Discrimination of Shifts in a Soil Microbial Community Associated with TNT-Contamination Using a Functional ANOVA of 16S rRNA Hybridized to Oligonucleotide Microarrays

LAURENT EYERS,1 JAMES C. SMOOT,1 LAURA M. SMOOT,1 CÉLINE BUGLI,5 HIDETOSHI URAKAWA,8 ZACK MCMURRY,1 SLIL SIRIPONG,2 SAID EL FANTROUSSI,1 PHILIPPE LAMBERT,9 SPIROS N. AGATHOS,2 AND DAVID A. STAHL*1

A functional ANOVA analysis of the thermal dissociation of RNA hybridized to DNA microarrays was used to improve discrimination between two soil microbial communities. Following hybridization of in vitro transcribed 16S rRNA derived from uncontaminated and 2,4,6-trinitrotoluene contaminated soils to an oligonucleotide microarray containing group- and species-specific perfect match (PM) probes and mismatch (MM) variants, thermal dissociation was used to analyze the nucleic acid bound to each PM-MM probe set. Functional ANOVA of the dissociation curves generally discriminated PM-MM probe sets when $T_D$ values (temperature at 50% probe-target dissociation) could not. Maximum discrimination for many PM and MM probes often occurred at temperatures greater than the $T_D$. Comparison of signal intensities measured prior to dissociation analysis from hybridizations of the two soil samples revealed significant differences in domain-, group-, and species-specific probes. Functional ANOVA showed significantly different dissociation curves for 11 PM probes when hybridizations from the two soil samples were compared, even though initial signal intensities for 3 of the 11 did not vary.

Introduction

Microarrays containing DNA oligonucleotide probes that target rRNA molecules were developed to detect single species in environmental matrices (1) and to evaluate the bacterial composition of complex environmental systems (2–6). Ideally, 16S rRNA microarrays should provide a direct characterization of the bacteria present in the system and information on their relative abundances. When nucleic acids isolated from pure cultures are examined with DNA microarrays, unambiguous hybridization patterns are generally observed, even between closely related bacterial species (7). However, when microarrays are used to evaluate environmental samples, the relative abundance of target and closely related nontarget molecules in the environment is unknown. Therefore, it is difficult to determine the extent of cross-hybridization events using signal intensity values at a single wash temperature (6).

Previous studies using thermal dissociation analyses (dissociation curves) were successful at discriminating hybridization events of perfect match (PM) and single mismatch (MM) duplexes. El Fantroussi et al. (6) visually evaluated (i.e., without statistical analysis) the extent of the discrimination between dissociation curves of PM and MM duplexes. Liu et al. (8) used $T_D$ values (temperature at which 50% of the probe-target duplexes remain intact) and obtained clear discrimination between PM and MM duplexes. Urakawa et al. (9) found that $T_D$ values did not effectively discriminate PM from MM duplexes. Therefore, they used a discrimination index (DI) to determine an optimum discrimination temperature defined as the temperature corresponding to the maximum product of difference and ratio of the signal intensities of PM and MM duplexes. However, these differences in signal intensities were not demonstrated to be statistically significant. In order to establish an appropriate statistical framework for analysis, Bugli et al. (manuscript submitted) developed a functional ANOVA calculator to analyze each dissociation curve and identify statistically different dissociation curves.

Changes in microbial communities of polluted soil samples are the subject of numerous investigations. 2,4,6-Trinitrotoluene (TNT) is frequently found as a soil pollutant at TNT-production facilities and at sites used for ammunition testing and destruction. Despite its recalcitrance to microbial degradation and its toxicity for many microorganisms, few reports describe the effect of TNT on soil microbial populations, and there are currently no reports providing extensive identification of bacterial populations in TNT-contaminated soils. Fuller and Manning (10) showed that Gram-positive bacteria are more sensitive to TNT than Gram-negative bacteria using heterotrophic plate counts and phospholipid fatty acid analyses. However, the identity of the bacteria was not resolved using these techniques. Eyers et al. (11) showed that the presence of TNT caused a dramatic alteration in soil populations using denaturing gradient gel electrophoresis of 16S rRNA gene amplicons, but sequencing of clones from the 16S rRNA gene amplicons was not performed in that study.

