APPLIED AND ENVIRONMENTAL MICROBIOLOGY, July 1996, p. 2501–2507 Vol. 62, No. 7
0099-2240/96/$04.00+0

A Computer-Simulated Restriction Fragment Length Polymorphism Analysis of Bacterial Small-Subunit rRNA Genes: Efficacy of Selected Tetrameric Restriction Enzymes for Studies of Microbial Diversity in Nature†

CRAIG L. MOYER,[1]* JAMES M. TIEDJE,[2] FRED C. DOBBS,[3] AND DAVID M. KARL[1]

Department of Oceanography, School of Ocean and Earth Science and Technology, University of Hawaii, Honolulu, Hawaii 96822[1]; Center for Microbial Ecology, Departments of Crop and Soil Sciences and of Microbiology, Michigan State University, East Lansing, Michigan 48824[2]; and Department of Oceanography, Old Dominion University, Norfolk, Virginia 23529[3]

Received 2 October 1995/Accepted 24 April 1996

An assessment of 10 tetrameric restriction enzymes (TREs) was conducted by using a computer-simulated restriction fragment length polymorphism (RFLP) analysis for over 100 proximally and distally related bacterial small-subunit (SSU) rRNA gene sequences. Screening SSU rDNA clone libraries with TREs has become an effective strategy because of logistic simplicity, commercial availability, and economy. However, the rationale for selecting the type and number of TREs has not been systematically evaluated. Our objective was to identify the optimal combination of TREs for RFLP screening of cloned SSU rRNA genes from undefined bacterial clone libraries. After computer-simulated TRE digestion, the resultant fragments were categorized on the basis of the frequency of different restriction fragment size classes. Three groups of distribution patterns for the TREs were determined and further examined via graphical exploratory data analysis. The RFLP size-frequency distribution data for each group of enzymes were then used to infer phylogenetic relationships via the neighbor-joining method. The resulting bootstrap values and the correct placement of node bifurcations were used as additional criteria to evaluate the efficacy of the selected TREs. These RFLP data were compared with known phylogenetic relationships based on SSU rRNA sequence analysis as defined by the Ribosomal Database Project. A heuristic approach testing random combinations of TREs showed that three or more TRE combinations detected >99% of the operational taxonomic units (OTUs) within the model data set. OTUs that remained undetected after three TRE treatments had a median sequence similarity of 96.1%. Of the 10 restriction enzymes examined, *Hha*I, *Rsa*I, and *Bst*UI (group 3) were the most efficacious at detecting and differentiating bacterial SSU rRNA genes on the basis of their ability to correctly classify OTUs. Group 3 TREs are therefore recommended for screening in studies using bacterial SSU rRNA genes as descriptors of in situ microbial diversity.

A current theme in ecological research, including microbial ecology, is to understand the composition, patterns, and dynamics of the diversity present in the plethora of habitats found on Earth. The development of molecular biological techniques has made possible the more comprehensive, rapid, and precise characterization of bacterial taxa from discrete habitats (for examples, see references 6, 14, 18, and 20). A cornerstone method in this approach has been the use of PCR to amplify small-subunit (SSU) rRNA genes either from microbial communities or directly from culturable isolates. The stepwise strategy used to define in situ microbial diversity often includes the following: (i) efficiently isolating the native DNA of the natural community, (ii) PCR amplifying SSU rRNA gene sequences with primers designed to react uniformly with group-specific taxa, (iii) screening clones for genetic variability, and (iv) using these detected variations to estimate genetic diversity and to select clones that are most important for subsequent sequencing for the assessment of phylogenetic relationships. This study explores the optimization of the screening step, since this step is critical to an efficient, economical, and informative evaluation of the extensive microbial diversity that appears to be present in most natural habitats.

Recent studies of microbial diversity have screened composite SSU rDNA clone libraries generated from bacterial communities using tetrameric (i.e., having a 4-bp recognition site) restriction enzymes (TREs) to identify putative operational taxonomic units (OTUs [2, 5, 11]). This approach has improved efficiency, but the rationale for selecting a particular restriction enzyme or set of restriction enzymes has not been systematically evaluated. Normally, the choice of restriction enzymes has been based solely on practical matters such as logistic simplicity, commercial availability, and economy. We assessed the efficacy of 10 commonly used TREs with a diverse spectrum of bacterial taxa, using computer-simulated restriction fragment length polymorphisms (RFLPs). These taxa spanned the entire *Bacteria* domain (sensu Woese [13, 23–25]) and were used to test the hypothesis that the decision for TRE selection could be based on each enzyme combination's efficacy in distinguishing among known bacterial taxa. This hypothesis is based on a given TRE's site specificity (i.e., cleavage sites), which determines the frequency and distribution of RFLPs produced from SSU rDNAs. It is this extrapolated site speci-

* Corresponding author. Present address: Center for Microbial Ecology, Michigan State University, A540 Plant and Soil Sciences Building, East Lansing, MI 48824-1325. Phone: (517) 353-9021. Fax: (517) 353-2917. Electronic mail address: cmoyer@ribo.cme.msu.edu.

† Sea Grant publication UNIHI-SEAGRANT-JC-96-21 and contribution 4101 from the University of Hawaii, School of Ocean and Earth Science and Technology.

ficity that ultimately dictated the selection of the most efficacious combination of TREs.

Restriction fragments from the computer-simulated digestion by individual or combinations of TREs were categorized into a series of size classes (on the basis of numbers of base pairs), and each bacterial taxon was determined to have either a presence or an absence of a restriction fragment, or fragments, in each of these size class categories. These RFLP data were then analyzed by computer algorithms to reconstruct the phylogeny of bacterial taxa and were compared with the well-defined phylogenetic affiliations based on each taxon's SSU rDNA sequence data. These intercomparisons of the RFLP-based and the DNA sequence-based phylogenetic reconstructions were used to assess the efficacy of each of the 10 TREs in differentiating among many (>100) bacterial taxa of known phylogeny. TRE-RFLP analysis should not be used to replace SSU rDNA sequencing but instead to more efficiently and effectively describe microbial diversity prior to reaching the sequencing step.

