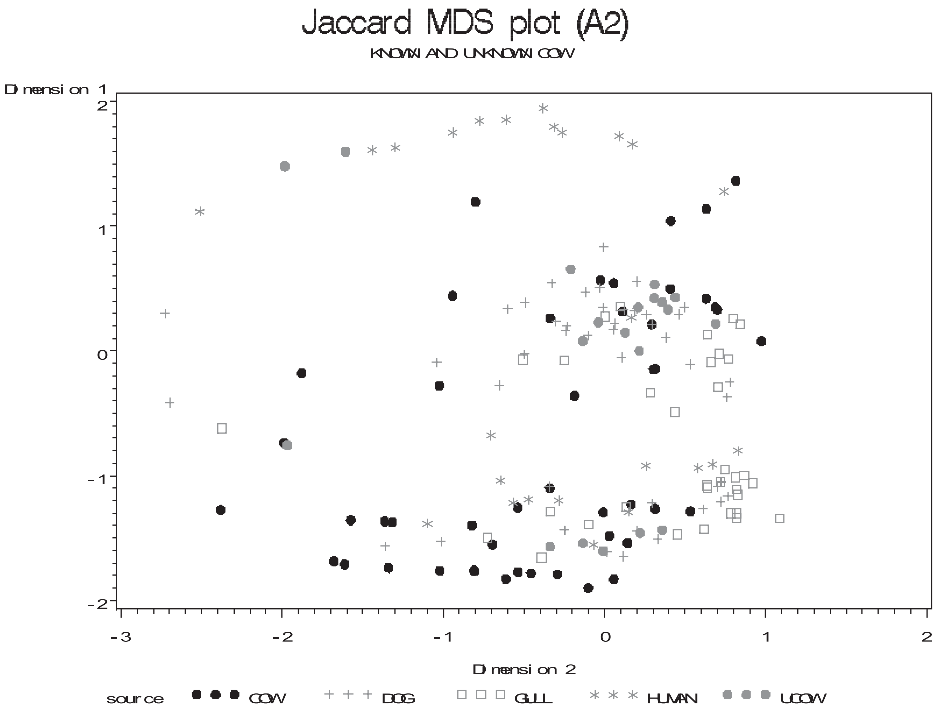
Visualizing fingerprint patterns in the library prior to classifying unknown samples identifies the degree of overlap among fingerprints from different sources, which in turn affects the potential of the statistical method. Many software packages (including SAS and BioNumerics) offer several alternatives for graphical representations of the data. Some of the most popular include canonical discriminant analysis and cluster analyses. When using these visual tools, it is important to note that both are dependent on the choice of distance or similarity measure. Canonical discriminant analyses uses Mahalanobis distances while cluster analyses may use alternative measures such as Jaccard, Dice, and simple matching. Visual display should reflect the distances employed by the particular statistical algorithm under investigation.

In addition to clustering trees and canonical plots where appropriate, we recommend multidimensional scaling (MDS) for visualizing distances among isolates for various measures of distance. MDS produces a "map" in two (or more) dimensions such that the distance between isolates best approximates the relative distances between isolates, allowing one to visualize proximity of isolates both within and among source groups. Therefore, MDS plots often allow one to see what measure of distance or similarity is likely to produce the most separation (or clustering).

MDS plots can also point to whether a statistical tool based on averaging or one based on proximity to singletons is most appropriate. In general, where source patterns are clustered about a central location, with limited overlap, the statistical methods based on averages, such as DA or AS tend to perform well. For sources where fingerprints were patchy and clustered around several locations, NN or MS techniques are often a better alternative. For example, the MDS plot for A2 using Jaccard distance shows that for cow, there are multiple subgroups within the known cow data and these overlap with other known sources (Figure 1). Therefore, maximum similarity was more appropriate than average similarity (%CC for MS = 90% and %CC for AS=0%). Of course the usefulness of MDS plots depends on how representative and complete the library is.