The general objective of this study was to further document the utility of DNA microarrays to resolve shifts in microbial community structure associated with TNT contamination. Specific objectives were to (i) compare hybridizations of in vitro transcribed 16S rRNA derived from uncontaminated soil and TNT-contaminated soil and (ii) assess the effectiveness of functional ANOVA to discriminate PM and MM duplexes compared to traditional approaches based on signal intensity measured prior to dissociation.
analysis and $T_9$ values. Functional ANOVA was used to compare thermal dissociation curves of PM probes hybridized with native RNA of *Pseudomonas putida* and in vitro transcribed 16S rRNA of *P. putida*, uncontaminated, and TNT-contaminated soil. To evaluate the specificity of each PM probe hybridized with RNA from soil samples, two single MM probes were designed for each PM probe. Dissociation curves generated from hybridizations of RNA from uncontaminated and TNT-contaminated soil samples were used to compare PM and MM probe performance and microbial composition of the two soil samples. The functional ANOVA was shown to resolve differences between most PM and MM probe sets, providing much greater resolution of microbial communities than could be achieved using only differences in signal intensity or $T_9$.

### Materials and Methods

**Sample Collection and Processing of Nucleic Acids.** *Pseudomonas putida* sp. JLR 11 was used because of its known association with TNT degradation (12) and because it is a suitable reference target to validate probes Univ907, Univ1390, Eub338, Eub927, and Gam42a. The strain was provided by Dr. J. L. Ramos (CSIC, Estacion Experimental del Zaidin, Spain). Samples were collected from TNT-contaminated soil (23.0 ± 2.2 g TNT/kg soil) and uncontaminated soil (i.e., TNT was not detected, detection limit of 1 mg TNT/kg soil) at a site in Bourges, France that was used for TNT destruction during the past 20 years. The uncontaminated soil, which was sampled a few meters from the contaminated soil, was unpolluted due to the presence of a protecting wall at the destruction field. Both samples were collected from the top 5-cm layer of soil, homogenized, and stored at 4 °C for 1 month until DNA extraction.

RNA used for microarray hybridizations was native RNA (i.e., total RNA—including rRNA, tRNA, and mRNA—harvested from cellular biomass at the time of lysis) isolated from *P. putida*, and non-native RNA was derived from in vitro transcription of the *P. putida* 16S rRNA gene or 16S rDNA gene sequences amplified from soil DNA. Native RNA of *P. putida* was isolated with the Qiagen RNeasy mini kit (Valencia, CA) following the manufacturer’s instructions. Total DNA from *P. putida* was extracted with lysozyme and sodium dodecylsulfate lysis followed by phenol extraction (13). DNA was extracted from soil samples with the UltraClean Soil DNA kit from Mo Bio (Carlsbad, CA) following the manufacturer’s instructions. DNA isolated from *P. putida* and soil samples was used as a template for PCR amplification of the 16S rDNA gene under conditions described in the Supporting Information. PCR products were purified with the Qiagen PCR purification kit (Qiagen, Inc.) following the manufacturer’s instructions. T7 RNA polymerase (Invitrogen, Inc., Carlsbad, CA, Cat. No. 18033-019) was used for in vitro transcription to produce 16S rRNA according to the manufacturer’s protocol. RNA samples were purified with the Qiagen RNeasy mini kit.

**Fragmentation and Labeling of RNA.** RNA was simultaneously fragmented and end-labeled following the protocol of Bavykin et al. (14) with modifications given in the Supporting Information.

**Microarray Fabrication and Hybridization.** CodeLink and gel-pad microarray formats were used in this study. The two formats are described in the Supporting Information. The gel-pad microarrays contained 29 different PM oligonucleotide probes (13–25 nucleotides in length) that were complementary to universal, domain-, group-, and species-specific rRNA sequences (Table S-1 in the Supporting Information). Fifteen of these 29 probes were spotted on the CodeLink microarrays (Table S-1). In addition to PM probes, single MM probes were printed on the gel-pad microarray. The design of PM and single MM probes was based on several criteria given in the Supporting Information. Hybridizations were conducted in triplicate for each sample, using 5 µg of RNA that was fragmented, labeled, hybridized, and washed as previously described (6), except that the hybridization time was reduced to 1 h.