In a recent computer-simulated analysis, the feasibility of typing bacteria with TRE pairs was demonstrated by using the hierarchical classification scheme (based on inferred phylogeny) of the Ribosomal Database Project (RDP). In addition, it was found that the confidence of identifying an unknown taxon is strongly correlated to the minimum pairwise relatedness (8). Another recent computer-simulated analysis of oligonucleotide hybridization potential across bacterial SSU rRNA sequences concluded that the majority of restriction site variations are the result of insertions and deletions rather than true restriction site polymorphisms (1). Therefore, the term RFLP used herein also encompasses these types of restriction fragment length variations.

MATERIALS AND METHODS

Selection of bacterial sequences in model data set. A total of 106 bacterial SSU rDNA sequences were selected from the RDP (10), with the following additions: *Moritella* sp. ANT-300 (10a), the representative hydrothermal vent clone isolates listed as Pele's Vents *Bacteria* operational taxonomic unit (PV*B* OTUs) ($n = 6$) (12), *Desulfitobacterium dehalogenans* (22), *Azoarcus denitrificans* (26), and *Desulfurella acetivorans* (15). This analysis focused on taxa spanning the domain *Bacteria*, including groups contained within each of the *Proteobacteria* subdivisions. Of the 106 bacterial sequences examined, 13 were contained in the α subdivision, 18 were contained in the β subdivision, 23 were contained in the γ subdivision, 15 were contained in the δ and ε subdivisions, 16 were contained in the gram-positive phylum, and 21 were distributed across several deeply rooted phyla of the *Bacteria*. Phylogenetic affiliations were as defined by the RDP through the use of maximum-likelihood analysis (10) and confirmed herein by neighbor-joining analysis for phylogenetic reconstruction. Sequences were aligned on the basis of primary and secondary structural considerations and were constructed by using the GDE multiple-sequence editor distributed through the RDP (10). SSU rDNA sequences from respective taxa were chosen on the basis of full sequence availability, multiple members being contained in a phylogenetically defined group (including proximally and distally related taxa), and the expected likelihood of the subset to represent an extreme level of community diversity.

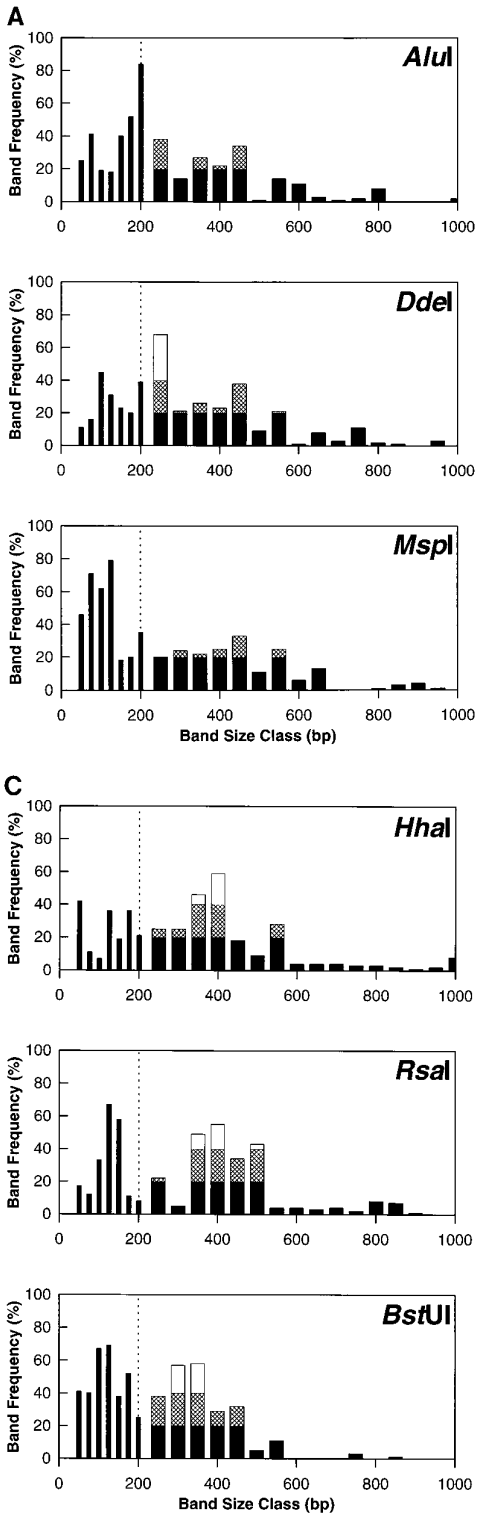
Computer generation of RFLPs. Computer-simulated RFLPs (with the various TRE combinations) were generated with the region of the SSU rDNA contained between the PCR priming sites described by Moyer et al. (11). These fragments corresponded to positions 49 through 1510 (*Escherichia coli* numbering system) for all the bacterial taxa examined. This region was chosen to give each bacterial taxon unambiguously alignable 5′ and 3′ SSU rDNA endpoints. Only sequence data which were contiguous between endpoints were used for computer-simulated RFLPs, thus disallowing the use of much of the current RDP database. The position of restriction sites for each TRE was identified from each bacterial taxon's SSU rDNA sequence by using the program Mapsort, contained in the Genetics Computer Group sequence analysis software package (version 7.3). The TREs used in the Mapsort program were as follows (sequence recognition sites indicated in parentheses): *Alu*I (AG'CT), *Bst*UI (CG'CG), *Dde*I (C'TNAG), *Hae*III (GG'CC), *Hha*I (GCG'C), *Hin*fI (G'ANTC), *Mbo*I ('GATC), *Msp*I (C'CGG), *Rsa*I (GT'AC), and *Taq*I (T'CGA). The properties of these 10 commonly used TREs have been previously well described (3). Known commercially available isoschizomers for *Bst*UI with *Acc*II and *Bsh*1236I; for *Hae*III with *Pal*I and *Bsu*RI; for *Hha*I with *Cfo*I and *Hin*PI; for *Mbo*I with *Nde*II, *Dpn*II, and *Sau*3AI; for *Msp*I with *Hpa*II and *Hap*II; for *Rsa*I with *Afa*I; and for *Taq*I with *Tth*HB8I also exist. The output from Mapsort was tabulated and placed into a spreadsheet program (Quattro Pro for Windows, version 6.0). The presence or absence of SSU rDNA restriction fragments contained within size categories was as follows: 50-bp increments from 1,400 to 250 bp (±25-bp precision), from 225 to 187.5 bp for the 200-bp category, 25-bp increments from 175 to 75 bp (± 12.5-bp precision), and from 62.5 to 45 bp for the 50-bp category. This process constituted the binning of the RFLP data. Because of the limitations of gel (i.e., agarose and acrylamide) electrophoresis, 45 bp was judged to be the lower limit of detection (this limitation also necessitated the binning of the RFLP data). The frequency and distribution of restriction sites across all bacterial taxa were calculated directly from these RFLP data. For each TRE, the generated distribution of size categories was represented graphically by an exploratory box plot analysis (21). The size range of ≥200 to ≤1,000 bp encompassed the majority of the diagnostic information. Such exploratory data analysis provided the basis for the subsequent grouping of TREs according to the restriction fragment size-frequency distributions generated for each TRE.

