APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Jan. 2005, p. 512–518 Vol. 71, No. 1
0099-2240/05/$08.00+0 doi:10.1128/AEM.71.1.512–518.2005

Methods To Increase Fidelity of Repetitive Extragenic Palindromic PCR Fingerprint-Based Bacterial Source Tracking Efforts

Wail M. Hassan, Shiao Y. Wang,* and Rudolph D. Ellender

Department of Biological Sciences, The University of Southern Mississippi, Hattiesburg, Mississippi

Received 16 April 2004/Accepted 23 August 2004

The goal of the study was to determine which similarity coefficient and statistical method to use to produce the highest rate of correct assignment (RCA) in repetitive extragenic palindromic PCR-based bacterial source tracking. In addition, the use of standards for deciding whether to accept or reject source assignments was investigated. The use of curve-based coefficients Cosine Coefficient and Pearson's Product Moment Correlation yielded higher RCAs than the use of band-based coefficients Jaccard, Dice, Jeffrey's x, and Ochiai. When enterococcal and *Escherichia coli* isolates from known sources were used in a blind test, the use of maximum similarity produced consistently higher RCAs than the use of average similarity. We also found that the use of a similarity value threshold and/or a quality factor threshold (the ratio of the average fingerprint similarity within a source to the average similarity of this source's isolates to an unknown) to decide whether to accept source assignments of unknowns increases the reliability of source assignments. Applying a similarity value threshold improved the overall RCA (ORCA) by 15 to 27% when enterococcal fingerprints were used and 8 to 29% when *E. coli* fingerprints were used. Applying the quality factor threshold resulted in a 22 to 32% improvement in the ORCA, depending on the fingerprinting technique used. This increase in reliability was, however, achieved at the expense of decreased numbers of isolates that were assigned a source.

The goal of bacterial source tracking is to identify sources of bacterial contamination in, for example, water or food products. One of the main uses of bacterial source tracking is to identify sources of fecal pollution in natural waters. Because the bacterium used as a tracer is part of the normal intestinal microbiota, it can be used as an indicator of the source of fecal pollution. With a DNA fingerprint library-based approach, the first step is to establish a collection of fingerprints of isolates from likely sources of fecal pollution. The fingerprints are grouped into library units according to the source of the isolate. The DNA fingerprints from water-isolated bacteria are then compared with the libraries of fingerprints to determine the most-likely origin of the isolate. Each isolate is assigned to an origin depending on how similar its fingerprint is to those in each library unit (2, 4, 5, 10, 12, 13).

There are several coefficients that can be used to calculate similarity, and the choice of which similarity coefficient to use can affect the outcome of the source tracking assignment. Cosine Coefficient and Pearson's Product Moment Correlation are curve-based coefficients that use both the presence or absence of DNA bands and the peak intensity of each band as variables, whereas Jaccard, Dice, Jeffrey's x, and Ochiai are band-based coefficients that consider only the presence or absence of DNA bands. Currently, there is no consensus on which coefficient results in more accurate source assignment. Pearson's Product Moment Correlation has been used to calculate similarities among repetitive extragenic palindromic (rep) PCR fingerprints (4, 12) and ribotypes (3). Cosine Coefficient has been used to calculate similarities among rep-PCR and pulsed-field gel electrophoresis fingerprints (10). The band-based coefficients Jaccard (5) and Dice (13) have been used to calculate similarities among rep-PCR fingerprints, and the latter has also been used to calculate similarities among ribotypes (7, 11) and among fingerprints generated by the amplification of the 16S-23S intergenic spacer region (13).

In addition to uncertainties concerning which similarity coefficient to use, there are no standards concerning acceptance of source assignments. When similarity coefficients are used, each environmental isolate is assigned a source depending on which library includes the DNA fingerprints to which its DNA fingerprint is most similar. Because each isolate is always assigned a source, without regard to the actual degree of similarity, the user must decide, using a priori assessments of the accuracy of the method, whether each source assignment is likely to be correct.

Currently, there are no published studies on the issue of reliability when source assignments are determined using similarity coefficients. Information is needed to better understand limitations of statistical analyses used for assigning sources of bacteria on the basis of DNA fingerprint patterns (1). In an attempt to evaluate the significance of source tracking on the basis of discriminant analysis, Whitlock et al. (18) suggested using the percentage of misassigned isolates in each source group calculated by Jackknife analysis as a lower limit for significance. When such an approach is used, each isolate from a collection of isolates from a known source (a library) is treated, one at a time, as an unknown and then identified by comparison to the remaining isolates. The proportion of isolates assigned to the correct source relative to the number of isolates in the library is the rate of correct assignment. Whitlock et al. (18) proposed that a source can be implicated in water pollution only when the percentage of environmental isolates assigned to it exceeds the percentage of library isolates

* Corresponding author. Mailing address: Department of Biological Sciences—USM, 118 College Dr. #5018, Hattiesburg, MS 39406-0001. Phone: (601) 266-6795. Fax: (601) 266-5797. E-mail: shiao.wang@usm.edu.