**Analysis of Images and Thermal Dissociation Curves.** Fluorescence signal intensities were recorded and dissociation curves were generated as described in the Supporting Information. Data of hybridizations to the gel-pad and CodeLink microarrays are available at the Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo) under respectively the accession numbers GSE3499 and GSE3525. Data from each hybridization were analyzed with the functional ANOVA calculator (available at http://stahl.ce.washington.edu) developed by Bugli et al. (manuscript submitted) with Matlab 7.0. Functional ANOVA is an adaptation of linear modeling to analyze curve data (15). Functional ANOVA uses penalized spline regressions to smooth data points of individual dissociation curves and obtain continuous curves. The functional ANOVA calculator was used to (i) obtain continuous curves, (ii) identify and reject outlier curves, (iii) normalize individual curves by setting the lowest intensity in a profile at a value of 0 and the highest intensity at 1, (iv) retrieve the $T_9$ value of individual curves, (v) average curves for each probe, (vi) calculate an interval of confidence for each average curve and obtain the p-values associated to the comparison of these curves, (vii) calculate, when the confidence bands of two average curves do not overlap, the maximum difference ($$\text{MAX}_{\text{DCSD}}$$, where DCSD stands for “dissociation curves significantly different”) in normalized signal intensities between the lower limit of the interval of confidence of the upper dissociation curve and the upper limit of the lower dissociation curve, (viii) calculate the temperature corresponding to the $$\text{MAX}_{\text{DCSD}}$$, and (ix) identify when signal intensities and $T_9$ values between probes were significantly different (unpaired tests described in the Supporting Information).

Calculation of confidence bands was performed with an $\alpha$ value of 0.05. Outlier dissociation curves were not included in the functional ANOVA model, calculation and comparison of $T_9$ values. Outlier curves were detected using leave-one-out error. The “leave-one-out error” is the error of the computed functional ANOVA model determined by successive single exclusion of each of the observed curves. The error was calculated by summing for each point the square difference between observed and estimated points determined with the model. A boxplot of the leave-one-out errors for each of the curves allowed the detection of extreme values of the errors and consequently outlier curves. The rejection of outliers often resulted in the calculation of different confidence intervals and $T_9$ values than without rejection.

**Results and Discussion**

**Discrimination of PM and MM Probes with Signal Intensity and $T_9$s.** Our ultimate analytical goal is to develop a combined experimental and statistical framework to determine sequence identity of nucleic acid hybridized to each array element (PM vs MM). For example, a common design feature of DNA microarrays has been the incorporation of mismatch variants of probes complementary to specific target sequences, setting a threshold of hybridization between match and mismatch variants as a basis to evaluate hybridization to the intended target sequence (16). However, given the complexity and undefined composition of natural microbial communities, we anticipate that such an assignment will not be possible using data collected at a single stringency of hybridization. For example, a highly abundant but closely related nontarget sequence could skew PM/MM hybridization values. Alternatively, unknown microbial species present
in the sample may have sequences complementary to a subset of the MM probes. Therefore, identifying hybridization events that deviate from expectations is of utmost importance for correct assignment. In this study, we used a new statistical package (functional ANOVA calculator developed by Bugli et al., manuscript submitted) to analyze data collected using nonequilibrium dissociation of rRNA hybridized to PM and MM probes (Table S-1) on oligonucleotide microarrays.

Initial analysis of collected dissociation curves revealed some curves that deviated from the general sigmoid shape of the majority of the curves. The shapes of these “outlier” curves differed from those reported by Pozhitkov et al. (17). Three classes of outliers were detected and are described elsewhere (Bugli et al., manuscript submitted). One hundred fifty-one individual curves from a total of 1548 were statistically identified as outliers and excluded from further analysis (Bugli et al., manuscript submitted). Ongoing studies have identified several contributing factors (movement of the microarray during data collection, sporadic occurrence of fluorescent particles, and the formation of bubbles within the hybridization chamber at high temperature) that are being alleviated by improvements in the experimental protocol and image analysis software (e.g., ref 18).