Estimating Predictive Success

Another tool for aiding in the choice of statistical analyses is to compare estimates of predicted success across each of the statistical methods. Among the readily available methods for estimating %CC, we recommend the jackknife. Assuming simple random sampling of the population, jackknifing provides unbiased estimates of correct classification and is readily available in both BioNumerics and SAS. In SAS, jackknife estimates are obtained using the cross validate option in PROC DISCRIM and is preferred over the resubstitution estimates (the default in PROC DISCRIM). In PROC DISCRIM one can also obtain jackknife estimates for percent correct classifications based on thresholds.



Known sources: COW, DOG,GULL, HUMAN
Unkown source: UCOW

Figure 1. Example MDS plot

Although not reported in this study, jackknife estimates were calculated using the standard software across each of the statistical methods for each library. For nearly all libraries, jackknife estimates of %CC were higher than those observed. Inflation may be due, in part, to library construction. In this study libraries depended on subsampling of isolates from individuals or groups of individuals within sources rather than on simple random sampling usually required for these standard jackknife procedures. Wiggins et al. (2003) suggest an alternative jackknife analysis to hold out entire feces instead of individual isolates that may be more appropriate for subsampling designs. Where estimation of predicted success differed substantially from that observed, test fingerprints were underrepresented in the library.

Assessing Library Representativeness

Because statistical algorithms used to identify sources rely on similarity to known patterns, it is crucial that libraries sufficiently represent the population of fingerprints within each source. We recommend a thorough investigation of library representativeness prior to classifying unknown sources of fecal contamination. Wiggins et al. (2003) offer several methods for assessing library representativeness. One method compares jackknife estimates of average rates of correct classification within each source to those estimated by the resubstitution method. Recall that the resubstitution method uses all isolates both to build the library and to predict source membership, thereby estimating how well the library can predict itself. A representative library, then, would provide jackknife estimates of average rates of correct classification comparable to those obtained by resubstitution. A second method for assessing library representatives, also suggested by Wiggins et al. is the "hold out" method, where a portion of the known samples are used to make up the library (i.e. a calibration data set) and the remaining portion is used to estimate average rates of correct classifications (i.e. a validation data set). Again average rates of correct classification are compared with those estimated via resubstitution.

Another potential statistic that may help to test for representativeness is the bootstrap analyses applied to the hold out method. By re-sampling with replacement the distribution against which the test is performed, we can get a better idea of how variable the rates of correct classification are within each source category. This is especially important when one is using maximum similarity to identify unknowns, since a single outlier can drastically affect the result. The magnitude of variability found in repeated simulations would inversely relate to the representativeness of the library. Further, large variabilities in %CC may indicate multiple clustering within each source population or provide evidence of sampling bias.

Managing Classification Ties

Ties occur when the classification rule inherent in the statistical method assigns a given fingerprint to more than one source group. Different software packages have different ways of handling ties and these mechanisms are often unknown or overlooked. In SAS's PROC DISCRIM, for example, the default places ties into an "other" category, allowing the researcher to decide on the appropriate action. Because SAS calculates %CC based on the entire sample (including ties), excluding these data from the final analyses requires the researcher to adjust reported rates of classification (i.e., the denominator). In contrast, BioNumerics' default systematically assigns ties according to the order in which the source groups are listed in the library. Identifications involving numerous ties are likely to result in severe bias. In this study we found a 30-percentage point increase in %CC for one researcher simply by removing the ties prior to applying MS in BioNumerics. Finally, BioNumerics' jackknife procedure for predicting %CC by default assigns ties to their own known source group, resulting in inflated estimates of predicted success. However, in the more recent version of BioNumerics (version 3.0) there is an option designed to reduce bias toward a single source by spreading ties equally among the source groups through random assignment.

Several options exist for handling ties. Selecting which option to use will depend on available auxiliary information and penalties associated with false positive or negative errors. One option is to exclude ties from final analyses. This option should be selected when positive identification of known sources may result in costly penalties relative to consequences of false negatives. For example, beaches can be incorrectly listed as unsafe for swimmers, resulting in severe penalties for sewage dischargers and lost revenue to surrounding business communities, though risk of illness may be low.

