BACTERIAL COMMUNITY DYNAMICS IN A PETROLEUM CONTAMINATED LAND TREATMENT UNIT INDICATE A DOMINANT ROLE FOR *FLAVOBACTERIUM* IN PETROLEUM HYDROCARBON DEGRADATION

A thesis presented to the
Faculty of the Biological Sciences Department
California Polytechnic State University, San Luis Obispo

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Biological Sciences

By
Christopher Wolff Kaplan
August 2002

APPROVAL PAGE

TITLE: Bacterial Community Dynamics in a Petroleum Contaminated Land Treatment Unit Indicate a Dominant Role for *Flavobacterium* in Petroleum Hydrocarbon Degradation

AUTHOR: Christopher Wolff Kaplan

DATE SUBMITTED: August 2002

Christopher Kitts

Advisor Signature

Raul Cano

Committee Member Signature

Andrew Schaffner

Committee Member Signature

ABSTRACT

Bacterial Community Dynamics in a Petroleum Contaminated Land Treatment Unit Indicate a Dominant Role for *Flavobacterium* in Petroleum Hydrocarbon Degradation

by

Christopher Wolff Kaplan

Bacterial community dynamics were investigated in a land treatment unit contaminated with petroleum hydrocarbons in the C10-C32 range. The treatment plot was monitored weekly for Total Petroleum Hydrocarbons (TPH), soil water content, nutrient levels, and aerobic heterotrophic bacterial counts. Weekly soil samples were analyzed with 16S rDNA Terminal Restriction Fragment (TRF) analysis to monitor bacterial community structure and dynamics during bioremediation. TPH degradation was rapid during the first 3 weeks and slowed for the remainder of the 24-week project. A sharp increase in plate counts was reported during the first 3 weeks indicating an increase in biomass associated with petroleum degradation. Principal Components Analysis (PCA) of TRF patterns indicated two sample clusters: one consisting of samples from the first 6 weeks, the other consisting of samples from the remainder of the study. TRF sets consisting of TRFs from multiple enzyme digests were associated with bacterial phylotypes. Two phylotypes, *Flavobacterium* and *Pseudomonas*, were dominant in TRF patterns from samples during the early period of the project and were positively correlated with TPH levels over the course of the study. These data suggest that bacteria in the *Flavobacterium* and *Pseudomonas* phylotypes are critical to effective degradation of petroleum at our site.

ACKNOWLEDGEMENTS

I would like to thank everyone at the Environmental Biotechnology Institute, past and present, who made working at the EBI a pleasurable and memorable experience that I will always cherish. I am grateful for the opportunity to work with Dr. Chris Kitts and Dr. Raul Cano; their direction and insight were always helpful and enlightening. I would like to thank my family for their support and encouragement of my academic career: past, present and future. Finally, I would like to thank my wife Johanna for her love and support during my graduate school career at Cal Poly.

This work was supported by the generous contributions of the UNOCAL Corporation whom I thank for the opportunity to conduct research at their site and for the support of their staff in accomplishing a common goal.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

In a companion report on this project (submitted for publication), we described biochemical data associated with bioremediation at the Guadalupe oil field, which occupies nearly 2,700 acres of the larger Guadalupe-Nipomo Dune Complex and is located on the Central California Coast in San Luis Obispo and Santa Barbara Counties. Due to the viscous nature of the oil at the site a light petroleum distillate, referred to as diluent, was pumped into the wells to thin the oil for more efficient removal. This diluent was inadvertently released into the environment over the years as pipes and storage tanks began to degrade. During site remediation, contaminated soil was stockpiled for eventual cleanup. Prior to treatment, the stockpiled soil contained an average total petroleum hydrocarbon (TPH) concentration of approximately 2,000 milligrams per kilogram (mg/kg). A pilot scale land treatment unit (LTU) was set up to investigate the feasibility of a full scale LTU. Soil at the site is coastal dune sand and contained negligible carbon, nitrogen and phosphorous. Therefore, basic nutrients consisting of phosphate, ammonia were added, and soil was periodically watered and tilled to a depth of 18 inches to aerate and mix nutrients. In a second LTU cell a complex carbon source was added and resulted in degradation kinetics similar to the control cell. Due to the similarity between the amended and control cells, only the control cell was investigated here.

Many studies have looked at the chemical degradation process associated with land treatment (Admon et al., 2001; Alexander, 2000; Olivera et al., 1998). A common phenomenon in land treatment is a two-phase pattern of degradation characterized by an initial fast degradation phase followed by a slow degradation phase. To explain the change in degradation rates it has been suggested that the initial fast degradation phase is mediated by bacterial utilization of bioavailable compounds and is governed by enzyme kinetics. In contrast, the slow phase is governed by the rate of petroleum dissolution from soil particles (Admon et al., 2001; Alexander, 2000). By the end of our 168-day pilot LTU project, 61% of the petroleum contamination was degraded, with 37% degraded during the first three weeks. The degradation rate during the first three weeks of the project was –0.0205 day^{-1}, an order of magnitude higher than that of the last 21 weeks of the project, which had a degradation rate of –0.0026 day^{-1}. These rates compare favorably with those reported for land farming of oily sludge from a petroleum refinery; although the degradation rates were slightly higher (–0.036 day^{-1}) during the early phase, and decreased less dramatically (–0.013 day^{-1}) during the late phase (Admon et al., 2001).

Although significant work has been published discussing bacterial community structure and degradation kinetics associated with bioremediation of environmental contaminants, few have focused on a detailed description of bacterial community dynamics during this process. A recent report, described the structure and dynamics of bacterial communities involved in bioremediation of crude oil (MacNaughton et al., 1999). In this study a few groups of bacteria were observed to increase in abundance in response to oil contamination, but the paucity of samples analyzed left gaps during the first three weeks when key events in bioremediation are known to occur (Alexander, 2000; Admon et al., 2001). In the current study, we present the characterization of an autochthonous bacterial community capable of degrading petroleum hydrocarbons after biostimulation by the addition of nitrogen and phosphorous nutrients. A combination of 16S rDNA Terminal Restriction Fragment (TRF) analysis of bacterial communities using multiple enzymes and a clone library constructed from study samples allowed monitoring of relative bacterial abundance and the identification of bacterial phylotypes associated with the phases of TPH degradation.

CHAPTER 2

Background

Land Treatment

Land treatment is an alluring method of remediation due to its effectiveness, low cost and minimal environment impact. Land treatment is a form of bioremediation whereby autochthonous soil bacteria convert petroleum hydrocarbons into H_2O, CO_2 and bacterial biomass. Three types of bioremediation are predominantly practiced: natural attenuation, biostimulation (addition of nutrients to stimulate organisms), and bioaugmentation (addition of contaminant degrading organisms). The simplest method of bioremediation to implement is natural attenuation. Natural attenuation is the natural process whereby bacterial communities naturally degrade contaminants in the environment. Contaminated sites are typically monitored for contaminant concentration during the process to assure that contamination is being removed. When nutrients are low and speed of contaminant degradation is an issue, biostimulation has been indicated in increased degradation (Venosa et al., 1996). Biostimulation is the process of providing bacterial communities with a favorable environment in which they can effectively degrade contaminants. The addition of nitrogen and phosphorous as well as aeration of soil have been indicated as speeding up the bioremediation process (Huesemann and Truex, 1996; Venosa et al., 1996). In cases where natural communities of degrading bacteria are not present or present at low levels, the addition of degrading communities to the contaminated environment, known as bioaugmentation, can speed up the process (Al-Awadhi et al., 1996). Although significant research is performed in this area, bioaugmentation is generally not practiced since introduced bacteria usually can't compete with autochthonous bacterial communities entrenched in their niches (MacNaughton et al., 1999).

Community analysis

Describing microbial community structure and dynamics is an important, yet daunting task due in part to the large number of bacteria that inhabit a sample, approximately 10^8/g in soil. On a global scale, bacteria communities play a critical role in the cycling of nutrients, such as carbon and nitrogen. Due to the vast importance of bacterial communities it is imperative to understand how they interact with their environment and changes to it.

Standard culture techniques have a limited ability to adequately describe microbial communities in soil. Culture techniques are typically time consuming, and cumbersome, requiring a battery of individual biochemical and nutritional tests to characterize each isolate which may require excessive incubation times for adequate growth. Most soil microorganisms are not easily grown in the laboratory, if they can be grown at all. Consequently, culture techniques grossly underestimate diversity, only describing approximately 0.3% of a community (Amann et al., 1995). To overcome the limitations of culture techniques, molecular techniques have been developed to describe microbial communities. The power of molecular techniques is that they can use the minute quantities of DNA extracted directly from microorganisms in the environment. To increase the amount of DNA available for analysis, molecular techniques employ Polymerase Chain Reaction (PCR) to amplify large quantities of DNA from template DNA extracted from a microbial community. A common target gene for PCR is the ribosome small subunit, or 16S rRNA gene. The 16S gene is particularly useful since it contains highly conserved regions that when targeted by primers can amplify an estimated 99% of the bacteria in a sample (Brunk et al., 1996). The 16S gene also contains several variable segments that serve as a basis for differentiating bacteria either through sequencing or restriction fragment analyses. Molecular techniques are not without their own biases. Extraction bias results from differential extraction of subsets of the same microbial community and further varies depending on the techniques used to lyse cells (Martin-Laurent et al., 2001). PCR bias results in the differential amplification of community members due factors such as primer hybridization (Brunk et al., 1996), annealing temperature, number of PCR cycles, amount of DNA in PCR (Suzuki and Giovannoni, 1996), rRNA copy number (Farrelly et al., 1995; Fogel et al., 1999), and chimeric amplicons (Wang and Wang, 1997). Despite these factors, molecular techniques have the clear advantage of being able to describe a larger portion of a microbial community than culture techniques.

Molecular techniques

Current molecular techniques for describing bacterial communities are varied in their methodology as in the information they provide, ranging from identification of specific organisms to fingerprinting of entire communities.

Cloning is a commonly practiced molecular method of describing bacterial communities due to its ability to identify large numbers of bacteria in a community in an efficient manner. In cloning, community DNA is amplified with PCR. The resulting amplicons are ligated into vectors that are used to transform host bacteria called clones. As clones grow and multiply, they replicate the plasmid thereby increasing the copy number of the ligated sequence. Plasmids are then extracted from clones and sequenced; sequences are compared to large sequence databases for purposes of identification of bacteria in the original community. The number of times a particular sequence is recovered can be used as a means of quantifying the abundance of the organisms from which the sequence was obtained (Hill et al., 2002).

Fluorescent In Situ Hybridization (FISH) has been likened to using cruise missiles to quantify specific bacteria in a community. The cruise missiles at the heart of the FISH technique are fluorescently labeled DNA probes designed to hybridize to specific DNA sequences. After hybridization, bacteria are viewed under a microscope with a light source that excites the fluorescent label causing it and the bacteria that contain it to fluoresce, allowing quantification of target bacteria relative to the entire community (Christensen et al., 1999).

Community fingerprinting is a commonly used molecular technique used to describe bacterial community structure and dynamics, but may also be used to identify bacteria within the community. Two forms of community fingerprinting are commonly practiced, Denaturing Gradient Gel Electrophoresis (DGGE) and Terminal Restriction Fragment (TRF). In DGGE, bacterial community DNA is amplified with a primer set in which one primer has a GC-clamp attached. The amplified DNA is loaded onto an acrylamide gel that has a gradient of denaturant (e.g. urea, formamide) which denatures the DNA as it migrates down the gel. When a DNA fragment denatures the GC-clamp remains double stranded causing the DNA fragment to stop its migration through the gel, thus differentiation DNA fragments based on their length and hydrogen bond strength. After a gel has been run bands of interest can be cut out and sequenced to identify the organism the DNA represents. TRF analysis differs from DGGE in that no GC-clamp is used and one primer has a fluorescent labeled attached which allows for visualization in an automated sequencer in which samples are run (Avaniss-Aghajani et al., 1994; Clement et al., 1998; Liu et al., 1997). The amplified community DNA is digested with a tetrameric restriction endonuclease that results in fragments of various lengths dependent on the 16S gene sequence variation of the bacteria in the community. The digested community DNA is loaded into the sequencer and results in a pattern containing a series of peaks with each peak represent one or more organisms in the community. In TRF only the terminal fragment is visualized while unlabeled fragments pass undetected, resulting in only one terminal fragment per organism. The resulting pattern is reproducible (Osborn et al., 1998) and digitally stored data can be easily compared with patterns from samples taken at different times or from other sites, a clear advantage over DGGE in which a new gel must be run for each new comparison. TRF has its origins in RFLP, which is used to create fingerprints of individual bacteria. RFLP samples are run on acrylamide gels in which all restriction fragments are visualized, resulting in a pattern of the fragments that serves as a basis for differentiating bacteria.

Microarray technology for analysis of bacterial communities is currently under development (Cho and Teidge, 2002; Small et al., 2001;Wu et al., 2001). Microarrays uses DNA probes attached to a slide to identify bacteria. Each array can have several hundred or thousand different probes per slide allowing for identification and quantification of many bacteria at once. Functional genes can also be detected on an array allowing determination of the functional potential of bacterial in a community. Although still in it preliminary stages of development, array technology has the ability to profoundly increase our knowledge of microbial ecology due to its ability to combine bacterial identification and quantification with high throughput screening.

Primer selection

Of the factors that influence a community fingerprint, primer selection is the most critical. By using the 16S rRNA gene a pattern reflecting the phylogenetic diversity of bacteria in a sample is produced. Different regions of the 16S gene can be used to select for different groups of prokaryotic organisms with different ranges of specificity (Brunk et al., 1996). When primers are targeted for functional genes, a pattern reflecting the metabolic potential of a community is produced. Patterns representing functional genes are typically sparse in comparison to 16S patterns since not all bacteria have the same functional genes. Functional genes typically targeted for analysis are genes known to control steps in the cycling of nitrogen and carbon (Braker et al., 2001).