Phylogenetic analyses of RFLP and DNA sequence data. All programs used in this study were taken from the PHYLIP phylogenetic analysis software package (version 3.5). Individual programs were SEQBOOT for data set bootstrapping, DNADIST for the calculation of evolutionary distances from DNA sequence data, NEIGHBOR for the neighbor-joining method of phylogenetic reconstruction, and CONSENSE for the calculation of a consensus tree. The program RESTDIST was used to compute distance matrix data from all restriction fragment size class data. PHYLIP and RESTDIST were provided by Joe Felsenstein; RESTDIST is planned to be released in a future version of the PHYLIP package. For the purpose of comparative phylogenetic analysis, DNA sequence data were limited to the comparison of highly to moderately conserved nucleotide positions that were unambiguously alignable in all sequences, corresponding to residues 101 to 183, 220 to 451, 482 to 839, 847 to 998, 1037 to 1130, and 1143 to 1440 (*E. coli* numbering system). Distance matrix calculations of corrected evolutionary distances were generated by the programs DNADIST and RESTDIST, using the Kimura (9) two-parameter model of sequence evolution. The algorithm of neighbor joining (16), which uses a stepwise approach of phylogenetic reconstruction rather than a heuristic search method, was used to compute the phylogeny for 100 independent bootstrapped data sets, and a consensus tree was generated from each of these data sets for each final tree examined. Bootstrap values (4) were categorized as ≥50 (phylogenetically correct for all nodes from all trees observed) or as $50 > x \geq 20$ (correct in the majority of nodes observed, with only the correct nodes being represented). The phylogenetic trees, based on both sequence data and 10-TRE-treatment RFLP data, were used as benchmarks for comparisons with trees generated from discrete groups of lesser-TRE treatments. The categorized bootstrap values were used as criteria to assess the ability of each group of TREs to successfully differentiate among taxa contained in the domain *Bacteria* on the basis of their representative SSU rDNAs.

Heuristic test for OTU detection using random TRE combinations. Heuristic additions of TREs were used to ascertain the number of enzymes required to distinguish among the 106 taxa in the model data set. RFLP data from each of the 10 single TREs and 15 random combinations of two, three, and four TREs were determined (i.e., a total of 55 treatments). These RFLP data were transformed to distance data by using RESTDIST and examined by using UPGMA dendrograms (both programs from PHYLIP). The number of OTUs detected was compared with the actual number of OTUs ($n = 106$). SSU rDNA sequence similarity values among bacterial taxa were estimated (using DNADIST as described above) for all undifferentiated sequence pairs, and the minimum and median sequence similarity values were calculated.

RESULTS

The computer-simulated RFLPs for the SSU rDNAs from over 100 taxa from the *Bacteria* domain resulted in a restriction fragment size-frequency distribution for each of the 10 TREs tested (Fig. 1). These data were analyzed by comparing each restriction fragment band size class generated for each TRE with that size class overall band frequency across all taxa. Initial observation revealed three groups of distribution patterns for band size classes of ≥200 to ≤1,000 bp. This observation was validated through an exploratory box plot analysis (21), which also demonstrated by inspection at least three groups of distribution patterns (data not shown). The first distribution pattern (group 1) was generated by the TREs *Alu*I, *Dde*I, and *Msp*I (Fig. 1A). The group 1 TRE distributions showed band frequencies in the 20 to 40% range for a majority of the size classes from 250 to 550 bp. There was only a single occurrence of a size class frequency that reached more than 40% in this group, which was the 250-bp size class for *Dde*I. The second type of distribution pattern (group 2) resulted