A second option for handling ties is to systematically assign ties to a likely source group based on auxiliary information, prior belief, or consequences of false identification. PROC DISCRIM has an option for specifying prior probabilities that will bias assignment of ties toward a particular source. For example, previous research may support a distribution of sources within the target area that favors one source over another. By specifying prior probabilities for source classification, one can weigh source assignment toward one particular source more than another. When reliable and current auxiliary data exist, we recommend using this option regardless. In addition to using prior information, systematic assignment of ties could be achieved visually through dendograms. This procedure is particularly useful for those sources containing multiple clusters or subtypes of fingerprint. Finally, ties can be assigned to a particular source group based on expert knowledge and/or familiarity with the study area. Samples obtained downstream from a dairy farm, for instance, are more likely to belong to cow source group than seagull.

Managing All-zero Patterns

In many cases, fingerprint profiles may yield all-zero patterns. In rep-PCR, all-zero patterns are usually attributed to measurement or laboratory error and are eliminated from further analyses. In ARA, however, all-zero patterns may result from bacteria that are not resistant to any of the antibiotics at the concentrations tested, and thus may contain useful discriminating characteristics. The all-zero patterns present a problem when using similarity measures which target only the number of "ones" that match in the binary data sets, ignoring all "zeros" (e.g. Jaccard and Dice). For example, the Jaccard similarity applied to two isolates whose fingerprints contain all zeros will have a similarity score of "0", yet they are identical. In fact, an all-zero isolate will have zero similarity with all isolates regardless of their pattern. We recommend several alternatives for matching all-zero patterns in ARA data. One alternative is to use exact matching if there is at least one other all-zero value in the library. If there is more than one source in the library that contains an all-zero pattern then one could apply those strategies previously mentioned for managing ties, including using auxiliary information and visualization. Alternatively, one could explore other similarity measures that takes into account zeros (e.g. Euclidean distance or simple matching). For the ARA data sets here we chose to report %CC after removing all-zero fingerprints prior to applying MS and AS to Jaccard similarity scores as there was little difference in %CC with all-zeros included.

Literature Cited

Bissonnette, G.K., Jezeski, J.J., McFeters, G.A. and Stuart, D. G. (1975). Influence of environmental stress on enumeration of indicator bacteria from natural waters. *Applied Microbiology* 29: 186-94.

Boualam, M., Mathieu, L., Fass, S., Cavard, J. and Gatel, D. (2002). Relationship between coliform culturability and organic matter in low nutritive waters. *Water Research* 36: 2618-26.

Carson, C.A., Shear, B.L., Ellersieck, M.R. and Asfaw, A. (2001). Identification of fecal Escherichia coli from humans and animals by ribotyping. *Applied and Environment Microbiology* 67: 1503-7.

Carson, C.A., Shear, B.L., Ellersieck, M.R. and Schnell, J.D. (2003). Comparison of ribotyping and repetitive extragenic palindromic-PCR for identification of fecal Escherichia coli from humans and animals. *Applied and Environment Microbiology* 69: 1836-9.

Dombek, P.E., Johnson, L.K., Zimmerley, S.T. and Sadowsky, M.J. (2000). Use of repetitive DNA sequences and the PCR to differentiate Escherichia coli isolates from human and animal sources. *Applied and Environment Microbiology* 66: 2572-7.

Gordon, D.M. (2001). Geographical structure and host specificity in bacteria and the implications for tracing the source of coliform contamination. *Microbiology* 147: 1079-85.

Gordon, D.M., Bauer, S. and Johnson, J.R. (2002). The genetic structure of Escherichia coli populations in primary and secondary habitats. *Microbiology* 148: 1513-22.

Hagedorn, C., Robinson, S.L., Filtz, J.R., Grubbs, S.M., Angier, T.A. and Beneau, R.B. (1999). Determining sources of fecal pollution in a rural Virginia watershed with antibiotic resistance patterns in fecal streptococci. *Applied and Environmental Microbiology* 65: 5522 - 5531.