Enzyme selection

Enzyme selection is another critical step in producing a TRF pattern since the conservation of cut sites for each enzyme varies between enzymes and changes for every gene. An enzyme with a cut site in a conserved portion of a gene with produce fewer peaks than an enzyme with cut sites in a variable region of a gene and have a poor ability to differentiate bacteria from different phylogenetic groups. Enzymes that cut less often typically have large peaks that represent broad groups of organisms. The practice of using multiple enzymes can resolve the problem since organisms that have the same TRF with one enzyme may not have the same TRF with another enzyme. A careful database analysis and digestion with many enzymes is a good way to evaluate the ability of different enzymes to create a good pattern. By analyzing the results of multiple separate enzyme digests, the results of each digests analysis can be used to corroborate the results of the others.

TRF analysis

TRF data reported by a sequencer consists of the size (base pairs), peak height, and peak area for each TRF peak in a pattern. Several methods of analyzing TRF data have been used since the technique was developed. One on the major differences is the use of either peak height or area to estimate TRF abundance. While widely used, peak height is problematic since peaks become wider and shorter as the fragments get longer resulting in an inaccurate estimate of larger TRFs abundances relative to shorter TRFs. Peak area results in an accurate estimate of TRF abundance since the area of a peak is a measure of both its height and width. Using Boolean datasets in which TRFs are treated as either present or not present is practiced less often presumably since it discards important information about TRF peak magnitude. Because the amount of DNA loaded onto a sequencer cannot be quantitatively controlled, the sum of all TRF peak areas in a pattern (total peak area) vary between TRF patterns. To compensate for this variation, it is necessary to normalize peak detection thresholds and peak areas. Peak detection threshold is normalized by creating an artificial detection threshold for each sample. The new threshold value is created by multiplying a pattern's relative DNA ratio (the ratio of total peak area in the pattern to the total peak area in the sample with the smallest total peak area) by 580 area units (the approximate area of a 50 fluorescent unit peak which represents the software detection threshold). TRF peaks with areas less than the new threshold value for a sample are removed from a data set (Table 1). Peak areas are then normalized by converting the value of each remaining peak area to parts per million of the new total area (Kaplan et al., 2001).

Table 1. Example of truncation procedure used to determine smallest observable peak for each sample in a dataset.

Sample value	Total area	Ratio of total peak area	Threshold
Smallest	200000	1:1	580*
Big	400000	2:1	1160
Bigger	2000000	10:1	5800

*Minimum detectable peak area with Genescan™ software.

Statistical analysis

From a statistical perspective, having more measured variables (TRFs) than observations (samples), poses a problem when attempting to determine the significance of groups within the dataset. To overcome the extreme dimensionality of a TRF dataset (more variables than observations) it is necessary to analyze the data with a statistical method that reduces the dimensionality so that meaningful inferences can be made. Principal Components Analysis (PCA) is a multivariate statistical method that determines the primary sources of variation in a dataset and reduces the dimensionality by creating linear combinations of variables that best preserves the overall variation in the dataset, referred to as Principal Components (PC). Samples are projected onto the PC vectors and are given scores for each PC while variables are given loading values that represent the amount of influence each variable has in making the separations along a particular PC. When used in the context of TRF data, loadings attributed to TRF represent the significance of a TRF, and the organisms the TRF represents, in separating samples along a PC. Two methods of PCA are used, covariance and correlation. Covariance PCA uses the raw data values to construct PCs, while correlation uses a dataset of scaled values (z-scores) produced by subtracting the mean and dividing by the standard deviation of a variable. Since the covariance method uses raw TRF areas, more significance is given to TRFs with large areas since the magnitude of a small percentage change in these TRFs is greater than a large percentage change in a small TRF. For example, a 10% change in TRF with an area of $1x10^6$ will result in a change of $1x10^5$ area units, while a 50% change in a TRF with an area of $1x10^5$ results in a change of only $5x10^4$ area units. Correlation PCA removes this bias by scaling the data so that changes in TRF area reflect the relative magnitude of changes in TRF area. Correlation PCA is intended for use with data that is measured on different scales so that one measurement with larger values does not dominate an analysis. For example, it is necessary to use a scaled dataset when measuring height in meters and weight in kilograms; in this example, weight in kilograms has larger measurement values and dominates an analysis. Using a scaled dataset in TRF analysis is a controversial topic since all peaks are measured on the same scale, yet some peaks may be smaller due to PCR induced biases that alter their actual abundance. A problem with correlation PCA is that it can give significance to small peaks that only change due to variation in machine measurements. Considering these factors it is therefore advisable to use both methods of PCA and consider the results in light of the method used to create them.

Database analysis

TRF analysis is a powerful tool for describing bacterial community structure and dynamics, but the information is also useful for identifying community members. As described above, each TRF in a pattern represents one or more organisms. When taken alone, an individual TRF could match with a large number of organisms with the same predicted TRF peak. By using multiple enzyme digests, it becomes possible to differentiate unrelated organisms with the same TRF in one pattern since they usually do not share the same peak in a digest with another enzyme. TRF peaks representing one organism, or a consistent set of organisms, should account for the same abundance of the community with each enzyme. If the abundance reported in each enzyme fluctuates dramatically either several organisms is present in a peak that separate out in other digests, or the peaks are unrelated. Clones matching TRFs with approximately the same abundance across enzyme digest were used to create a TRF set representing the organism the clone represents. Further validation of a TRF set can be achieved by following TRF sets in different samples or across a time series. While creating a clone library is useful, it is not necessary for 16S TRFs since vast amounts of sequence information exist from which databases of predicted 16S TRFs for thousands of organisms can be created.

CHAPTER 3

Materials and Methods

Sample Storage and DNA Extraction

One soil sample was taken from the large pile of soil that was treated in the LTU. Five soil samples were taken from the LTU on a weekly basis after initiation of treatment until the 15th week, after which samples were taken at the 18th and 26th week. One soil sample was taken from the stockpile of soil used to fill the LTU. Soil samples were collected and stored on site in a freezer until transferred to our lab where they were stored at –80°C. The five soil samples from each weekly sampling were combined and mixed thoroughly. Five replicate soil extractions were performed on combined soil. In a sterile weigh boat, 1 g of soil was weighed out and transferred into MoBio® bead lysis tubes (Solano Beach, CA). The protocol given in the Mo Bio® kit was followed for the extraction process with the following exception: cells were lysed in the Bio 101 FP-120 FastPrep machine (Carlsbad, CA) running at 5.0 m/s for 45 seconds. The isolated DNA was visualized by agarose gel electrophoresis and quintuplet extractions were combined from each soil sample before PCR. The combined DNA was quantified UV spectrophotometry.

Polymerase Chain Reaction

Amplification of the template DNA was performed by using the 16S rDNA labeled primers 46f (5'-GCYTAACACATGCAAGTCGA), and 536r (5'-GTATTACCGCGGCTGCTGG). Reactions were carried out in triplicate with the following reagents in 50 µl reactions: template DNA, 10 ng; 1X Ampli*Taq* Gold Buffer (Applied Biosystems, Fremont, California, USA); dNTPs, $3x10^{-5}$ mmols; bovine serum albumin, $4x10^{-2}$ µg; $MgCl_2$, $1.75x10^{-4}$ mmols; 46f, $1x10^{-5}$ mmols; 536r, $1x10^{-5}$ mmols; Ampli*Taq* Gold DNA polymerase (Applied Biosystems), 1.5U. Reaction temperatures and cycling for samples were as follows: 95^{o}C for 2 min, 35 cycles of 94^{o}C for 2 min, 46.5^{o}C for 1 min, 72^{o}C for 1 min, followed by 72^{o}C for 10 min. Products were visualized by agarose gel electrophoresis and any inconsistent or unsuccessful reactions were discarded. Primers were removed and amplicons concentrated with the Mo Bio® PCR Clean-Up kit according to the normal protocol. The combined amplicons were quantified by UV spectrophotometry.

Amplicon Digestion

Restriction enzyme reactions contained 75 ng of labeled DNA, and 1.5 Units of restriction endonuclease, *Dpn*II, *Hae*III, or *Hha*I, (New England Biolabs, Beverly, MA, USA) in the manufacturer's recommended reaction buffers. Reactions were incubated for 2 hours at 37°C followed by a 20 minute 65°C denature step. Digested DNA was purified by ethanol precipitation.

TRF Size Determination

The precipitated DNA was dissolved in 9 µl of Hi-DI formamide (Applied Biosystems), with 0.5 µl of Genescan Rox 500 (Applied Biosystems) and Rox 550-700 (BioVentures, Murfreesboro, TN, USA) size standards. The DNA was denatured at 95°C for 4 minutes and snap-cooled for 10 minutes in an ice slurry. Samples were run on an ABI Prism™ 310 Genetic Analyzer at 15 kV and 60°C. TRF sizing was performed using

Genescan^{TM} 3.1.2 software with Local Southern method and heavy smoothing (Applied Biosystems).

TRF Data Analysis

Sample data consisted of the peak area for each TRF peak in a TRF pattern. TRF data was normalized before analysis as discussed in Kaplan et al. (2001). TRF patterns from all samples were analyzed using covariance and correlation Principal Components Analysis (PCA) and Agglomerative Hierarchical Cluster Analysis (AHCA). All analyses were performed on normalized data sets consisting of sample name, TRF length, and TRF peak area using S-Plus 6 (Insightful, Seattle WA). TRF data from *Dpn*II, *Hae*III and *Hha*I digested samples were combined into a composite dataset for analysis with PCA and AHCA. Covariance and correlation PCA were used in this analysis, but only covariance data is presented since both methods produced similar results. Clusters described in this analysis were determined using AHCA with complete linkage. A Loess (nonparametric local regression) line was generated using Mintab 13 (Minitab, Inc.) to approximate the trends present in diversity indices.

16S rDNA Cloning and Phylogenetics

Two soil samples (day 14 and 56) were used for constructing a bacterial 16S rDNA clone library. Bacterial communities were amplified as stated above except an unlabeled forward primer was used. PCR product from these samples was purified using the Mo Bio® PCR Clean-Up kit as stated above. Cleaned PCR product was then ligated into the pCR 2.1 vector provided the Original TA cloning kit as directed by the manufacturer (Invitrogen, Carlsbad, California, USA). Ligated vector was then used to transform Epicurian Coli® XL10-Gold® Ultracompetent Cells (Stratagene, La Jolla, California, USA) according to manufacturer's protocol. Cells were plated onto ampicillin/IPTG containing media and grown overnight. White colonies were picked and grown in TB containing ampicillin overnight. Cells were pelleted and plasmids extracted using Quantum Prep HT/96 Plasmid Miniprep Kits as directed by the manufacturer (Bio-Rad, Hercules, CA, USA). Sequencing reactions (10µl) contained: DNA, 4µl; primer, 1.6e10^{-5} mmol; ABI Big Dye (Applied Biosystems), 4µl; PCR water, 0.4µl. Samples were run on an ABI 377 DNA sequencer and the resulting sequences analyzed using SeqMan™II in the (DNAStar, Madison, WI, USA). Clone sequences were also analyzed with Chimera Check on the RDP website (Maidak et al., 2001). Non-chimeric sequences were tentatively identified using a BLAST search on the National Center for Biotechnology Information webpage (http://www.ncbi.nlm.nih.gov/BLAST). The BLAST search matches with the highest BLAST scores were used in creating phylogenetic trees. Reference organisms and clones were aligned using ClustalX (Thompson et al., 1994) and phylogenetic trees were constructed from the aligned sequences using Seqboot, DNADIST, DNAPARS, DNAML and Consense in the Phylip v3.6a2.1 package (Felsenstein, 1989). Agreement between trees created with different methods served as a basis for evaluating accurate recreation of phylogenetic structure. Phylogenetic trees used in this paper were the result of resampling 100 jackknifed datasets with the DNAML maximum likelihood algorithm. Trees were visualized using Treeview v.1.6.5.

Database Matching of TRF Peaks

A database was created containing predicted TRFs for clones in this study, and ~30,000 16S rDNA sequences from the Ribosomal Database Project (Maidak et al., 2001) and GenBank; all sequences were generated using *in silico* PCR with primers 46f and 536r, and digestion with every commercially available restriction enzyme. Bacterial phylotypes were identified by associating TRFs present in community TRF patterns with database predicted TRFs. To facilitate a more precise association between phylotypes and TRF peaks, TRF peaks were first grouped into TRF sets. Each TRF set consisted of three TRF peaks, one from each enzyme digest, which had similar temporal profiles and PCA loadings. In PCA, loadings indicate the importance of a particular variable (i.e. TRF) to the separation along a principal component (PC). TRFs with large loadings along a PC indicate a substantial influence of these TRFs in separating samples along that PC.

TRF sets were then compared to predicted TRF peaks of clones and public database sequences to generate phylotype associations. Differences are commonly reported between observed TRF lengths and those predicted from sequence analysis (Clement et al., 1998; Kitts, 2001; Kaplan et al., 2001). This was compensated for by correcting for dye-based differences in the migration of ROX-labeled standard peaks and 6-FAM-labeled sample peaks (Appendix A). In addition, observed TRFs (sample) were allowed to be within +/– 2 base pairs (~4 standard deviations) of the predicted TRFs (database).

CHAPTER 4

Results

Companion Study Results

The petroleum contaminant in this study was in the C10 to C32 range and was well weathered after 30 years in the soil. No BTEX or PAH compounds were detected at any time during the study and lighter chain compounds were not present so it is suspected that a limited amount of TPH was lost to evaporation. As discussed in the introduction, an abrupt change from a fast to a slow phase of TPH degradation was observed on the third week of LTU operation. This change was not associated with changes in ambient temperature or soil moisture (Figures 1A and B). Aerobic heterotrophic bacterial (AHB) counts showed a large increase in bacterial biomass in the first three weeks of the study, from an average of 1.65×10^7 to 1.30×10^8. After day 21, AHB counts decreased to initial levels until day 42 when they began to climb again, eventually reaching 1.00×10^8 at the end of the study (Figures 1C and D).

Bacterial DiversityDynamics

16S rDNA TRF patterns were created by digesting with the three tetrameric restriction endonucleases (*Dpn*II, *Hae*III and *Hha*I). TRF patterns from *Dpn*II and *Hae*III digestions had similar numbers of TRFs on average (66.1 and 64.3 respectively), while *Hha*I had far fewer (51.7). Shannon-Weaver diversity index (H') and Simpson Dominance index (SI') were calculated for TRF patterns generated with each restriction enzyme. Results for *Dpn*II and *Hae*III showed similar trends in H' and SI', while *Hha*I results differed slightly. H' and SI' are shown for *Dpn*II because it produced patterns with the largest number of TRFs and produced results similar to *Hae*III (Figure 2). TRF diversity (H') increased after the third week and was followed by a plateau in diversity that lasted until the end of the study. In contrast, dominance decreased after the third week and was followed by a plateau in dominance for the remainder of the study. This suggests that during the first three weeks of the study a few TRFs, which dominated the community patterns at the beginning of the study slowly decreased in abundance, while less abundant and newly detected TRFs began to rise (Figure 3).