TABLE 1. Sources and numbers of fecal samples and bacterial isolates (before and after exclusion of clonal isolates) used in the study

Animal source	No. of fecal samples (no. of isolates)[a]	
	Escherichia coli	Enterococci
Humans	59 (209/108)	65 (186/100)
Cows	71 (302/202)	105 (253/170)
Deer	48 (137/102)	86 (184/126)
Dogs	64 (133/100)	80 (131/107)
Chicken	69 (385/202)	113 (654/399)
Gulls	107 (248/153)	81 (176/118)
Total	418 (1,414/867)	530 (1,584/1,020)

[a] Numbers in parentheses represent the number of isolates before exclusion of clonal isolates/number of isolates after exclusion of clonal isolates.

misassigned to it by Jackknife analysis. Although this approach can be useful when libraries with good representation of the diversity of isolates in the environmental site being studied are used, it is less useful when libraries with poor representation are used or in the presence of isolates contributed by nonlibrary sources. Therefore, in computer-assisted, library-based bacterial source tracking efforts, the choice of which similarity coefficient to use for DNA fingerprint comparisons and what similarity threshold to use in deciding whether each source assignment is to be accepted are important issues that need further investigations to improve the reliability of source assignments.

In the present study, six different similarity coefficients were compared in terms of their rates of correct assignment (RCAs). In addition, statistical options to improve the reliability of source assignments on the basis of the use of similarity coefficients were investigated. These options include the choice of how the similarity values are used and the effect of the use of a threshold similarity value and quality factor on improvement of the reliability of source assignments.

MATERIALS AND METHODS

Sample collection. Fecal samples were collected from humans, cows, deer, dogs, chicken, and gulls (Table 1). A total of 784 fecal samples were used in the study (Table 1). Some samples did not yield *Escherichia coli*, while others did not yield *Enterococcus* spp. Thus, the numbers of fecal samples and isolates were not the same for the two indicator organisms.

Eighty-four human fecal samples were used in the present study. A total of 46 rectal swab samples were collected from human volunteers at the University of Southern Mississippi; of these, 42 were collected using a BBL CultureSwab collection and transport device with Cary-Blair transport medium (BD Diagnostics Systems, Sparks, Md.), and 4 were collected with sterile swabs and suspended in fetal bovine serum containing 10% dimethyl sulfoxide (DMSO). Thirty-eight samples were collected at a local hospital in Laurel, Mississippi. Fifty-nine samples yielded *E. coli*, and 65 yielded *Enterococcus* spp. (Table 1). Samples collected using CultureSwabs were refrigerated and processed within 24 h after collection. Samples collected in fetal bovine serum were frozen immediately after collection. Hospital samples were collected in sterile cups shipped to our laboratory within 24 h and used within another 24 h.

Cow, chicken, and deer samples were collected from south and central Mississippi. All of the fecal samples were collected from individuals except for some of the cow and chicken samples. Four of the cow samples were composite samples from several cows at the same farm. A total of 86 of the chicken samples were litter samples from commercial chicken farms, while 27 were obtained from cloacal swabs from individual chickens. Although some cow and chicken cloacal samples were collected and transported using CultureSwabs, the majority of cow and chicken litter samples, as well as all deer samples, were collected and frozen without additives at their respective collection sites across the State. Dog fecal samples from veterinary adoption centers and humane societies in Hattiesburg and Gulfport, Mississippi, and gull fecal samples from beaches along the Mississippi Gulf coast were collected using CultureSwabs.

Bacteria isolation. Fecal samples were streaked on mTEC (Difco) and mEI plates for the isolation of *E. coli* and enterococci, respectively (15). mTEC plates were incubated at 37°C for 2 to 4 h and then at 44.5°C for 18 to 24 h. Yellow colonies were picked and confirmed using standard microbiological methods. Isolates that lacked phenylalanine deaminase, that produced indole from tryptophan, that were unable to utilize sodium citrate as a sole carbon source, and that fermented glucose through a mixed-acid fermentation pathway (but not a butanediol pathway) were considered to be *E. coli*. mEI plates were incubated at 41°C for 24 to 36 h. Colonies that formed blue halos were presumed to be enterococci. Confirmation was performed by testing each isolate for growth at 45°C and in the presence of 6.5% sodium chloride at 37°C and for esculin hydrolysis. Among the isolates picked, 89.5 and 73.1% were confirmed to be *E. coli* and enterococci, respectively.

rep-PCR and BOX-PCR. rep-PCR (8, 17) was performed using a modified method of Rademaker and DeBruijn (12). Isolates were grown at 37°C in brain heart infusion for 12 to 16 h. Cells harvested from 0.5 and 1.0 ml of broth for *E. coli* and enterococci, respectively, were washed twice with 0.5 ml of sterile deionized water. The resulting pellets were resuspended in deionized sterile water (0.5 and 0.25 ml for *E. coli* and enterococci, respectively) and stored frozen at −20°C until use as a template for PCR. DNA amplification reactions were performed with a 10-μl reaction mixture that consisted of 1 μl of cell suspension and 9 μl of PCR master mix. The BOX-PCR (9, 16) master mix contained 2 μM primer (BOX A1R [CTA CGG CAA GGC GAC GCT GAC G]), 1 mM deoxynucleoside triphosphates, 4.5 mM $MgCl_2$, 1× buffer provided by the manufacturer of the DNA polymerase, and 0.4 units of JumpStart *Taq* DNA polymerase (Sigma, St. Louis, Mo.). Thermal cycling started with 2 min at 95°C followed by 35 cycles of 94°C for 3 s, 92°C for 30 s, 50°C for 1 min, and 65°C for 8 min. A final extension step was performed at 65°C for 8 min after completion of the 35 cycles. The REP-PCR (6, 9, 14) master mix contained 3 μM of each of two primers (REP 1R [III ICG ICG ICA TCI GGC] and REP 2I [ICG ICT TAT CIG GCC TAC]), 1 mM deoxynucleoside triphosphates, 2.5 mM $MgCl_2$, 1× buffer, and 0.4 units of JumpStart *Taq* DNA polymerase. The thermal cycling protocol for REP-PCR was the same as that for BOX-PCR except that 40°C was used instead of 50°C for primer annealing.