Significant differences in initial signal intensities between PM and their respective c-MM and/or g-MM variants were achieved for 24 PM probes using nucleic acid derived from uncontaminated soil and for 23 PM probes with contaminated soil (Table S-2). Thus, initial signal intensity data provided some discriminatory information about the two soil communities. However, dissociation analysis provided the basis for a higher resolution comparison of the two communities. For instance, comparison of dissociation curves from in vitro transcribed 16S rRNA of \textit{P. putida} with in vitro transcribed 16S rRNA of the uncontaminated and TNT-contaminated soil sample showed no differences in the shape of the curves for the \textit{Bacteria} probe Eub338 (Figure 1A). This and previous studies (6, 8) demonstrate that a robust and reproducible reference curve can be established for each probe.

Thermal dissociation generates a large data set for each probe, and the analysis of several hundred microarray elements requires computational processing to summarize the data. One method of data reduction is the determination of $T_d$ for each probe. Although the strategy used in designing single MM probes was to have a 4–6°C difference in $T_d$ between PM and MM probes, discrimination was not achieved for all probes using RNA from the complex soil communities, and the extent of discrimination differed between samples. Significant differences between PM probes and their respective c-MM and/or g-MM variants were achieved for 21 PM probes using nucleic acid derived from uncontaminated soil and for only 16 PM probes with contaminated soil (Table S-3). In addition, $T_d$s of PM probes

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure1.png}
\caption{Dissociation curves analysis. Average normalized dissociation curves of probe Eub338 hybridized with in vitro transcribed 16S rRNA of \textit{P. putida} (black triangles, $n=18$), uncontaminated (blue squares, $n=18$), and TNT-contaminated soil sample (pink circles, $n=18$) to CodeLink microarrays (A). Average normalized dissociation curves of probe Eub338 (blue lozenges, $n=12$), Eub338-g11c (green lozenges, $n=12$), and Eub338-t7g (red lozenges, $n=12$) hybridized with in vitro transcribed 16S rRNA of uncontaminated soil sample to gel-pad microarrays (B). Error bars represent the standard deviation. Average normalized dissociation curves with functional ANOVA analysis of probe Eub338 hybridized with in vitro transcribed 16S rRNA of \textit{P. putida} (black line), uncontaminated (blue line), and TNT-contaminated soil sample (pink line) to CodeLink microarrays (C). Average normalized dissociation curves with functional ANOVA analysis of probe Eub338 (blue line), Eub338-g11c (green line), and Eub338-t7g (red line) hybridized with in vitro transcribed 16S rRNA of uncontaminated soil sample to gel-pad microarrays (D). The number of individual dissociation curves of C and D are identical to A and B, respectively. Confidence bands are indicated by noncontinuous lines. The big and small arrows represent, respectively, the $\text{MAX}_{\text{CSD}}$ of Eub338/Eub338-g11c and Eub338/Eub338-t7g.}
\end{figure}
The functional ANOVA calculator revealed significant differences for practically all the PM-MM probes sets when hybridized with uncontaminated and contaminated soil samples, with MAXDSCD values ranging from 0.008 to 0.230 (Table 1). Twenty-four PM probes were discriminated from their respective c-MM and/or g-MM probes using nucleic acid probes.