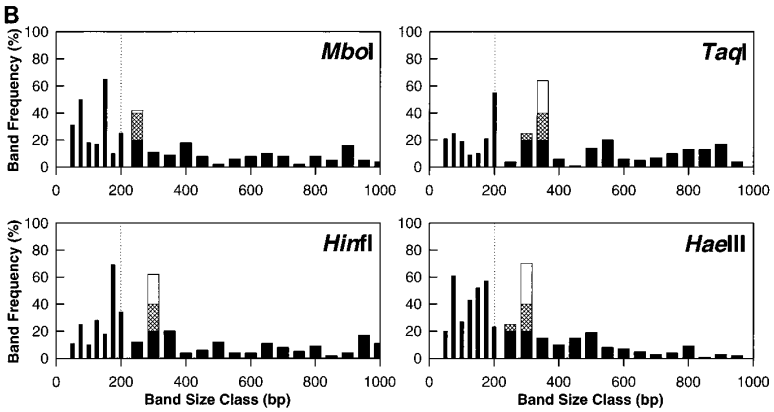


FIG. 1. Restriction fragment size-frequency distributions cumulative across all taxa for TREs *Alu*I, *Dde*I, and *Msp*I (group 1) (A), *Mbo*I, *Hin*fI, *Taq*I, and *Hae*III (group 2) (B), and *Hha*I, *Rsa*I, and *Bst*UI (group 3) (C). Vertical dotted lines show the threshold at 200 bp for restriction fragment band size classes. For restriction fragment band size classes from 250 to 1,000 bp, shading of histogram bars shows band frequencies at 20 to 40%.

from the TREs *Mbo*I, *Hin*fI, *Taq*I, and *Hae*III. The group 2 distributions showed a greater frequency of occurrence of the larger band size classes, ranging from 400 to 1,000 bp, but with none of these size classes present at a band frequency >20% (Fig. 1B). In addition, a single characteristic size class from 250 to 350 bp obtained a band frequency >25% (in all cases reaching >40%) for each of the group 2 TREs. The third type of distribution pattern (group 3) was produced by the TREs *Hha*I, *Rsa*I, and *Bst*UI. This pattern showed an increased level of band frequency for the size classes ranging from 250 to 550 bp (Fig. 1C). These band frequency levels were elevated to between 40 and 60% for at least two size classes and to >20% for five size classes within the 250-to-550-bp range for each of the group 3 TREs. The group 3 distribution patterns displayed more occurrences in the >200-bp band size classes than the remaining TREs examined (i.e., the highest levels of band frequency).

The average number of restriction sites per taxon for each of the 10 TREs examined along with the corresponding standard deviations was estimated across all taxa (Table 1). The results showed a range of an average of 3.71 to 5.65 sites per taxon, with an overall average across all TREs of 4.47. The mean numbers of restriction sites per taxon were 4.66, 4.10, and 4.77 for groups 1, 2, and 3, respectively.

A phylogenetic tree was generated with SSU rDNA sequence data (Fig. 2) from all the taxa examined and compared with phylogenetic trees generated from RFLP data derived from single- and multiple-TRE treatments (data not shown). In all cases, the neighbor-joining distance method was em-

TABLE 1. Summary of restriction sites per taxon for TREs

TRE	TRE group[a]	Mean no. of restriction sites per taxon (SD)
*Alu*I	1	4.54 (1.19)
*Dde*I	1	4.24 (0.98)
*Msp*I	1	5.21 (1.12)
*Mbo*I	2	3.97 (1.34)
*Hin*fI	2	3.91 (0.91)
*Taq*I	2	3.71 (0.94)
*Hae*III	2	4.80 (1.21)
*Hha*I	3	4.17 (1.09)
*Rsa*I	3	4.48 (0.93)
*Bst*UI	3	5.65 (1.12)
All TREs		4.47 (0.61)

[a] As determined by the restriction fragment size-frequency distributions across all taxa.

ployed. As anticipated, the sequence-based consensus tree generated in this study agreed with the accepted phylogenetic relationships for all representative taxa examined (on the basis of SSU rDNA phylogeny). The sequence-based phylogenetic tree demonstrated that SSU rDNA sequence data were sufficient to yield bootstrap values of ≥50 at the majority of nodes and of $50 > x \geq 20$ at all remaining nodes (Fig. 2). This result was then compared with the RFLP-based phylogenetic tree that was constructed with the entire data set from all 10 TREs (Table 2). This (best-case scenario) RFLP-based phylogenetic tree demonstrated the generation of 25 bootstrap values of ≥50 and an additional 23 bootstrap values of $50 > x \geq 20$. This RFLP-based phylogenetic tree was used as a metric to ascertain the success of the individual groups of restriction enzymes in predicting correct phylogenetic outcomes with RFLP data subsets from TRE groups 1, 2, and 3 (Table 2).

The data from any single TRE or combination of two TREs were insufficient to generate reproducible bootstrap values or accurate phylogenetic relationships (data not shown). When the RFLP data from three or more TRE combinations were examined, the results were sufficient to yield phylogenetic trees with reproducible bootstrap values at phylogenetically correct node bifurcations (Table 2). The group 2 tree was ultimately constructed without using the restriction fragment data from *Hae*III, as this TRE was the least phylogenetically informative of the group, having the highest average number of restriction sites per taxon (Table 1) and the greatest abundance of <200-bp band size classes (Fig. 2). The premise that *Hae*III was the least informative group 2 TRE was confirmed by the addition of *Hae*III restriction data and the sequential deletion of each of the other group 2 TREs' restriction data from the analysis, a process which in every case yielded lower numbers of descriptive bootstrap values (data not shown). Eliminating *Hae*III was also necessary so that the final analysis would compare normalized RFLP data sets, each combined from three-TRE treatments. The TREs contained in group 3 (*Hha*I, *Rsa*I, and *Bst*UI) demonstrated the largest number of descriptive bootstrap values of the three groups of TREs tested at both the ≥50 and $50 > x \geq 20$ bootstrap intervals (Table 2).