Jackknife analysis. The effect of having clonal isolates in fingerprint libraries on RCAs was examined by performing Jackknife analysis both before and after their removal by use of BOX and REP fingerprints of enterococcal and *E. coli* isolates. Clonal isolates were defined in the present study as isolates with identical fingerprints obtained from the same sample. Removal of clonal isolates was performed for Jackknife analysis only.

Jackknife analysis was also used to compare the RCAs generated using six similarity coefficients (Cosine Coefficient, Pearson's Product Moment Correlation, Jaccard, Dice, Jeffrey's x, and Ochiai). The fingerprints used were produced by BOX-PCR using enterococcal isolates with clonal isolates removed. All Jackknife analyses were performed using BioNumerics version 3.0 (Applied Maths, Sint-Martens-Latem, Belgium). Pattern optimization (i.e., the percentage of pattern shift within which the software looks for the best match) was set at 5%, and band tolerance (i.e., the maximum gel migration difference for any pair of bands to be considered matching) was set at 2% with a 2% gradual tolerance increase towards the bottom of the gel.

Discriminant analysis and multivariate analysis of variance. Enterococcal isolates from human, cow, deer, dog, chicken, and gull fecal samples were classified into groups according to their sources. Discriminant analysis was used to show the separation between these predefined groups on the basis of their BOX fingerprints. The first and second discriminants were plotted on the *x* and *y* axes, respectively, generating a two-dimensional plot showing the separation of isolates from six sources. Multivariate analysis of variance was performed accounting for the covariance structure to evaluate the significance of discriminant analysis. The *P* value indicated the probability of obtaining equivalent separation results among isolates of different sources due to random classification of isolates. The probability of obtaining the same level of discrimination, assuming that all isolates were obtained from a homogeneous population (i.e., the effect of grouping by source was insignificant), is indicated by the Wilkinson's likelihood for normal distribution (*L*). Low *P* and *L* values indicate significant discrimination by source group.

Identification libraries and the blind test. The RCAs using each of the six similarity coefficients listed above were also determined using a blind test. First, libraries containing rep-PCR fingerprints of enterococcal and *E. coli* isolates from each known animal source were constructed. Each library consisted of five units, one for each known source: human, cow, deer, dog, and chicken. These

TABLE 2. Comparison of the RCAs of enterococcal isolates obtained before and after excluding clonal isolates

Animal source	% RCA (no. of isolates used)[a]:	
	With clonal isolates included	With clonal isolates removed
Humans	97 (186)	87 (100)
Cows	90 (253)	82 (170)
Deer	95 (184)	88 (126)
Dogs	79 (131)	71 (107)
Chicken	96 (654)	89 (399)
Gulls	84 (176)	59 (118)
Overall	92 (1,584)	82 (1,020)

[a] The isolates were fingerprinted using BOX-PCR, similarities were calculated using Cosine Coefficient, and the RCAs were calculated using Jackknife analysis.

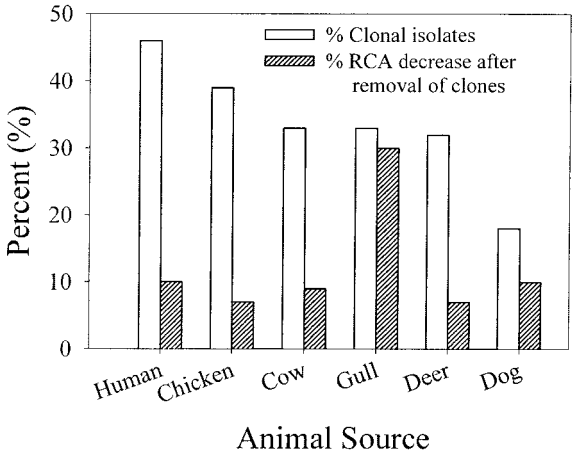

FIG. 1. The percentages of clonal isolates among enterococci isolated from different animal sources and the decreases in RCA that resulted from their removal.

libraries were then used as reference to determine the most likely animal source of new isolates in the blind test. The enterococcal fingerprint library contained 762 isolates (67 human, 141 cow, 99 deer, 103 dog, and 352 chicken). All were analyzed by BOX-PCR, but only 458 (40 human, 118 cow, 84 deer, 71 dog, and 145 chicken) were analyzed by REP-PCR. The *E. coli* library contained 514 isolates (65 human, 136 cow, 39 deer, 142 dog, and 132 chicken), and all were analyzed by both BOX- and REP-PCR.