Bacterial Community Dynamics

TRF patterns from early and late samples were very different based on H', SI', visual and ANOVA of PC1 scores ($p < 0.05$, Figures 2 and 3). Coincidently, a change in ambient temperature occurred about the seventh week of operation. The average temperature before and after the seventh week was 24.4^{O}C and 16.7^{O}C respectively (Figures 1A and B). PCA and AHCA showed two major groups: one group consisted of samples from early in the study (day 0 to 42) while the second group consisted of samples from later in the study (day 49 to 168). In PCA, the separation between early and late clusters occurred along PC1, which explained 50.1% of the variation in the data (Figure 4).

AHCA further distinguished the two large groups into five smaller clusters that included three temporal shifts (Figure 4). Clusters 1 and 2 represent communities present during the fast degradation phase. Day 0 represents a community baseline for this study due to the temporal relationship of the samples. The second cluster (days 7 to 21)

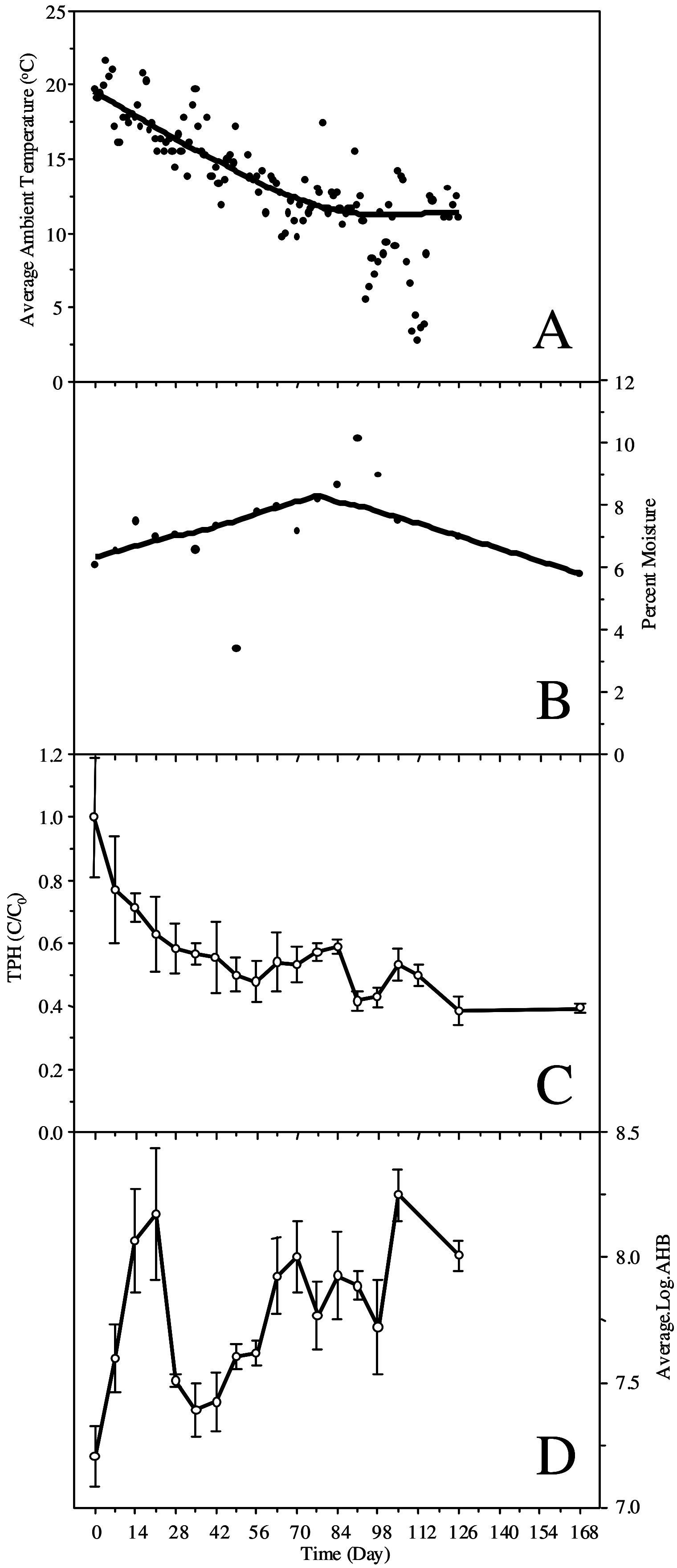


Figure 1. A) Average ambient temperature decreased slowly over the course of the study. **B**) Soil moisture remained relatively constant throughout the study. **C**) Average relative TPH concentration in the LTU during treatment. TPH concentration decreased quickly during the first three weeks of treatment and was followed by slow degradation for the remainder of LTU operations. **D**) Aerobic heterotrophic bacterial counts from soil samples during land treatment. Bacterial counts increase dramatically during the first three weeks. After a decrease in numbers after day 21 bacterial counts slowly increased of the remainder of the project.

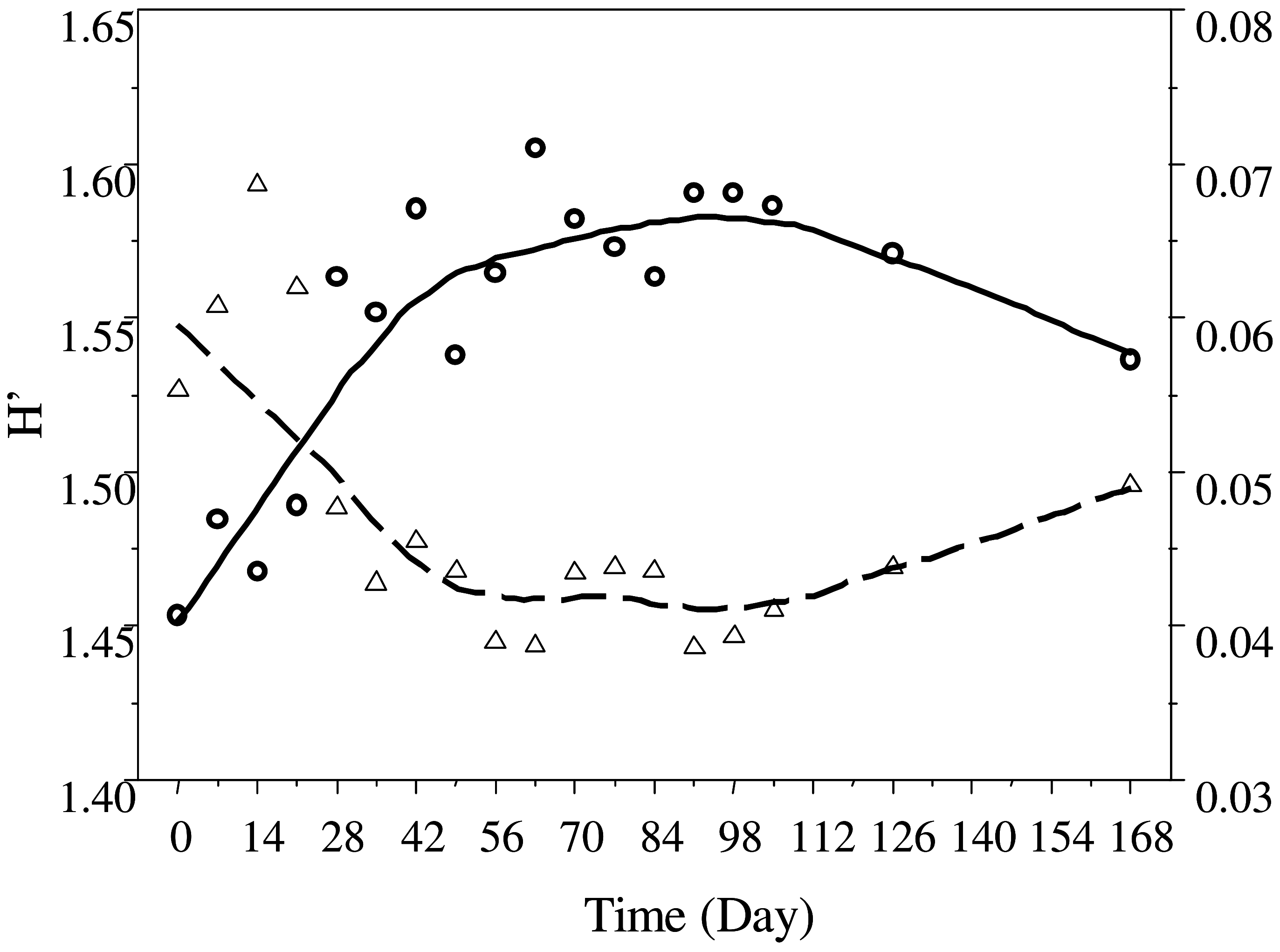


Figure 2. Shannon-Weiner Diversity Index (circles, solid line) and Simpson Dominance Index (triangles, dashed line) based on *Dpn*II digested samples with Loess fitted curves showing low bacterial diversity and high dominance before day 21. After day 21 bacterial diversity increases and dominance decreases, remaining relatively unchanged until the end of the study.

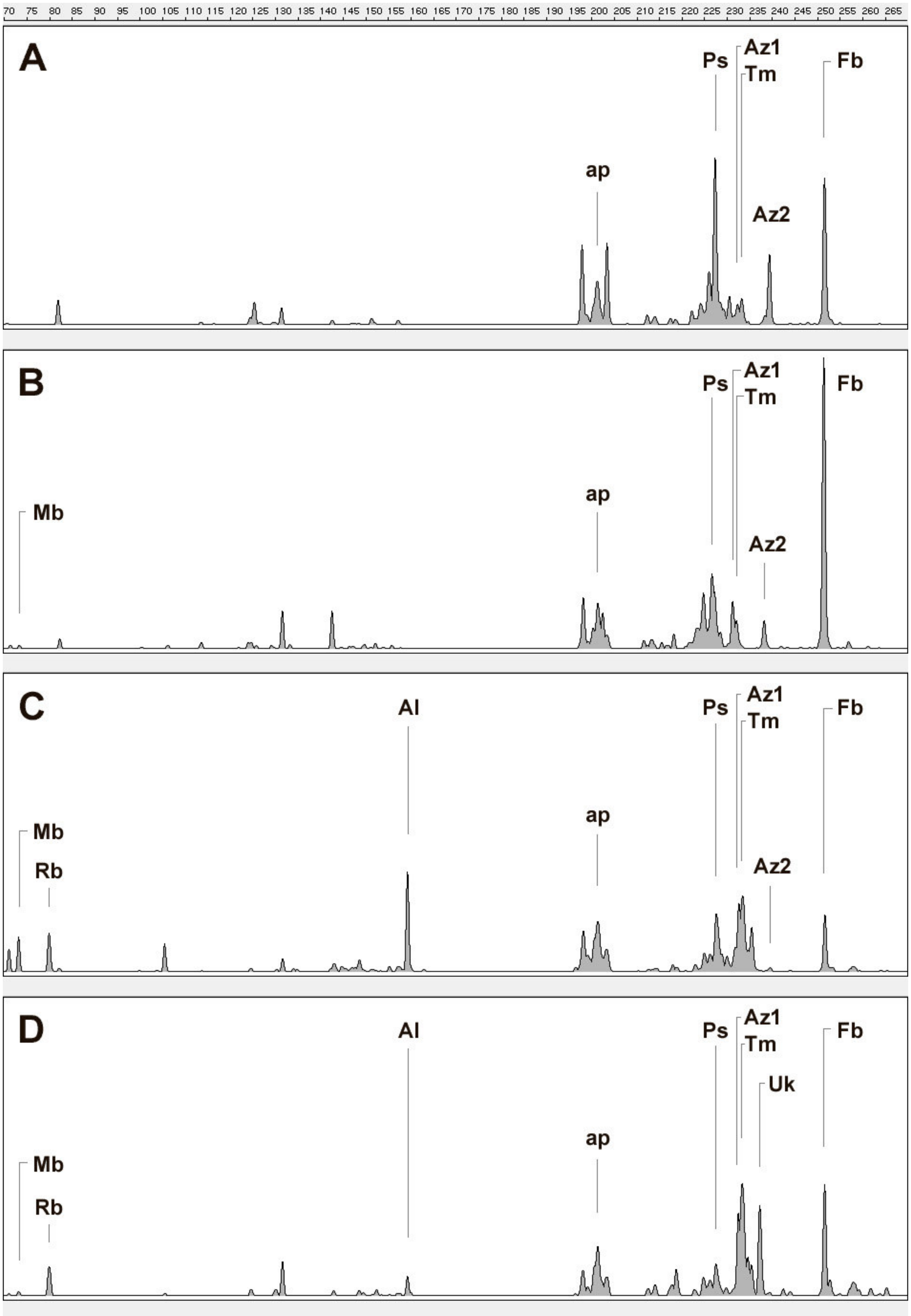


Figure 3. TRF patterns of LTU soil samples, **A**) day 0, **B**) day 14, **C**) day 56, **D**) day 91. Major phylotypes followed in this study are labeled: ap, alpha-proteobacteria; Al, *Alcaligenes*; Az1, *Azoarcus* 1; Az2, *Azoarcus* 2; Fb, *Flavobacterium*; Mb, *Microbacterium*; Ps, *Pseudomonas*; Rb, *Rhodanobacter*; Tm, *Thermomonas*; Uk, Unknown.