Harwood, V.J., Whitlock, J. and Withington, V. (2000). Classification of antibiotic resistance patterns of indicator bacteria by discriminant analysis: use in predicting the source of fecal contamination in subtropical waters. *Applied and Environment Microbiology* 66: 3698-704.

Lleo, M.M., Bonato, B., Tafi, M.C., Signoretto, C., Boaretti, M. and Canepari, P. (2001). Resuscitation rate in different enterococcal species in the viable but non-culturable state. *Journal of Applied Microbiology* 91: 1095-102.

McLellan, S.L., A.D. Daniels, and Alissa K. Salmore. (2003), Genetic characterization of *Escherichia coli* populations from host sources of fecal pollution by using DNA fingerprinting. *Applied and Environment Microbiology* 69: 2587-594.

Scott, T.M., Parveen, S., Portier, K.M., Rose, J.B., Tamplin, M. L., Farrah, S. R., Koo, A. and Lukasik, J. (2003). Geographical variation in ribotype profiles of Escherichia coli isolates from humans, swine, poultry, beef, and dairy cattle in Florida. *Applied and Environment Microbiology* 69: 1089-92.

Scott, T.M., Rose, J.B., Jenkins, T.M., Farrah, S.R. and Lukasik, J. (2002). Microbial source tracking: current methodology and future directions. *Applied and Environment Microbiology* 68: 5796-803.

Simpson, J.M., J.W. Santo Domingo and D.J. Reasoner. (2002) Microbial source tracking: state of the science. *Environmental Science & Technology* 36: 5280-5288

Versalovic, J., Koeuth, T., and Lupski, J.R. (1991). Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. *Nucleic Acids Research* 24, 6823-6831.

Versalovic, J.,M. Schneider, F.J. de Bruijn, and J.R. Lupski. (1994). Genomic fingerprinting of bacteria using repetitive sequence based polymerase chain reaction. *Methods in Molecular and Cellular Biology* 5:25-40.

Wheeler, A.L., Hartel, P.G., Godrey, D.G., Hill, J.L. and Segars, W.I. (2002). Potential of *Enterococcus faecalis* as a human fecal indicator for microbial source tracking. *Journal of Environmental Quality* 31: 1286-1293.

Whitlock, J.E., Jones, D.T. and Harwood, V.J. (2002). Identification of the sources of fecal coliforms in an urban watershed using antibiotic resistance analysis. *Water Research* 36: 4273-82.

Wiggins, B.A. (1996). Discriminant analysis of antibiotic resistance patterns in fecal streptococci, a method to differentiate human and animal sources of fecal pollution in natural waters. *Applied and Environment Microbiology* 62: 3997-4002.

Wiggins, B.A., Andrews, R.W., Conway, R.A., Corr, C.L., Dobratz, E.J., Dougherty, D.P., Eppard, J.R., Knupp, S.R., Limjoco, M.C., Mettenburg, J.M., Rinehardt, J.M., Sonsino, J., Torrijos, R.L. and Zimmerman, M.E. (1999). Use of antibiotic resistance analysis to identify nonpoint sources of fecal pollution. *Applied and Environment Microbiology* 65: 3483-6.

Wiggins, B.A., P. W. Cash, W.S. Creamer, S.E. Dart, P.P. Garcia, T.M. Gerecke, J. Han, B.L. Henry, K.B. Hoover, E.L. Johnson, K.C. Jones, J.G. McCarthy, J.A. McDonough, S.A. Mercer, M.J. Noto, H. Park, M.S. Phillips, S.M. Purner, B.M. Smith, E.N. Stevens and A.K. Varner. (2003) Representativeness testing of multi-watershed libraries using antibiotic resistance analysis. *Applied and Environment Microbiology* 69: 3399-405.

Acknowledgements

I would like to thank J. Griffith for all of his efforts that led to the success of the MST study. He provided me with both a better understanding of library based methods as well as a deep appreciation for sample collection and preparations. I would also like to thank D. Stoekel for asking the really difficult questions. Finally I would like to thank S. Weisberg for his helpful comments relating to the organization and writing of the manuscript.