Isolates used as blind samples were obtained from feces of animals in the same general population as those used to obtain isolates for rep-PCR fingerprint library construction. A total of 131 enterococcal isolates (28 human, 29 bovine, 27 deer, and 47 chicken) and 130 *E. coli* isolates (19 human, 43 bovine, 17 deer, and 51 chicken) were analyzed using BOX-PCR for the blind test. A total of 96 enterococcal isolates (12 human, 28 bovine, 23 deer, and 33 chicken) and the same 130 *E. coli* isolates were analyzed using REP-PCR for the blind test. Source assignments were made using Cosine Coefficient to calculate similarity matrices, and average RCAs were compared using both maximum and average similarity options.

Setting "similarity value" and "quality factor" thresholds. A similarity threshold was determined for each indicator organism-fingerprinting technique combination. Cosine Coefficient was used to calculate similarity matrices. The threshold was determined by dividing the sum of the average similarity values of the correctly and the incorrectly assigned isolates by 2. In other words, the threshold was the midpoint between the average similarities of the correctly and incorrectly assigned isolates. When this method was used, the similarity threshold values were 90, 90.8, and 89.2% for enterococcal BOX, REP, and BOX-REP combined fingerprints, respectively. The combined fingerprints were generated electronically using BOX and REP fingerprints. The threshold values for *E. coli* BOX, REP, and combined fingerprints were 91.5, 90.1, and 87.7%, respectively.

A quality factor was also used as a threshold for determining the reliability of source assignments. A quality factor is generated by BioNumerics for each unknown as the unknown is assigned to an animal source. This value is calculated by dividing the average pairwise similarity of all fingerprints in the source group by the average pairwise similarity of the unknown with each of the library's component isolates. Assignments with a quality factor of 1.0 or less (B or better) were accepted, while those with a quality factor of more than 1.0 (C, D, or E) were considered unidentifiable.

RESULTS AND DISCUSSION

The effect of clonal isolates on Jackknife RCAs. The presence of clonal isolates artificially inflates RCAs calculated by Jackknife analysis. The overall RCA (ORCA) among 1,584 enterococcal isolates, including clonal isolates, was 92%. However, the ORCA of the remaining 1,020 isolates after removing clonal isolates was 82% (Table 2).

The decrease in RCA as a result of the removal of clonal isolates differed among animal sources, ranging from a 7% decrease among deer and chicken isolates to 25% for gull isolates. The decrease in RCA did not correlate with the number of clonal isolates removed from each source (Fig. 1). For example, the human and dog groups contained 46 and 18% clonal isolates, respectively, but removing clonal isolates from either group resulted in a 10% decrease in the RCA. On the other hand, the cow and gull groups contained 33% clonal isolates each, and removing clonal isolates resulted in a 9 and 25% decrease in the RCA, respectively. The observation that the RCA of gull isolates was the most susceptible to artificial inflation due to the presence of clonal isolates may have resulted from their lowest RCA among isolates of other sources (see below). This may be due to the scavenger feeding strategy of gulls. Gulls were frequently observed feeding and bathing at a sewage treatment lagoon near the study site. Our hypothesis is that by feeding on food or drinking water already contaminated by enterococci from other animal sources, gull isolates can be more easily confused with those from other animal sources, hence, their low RCA. Therefore, providing perfect matches to some of the isolates, by including clonal isolates, would artificially inflate the RCA of gull isolates to the greatest degree.

The presence of clonal isolates inflates the RCA in Jackknife analysis because members of a clone produce almost identical fingerprints. In Jackknife analysis, isolates are divided into groups depending on their sources. Then, individual isolates are sequentially removed, one at a time, from the collection of isolates under study, treated as unknowns, and assigned to sources on the basis of fingerprint similarity. When clonal isolates are present, the isolate treated as the unknown is always assigned to the library source where another member of the same clone is present. Because Jackknife analysis can be used to validate the reliability of the source tracking method or used to determine isolate overlap among potential sources, it is important for users to be aware of this potential source of error. It is also important to know, however, that this potential source of error is concerned specifically with Jackknife analysis and not with library-based source tracking.

The effect of using different similarity coefficients on RCAs. A comparison of Jackknife RCAs of 1,020 enterococcal isolates on the basis of their BOX fingerprints indicated that the highest RCAs were obtained using curve-based similarity coefficients. The ORCA was 82% with both Pearson's Product

TABLE 3. Comparison of the RCAs of enterococcal isolates source assigned using six similarity coefficients[a]

Animal source	% RCA by:					
	Pearson's	Cosine	Jaccard	Dice	Jeffrey's x	Ochiai
Human	86	87	77	77	76	76
Cow	82	82	79	79	76	79
Deer	86	88	89	89	87	87
Dog	73	71	58	58	60	57
Chicken	88	89	87	87	88	87
Gull	59	59	54	54	54	55
Overall	82	82	78	78	78	78
Standard deviation	11.2	12.0	14.7	14.7	13.9	14.3

[a] Source assignments of the 1,020 isolates were made using their BOX-PCR fingerprints, and the RCAs were calculated using Jackknife analysis.

TABLE 4. Comparison of the RCAs obtained by use of average versus maximum similarity as a basis for bacterial source assignments

Bacterial species and animal source	RCA by[a]:			
	BOX fingerprinting		REP fingerprinting	
	Average similarity	Maximum similarity	Average similarity	Maximum similarity
Enterococcus spp.				
Humans	71	68	75	50
Cows	79	63	11	71
Deer	56	74	78	57
Chicken	64	89	67	79
Overall	64	76	54	69
Standard deviation	9.8	11.3	31.5	13.2
E. coli				
Humans	5	32	16	32
Cows	79	86	72	74
Deer	35	41	0	35
Chicken	71	69	47	78
Overall	59	65	45	65
Standard deviation	34.2	24.9	32.1	24.6

[a] The RCAs were calculated from the blind test.