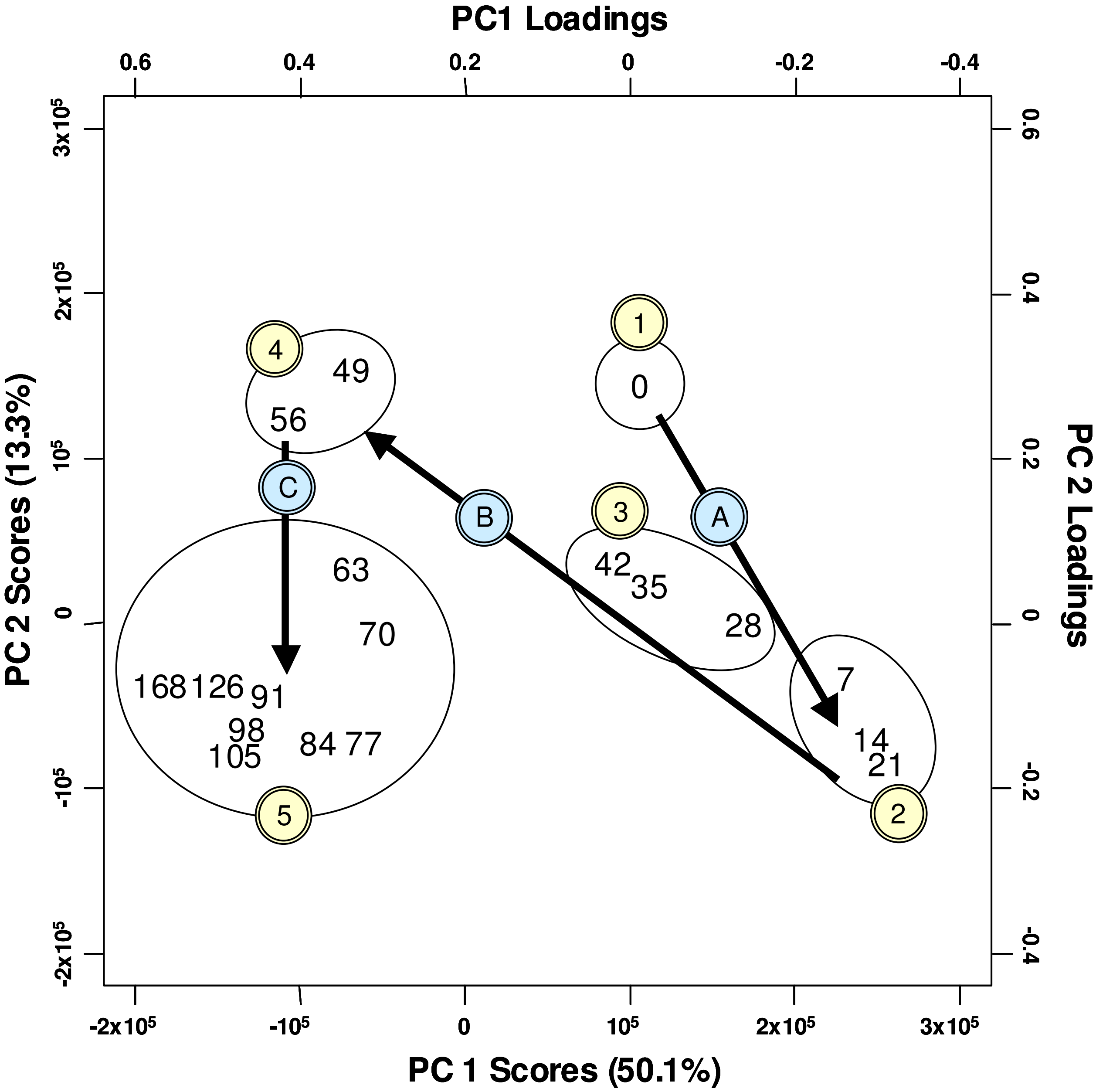


Figure 4. Principal component analysis of combined enzyme TRF data from land treatment unit samples. 1) Day 0 of the study was dominated by *Flavobacterium*, *Pseudomonas*, and *Azoarcus* 2 phylotypes which accounted for 10% and 8.5%, and 3.2% of the bacterial community respectively. A) A shift in the bacterial community from day 7 to 21 corresponds with *Flavobacterium* peaking in abundance as the study reached day 14. 2) *Flavobacterium* comprised 20% of the bacterial community on day 14. B,3) Another shift in the bacterial community occured after day 21 until day 49, which corresponded with a decrease in abundance of phylotypes *Flavobacterium* , *Pseudomonas* and *Azoarcus* 2 and increased abundance of *Thermomonas*. 4) On day 56 both *Alcaligenes* and *Microbacterium* increased in relative abundance. C) From day 56 until the end of the study *Thermomonas* , *Rhodanobacter* and Unknown increased in abundance. 5) In last 100 days the bacterial community were more evenly distributed than the beginning of the study (Figure 2).

represents samples taken during the fast TPH degradation phase. The first temporal shift occurred from day 0 through day 21, which coincided with fast TPH degradation and increasing bacterial counts, indicating a bloom of bacteria associated with TPH degradation (Figures 1C and D). The second temporal shift began on day 28 and ended on day 49, which correlates with a decrease in bacterial counts during this period. Clusters 3 and 4 represent communities in transition after the fast degradation phase. The third shift occurred after day 56 and continued through the final samples (days 63 to 168), which formed one large cluster.

Phylogenetic Analysis of Bacterial Clones from Land Treatment Unit

Days 14 and 56 were chosen to represent early and late samples in a combined clone library consisting of 115 clones from two samples (day 14, 63 clones; day 56, 52 clones) was analyzed for phylogeny. LTU clones consisted of four large groups: Cytophaga-Flavobacterium-Bacteroides (CFB), α-proteobacteria, β-proteobacteria, and γ-proteobacteria (Table 2). Two smaller groups were also represented: ε-proteobaceria, and Gram positives (Figures 5 A-E). CFB clones comprised the largest group of clones in the study, although most came from day 14. A majority of the CFB clones were identified as *Flavobacterium* spp. (92.3%), of these 83.3% shared the same TRF peaks with all three enzymes. α-, β- and γ-proteobacteria accounted for the majority of the remaining clones. Four ε-proteobacteria clones came from day 14 while none were present in day 56. Eight gram positive clones came from day 56, while none were present in day 14. Clones in the α-proteobacteria group were not closely related to other α-proteobacteria clones or database sequences, indicating a broad diversity in this group. Despite this diversity, the abundance of these clones throughout the study indicates a potentially important role for this group in the soil. The β-proteobacteria clones had two major groups: group 1 associated with *Azoarcus* spp. while group 2 associated with *Alcaligenes* spp. and *Bordatella sp*. The γ-proteobacteria clones had a few significant clusters. Two clones (LTU00356 and LTU01856) showed a close relationship to *Rhodanobacter lindaniclasticus*, a recently described lindane degrader (Nalin et al., 1999). A large group of clones associated with *Thermomonas heamolytica*, yet this group was not well defined and was also closely associated with *Stenotrophomonas maltophilia* and *Xanthomonas sacchari*. Clones LTU00856 and LTU08856 were closely associated with *Pseudomonas* spp., a genus known to degrade petroleum. Another cluster of clones including LTU005, LTU024, LTU01456, LTU07556 were loosely associated with *Nitrosococcus oceani*, an ammonia oxidizer. The largest cluster of clones in γ-proteobacteria was associated with methane oxidizing genera, *Methylococcus* and *Methylobacter* (LTU017, LTU034, LTU036, LTU071 and LTU094). Within the CFB clones, a large group containing 24 clones was associated with *Flavobacterium* spp. (Figure 5D), a known petroleum degrader (Atlas and Bartha, 1972). The two other clones in the CFB cluster were associated with *Bacteroides* spp. (LTU047, LTU090). Clones within the Gram + group were associated with *Microbacterium* spp. (LTU00156 and LTU002356), *Planktomyces* sp. (LTU05356 and LTU07056), and *Neochlamydia hartmannellae* (LTU02956, LTU08556, LTU09456, and LTU09656). *Neochlamydia* spp. have been reported as endoparasites of amoebae and may indicate the presence of microeukaryotes in the LTU soil (Horn et al., 2000).

Table 2. Number of clones in library representing each phylotype from days 14 and 56 of the LTU project.

	Number clones and (%)	
Phylotype	**Day 14**	**Day 56**
Flavobacterium	21 (33.3)	3 (5.8)
Pseudomonas	2 (3.2)	4 (7.79)
Azoarcus 2	3 (4.8)	0 (0)
Azoarcus 1	3 (4.8)	2 (3.8)
Bacteroides	2 (3.2)	0 (0)
Microbacterium	0 (0)	1 (1.9)
Rhodanobacter	0 (0)	2 (3.8)
Thermomonas	5 (7.9)	5 (9.6)
Alcaligenes	0 (0)	12 (23.1)
Total Clones	63 (100)	52 (100)

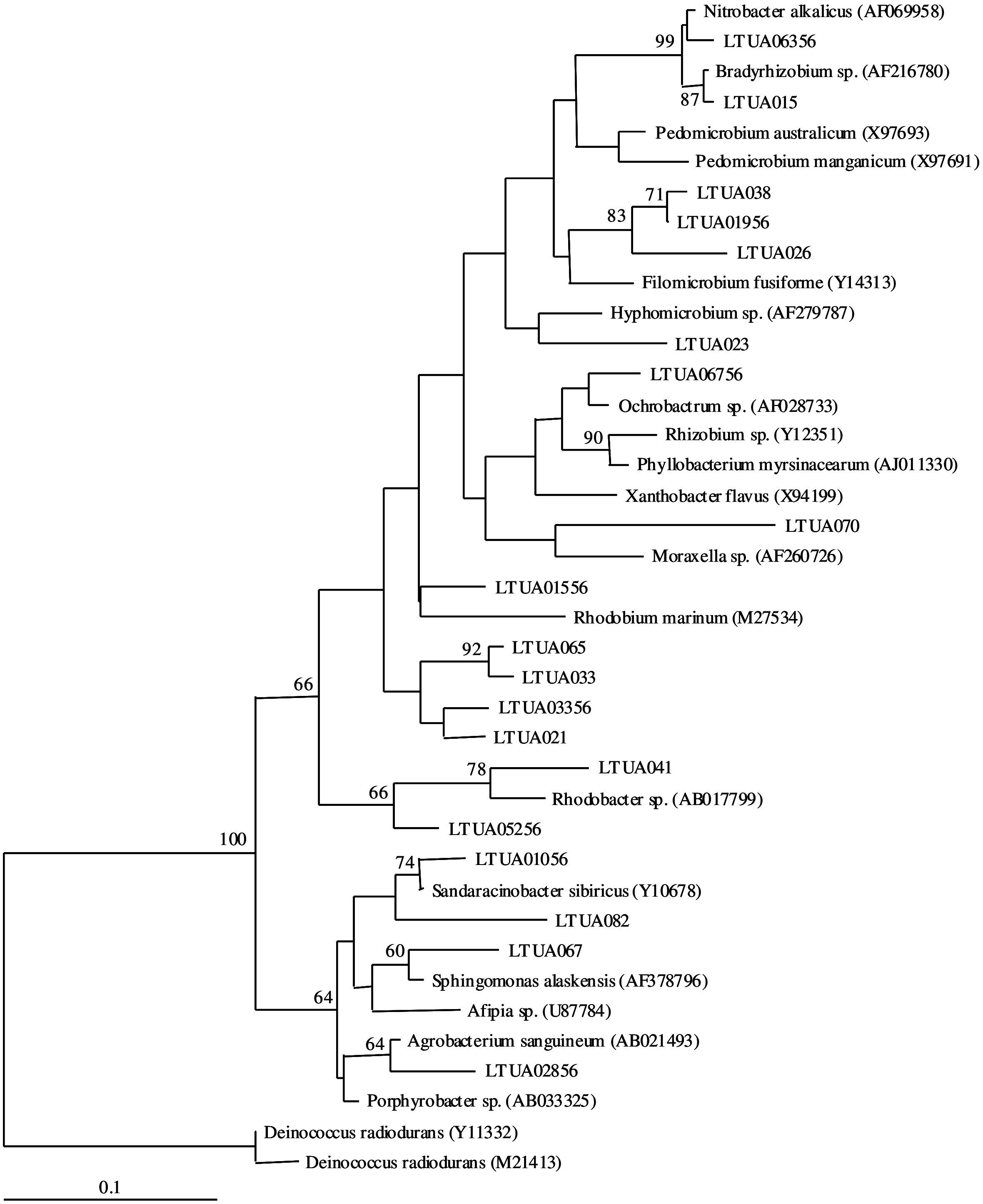

Nitrobacter alkalicus (AF069958)
99
LTUA06356
Bradyrhizobium sp. (AF216780)
87
LTUA015
Pedomicrobium australicum (X97693)
Pedomicrobium manganicum (X97691)
71
LTUA038
83
LTUA01956
LTUA026
Filomicrobium fusiforme (Y14313)
Hyphomicrobium sp. (AF279787)
LTUA023
LTUA06756
Ochrobactrum sp. (AF028733)
90
Rhizobium sp. (Y12351)
Phyllobacterium myrsinacearum (AJ011330)
Xanthobacter flavus (X94199)
LTUA070
Moraxella sp. (AF260726)
LTUA01556
Rhodobium marinum (M27534)
92
LTUA065
LTUA033
66
LTUA03356
LTUA021
78
LTUA041
66
Rhodobacter sp. (AB017799)
100
LTUA05256
74
LTUA01056
Sandaracinobacter sibiricus (Y10678)
LTUA082
60
LTUA067
Sphingomonas alaskensis (AF378796)
64
Afipia sp. (U87784)
64
Agrobacterium sanguineum (AB021493)
LTUA02856
Porphyrobacter sp. (AB033325)
Deinococcus radiodurans (Y11332)
Deinococcus radiodurans (M21413)
0.1

Figure 5A. Phylogenetic tree of Alpha-proteobacteria LTU clones constructed using maximum likelihood algorithm.