**Discrimination of PM and MM Probes with Functional ANOVA Analysis.** The functional ANOVA was more efficient than the $T_d$ approach in determining when dissociation curves significantly differed. This is due to the fact that functional ANOVA compares signal intensities of dissociation curves at each temperature, whereas $T_d$s only take into account signal intensities at 50% dissociation. Examples of the curves produced by the functional ANOVA calculator are shown in Figure 1C,D. The comparisons of the dissociation curves produced from the hybridizations of different samples to probe Eub338 showed no significant difference with CodeLink (Figure 1C) and gel-pad microarrays (Table 1). The absence of significant differences was also observed with these samples hybridized to probes Univ907, Univ1390, and Eub927 with gel-pad (Table 1) and CodeLink microarrays (results not shown). When the PM and MM Eub338 probes were compared on gel-pad microarrays, significant differences in the dissociation curves were detected between probes Eub338 and Eub338-g11c and Eub338-t7g hybridized with the uncontaminated soil sample (Figure 1D) as shown by nonoverlapping confidence bands. In addition to $p$-values (Table 1), MAXDCSD values were calculated to provide an indication of the extent of discrimination between the 95% confidence intervals (e.g., a MAXDCSD Value of 0.10 indicates a 10% maximum difference in normalized dissociation curves as they start at 1 and finish at 0). MAXDCSD values were 0.207 and 0.124 for the Eub338/Eub338-g11c and Eub338-t7g probe sets, respectively (Table 1). Although not demonstrated, the extent of discrimination may be related to the number of mismatches and/or the presence of multiple targets contributing to the shape of the curve.
between probes Alf968 and Alf968-g7c was 37.5°C, whereas their respective \( T_{\text{m}} \)s were 37.3 and 36.6°C (Table S-3 and Table 1). For the probes that were significantly different as determined by both functional ANOVA and \( T_{\text{m}} \) discrimination often occurred at higher temperatures than the respective \( T_{\text{m}} \)s (e.g., with the uncontaminated soil sample, the temperature of the MAXDCSD between probes Alf968 and Alf968-g7c was 47.5°C, whereas their respective \( T_{\text{m}} \)s were 37.3 and 35.2°C). The functional ANOVA calculator was also used to examine the presumptive cross hybridization of the 23S rRNA probe Gam42a with in vitro transcribed 16S rRNA of \( P. \ putida \). Dissociation curves from the 23S rRNA probe Gam42a hybridized to native \( P. \ putida \) rRNA and the in vitro transcribed 16S rRNA molecule of \( P. \ putida \) were significantly different (MAXDCSD value of 0.469). The 23S rRNA molecule from native \( P. \ putida \) rRNA is a suitable reference standard because it is a MM target to the Gam42a probe. Of note, the dissociation curves of the Gam42a probe hybridized with in vitro transcribed 16S rRNA of uncontaminated and TNT-contaminated soil samples were also significantly different than the dissociation curve obtained with native \( P. \ putida \) rRNA (MAXDCSD value of 0.428 and 0.441, respectively).

**Comparison of 16S rRNA Profiles of Uncontaminated and TNT-Contaminated Soils.** Significant differences in the nucleic acids amplified from the uncontaminated and TNT-contaminated soil samples were detected when comparing initial signal intensity and dissociation curves. Initial signal intensities showed that several domain-, group-, and species-specific PM probes were significantly different between in vitro transcribed 16S rRNA of uncontaminated and TNT-contaminated soils (Figure 2). The signal intensities of ten PM probes were significantly lower with the TNT-contaminated soil sample target sequences (Table 1). Eleven PM/MM probe sets revealed subtle differences between the contaminated and uncontaminated sample target sequences (Table 1). Eleven PM/MM probe sets were detected as significantly different with one of the soil samples but not the other one. In addition, two PM/MM probe sets (i.e., Alf968/Alf968-g7c and Alf968/Alf968-c13g) had temperatures more than 5°C different between the two soil samples at the point of the MAX DCSD (Table 1).

Several parameters may explain the observed differences between the two soil samples: presence or absence of TNT, spatial heterogeneity (e.g., pH, granulometry and carbon content), and PCR bias. Several authors (19, 20) have demonstrated that PCR-amplification bias is low when the sample has diverse templates. In another study (11), 16S rRNA gene PCR amplicons from the same soil samples were analyzed with denaturing gradient gel electrophoresis (DGGE). The number of bands identified in these two soil samples (more than 15 bands) indicated that the samples contained highly diverse templates. Therefore, we suggest that the influence of PCR bias was probably small. The position of the bands differed between the two soil samples, and three brighter bands were present in the TNT-contaminated but not the uncontaminated soil sample (11). These results corroborate the microarray results that differences in amplicon sequences existed between the two soil samples. DGGE was also used to characterize 9 additional soils sampled at the same TNT-destruction field (George et al., manuscript in preparation). Three to seven brighter bands belonging to
Proteobacteria were present in all TNT-contaminated soils but not in the uncontaminated soils. Several characteristics of these soils were investigated (pH, granulometry, and carbon content), but the differences in DGGE fingerprints between uncontaminated and TNT-contaminated soil samples were explained by the absence or presence of TNT. The influence of TNT in soil microbial population was also supported by additional results of George et al. (manuscript in preparation) who artificially contaminated a pristine soil with TNT and found a rapid selection of Pseudomonas aeruginosa.