In a heuristic approach, multiple series of TRE combinations were used to determine the number of enzymes required to distinguish among OTUs. This approach tested the ability of single TREs ($n = 10$), as well as those of 15 random combinations of two, three, and four TRE treatments, to distinguish among the taxa contained in the model data set (Fig. 3). The median (i.e., midpoint unaffected by outliers) and minimum (i.e., maximum distance) sequence similarity values served as first-order metrics for estimating the level of relatedness among taxa undifferentiated after treatment with a given series of TRE combinations. This exercise demonstrated the ability of three-TRE and four-TRE combinations to detect >99% of OTUs in the model data set. The median sequence similarity for three-TRE and four-TRE combinations among undetected OTUs was 96.1 and 97.4%, respectively, whereas the minimum sequence similarity was 86.5 and 96.1%, respectively. These data also showed that for random combinations of TREs, the standard deviation about the mean decreased significantly until the third and fourth TREs were added (Fig. 3). A similarity value of 96.1% between OTUs after group 3 TRE treatment was also ascertained. These observations, together with the initial RFLP size-frequency distribution patterns, were used as the basis for the further examination of the RFLP-based data through phylogenetic analysis using the three TRE groups.

The percent successful phylogenetic affiliations from the RFLP-based trees are presented in Table 3. The entire 10-TRE data set was the most efficacious at describing phylogenetic relationships with respect to node bifurcations in addition to bootstrap values, as no TRE group yielded a greater percentage of successful affiliations (i.e., 91% across all taxa). Of the three TRE groups, group 3 was the most efficacious overall at predicting correct phylogenetic node bifurcations. TRE group 1 showed a greater percentage of successful affiliations for the α subdivision of the *Proteobacteria*, i.e., 100%. TRE group 2 showed a greater percentage of successful affiliations for both the gram-positive phylum and the β subdivision of the *Proteobacteria*, i.e., 81 and 94%, respectively. TRE group 3 was the most successful in describing the phylogenetic affiliations for the deeply rooted bacterial phyla and the δ and ε subdivisions of the *Proteobacteria* (67 and 87%, respectively), in addition to being the most successful overall for all the taxa examined (80 versus 67 and 73% for TRE groups 1 and 2, respectively).

DISCUSSION

The TRE-RFLP strategy provides an estimation of OTU community structure and diversity as well as presumptive OTU identification, which at the screening step can facilitate the prioritization of sequencing efforts. Restriction site variations have several appropriate characteristics which make them suitable for use in both the differentiation of SSU rDNAs and phylogenetic reconstructions. RFLP data, though providing less direct information on the evolution of DNA sequences, are easier to obtain and more economical than complete SSU rRNA gene sequences. This is still the case today for many laboratories worldwide, even with the current advances in DNA sequencing technology. Three characteristics which RFLP and sequencing methods have in common are that (i) character states can be scored unambiguously, (ii) a large number of characters can be scored for each taxon, and (iii) information on both the extent and the nature of divergence between two DNA sequences is provided (7). A critical characteristic that differs between RFLP and DNA sequence analyses relates to the asymmetry with respect to the evolution of restriction sites versus nucleotide positions. The 4-bp recognition site for a TRE will be inactivated by any 1 of 12 different nucleotide substitutions, whereas if the sequence differs from the recognition site at only a single site, then only one substitution can occur to produce that restriction site. As a consequence, convergent losses of a restriction site are more likely than convergent gains, and the ratio of convergent losses to convergent gains increases as taxa become more divergent (19). These characteristics must be considered when deciding which types of algorithms are best used to ascertain phylogenetic relationships with RFLP data (7).

This study has focused on specific criteria to evaluate which TREs accurately detect and differentiate among bacterial taxa (OTUs) on the basis of their representative SSU rDNAs isolated from naturally occurring microbial communities. The specific criteria were (i) analysis of restriction fragment band size-class frequency distributions and (ii) the phylogenetic reliability of three groups of TREs based on their bootstrap values with respect to OTU classification. Other criteria may also be useful in deciding among TREs (e.g., the ability to differentiate closely related SSU rDNA clones). This latter criterion might best be satisfied with the TREs *Bst*UI, *Msp*I, and *Hae*III, solely on the basis of their high average number of restriction sites per taxon (Table 1), as is represented by a high frequency of restriction fragments in the <200-bp band size classes (Fig. 1). Although, screening with high-frequency-cutting TREs increasingly results in the erroneous presumptive identification of an OTU. Using more than three TREs has the effect of increasing the selectivity to the point that the specific