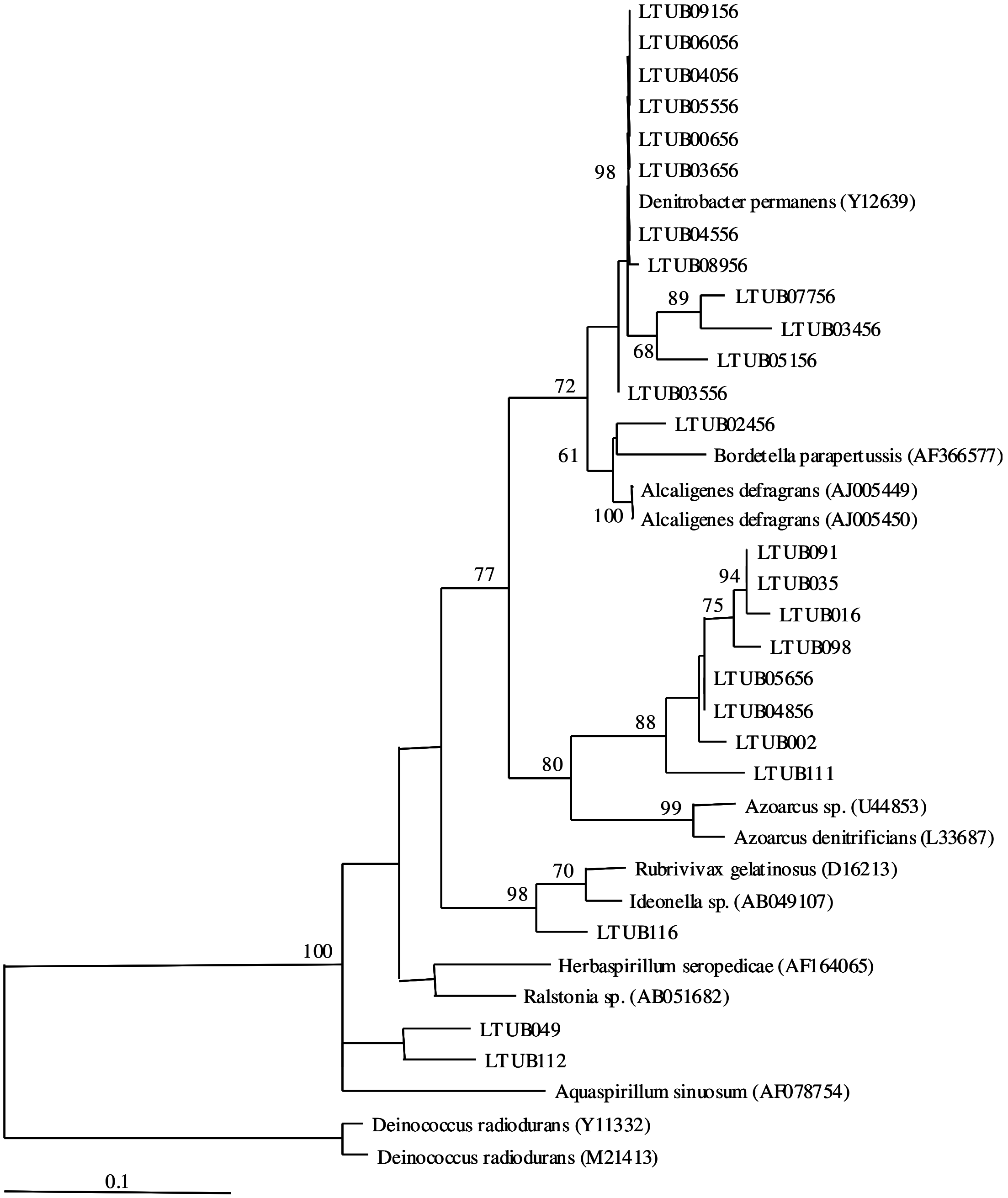

LTUB09156
LTUB06056
LTUB04056
LTUB05556
LTUB00656
98
LTUB03656
Denitrobacter permanens (Y12639)
LTUB04556
LTUB08956
89
LTUB07756
LTUB03456
68
LTUB05156
72
LTUB03556
LTUB02456
61
Bordetella parapertussis (AF366577)
Alcaligenes defragrans (AJ005449)
100
Alcaligenes defragrans (AJ005450)
LTUB091
77
94
LTUB035
75
LTUB016
LTUB098
LTUB05656
LTUB04856
88
LTUB002
80
LTUB111
99
Azoarcus sp. (U44853)
Azoarcus denitrificians (L33687)
70
Rubrivivax gelatinosus (D16213)
98
Ideonella sp. (AB049107)
LTUB116
100
Herbaspirillum seropedicae (AF164065)
Ralstonia sp. (AB051682)
LTUB049
LTUB112
Aquaspirillum sinuosum (AF078754)
Deinococcus radiodurans (Y11332)
Deinococcus radiodurans (M21413)
0.1

Figure 5B. Phylogenetic tree of Beta-proteobacteria LTU clones constructed using maximum likelihood algorithm.

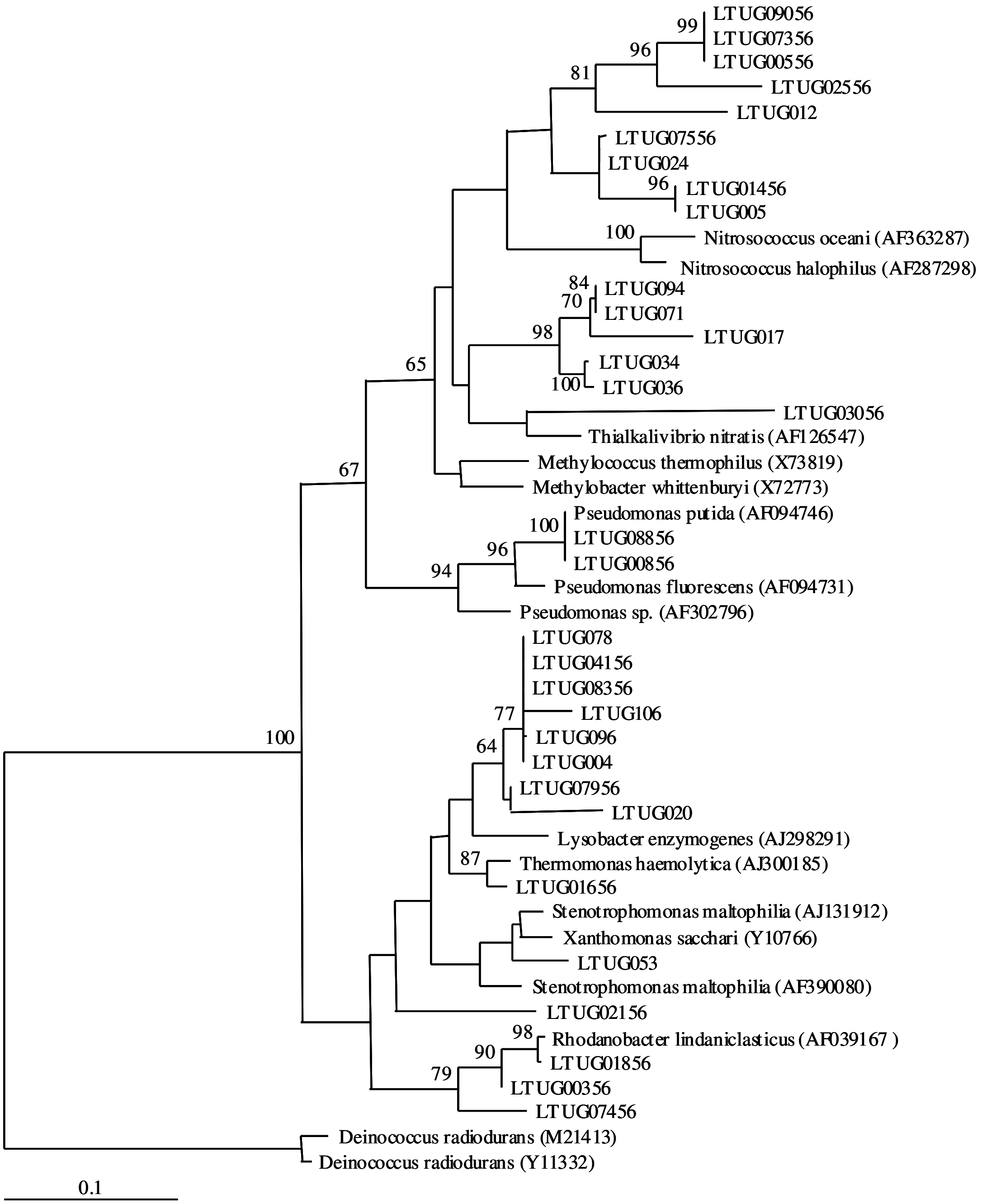

99
96
81
LTUG09056
LTUG07356
LTUG00556
LTUG02556
LTUG012
LTUG07556
LTUG024
96
LTUG01456
LTUG005
100
Nitrosococcus oceani (AF363287)
Nitrosococcus halophilus (AF287298)
84
70
LTUG094
LTUG071
98
LTUG017
LTUG034
100
LTUG036
65
LTUG03056
Thialkalivibrio nitratis (AF126547)
67
Methylococcus thermophilus (X73819)
Methylobacter whittenburyi (X72773)
100
Pseudomonas putida (AF094746)
LTUG08856
96
LTUG00856
94
Pseudomonas fluorescens (AF094731)
Pseudomonas sp. (AF302796)
LTUG078
LTUG04156
LTUG08356
77
LTUG106
100
64
LTUG096
LTUG004
LTUG07956
LTUG020
Lysobacter enzymogenes (AJ298291)
87
Thermomonas haemolytica (AJ300185)
LTUG01656
Stenotrophomonas maltophilia (AJ131912)
Xanthomonas sacchari (Y10766)
LTUG053
Stenotrophomonas maltophilia (AF390080)
LTUG02156
98
Rhodanobacter lindaniclasticus (AF039167)
90
LTUG01856
79
LTUG00356
LTUG07456
Deinococcus radiodurans (M21413)
Deinococcus radiodurans (Y11332)
0.1

Figure 5C. Phylogenetic tree of Gamma-proteobacteria LTU clones constructed using maximum likelihood algorithm.

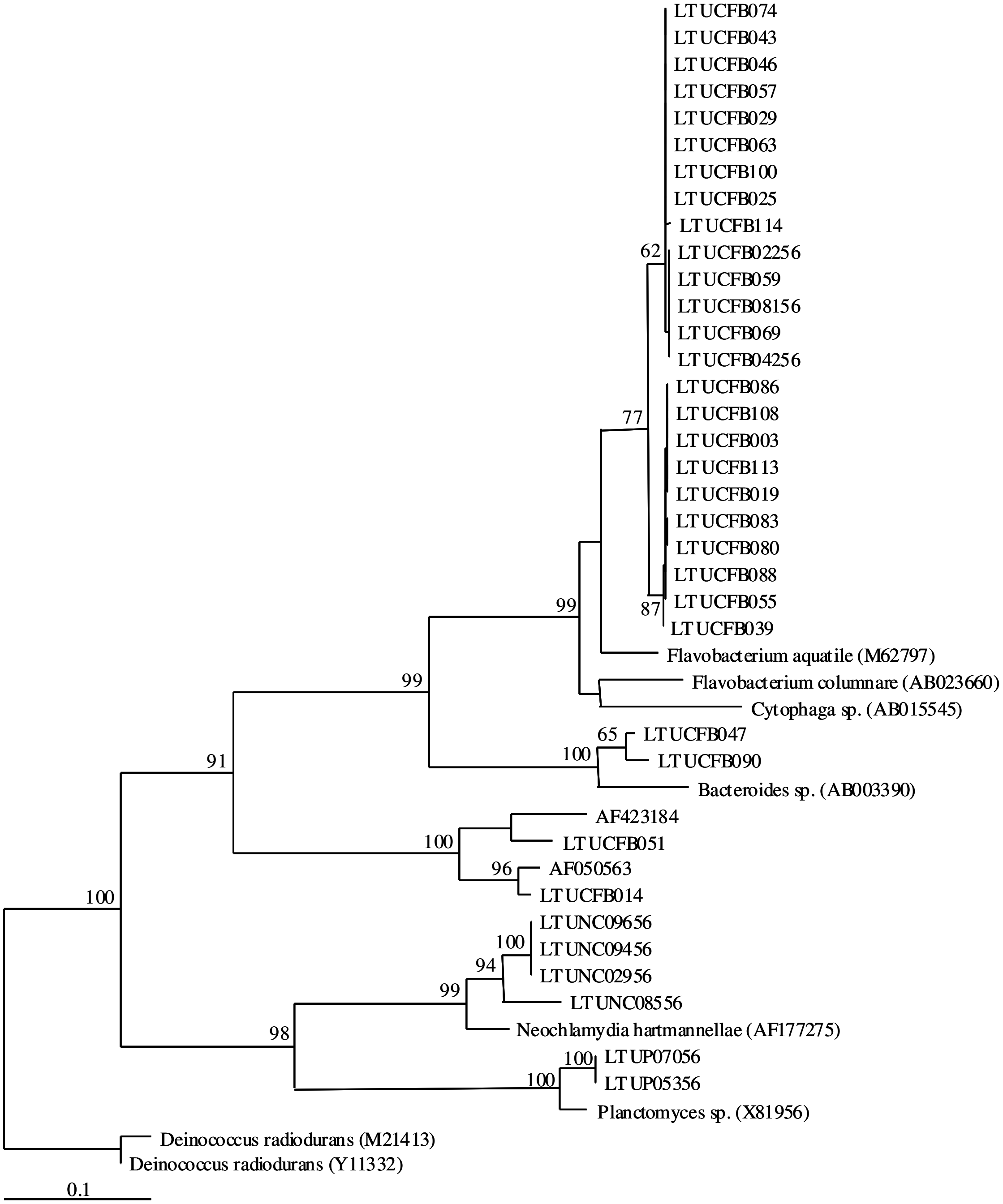


Figure 5D. Phylogenetic tree of Cytophaga/Flavobacterium/Bacteroides LTU clones constructed using maximum likelihood algorithm.