Fuller and Manning (21) showed that TNT was more toxic to Gram-positive than Gram-negative bacteria. A significant decrease in signal intensity for the probes Hgc69a and Lowg353 targeting Gram-positive bacteria was observed for the contaminated soil sample (Figure 2; Table S-2), although the magnitude of this difference was relatively small. The significant increase in signal intensity from the probe targeting Bacillus (Bacil1136) in the contaminated soil and the difference in the dissociation curves for the Hgc69a probe between the uncontaminated and contaminated soil sample further suggests a possible shift within the Gram positive community. However, detection of specific microbial populations sensitive or resistant to the toxicity of TNT is limited by the relatively small number of specific probes (genus or species) on this prototype microarray. Ultimately, improvements in microarray technology will be needed to not only identify differences among samples (e.g., how differences in amplicon diversity and abundance influence signal intensities and dissociation curves) but also to provide sufficient sequence information to identify specific target sequences and to determine appropriate stringency temperatures to compare signal intensities of different environmental samples.

Previous studies have documented the general utility of using dissociation analysis to provide additional information about the sequence composition of RNA hybridized to microarrays. However, those studies had not presented a fully elaborated experimental and statistical approach for the generation and analysis of dissociation data sets. This study has more fully developed the experimental and analytical approach. Functional ANOVA was shown to resolve differences between most perfect-match and single mismatch probe sets, providing for much greater resolution of microbial communities inhabiting TNT contaminated and uncontaminated soils than could not be achieved by using either differences in signal intensity or $T_m$. Experimental replication demonstrated that dissociation curves were highly reproducible (e.g., probe Eub338, Figure 1C). The ability to clearly resolve single nucleotide mismatch variants and to identify dissociation curves that conform with a specific target reference are essential to achieving the ultimate analytical goal of a robust assignment of identity to nucleic acid hybridized to each array element. Although as yet there is no established theoretical or analytical framework for using DNA microarrays to fully resolve complex and undefined mixtures of nucleic acid sequences, we anticipate that the development of a robust statistical framework that captures the additional information of duplex composition available using dissociation analysis will provide a critical foundation for advancing this important area of environmental biotechnology.

**Acknowledgments**

This work was supported by a short-term visiting fellowship from the Belgian French Community to L.E. Additional support was received from NIH/NIDCR (U01 DE14955) and US DARPA (DABT63-99-1-0008) to D.A.S. and NASA (MSMT-2004-0045-0066) to D.A.S. and J.C.S. We thank J. L. Ramos for providing P. putida sp. strain JRL11, E. M. Starke, P. A. Noble, and A. Pozhitkov for critical review of the manuscript. We also thank J. Jackman and M. Theodore of Johns Hopkins University Applied Physics Laboratory for assistance throughout this study.

**Supporting Information Available**

PCR amplification of the 16S rRNA gene; fragmentation and labeling of RNA samples; description of microarray formats and design of PM and MM probes; recording of signal intensities and generation of dissociation curves; description of unpaired tests to compare signal intensities and $T_m$ values; characteristics of PM and MM probes (Table S-1); initial signal intensities of PM and MM probes (Table S-2); $T_m$ values of PM and MM probes (Table S-3); and images of representative hybridizations (Figure S-1). This material is available free of charge via the Internet at http://pubs.acs.org.

**Literature Cited**


Received for review April 14, 2006. Revised manuscript received July 19, 2006. Accepted July 27, 2006.

ES0609093