Moment Correlation Coefficient and Cosine Coefficient. When band-based coefficients were used, the ORCA was 78% for each of the four coefficients. Although the RCAs differed among animal sources with each similarity coefficient, they were less variable using curve-based coefficients. The standard deviations for the ORCAs were 11.2 and 12.0% using Pearson's Product Moment Correlation Coefficient and Cosine Coefficient, respectively, but ranged from 13.9 to 14.7% for the four band-based coefficients (Table 3). These results suggest that the use of curve-based coefficients is preferred over the use of band-based coefficients for source tracking in our study area with BOX fingerprinting. Additional data, from other study areas and with other DNA fingerprinting protocols, are needed to ascertain whether the superiority of curve-based coefficients is a general rule.

The effect of using maximum versus average similarity on RCAs. The ORCAs obtained with the blind test were consistently higher using maximum similarity for source assignments than using average similarity (Table 4). This was the case regardless of the target organism (*E. coli* or *Enterococcus* spp.) or the fingerprinting technique used (BOX-PCR or REP-PCR). Although the ORCAs obtained using maximum similarity were consistently higher, the individual RCAs among animal sources were frequently but not always higher when maximum similarity was used. Among the 16 pair-wise comparisons (4 animal sources × 2 fingerprinting methods × 2 bacterial indicators), average similarity yielded higher RCAs in five cases (human and cow enterococci analyzed by BOX-PCR, human and deer enterococci analyzed by REP-PCR, and chicken *E. coli* analyzed by BOX-PCR). A striking difference between using average and maximum similarity is the large gap in the RCAs for some animal sources. For example, among cow enterococci analyzed using REP-PCR, the RCA was only 11% using average similarity but was 71% using maximum similarity. Among deer *E. coli* isolates, the RCA was 0% using average similarity but 35% using maximum similarity.

When maximum similarity, an epidemiological approach, is used, an isolate is assigned to the source group containing the best-matching isolate in the identification library. When average similarity, a population genetics approach, is used, an isolate is assigned to the source group with which it shares the highest average similarity. Consequently, the use of maximum similarity would be advantageous when the within-source group fingerprints are divergent, while the use of average similarity would be advantageous when the fingerprints within source groups are closely related; our results imply that the former is the case for the source animal populations in our study area. The implied genetic diversity was reflected in the complex cluster analysis dendrogram of the isolates. Isolates from one source often clustered with isolates from a different source (data not shown). This overlap is apparent in the two-dimensional plot generated using discriminant analysis (Fig. 2). The relative contributions of discriminants to total discrimination in descending order from discriminant 1 through 5 were 42, 21, 18, 11, and 7%. The *P* value associated with each of the discriminants was 0.001%, indicating that the likelihood of obtaining the same level of discrimination by use of random grouping of isolates is remote. The *L* values for discriminants 1 through 5 were 0.0904, 0.2166, 0.3656, 0.5849, and 0.8053, respectively. In other words, the probability that isolates from all sources belong to a homogeneous population calculated using the first discriminant to distinguish groups (i.e., the probability that grouping by source by use of the first discriminant is insignificant) equals 9%. These results indicate that the grouping of enterococcal isolates by animal source is significant. However, the loose clustering of isolates in the discriminant analysis plot (Fig. 2) shows diversity among fingerprints which supports the use of maximum similarity in source tracking.

In addition to resulting in higher RCAs in general, the use of maximum similarity also resulted in less-variable RCAs among animal sources compared to the results seen with RCAs obtained using average similarity. The standard deviations of RCAs among enterococcal isolates as determined on the basis of their REP fingerprints were 13.2 and 31.5% by use of maximum and average similarity, respectively (Table 4). Among *E. coli* isolates, they were 24.9 and 34.2%, respectively, as deter-

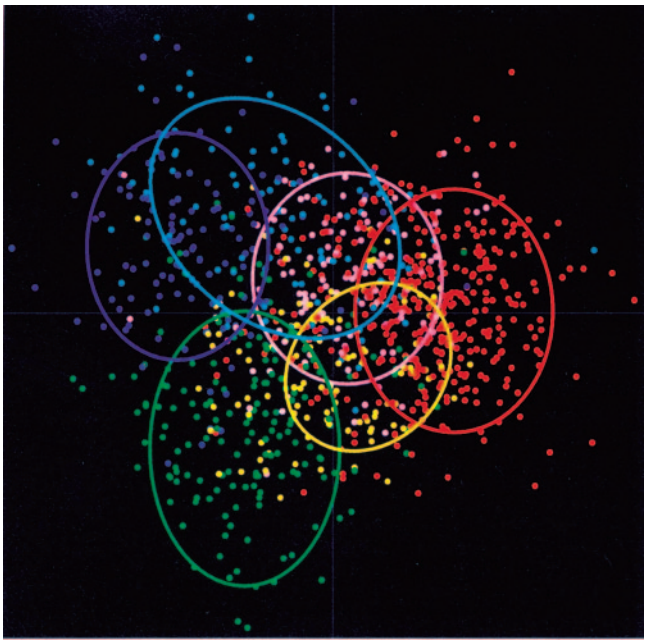

FIG. 2. A two-dimensional plot using discriminant analysis showing the separation of isolates on the basis of BOX fingerprints. The plot was generated by plotting the first discriminant (contributing 42% of total discrimination) on the *x* axis and the second discriminant (contributing 21% of total discrimination) on the *y* axis. The *P* value was 0.001 for both discriminants, while the *L* values were 0.0904 and 0.2166 for the first and second discriminants, respectively. Human isolates are shown in cyan, cow isolates are shown in green, deer isolates are shown in blue, dog isolates are shown in yellow, chicken isolates are shown in red, and gull isolates are shown in pink.