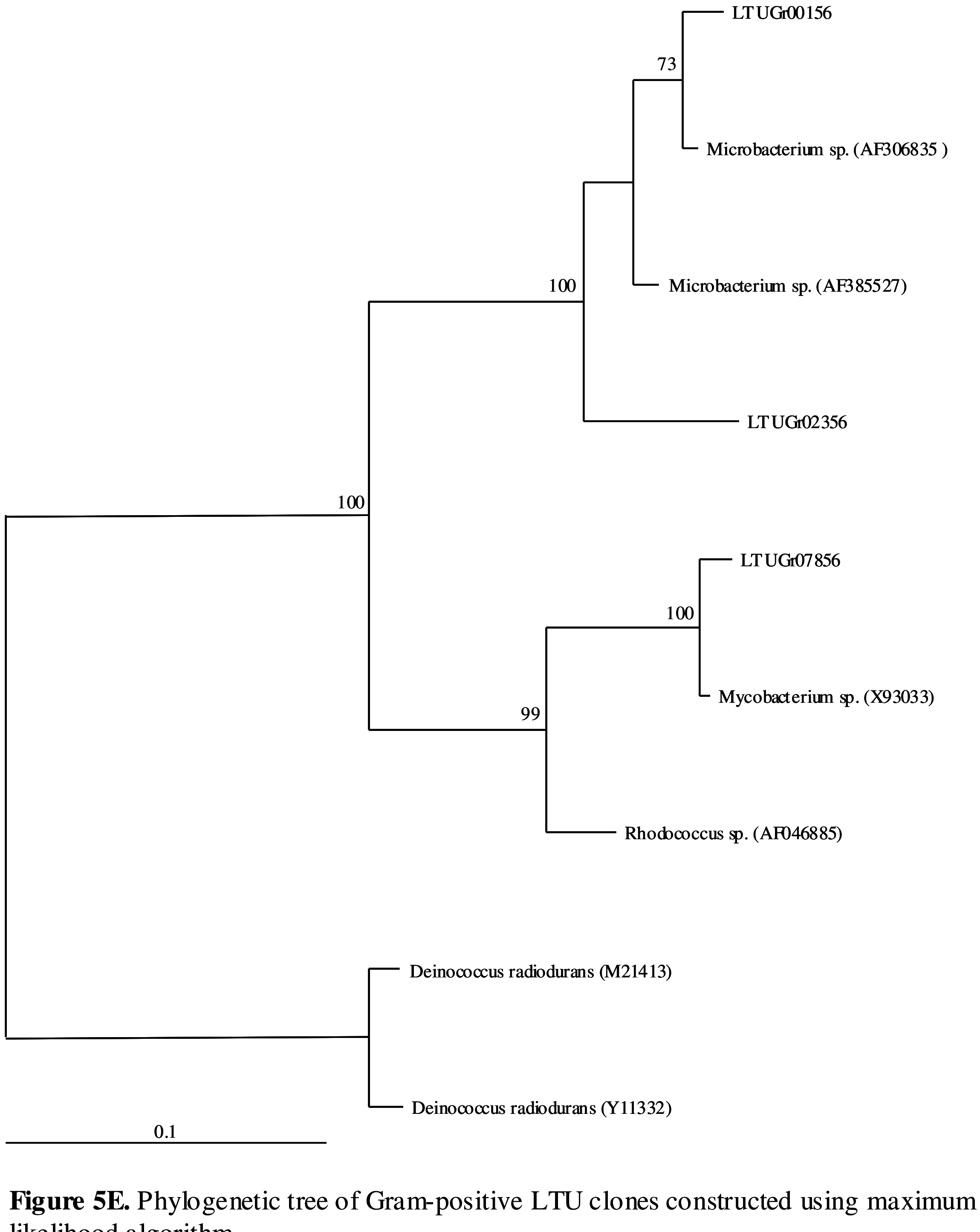

LTUGr00156
73
Microbacterium sp. (AF306835)
100
Microbacterium sp. (AF385527)
LTUGr02356
100
LTUGr07856
100
Mycobacterium sp. (X93033)
99
Rhodococcus sp. (AF046885)
Deinococcus radiodurans (M21413)
Deinococcus radiodurans (Y11332)
0.1

Figure 5E. Phylogenetic tree of Gram-positive LTU clones constructed using maximum likelihood algorithm.

Dynamics of Dominant Phylotypes During Land Treatment

To better understand temporal shifts present in PCA (Figure 4), TRF peaks were associated with bacterial phylotypes (see Materials and Methods for procedure, Figure 6, Table 3).

Eleven phylotpyes were generated and tracked in the context of the whole community by averaging their abundance across all three enzyme digests (Figure 7). These 11 phylotypes, while only representing an average of 16% of the peaks in any TRF pattern, accounted for an average of 40% of the total area in TRF patterns. This makes the average TRF peak area in a phylotype (4%) four times larger when compared to the average peak not included in a phylotype (1%).

Three phylotypes had large abundance during the early phase of the LTU project. TRF set 1, which was associated with *Flavobacterium* clones (Table 3), had large positive loadings along PC1 indicating these TRFs had a large influence on the separation of early and late samples. TRF set 1 also had large negative loadings along PC2 indicating the importance of these peaks in the separation of cluster 2 (days 7 to 21) from cluster 1 (day 0) and cluster 3 (days 28 to 42) in the early samples. *Flavobacterium* TRF peak area increased dramatically during the first 21 days peaking at a high of 19.7%. *Flavobacterium* peak area slowly declined after day 14, reaching a low 3.7% on day 49 and 5.3% by the end of the study. The abundance of *Flavobacterium* in the LTU as depicted by TRF area is also reflected in the number of *Flavobacterium* clones sequenced (Table 2). Interestingly, *Flavobacterium* was detected at very low levels in a pretreatment sample (0.1%) in contrast to its large abundance throughout the rest of the project (Figure 7). TRF set 2, which was associated with *Pseudomonas* clones (Table 3),

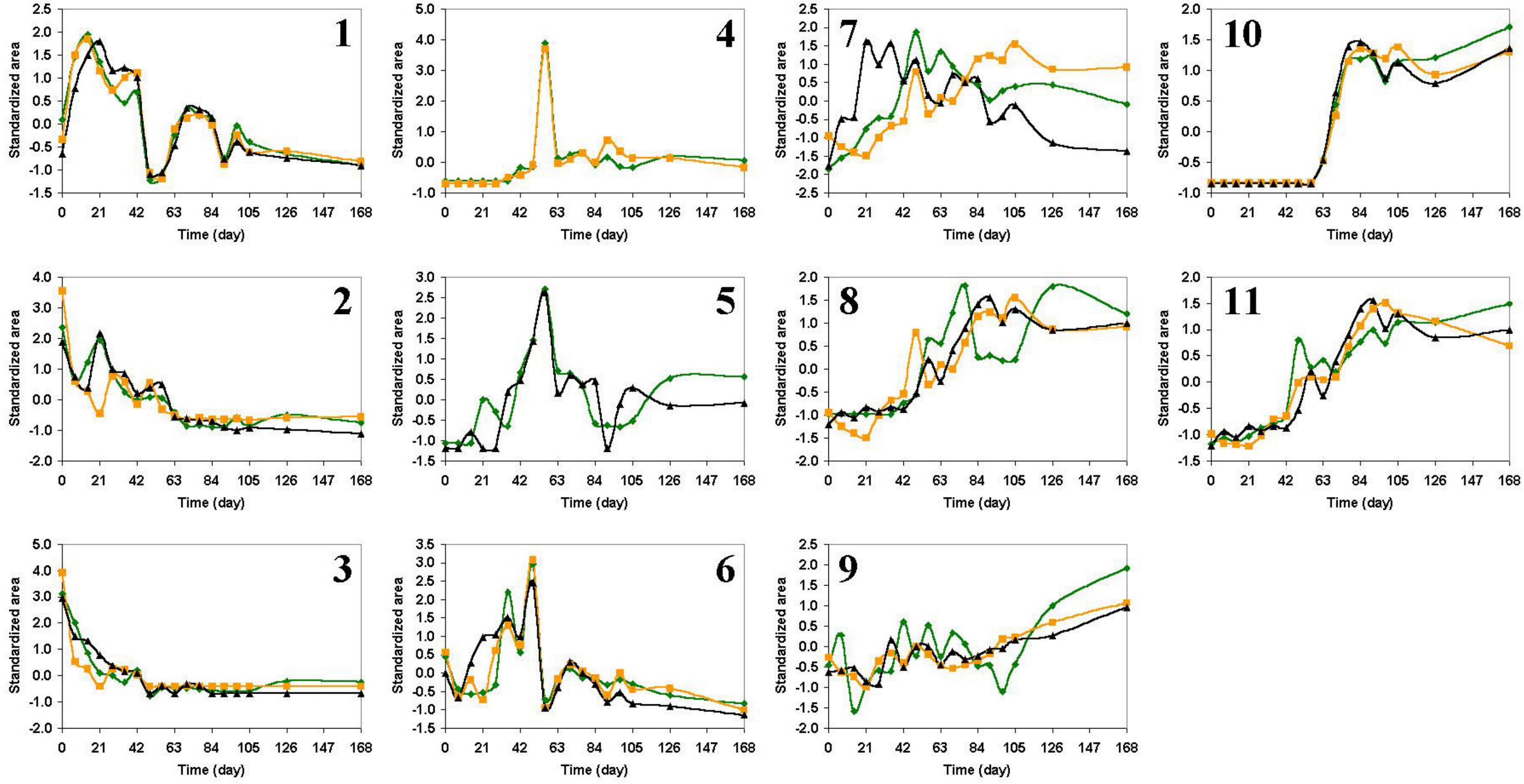


Figure 6. Relative abundance of phylotypes based on TRF peak sets from *Dpn*II, *Hae*III, and *Hha*I digested community samples. TRF sets: 1) *Flavobacterium* 2) *Pseudomonas* 3) *Azoarcus* 2 4) *Alcaligenes* 5) *Bacteroides* 6) *Microbacterium* 7) Alpha-proteobacterium 8) *Azoarcus* 1 9) *Rhodanobacter* 10) Unknown 11) *Thermomonas*. Samples digested with DpnII, HaeIII, and HhaI are labeled Green, Orange, and Yellow respectively.

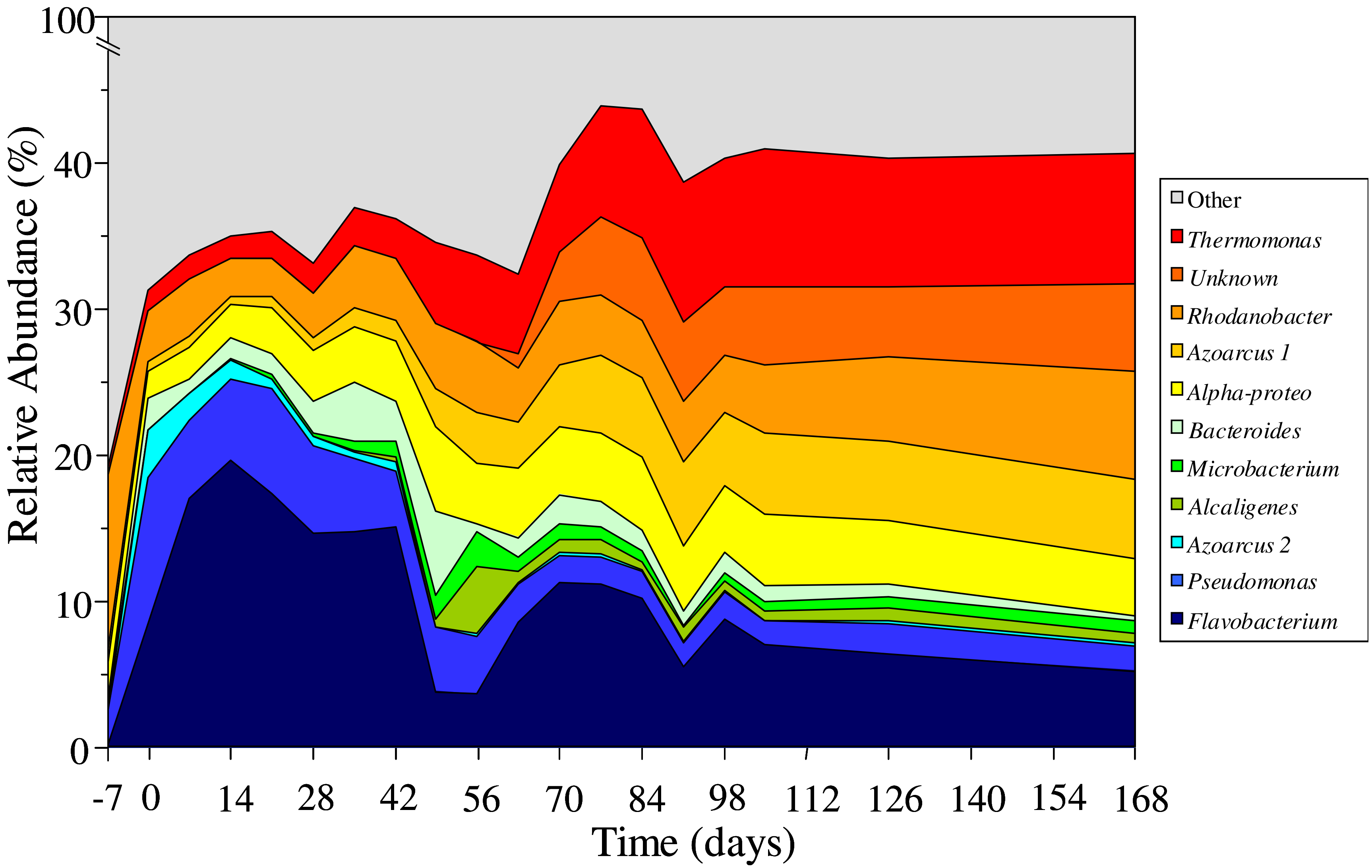


Figure 7. Relative abundance of bacterial community members based on average TRF peak areas from samples digested with *Dpn*II, *Hae*III and *Hha*I. Three phylotypes dominated the early phase of the study: Flavobacterium , Pseudomonas , and Azoarcus 2. As the petroleum content to the LTU declined so did the abundance of the early phylotypes. Day 56 witnessed a bloom in Alcaligenes and Microbacterium. The late phase of the LTU saw an increase in the abundance of Thermomonas , Unknown and Rhodanobacter Throughout the treatment Azoarcus 1 and alpha-proteobacteria were present in large numbers. Dominance of the bacterial community decreased as the LTU progressed suggesting a more evenly distributed bacterial community.