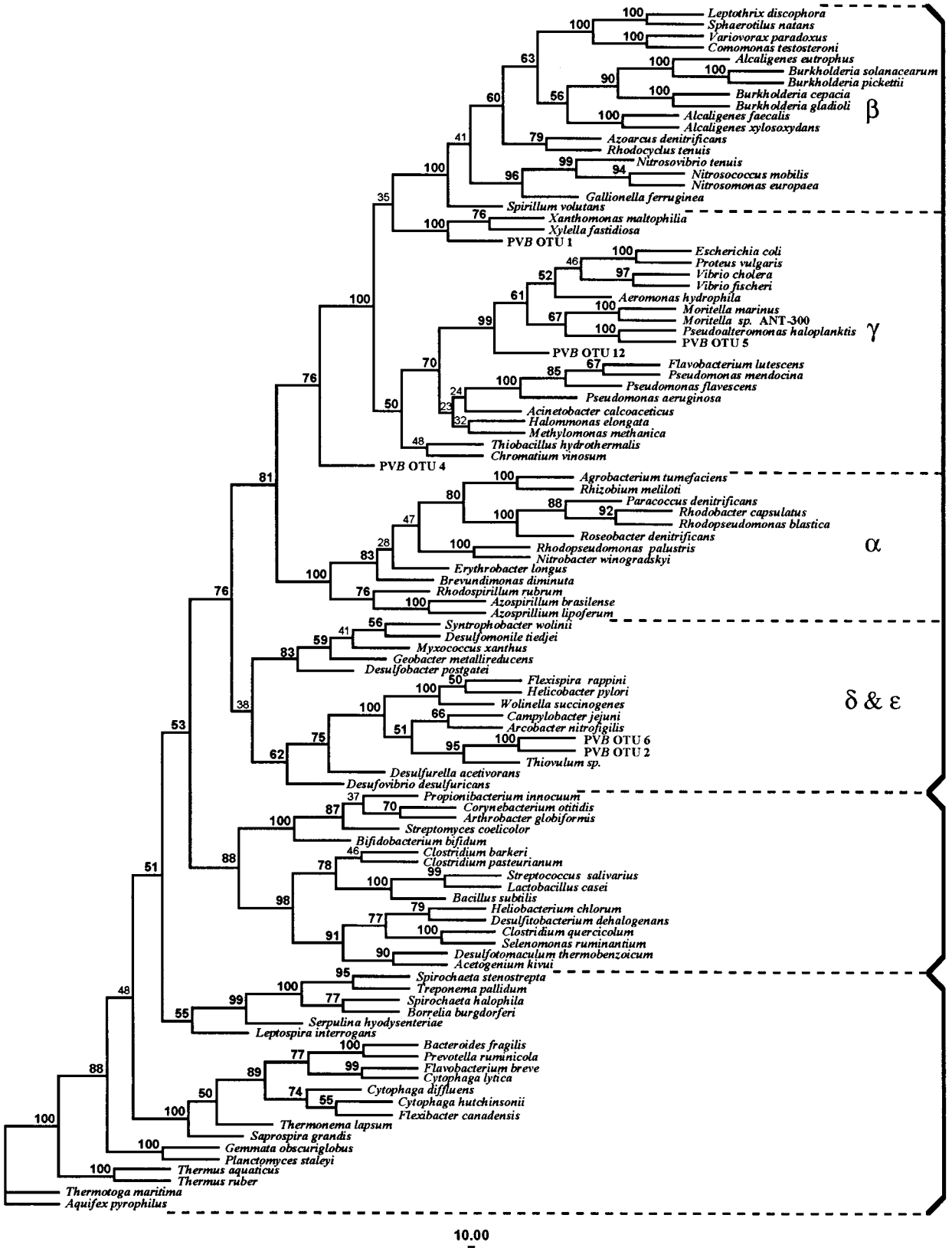


FIG. 2. Phylogenetic consensus tree based on SSU rDNA sequence data for all bacterial taxa examined by the neighbor-joining distance method. Bootstrap values of ≥50 are shown in boldface type. Bootstrap values of $50 > x \geq 20$ are also shown. The scale bar represents 10 fixed mutations per 100 nucleotide positions.

TABLE 2. Phylogenetically estimated bootstrap values for TRE groups

Bootstrap value	No. of occurrences in:			
	10-TRE treatment	TRE group 1	TRE group 2	TRE group 3
≥50	25	4	3	**7**[a]
$50 > x \geq 20$	23	12	16	**23**

[a] Boldface type indicates the highest number of occurrences among the three TRE groups.

type of TRE used is no longer as important a factor in the detection of OTUs. However, when screening with three or fewer TREs is performed, the specific TREs selected are of importance with regard to maximizing the differentiation of SSU rDNA sequences (Fig. 3) as well as optimizing the presumptive identification (Table 3).

The TREs contained in group 3 (*Hha*I, *Rsa*I, and *Bst*UI) were superior for the detection and differentiation of bacterial taxa based on criteria of RFLP size-frequency distribution patterns. This result is most likely because group 3 TREs yielded the highest frequency (up to 60% across all taxa) of restriction fragment band size classes in the 250-to-550-bp range (Fig. 1C). Of the group 3 TREs, *Hha*I and *Rsa*I are superior choices, because *Bst*UI has a significantly greater average number of restriction sites per taxon (5.65 versus 4.17 and 4.48 for *Hha*I and *Rsa*I, respectively [Table 1]) and because *Bst*UI has much higher frequency values for the band size classes of <200 bp (Fig. 1C). Group 3 TREs had the greatest group average number of restriction sites per taxon,

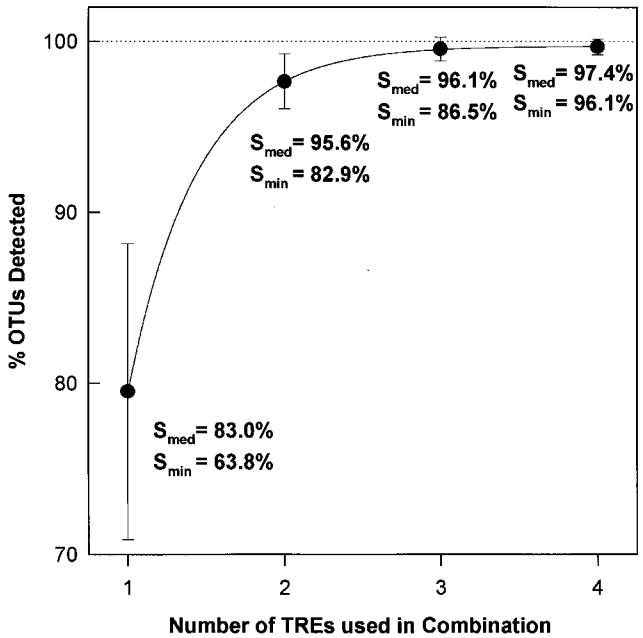


FIG. 3. Number of TREs used singly or in combination (i.e., 10 single-TRE treatments plus 15 random combinations of two-, three-, and four-TRE treatments) to detect OTUs from the model SSU rDNA data set. Error bars indicate standard deviations about the mean. Sample sizes for OTUs that remained undetected were 220, 37, 7, and 5, corresponding to the one-, two-, three-, and four-TRE treatments, respectively. S_{med} and S_{min} indicate the median and minimum sequence similarities, respectively, among OTUs that remained undetected for each series of TRE treatments. An exponential rise to a theoretical maximum (i.e., 100% of OTUs detected) is shown by the line that best fits these data.