mined on the basis of their BOX fingerprints. When REP fingerprints were used, the standard deviation of RCAs among *E. coli* isolates were 24.6 and 32.1% by use of maximum and average similarity, respectively (Table 4). The only case in which the standard deviation of RCAs was higher using maximum similarity than using average similarity was in the case of RCAs of enterococci determined using BOX-PCR fingerprints, where they were 9.8% using average similarity and 11.3% using maximum similarity. These results suggest that the use of maximum similarity in future DNA fingerprint library-based bacterial source tracking efforts is warranted. The use of maximum similarity results in RCAs that are not only on average higher but also more consistent among different animal sources, regardless of the target organism (*E. coli* or *Enterococcus* spp.) or the fingerprinting technique used (BOX-PCR or REP-PCR).

The effect of using a similarity value threshold on RCAs. By classifying as unidentified those isolates that did not meet the similarity threshold requirement, the RCAs among isolates assigned a source improved significantly. The ORCA among enterococci fingerprinted by BOX-PCR increased from 76 to 87% (Table 5). In addition, the number of isolates assigned incorrectly to a source decreased from 24 to 9%. The drawback, however, is that the proportion of isolates that were assigned to a source decreased from 100 to 69%, and 16% of the isolates assigned correctly to a source when a threshold was not used are designated as unidentifiable when a threshold was used. Similar results were obtained using enterococcal REP fingerprints, where the ORCA increased from 63 to 80% after the threshold was applied. When combined BOX-REP fingerprints were used, the RCAs before and after application of the similarity threshold were 77 and 89%, respectively. With *E. coli*, the ORCAs increased from 65 to 70% for isolates fingerprinted using BOX-PCR, from 65 to 82% for isolates fingerprinted using REP-PCR, and from 69 to 89% for isolates fingerprinted using combined BOX- and REP-PCR fingerprints (Table 5). However, note that these increases in ORCA are achieved at a cost. A total 13 to 29% of the isolates formerly assigned to a correct source when a threshold was not used are classified as unidentifiable when a threshold is used.

The use of a similarity threshold in deciding whether to accept source assignments increased the RCA for each of the six similarity coefficients in a blind test. Its use also had a significant effect on the proportion of isolates that were iden-

TABLE 5. Effect of using a similarity threshold during source assignments on the ORCA, the percentages of isolates correctly assigned, and the percentages incorrectly assigned[a]

Fingerprint type	Similarity threshold	ORCA (%)[a]	% Assigned to an animal source	% Correctly assigned	% Incorrectly assigned
Enterococcus spp.					
BOX	None	76	100	76	24
BOX	90.0%	87	69	60	9
REP	None	63	100	63	37
REP	90.8%	80	63	50	13
BOX + REP	None	77	100	77	23
BOX + REP	89.2%	89	67	59	8
E. coli					
BOX	None	65	100	65	35
BOX	91.5%	70	52	36	15
REP	None	65	100	65	35
REP	90.1%	82	55	45	10
BOX + REP	None	69	100	69	31
BOX + REP	87.7%	89	56	50	6

[a] The ORCA values represent the percentages of isolates among those assigned a source that were assigned correctly. The percent correctly assigned and percent incorrectly assigned are the percentages of isolates that were correctly and incorrectly assigned among all isolates, respectively (including those not assigned to an animal source).

TABLE 6. Effect of using a similarity threshold on the RCAs of enterococcal isolates, the proportion of isolates assigned a source, and the proportion of isolates assigned to the correct source[a]

Animal source	RCA by:											
	Pearson's (%)		Cosine (%)		Jaccard (%)		Dice (%)		Jeffrey's x (%)		Ochiai (%)	
	Without threshold	With threshold	Without threshold	With threshold	Without threshold	With threshold	Without threshold	With threshold	Without threshold	With threshold	Without threshold	With threshold
Humans	71	75	68	75	71	88	71	88	68	88	71	88
Cows	69	75	69	75	76	73	76	73	79	82	76	82
Deer	63	79	66	78	67	71	67	83	67	71	63	71
Chicken	89	97	89	97	85	94	85	93	87	93	87	93
Overall	76	85	76	85	76	84	76	85	77	85	76	85
Identifiable	100	67	100	71	100	34	100	30	100	31	100	31
Percent correctly assigned	76	57	76	60	76	28	76	25	77	26	76	26
Percent incorrectly assigned	24	10	24	11	24	5	24	5	23	5	24	5

[a] Source assignments were made on the basis of BOX-PCR fingerprints. The RCAs were calculated using a blind test with and without a threshold value. The bottom two rows show the proportions of total tested isolates that were correctly or incorrectly assigned to an animal source.