Table 3. TRF and clone data used to identify TRF sets and assign bacterial phylotypes

		TRF (base pairs)			PCA Loadings[5]		
TRF set	**Phylotype**	**Enzyme**	**Predicted[1]**	**Observed**	**PC 1**	**PC 2**	**Matching Clones**
1	Flavobacterium	*Dpn* II	253	252	0.28	-0.25	003, 019, 025, 029, 039, 043, 046, 057, 059, 063, 069, 074, 080, 083, 086, 088, 100, 108, 113, 114, 02256, 04256, 08156
		Hae III	479	473	0.38	-0.30	
		Hha I	52	49	0.21	-0.19	
2	Pseudomonas	*Dpn* II	229	228	0.18	0.14	005, 024, 00856, 01456, 07556, 08856
		Hae III	162	160	0.07	0.16	
		Hha I	169	167	0.13	0.12	
3	Azoarcus 2	*Dpn* II	241	239	0.05	0.04	016, 035, 091
		Hae III	174	173	0.01	0.03	
		Hha I	181	179	0.04	0.02	
4	Alcaligenes	*Dpn* II	160	159	-0.04	0.06	00656, 03456, 03656, 04056, 04556, 05156, 05556, 06056, 07756, 08956, 09156
		Hae III	113	111	-0.03	0.03	
		Hha I	29	ND	ND	ND	
6	Microbacterium	*Dpn* II	76	73	-0.02	0.03	07856
		Hae III	29	ND	ND	ND	
		Hha I	331	330	-0.03	0.04	
5	Bacteroides	*Dpn* II	455	451	0.00	0.14	047, 090
		Dpn II	455	453	0.01	-0.02	
		Hae III	226	225	0.01	0.11	
		Hha I	63	60	0.03	0.05	
7	Alpha-proteobacteria	*Dpn* II	203	201	-0.05	0.03	01556
		Hae III	155	154	-0.04	0.00	
		Hha I	434	430	-0.07	0.02	
8	Azoarcus 1	*Dpn* II	233	232	-0.08	0.04	002, 098, 111, 04856, 05656
		Hae III	166	167[2]	-0.10	-0.04	
		Hha I	173	172	-0.04	0.00	
9	Rhodanobacter	*Dpn* II	83	80	-0.08	-0.05	00356, 01856
		Hae III	168	167[2]	-0.10	-0.04	
		Hha I	175	173[3]	-0.20	-0.21	
10	Unknown	*Dpn* II	239[4]	237	-0.15	-0.22	NC
		Hae III	223[4]	224	-0.13	-0.20	
		Hha I	194[4]	193	-0.12	-0.18	
11	Thermomonas	*Dpn* II	235	233	-0.23	-0.08	004, 020, 078, 096, 04156, 07956, 08356
		Hae III	221	220	-0.17	-0.11	
		Hha I	175	173[3]	-0.20	-0.21	

[1] Predicted from clone sequences

[2] HaeIII 167 represents Rhodanobacter and Azoarcus 1

[3] HhaI 173 represents Thermomonas and Rhodanobacter

[4] Predicted from database sequences

[5] Standard Deviation = 0.04

NC - No clone

ND - Not detected (Predicted TRF outside detection range of 36-600 base pairs) had large positive loadings along both PC1 and PC2 indicating influence on separating day 0 from days 7 to 42 in the early samples. *Pseudomonas* TRF peak area showed a rapid decrease from an average of 10.0% on day 0 to 3.9% on day 49. The decreasing trend ended after day 49 with *Pseudomonas* returning to pretreatment levels, approximately 2% of total peak area. TRF set 3, which was associated with *Azoarcus* 2 clones had a similar trend, decreasing from 3.2% at day 0 to 0.7% by day 21 (Figure 7). Trends in TRF peak area for *Flavobacterium*, *Pseudomonas*, and *Azoarcus* 2, were similar to the trend in TPH concentration during the study. Regression analysis showed that the area of TRFs assigned to *Flavobacterium* spp. from day 7 to day 168 had a positive correlation with relative TPH concentration ($p < 0.05$). The *Pseudomonas* and *Azoarcus* 2 TRF peak areas also had positive correlations with TPH throughout the study ($p < 0.01$).

Four TRF peak sets had large abundance during the late phase of the LTU project. TRF set 11, associated with *Thermomonas* clones (Table 3), had the largest negative loadings along PC1 and large negative loadings along PC2. TRF sets 9 and 10, associated with *Rhodanobacter* clones and Unknown, also had large negative loadings along PC1 and PC2. The large negative loadings along both PCs indicate their importance in separating cluster 5 (days 63 to 168) from the rest of the samples. TRF set 8, associated with *Azoarcus* 1 clones, had negative loadings along PC1 indicating its importance to the separation of late samples from early samples. *Thermomonas* and *Rhodanobacter* TRF peak area had an increasing trend beginning with a low of 1.4% and 0.7% respectively on day 0, to a high of 9.6% and 5.7% on day 91. TRF set 10, labeled

"Unknown" since it lacked matching clones, remained undetected until day 63 (1.0%) then dramatically increased to a high of 6.1% by the end of the study. *Azoarcus* 1 TRF peak area gradually increased from a low of 1.8% at day 0 to a high of 5.7% on day 49, after which its abundance remained relatively constant. Regression analysis showed that TRF peak area associated with *Thermomonas*, Unknown, and *Rhodanobacter* had negative correlations with TPH ($p < 0.05$).

TRF sets 4 through 7 all had relatively small loadings indicating a minimal influence of these TRF sets on the creation of sample clusters. However, abundance profiles and association with clones indicated that these TRF sets represented dominant members of the community. TRF sets 4, 5 and 6, associated with *Bacteroides*, *Microbacterium*, and *Alcaligenes* clones respectively, had high abundance during the transition from early to late samples, days 28 to 56. Assigning a phylotype was most difficult with TRF set 7. Four clones had TRFs in common with two of the enzymes in TRF set 7 (021, 02856, 03356, 05256), while only one clone (01556) had all three TRFs in the set. In addition, the 5 clones did not cluster on the α-proteobacteria phylogenetic tree. This made tracking TRF set 7 difficult because inconsistent overlap within the phylotype disrupted the abundance profiles of some enzymes (Figure 6). In spite of these difficulties, TRF set 7 served as a proxy for tracking α-proteobacteria during the study. TRF set 7 was most abundant in the pretreatment sample (12.0%), but decreased dramatically to 3.4% on day 0. As the study progressed, there was a slow increase in α-proteobacteria TRF peak area, reaching 7.4% at the end of the project. The consistent presence of these organisms in the LTU attests to their pervasive nature in the contaminated soil community.

CHAPTER 5

Discussion

Bacterial Community Dynamics in Relation to Petroleum Concentration

The key elements in the land treatment process are the bacteria in the soil whose cellular machinery is responsible for bioconversion of contaminants. Previous studies have indicated that bacterial communities exposed to contamination adapt to degrade the contamination (Abed et al., 2002; MacNaughton et al., 1999). The contamination at our site existed undetected for an estimated 10 to 30 years, which gave ample time for bacterial communities to adapt and take advantage of the readily available carbon sources in petroleum. Conditions created in the LTU provided an environment conducive to the rapid degradation of available TPH. A two-phase pattern of petroleum degradation was observed in the project and is typical in contaminated soils undergoing land treatment (Admon et al., 2001; Alexander, 2000; Salanitro et al., 1997). Although the amount of TPH at this site was low when compared to other studies, Admon et al. (2001) showed that the level of contamination does not affect this two-phase pattern. Current data support the theory that the fast phase of petroleum degradation is limited by the microbial degradation rate of free TPH, while the slow phase is limited by the much slower desorption rate of soil sequestered TPH.

Dominant members of the bacterial community were tracked during the project using a clone library and TRF analysis. TRF patterns separated into five clusters that reflected the TPH degradation phases and trends in AHB counts, SI' and H' (Figures 1C, 1D, 2, and 4). Clusters 1 and 2 were associated with the fast degradation phase, increasing AHB counts, high SI' and low H' during the first 21 days of the project, indicating a bloom of fast growing petroleum degraders in the bacterial community. Clusters 3 and 4 reflected a transition out of the fast degradation phase that began with an abrupt decrease in AHB counts and SI', an abrupt increase H' and ended with a change in ambient temperature (Figures 1A and B). Cluster 5 was associated with the stable slow degradation phase where SI' was low and H' was high and AHB counts increased slowly, indicating the establishment of a stable, slowly growing community fed by the slow desorption of petroleum from the soil.

Because TPH was the only C-source available in this soil the dominant phylotypes should have some relationship to TPH degradation. Dominant bacterial phylotypes were tracked as TRF sets consisting of TRFs from each of the three enzyme digests (Table 3). TRF sets 1, 2, and 3 were most abundant during the early phase associated with clusters 1 and 2. TRF sets 4 through 7 were most abundant during the transition phase, clusters 3 and 4, although their overall abundance was low throughout the study. TRF sets 8 through 11 were most abundant during the late phase associated with cluster 5.

Significance of Bacterial Phylotypes in LTU

TRF set 1, *Flavobacterium*, was of particular interest in this study due to a high abundance during the first few weeks and a correlation with TPH levels (Figure 6). A rapid increase in bacterial counts was observed during the first 21 days, the same period as *Flavobacterium* increased in abundance, suggesting that these bacteria contributed to the increase in counts during this period. The rapid increase in *Flavobacterium* was most likely due to favorable conditions in the soil established by the addition of nutrients and aeration of the LTU. *Flavobacterium* were less abundant in pretreatment stockpile where nutrients were low and oxygen was limited. The presence of *Flavobacterium* in the community is predictable since this genus has a body of work supporting its importance in the bioremediation of hydrocarbon-contaminated soils. As one of the first reported bacterial isolates capable of degrading petroleum products, *Flavobacterium* is well known as a petroleum degrader (Atlas and Bartha, 1972). Atlas and Bartha showed that a *Flavobacterium* spp. was capable of degrading 57% of light crude oil in 12 days in an aerobic microcosm experiment. Species within the genus *Flavobacterium* have been shown to degrade petroleum and other chemicals under aerobic conditions. Recent studies have found *Flavobacteria* spp. capable of degrading Fluorobenzene (Carvalho et al., 2002), chlorinated hydrocarbons (Chaudhry and Huang, 1998; Mannisto et al., 1999), phenol (Whiteley and Bailey, 2000), polychlorinated biphenyls (Rojas-Avelizapa et al., 1999), nylon (Kato et al., 1995), and parathion (Mulbry et al., 1987). MacNaughton et al. showed an increase in *Flavobacterium* abundance during an artificial oil spill, attesting to its ability to respond to hydrocarbons in the environment (1999). These results suggest that the aerobic conditions provided by extensive tilling of the soil contributed to the stimulation of these bacteria, thus initiating an increase in counts and a decrease in TPH levels.

TRF set 2, a *Pseudomonas* phylotype including the closely related genera *Pseudomonas*, *Nitrosococcus*, *Methylococcus* and *Methylobacter*, was abundant during the fast degradation phase and correlated with TPH concentration. The presence of *Pseudomonas* in the LTU is unsurprising since this genus is well described as a petroleum degrader. Bacteria in the genera *Pseudomonas* have an ability to utilize a diverse range of substrates including those found in petroleum (Greene et al., 2000, Esteve-Nunez et al., 2001). A large amount of work has been performed on the alkane oxidation genes in *Pseudomonas*, which allow bacteria with these genes to grow on alkanes as a sole carbon source (Chakrabarty et al., 1973). *Methylococcus* and *Methylobacter* may have a role in TPH degradation, while the presence of *Nitrosococcus* may be in response to ammonia addition at the beginning of the study. The breadth of genera present in this phylotype makes identification of a specific bacterium difficult. Two clones associated with this TRF set matched well with the fluorescent pseudomonad cluster. Several TRF set 2 associated clones were linked to other genera in this phylotype although bootstrap values were poor (Figure 5C, Table 3).

TRF set 3, *Azoarcus* 2 phylotype, showed a trend similar to TRF sets 1 and 2. Evidence suggests that *Azoarcus* spp. have a role in the degradation of BTEX compounds (Fries et al., 1994; Pelz et al., 2001), suggesting they may play a role in TPH degradation. The role of *Azoarcus* 2 in the LTU is interesting since it decreased in abundance, while TRF set 8, *Azoarcus* 1, increased. It appears that these two phylotypes of *Azoarcus* have different growth requirements.

TRF sets 4 through 6, associated with *Bacteroides*, *Microbacterium* and *Alcaligenes* respectively, remained relatively stable throughout the bioremediation process except for an increase in abundance during the transition phase, (i.e. clusters 3 and 4). The stable numbers of these bacteria and lack of response to petroleum concentration, as observed in phylotypes 1 through 3, may be deceiving since these genera have been reported as petroleum hydrocarbon degraders at other contaminated sites (Greene et al., 2000). The ability of *Alcaligenes* and *Microbacterium* to degrade petroleum products have been previously documented (Lai and Khanna, 1996; Yeom and Daugulis, 2001; Greene et al., 2000; et al., 1994; Baggi and Zangrossi, 1999), but due to their low abundance, their role in degradation was probably less than *Flavobacterium* or *Pseudomonas*.

Three of the slow degradation phase, cluster 5, phylotypes were associated with either new genera (TRF sets 9 and 11, *Rhodanobacter* and *Thermomonas*, respectively) giving little or no information about their physiology or with unidentified rhizosphere bacterial sequences from GenBank (TRF set 10, Unknown). Both *Rhodanobacter* and *Thermomonas* live in aerobic conditions and have a similar phylogeny to *Pseudomonas*, *Xanthomonas* and *Stenotrophomonas*, all known hydrocarbon degraders. Nalin et al. reported isolating a strain of *Rhodanobacter lindaniclasicus* able to degrade lindane (1999). TRF set 10, Unknown, was interesting due to its strong association with the late phase of land treatment. Cloning from a sample where Unknown was more abundant would allow sequence identification for this phylotype from the LTU.

CHAPTER 6

Conclusions

The significance of this work is the description of the dynamics of dominant bacterial phylotypes correlated with the degradation of weathered petroleum hydrocarbons in the C10 to C32 range during land treatment. Based on our analysis of the bacterial community it appears that a major portion of petroleum degradation is carried out by a few species represented by *Flavobacterium* and *Pseudomonas* phylotypes. Once these bacteria had depleted the readily available free petroleum in the soil, they decreased in abundance and remained at a level were their numbers could be sustained and continued to degrade petroleum as it was slowly released from the soil particles. As the dominant petroleum degraders waned, other petroleum degraders present at lower levels increased in abundance. The phylotypes associated with slow degradation in this study represent a group of poorly described bacteria phylogenetically related to known hydrocarbon degraders, but with no previous evidence showing their ability to perform such activities.