TABLE 3. Percent successful phylogenetic affiliations across TRE groups as determined by phylogenetic analysis of RFLP data

Phylogenetic group	% Successful affiliations			
	10-TRE	TRE group 1	TRE group 2	TRE group 3
Deeply rooted phyla	76	52	52	**67**[a]
Gram-positive phylum	81	50	**81**	75
Proteobacteria				
α subdivision	100	**100**	77	92
β subdivision	94	78	**94**	83
γ subdivision	96	74	**83**	**83**
δ and ε subdivisions	100	53	47	**87**
All taxa	91	67	73	**80**

[a] Boldface type indicates the highest percentage value(s) among the three TRE groups.

i.e., 4.77, which when combined with supporting RFLP size-frequency distribution patterns and phylogenetic reliability estimates indicates a possible synergistic effect among these TREs with respect to bacterial OTU detection and presumptive identification.

The criterion of phylogenetic reliability for group 3 TREs is based on their bootstrap values and phylogenetic affiliations (Tables 2 and 3, respectively). However, under more specific circumstances other TRE combinations might be more useful, such as the group 1 TREs used with clone libraries known to contain members of the α subdivision of the *Proteobacteria* (Table 3). As described previously, the phylogenetic bootstrap values were not reproducible until at least three TREs were used. A similar result allowing for increased levels of OTU detection (i.e., >99%) was demonstrated for random combinations of at least three TREs (Fig. 3). These were independent results, since the first was based on the bootstrapping of the data after simulated RFLP treatments and the second was based on the raw or unbootstrapped data after treatment with random combinations of TREs.

The level of SSU rDNA variation allowed by TRE-RFLP screening was investigated through the heuristic additions of TRE treatments. Three or more TREs were required to sufficiently differentiate among OTUs (Fig. 3). Previously, this estimate was made through the simultaneous use of two TREs which were found sufficient to distinguish between closely related SSU rDNAs (e.g., PV*B* OTUs 1 and 11, which varied by a single *Rsa*I site and contained seven variable nucleotide positions across the entire SSU rRNA gene [11, 12]). The ability of single-TREs and multiple-TRE combinations to detect OTUs from a bacterial SSU rDNA clone library with known diversity was described by an exponential increase approaching a theoretical maximum of 100% (Fig. 3). Additionally, this procedure allowed for the examination of the sequence similarity values among the undetected OTUs, which demonstrated an increased level of relatedness as more TREs were used in combination. Median sequence similarity was used as a midpoint value unaffected by outliers to describe the point of central tendency, whereas minimum sequence similarity was used as a conservative estimate of the maximum divergence among undetected OTUs. Group 3 TREs showed minimum similarity of 96.1%, which was the value found across all random four-TRE treatments (Fig. 3). For comparison, organisms with sequence similarities of ≤97.0% belong to different species as they are not known to have DNA-DNA reassociation values of <70%, a key criterion used to define species (17). This comparison supports our primary conclusion that the most efficacious results for screening clone libraries of SSU

rDNAs from unknown bacterial taxa would be through the use of the group 3 TREs *Hha*I, *Rsa*I, and *Bst*UI, since these enzymes have been shown to be superior for accurately differentiating among many (>100) diverse bacterial taxa.

Enhanced resolution of restricted DNA fragments has the potential of increasing the effective separation and the accuracy of presumptive identification of closely related OTUs. Metaphore (FMC Bioproducts, Rockland, Maine) agarose, which allows resolution to ~50 bp, is currently used in our protocol. We are currently experimenting with increasing the resolution limit through the use of a high-pressure capillary electrophoresis system which has a lower detection limit (~5 bp) and the additional advantage of high throughput. The result will directly affect the binning of the RFLP data, because more precise estimates of DNA restriction fragment sizes can be obtained. This methodological improvement will allow an increased rate of OTU detection (in addition to lowering the standard deviation and raising the minimum and median sequence similarity values) than is currently achieved with agarose gels (Fig. 3). High-pressure capillary electrophoresis technology may allow initial screening of bacterial SSU rDNA libraries with as few as two TREs with the same level of OTU separation and presumptive identification as is achieved by three or four TREs with agarose gels. At the very least, high-pressure capillary electrophoresis will increase our level of accuracy with the group 3 TREs.

In a broader sense, this simulation has demonstrated that the choice of TREs can enhance the information gained at the bacterial SSU rDNA clone screening step based on site specificity. The most effective way to determine which combination of TREs optimally identifies among in situ bacterial SSU rDNAs would be to examine every taxon with each TRE. This approach is impossible as the current databases do not include sequence information from all extant taxa. In addition, the sequence data from known bacterial SSU rRNA genes are largely incomplete as many investigators focus their efforts on partial sequences. Finally, we do not know the full extent of the microbial diversity that exists in nature, as every new habitat examined yields novel phylotypes. Therefore, a first-order approximation of diversity is necessary. This goal was achieved through an improved TRE-RFLP screening strategy. The improved screening of bacterial SSU rDNA clones will enhance and facilitate the exploration of microbial diversity from various habitats.

ACKNOWLEDGMENTS

This project was funded by a grant from the National Oceanic and Atmospheric Administration, project R/OM-8, which is sponsored by the University of Hawaii Sea Grant College Program (SOEST), under Institutional Grant NA36RG0507 from the National Oceanic and Atmospheric Administration Office of Sea Grants; by the National Oceanic and Atmospheric Administration-National Undersea Research Program, Department of Commerce (to F. C. D. and D. M. K.); and by a Research and Training Revolving Fund Award from the University of Hawaii Research Council (to D. M. K.). This project was also funded by NSF grant BIR-9120006 (to J. M. T.).