tifiable in terms of animal origin and of the proportion identified correctly. Among the 131 enterococcal isolates analyzed in blind testing by BOX-PCR, the ORCAs were 76 to 77% for the six similarity coefficients when similarity thresholds were not used (Table 6). When source assignments were accepted only when similarity thresholds were met, the ORCAs increased to 84 to 85%. Although the ORCAs were equivalent among the six similarity coefficients, the proportion of isolates that were correctly assigned to an animal source was significantly higher by use of curve-based coefficients than by use of band-based coefficients (Fig. 3). When the curve-based coefficients Pearson's and Cosine Coefficient were used, 57 and 60%, respectively, of the isolates were correctly assigned. When the band-based coefficients Jaccard, Dice, Jeffery's x, and Ochiai were used, only 28, 25, 26, and 26%, respectively, of the total isolates were correctly assigned (Table 6 and Fig. 3).

The effect of using a quality factor threshold on RCAs. Applying the quality threshold criterion to source assignments also resulted in significant improvements in the RCAs, regardless of whether BOX, REP, or combined BOX-REP fingerprints were used. The ORCA increased from 74 to 98% with BOX-PCR fingerprints, from 63 to 77% with REP-PCR fingerprints, and from 77 to 97% with combined BOX-REP fingerprints (Table 7). This increase in source assignment accuracy was obtained at the expense of efficiency. The percentage of isolates assigned a source decreased from 100% when all source assignments were accepted to 43, 50, and 32% for BOX, REP, and combined BOX-REP fingerprints, respectively, when a quality threshold was used. However, mistakes in source identification also decreased dramatically. When a quality threshold was not used for BOX, REP, and combined BOX-REP fingerprints, the percentages of isolates assigned to an incorrect source were 26, 37, and 23%, respectively (Table 7). When a quality threshold was used, these values decreased to 1, 11, and 1%, respectively.

Although the usefulness of a similarity value and/or a quality factor threshold to improve the reliability of bacterial source assignments appears obvious, the reduction in the proportion of isolates that can be assigned to a source can be a concern. If the proportion of isolates classified as unidentified is large, a significant source of fecal pollution may remain hidden. This would occur if the fingerprint libraries used were of insufficient

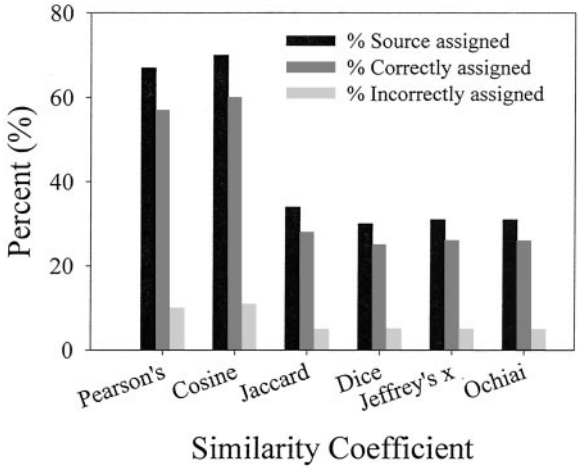

FIG. 3. The effect of curve-based and band-based similarity coefficients on the percentages of enterococcal isolates assigned to an animal source and the percentages of isolates assigned correctly and incorrectly. Source assignments were made on the basis of their BOX-PCR fingerprints (n = 131).

TABLE 7. Effect of applying a quality factor threshold on the ORCA, the percentages of enterococcal isolates correctly assigned, and the percentages incorrectly assigned[a]

Fingerprint type	Quality factor threshold	ORCA	% Assigned to an animal source	% Correctly assigned	% Incorrectly assigned
BOX	None	74	100	74	26
BOX	1.0	98	43	42	1
REP	None	63	100	63	37
REP	1.0	77	50	38	11
BOX + REP	None	77	100	77	23
BOX + REP	1.0	97	32	31	1

[a] Source assignments were only accepted when the quality factor was equal to or less than 1.0. Similarity matrices were calculated using Cosine Coefficient.

size. The proportion of unknown isolates that can be classified given a certain similarity value or quality factor threshold may be useful in future studies in determining whether libraries of sufficient size have been achieved for reliable bacterial source tracking efforts.

In conclusion, results from the present study indicate that (i) the use of curve-based coefficients (e.g., Cosine Coefficient and Pearson's Product Moment Correlation) results in higher ORCAs than the use of band-based coefficients (e.g., Jaccard and Dice); (ii) the removal of clonal isolates is essential for the proper calculation of RCAs by Jackknife analysis; (iii) the use of maximum, as opposed to average, similarity yields higher ORCAs; and (iv) the application of a similarity value or a quality factor threshold for source assignment improves the ORCA, but this is achieved at the expense of the total numbers of isolates assigned a source.

ACKNOWLEDGMENTS

We thank Jim Lipe and Damon Nowlin of the Mississippi Department of Agriculture and Commerce for coordinating the collection of cow fecal samples, Jim Watson, Betty Roberts, Jerry Jones, and Keith Rusell of the Mississippi Board of Animal Health for the chicken fecal samples, Christopher Alonzo and Larry Castle of the Mississippi Department of Wildlife, Fisheries, and Parks for the deer fecal samples, and Sabrina Bryant for some of the human fecal samples. We also thank Joel Brumfield, Kimberly Peterson, Mary Phares, and Brian Robinson for their assistance in the laboratory.