We also thank Arturo Massol Deyá and Duane Meeter for their thoughtful discussions and insights regarding molecular phylogeny and exploratory data analysis, respectively.

REFERENCES

1. **Brunk, C. F., E. Avaniss-Aghajani, and C. A. Brunk.** 1996. A computer analysis of primer and probe hybridization potential with bacterial small-subunit rRNA sequences. Appl. Environ. Microbiol. **62:**872–879.
2. **DeLong, E. F., D. G. Franks, and A. L. Alldredge.** 1993. Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. Limnol. Oceanogr. **38:**924–934.
3. **Dowling, T. E., C. Moritz, and J. D. Palmer.** 1990. Nucleic acids II: restriction site analysis, p. 250–317. *In* D. M. Hillis and C. Moritz (ed.), Molecular systematics. Sinauer Assoc., Inc., Sunderland, Mass.
4. **Felsenstein, J.** 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:**783–791.
5. **Haddad, A., F. Camacho, P. Durand, and S. C. Cary.** 1995. Phylogenetic characterization of the epibiotic bacteria associated with the hydrothermal vent polychaete *Alvinella pompejana*. Appl. Environ. Microbiol. **61:**1679–1687.
6. **Holben, W. E., and J. M. Tiedje.** 1988. Applications of nucleic acid hybridization in microbial ecology. Ecology **69:**561–568.
7. **Holsinger, K. E., and R. K. Jansen.** 1993. Phylogenetic analysis of restriction site data. Methods Enzymol. **224:**439–455.
8. **Kim, J., J. R. Cole, E. Torng, and S. Pramanik.** Inferring relatedness of a macromolecule to a sequence database without sequencing. *In* Intelligent systems for molecular biology '96. Proceedings of the fourth international conference on computational biology, in press. AAAI & MIT Press, Cambridge, Mass.
9. **Kimura, M.** 1980. A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16:**111–120.
10. **Maidak, B. L., N. Larsen, M. J. McCaughey, R. Overbeek, G. J. Olsen, K. Fogel, J. Blandy, and C. R. Woese.** 1994. The Ribosomal Database Project. Nucleic Acids Res. **22:**3485–3487.

10a. **Moyer, C. L.** Unpublished data.

11. **Moyer, C. L., F. C. Dobbs, and D. M. Karl.** 1994. Estimation of diversity and community structure through restriction fragment length polymorphism distribution analysis of bacterial 16S rRNA genes from a microbial mat at an active, hydrothermal vent system, Loihi Seamount, Hawaii. Appl. Environ. Microbiol. **60:**871–879.
12. **Moyer, C. L., F. C. Dobbs, and D. M. Karl.** 1995. Phylogenetic diversity of the bacterial community from a microbial mat at an active, hydrothermal vent system, Loihi Seamount, Hawaii. Appl. Environ. Microbiol. **61:**1555–1562.
13. **Olsen, G. J., C. R. Woese, and R. Overbeek.** 1994. The winds of (evolutionary) change: breathing new life into microbiology. J. Bacteriol. **176:**1–6.
14. **Pace, N. R., D. A. Stahl, D. J. Lane, and G. J. Olsen.** 1986. The analysis of natural microbial populations by ribosomal RNA sequences. Adv. Microb. Ecol. **9:**1–55.
15. **Rainey, F. A., R. Toalster, and E. Stackebrandt.** 1993. *Desulfurella acetivorans*, a thermophilic, acetate-oxidizing and sulfur-reducing organism, represents a distinct lineage within the *Proteobacteria*. Syst. Appl. Microbiol. **16:** 373–379.
16. **Saitou, N., and M. Nei.** 1987. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:**406–425.
17. **Stackebrandt, E., and B. M. Goebel.** 1994. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. Int. J. Syst. Bacteriol. **44:**846–849.
18. **Stahl, D. A., B. Flesher, H. R. Mansfield, and L. Montgomery.** 1988. Use of phylogenetically based hybridization probes for studies of ruminal microbial ecology. Appl. Environ. Microbiol. **54:**1079–1084.
19. **Templeton, A. R.** 1983. Convergent evolution and non-parametric inferences from restriction fragment and sequence data, p. 151–179. *In* B. Weir (ed.), Statistical analysis of DNA sequence data. Marcel Dekker, New York.
20. **Tiedje, J. M.** 1995. Approaches to the comprehensive evaluation of prokaryotic diversity of a habitat, p. 73–87. *In* D. Allsopp, D. L. Hawksworth, and R. R. Colwell (ed.), Microbial diversity and ecosystem function. CAB International North America, Tucson, Ariz.
21. **Tukey, J. W.** 1977. Exploratory data analysis. Addison-Wesley, Reading, Mass.
22. **Utkin, I., C. R. Woese, and J. Wiegel.** 1994. Isolation and characterization of *Desulfitobacterium dehalogenans* gen. nov., sp. nov., an anaerobic bacterium which reductively dechlorinates chlorophenolic compounds. Int. J. Syst. Bacteriol. **44:**612–619.
23. **Winker, S., and C. R. Woese.** 1991. A definition of the domains *Archaea, Bacteria* and *Eucarya* in terms of small subunit ribosomal RNA characteristics. Syst. Appl. Microbiol. **14:**305–310.
24. **Woese, C. R.** 1994. There must be a prokaryote somewhere: microbiology's search for itself. Microbiol. Rev. **58:**1–9.
25. **Woese, C. R., O. Kandler, and M. L. Wheelis.** 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA **87:**4576–4579.
26. **Zhou, J., M. R. Fries, J. C. Chee-Sanford, and J. M. Tiedje.** 1995. Phylogenetic analyses of a new group of denitrifiers capable of anaerobic growth on toluene and description of *Azoarcus tolulyticus* sp. nov. Int. J. Syst. Bacteriol. **45:**500–506.