Financial support that made the research possible was provided by the Mississippi Departments of Agriculture and Commerce and Environmental Quality, the U.S. Environmental Protection Agency Gulf of Mexico Program (EPA MS97449202), and the U.S. Coastal Impact Assistance Program (NOAA 17OZ2171 Project MS.R.17).

REFERENCES

1. **Albert, J. M., J. Munakata-Marr, L. Tenorio, and R. L. Siegrist.** 2003. Statistical evaluation of bacterial source tracking data obtained by rep-PCR DNA fingerprinting of *Escherichia coli*. Environ. Sci. Technol. **37:**4554–4560.
2. **Aslam, M., F. Nattress, G. Greer, C. Yost, C. Gill, and L. McMullen.** 2003. Origin of contamination and genetic diversity of *Escherichia coli* in beef cattle. Appl. Environ. Microbiol. **69:**2794–2799.
3. **Carson, C. A., B. L. Shear, M. R. Ellersieck, and A. Asfaw.** 2001. Identification of fecal *Escherichia coli* from humans and animals by ribotyping. Appl. Environ. Microbiol. **67:**1503–1507.
4. **Carson, C. A., B. L. Shear, M. R. Ellersieck, and J. D. Schnell.** 2003. Comparison of ribotyping and repetitive extragenic palindromic-PCR for identification of fecal *Escherichia coli* from humans and animals. Appl. Environ. Microbiol. **69:**1836–1839.
5. **Dombek, P. E., L. K. Johnson, S. T. Zimmerley, and M. J. Sadowsky.** 2000. Use of repetitive DNA sequences and the PCR to differentiate *Escherichia coli* isolates from human and animal sources. Appl. Environ. Microbiol. **66:**2572–2577.
6. **Gilson, E., J. M. Clement, D. Brutlag, and M. Hofnung.** 1984. A family of dispersed repetitive extragenic palindromic DNA sequences in *E. coli*. EMBO J. **3:**1417–1421.
7. **Hartel, P. G., J. D. Summer, and W. I. Segars.** 2003. Deer diet affects ribotype diversity of *Escherichia coli* for bacterial source tracking. Water Res. **37:**3262–3268.
8. **Lupski, J. R., and G. M. Weinstock.** 1992. Short, interspersed repetitive DNA sequences in prokaryotic genomes. J. Bacteriol. **174:**4525–4529.
9. **Martin, B., O. Humbert, M. Camara, E. Guenzi, J. Walker, T. Mitchell, P. Andrew, M. Prudhomme, G. Alloing, R. Hakenbeck, D. A. Morrison, G. J. Boulnois, and J. P. Claverys.** 1992. A highly conserved repeated DNA element located in the chromosome of *Streptococcus pneumoniae*. Nucleic Acids Res. **20:**3479–3483.
10. **McLellan, S. L., A. D. Daniels, A. K. Salmore.** 2003. Genetic characterization of *Escherichia coli* populations from host sources of fecal pollution by using DNA fingerprinting. Appl. Environ. Microbiol. **69:**2587–2594.
11. **Parveen, S., K. M. Portier, K. Robinson, L. Edmiston, and M. L. Tamplin.** 1999. Discriminant analysis of ribotype profiles of *Escherichia coli* for differentiating human and nonhuman sources of fecal pollution. Appl. Environ. Microbiol. **65:**3142–3147.
12. **Rademaker, J., and F. DeBruijn.** 1997. Characterization and classification of microbes by REP-PCR genomic fingerprinting and computer assisted pattern analysis. *In* DNA markers: protocols, applications, and overviews. Wiley-Liss, Inc., New York, N.Y.
13. **Seurinck, S., W. Verstraete, and S. D. Sicilinano.** 2003. Use of 16S-23S rRNA intergenic spacer region PCR and repetitive extragenic palindromic PCR analyses of *Escherichia coli* isolates to identify non-point fecal sources. Appl. Environ. Microbiol. **69:**4942–4950.
14. **Stern, M. J., G. F. Ames, N. H. Smith, E. C. Robinson, and C. F. Higgins.** 1984. Repetitive extragenic palindromic sequences: a major component of the bacterial genome. Cell **37:**1015–1026.
15. **U.S. Environmental Protection Agency.** 2000. Improved enumeration methods for the recreational water quality indicators: enterococci and *Escherichia coli.* U.S. Environmental Protection Agency, Office of Science and Technology, Washington, DC 20460 [Online.] http://www.epa.gov/nerlcwww/RecManv.pdf. Accessed 12 April 2004.
16. **van Belkum, A., M. Sluijter, R. D. Groot, H. Verbrugh, and P. W. M. Hermans.** 1996. Novel BOX repeat PCR assay for high-resolution typing of *Streptococcus pneumoniae* strains. J. Clin. Microbiol. **34:**1176–1179.
17. **Versalovic, J., T. Koeuth, and J. Lupski.** 1991. Distribution of repetitive DNA sequences in eubacteria and application to fingerprinting of bacterial genomes. Nucleic Acids Res. **19:**6823–6831.
18. **Whitlock, J. E., D. T. Jones, and V. J. Harwood.** 2002. Identification of the sources of fecal coliforms in an urban watershed using antibiotic resistance analysis. Water Res. **36:**4273–4282.