Land treatment involved extensive soil tilling, drying/wetting events and nutrient addition during the study that likely contributed to the disruption of community structure leading to a narrowed bacterial community dominated by *Flavobacterium* and *Pseudomonas* phylotypes. Shannon-Weaver and Simpson Dominance indices corroborate this interpretation of events showing high dominance and low diversity during the fast degradation phase (Figure 2). It appears *Flavobacterium* in particular was enriched by these activities as demonstrated with its increased abundance during the first three weeks of LTU operation. The enrichment of *Flavobacterium* in the LTU may prove critical to bioremediation at our site, with a slowing or absence of petroleum degradation in their absence. In fact, preliminary data from another land treatment project at the same site indicates that low abundance of the *Flavobacterium* and *Pseudomonas* phylotypes results in slow TPH degradation.

REFERENCES

Abed, R. M. M., N. M. D. Safi, J. Köster, K. de Beer, Y. El-Nahhal, J. Rullkötter, and F. Garcia-Pichel. 2002. Microbial Diversity of a Heavily Polluted Microbial Mat and Its Community Changes following Degradation of Petroleum Compounds. Appl. Environ. Mirobiol. 68:1674-1683.

Admon, S., M. Green, Y. Avnimelech. 2001. Biodegradation Kinetics of Hydrocarbons in Soil during Land Treatment of Oily Sludge. Bioremediation Journal 5:193-209.

Al-Awadhi, N., R. Al-Daher, A. ElNavavy, and M. T. Balba. 1996. Bioremediation of Oil-Contaminated Soil in Kuwait: Landfarming to Remediate Oil-Contaminated Soil. Journal of Soil Contamination. 5:243-260.

Alexander, M. 2000. Aging, Bioavailability, and Overestimation of Risk from Environmental Pollutants. Environmental Science and Technology 34:4259-4265.

Amann, R. I., W. Ludwig., and K.-H. Schleifer. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. 59:143-169.

Atlas, R. M., and R. Bartha. 1972. Degradation and Mineralization of Petroleum by Two Bacteria Isolated from Coastal Wasters. Biotechnology and Bioengineering 14:297-308.

Avaniss-Aghajani, E., Jones, K., Chapman, D. and Brunk, C. 1994. A Molecular Technique for Identification of Bacteria Using Small Subunit Ribosomal RNA Sequences. BioTechniques 17, 144-149.

Baggi, G., and M. Zangrossi. 1999. Degradation of Chlorobenzoates in Soil Suspensions by Indigenous Populations and a Specialized Organism: Interaction between Growth and Non-growth Substrates. FEMS Microbial Ecology 29:311-318.

Braker, G., H. L. Ayala-del-Rio, A. H. Devol, A. Fesefeldt, and J. M. Tiedje. 2001. Community structure of denitrifiers, Bacteria, and Archaea along redox gradients in Pacific Northwest marine sediments by terminal restriction fragment length polymorphism analysis of amplified nitrite reductase (nirS) and 16S rRNA genes. Appl. Environ. Microbiol. 67:1893-1901.

Brunk, C. F., E. Avaniss-Aghajani, and C. A. Brunk. 1996. A Computer Analysis of Primer and Probe Hybridization Small-Subunit rRNA Sequences. Applied and Environmental Microbiology 62:872-879.

Carvalho, M. F., C. C. T. Alves, M. I. M. Ferreira, P. De Marco, and P. M. L. Castro. 2002. Isolation and Initial Characterization of a Bacterial Consortium Able To Mineralize Fluorobenzene. Applied and Environmental Microbiology 68:102-105.

Chakrabarty, A. M., G. Chou, and I. C. Gunsalus. 1973. Genetic regulation of octane dissimulation plasmids in Pseudomonas. Proc. Natl. Acad. Sci. USA 70:1137-1140.

Chaudhry, G. R., and G. H. Huang. 1998. Isolation and Characterization of a New Plasmid from a Flavobacterium sp. which Carries the Genes for Degradation of 2,4-dichlorophenoxyacetate. Journal of Bacteriology 170:3897-3902.

Cho, J-C. and J M. Teidhe. 2002. Quantitative Detection of Microbial Genes by Using DNA Microarrays. Appl. Environ. Microbiol. 68:1425-1430.

Christensen, H., M. Hansen, and J. Sørensen. 1999. Counting and size classification of active soil bacteria by fluorescence in situ hybridization with an rRNA oligonucleotide probe. Appl. Environ. Microbiol. 65:1753-1761.

Clement, B. G., L. E. Kehl, K. L. DeBord and C. L. Kitts. 1998. Terminal Restriction Fragment Patterns (TRFPs), a Rapid, PCR-Based Method for the Comparison of Complex Bacterial Communities. Journal of Microbiological Methods. 31:135-142.

Esteve-Nunez, A., A. Caballero, and J. L., Ramos. 2001. Biological Degradation of 2,4,6-Trinitrotoluene. Microbiol. Mol. Biol. Rev. 65:335-352

Farrelly, V. F. A. Rainey, and E. Stackebrandt. 1995. Effect of Genome Size and rrn Gene Copy Number on PCR Amplification of 16S rRNA Genes from a Mixture of Bacterial Species. Applied and Environmental Microbiology 61:2798-2801.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics. 5: 164-166.

Fogel, G. B., C. R. Collins, J. Li, and C. F. Brunk. 1999. Prokaryotic genome size and SSU rDNA copy number: estimation of microbial relative abundance from a mixed population. Microb Ecol. 38:93-113.

Fries, M. R., J. Zhou, J. Chee-Sanford, and J. M. Tiedje. 1994. Isolation , Characterization, and Distribution of Denitrifying Toluene Degraders from a Variety of Habitats. Appl. Environ. Microbiol. 60:2802-2810.

Greene, E. A., J. G. Kay, K. Jaber, L. G. Stehmeier, and G. Voodouw. 2000. Composition of Soil Microbial Communities Enriched on a Mixture of Aromatic Hydrocarbons. Appl. Environ. Mirobiol. 66:5282-5289.

Hill, J. E., R. P. Seipp, M. Betts, L. Hawkins, A. G. Van Kessel, W. L. Crosby, and S. M. Hemmingsen. 2002. Extensive Profiling of a Complex Microbial Community by High-Throughput Sequencing. Appl. Environ. Microbiol. 68:3055-3066.

Horn, M., M. Wagner, K.-D. Muller, E. N. Schmid, T. R. Fritsche, K.-H. Schleifer, and R. Michel. 2000. Neochlamydia hartmannellae gen. nov, sp. nov. (Parachlamydiaceae), an endoparasite of the amoeba Hartmannella vermiformis. Microbiol. 146:1231-1239.

Huesemann, M. H. and M. J. Truex. 1996. The Role of Oxygen Diffusion in Passive Bioremediation of Petroleum Contaminated Soils. Journal of Hazardous Materials. 51:93-113.

Kaplan, C. W., J. C. Astaire, M. E. Sanders, B. S. Reddy, and C. L. Kitts. 2001. 16S Ribosomal DNA Terminal Restriction Fragment Pattern Analysis of Bacterial Communities in Feces of Rats Fed Lactobacillus acidophilus NCFM. Applied and Environmental Microbiology 67:1935-1939.

Kato, K., K. Ohtsuki, Y. Koda, T. Maekawa, T. Yomo, S. Negoro, and I. Urabe. 1995. A Plasmid Encoding Enzymes for Nylon Oligomer Degradation: Nucleotide Sequence and Analysis of pOAD2. Microbiology 141:2585-2590.

Kitts, C. L. 2001. Terminal Restriction Fragment Patterns: A Tool for Comparing Microbial Communities and Assessing Community Dynamics. Current Issues in Intestinal Microbiology 2:17-25.

Lai, B., and S. Khanna. 1996. Degradation of Crude Oil by Acintobacter calcoaceticus and Alcaligenes oderans. Journal of Applied Bacteriology. 81:355-362.

Liu, W., T. L. Marsh, H. Cheng and L. Forney. 1997. Characterization of Microbial Diversity by Determining Terminal Restriction Fragment Length Polymorphisms of Genes Encoding 16S rRNA. Applied and Environmental Microbiology. 63:4516-4522.

MacNaughton, S. J., J. R. Stephen, A. D. Venosa, G. A. Davis, Y. Chang and D. C. White. 1999. Microbial Population Changes during Bioremeditaion of an Experimental Oil Spill. Applied and Environmental Microbiology. 65:3566-3574.

Maidak B. L., J. R. Cole, T. G. Lilburn, C. T. Parker Jr., P. R. Saxman, R. J. Farris, G.M. Garrity, G. J. Olsen, T. M. Schmidt, and J. M. Tiedje. 2001. The RDP-II (Ribosomal Database Project). Nucleic Acids Research 29:173-174.

Mannisto, M. K., M. A. Tiirola, M. S. Salkinoja-Salonen, M. S. Kulomaa, J. A. Puhakka. 1999. Diversity of Chlorophenol-Degrading Bacteria Isolated from Contaminated Boreal Groundwater. Archives of Microbiology 171:189-97.

Marsh, T. L. 1999. Terminal Restriction Fragment Length Polymorphism (T-RFLP): an Emerging Method for Characterizing Diversity Among Homologous Populations of Amplification Products. Current Opinion in Microbiology 2:323-327.

Martin-Laurent, F., L. Philippot, S. Hallet, R. Chaussod, J. C. Germon, F. Soulas, and G. Catroux. 2001. DNA extraction from soils: old bias for new microbial diversity analysis methods. Appl. Environ. Mirobiol. 67:2354-2359.

Moeseneder, M. M., C. Winter, J. M. Arrieta and G. J. Herndl. 2001 Terminal-restriction fragment length polymorphism (T-RFLP) screening of a marine archaeal clone library to determine the different phylotypes. J. Microbiol Methods. 44:159-172.

Mulbry, W. W., P. C. Kearney, J. O. Nelson, J. S. Karns. 1987. Physical Comparisons of Parathion Hydrolase Plasmids from Pseudomonas diminuta and Flavobacterium sp. Plasmid 18:173-177.

Nalin, R., P. Simonet, T. M. Vogel, and P. Normand. 1999. Phodanobacter lindaniclasticus gen. Nov., sp. nov., a lindane-degrading bacterium. 49:19-23.

Olivera, F. L., R. C. Loehr, B. C. Coplin, H. Eby, M. T. Webster. 1998. Prepared Bed Land treatment of Soils Containing Diedel and Crude Oil Hydrocarbons. Journal of Soil Contamination 7:657-674.

Osborn, A. M., E. R. B. Moore, and K. N. Timmis. 2000. An Evaluation of Terminal-Restiction Fragment Length Polymorphism (T-RFLP) Analysis for the Study of Microbial Community Structure and Dynamics. Environmental Microbiology 2:39-50.

Pelz, O., A. Chatzinotas, N. Anderson, S. M. Bernasconi, C. Hesse, W.-R. Abraham, and J. Zeyer. 2001. Use of Isotopic and Molecular Techniques to Link Toluene Degradation in Denitrifying Aquifer Microcosms to Specific Microbial Populations. Arch. Microbiol. 175:270-281.

Rojas-Avelizapa, N. G., R. Rodriguez-Vazquez, R. Enriquez-Villanueva, J. Martinez-Cruz, and H. M. Poggi-Varaldo. 1999. Transformer Oil Degradation by an Indigenous Microflora Isolated from a Contaminated Soil. Resouces, Conservation and Recycling 27:15-26.

Sakano, Y., K. D. Pickering, P. F. Strom, and L. J. Kerkhof. 2002. Spatial Distribution of Total, Ammonia-Oxidizing, and Denitrifying Bacteria in Biological Wastewater Treatment Reactors for Bioregenerative Life Support. Appl. Environ. Microbiol. 68: 2285-2293.

Salanitro, J. P., P. B. Dorn, M. H. Huesemann, K. O. Moore, I. A. Rhodes, L. M. Rice Jackson, T. E. Vipond, M. M. Western, and H. L. Wisniewski. 1997. Crude Oil Hydrocarbon Bioremediation and Soil Ecotoxicity Assessment. Environmental Science and Technology. 31:1769-1776.

Small, J., D. R. Call, F. J. Brockman, T. M. Straub, and D. P. Chandler. 2001. Direct detection of 16S rRNA in soil extracts by by using oligonucleotide microarrays. Appl. Environ. Microbiol. 67:4708-4716.

Suzuki, M. T., and S. J. Giovannoni. 1996. Bias Caused by Template Annealing in the Amplification of Mixtures of 16S rRNA Genes by PCR. Applied and Environmental Microbiology 62:625-630.

Thompson, J. D., Higgins, D. G. and Gibson, T. J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673-4680.

Venosa, A. D., M. T. Suidan, B. A. Wrenn, K. L. Strohmeier, J. R. Haines, B. L. Eberhart, D. King and E. Holder. 1996. Bioremediation of an Experimental Oil Spill on the Shoreline of Deleware Bay. Environmental Science and Technology. 30:1764-1775.

Wang, G. C.-Y., and Y. Wang. 1997. Frequency of Formation of Chimeric Molecules as a Consequence of PCR Coamplification of 16S rRNA Genes form Mixed Bacterial Genomes. Applied and Environmental Microbiology 63:4645-4650.

Whiteley A. S., and M. J. Bailey. 2000. Bacterial Community Structure and Physiological State within an Industrial Phenol Bioremediation System. Applied and Environmental Microbiology 66:2400-2407.

Wu, Liyou, D. K. Thompson, G. Li, R. A. Hurt, J. M. Tiedje, and J. Zhou. 2001. Development and evaluation of functional gene arrays for detection of selected genes in the environment. Appl. Environ. Microbiol. 67:5780-5790.

Yeom, S-H., and A. J. Daugulis. 2001. Benzene Degradation in a Two-Phase Partitioning Bioreactor by Alcaligenes xylosoxidans Y234. Process Biochemistry 36:765-772.

APPENDIX A

Variation between Observed Terminal Restriction Fragments is Dependent on True TRF Length and Sequence Composition

Abstract

Terminal Restriction Fragment (TRF) pattern analysis has become a widely used and informative tool for studying microbial communities. A variance between sequence-determined TRF length and observed TRF length, referred to here as “TRF drift”, has been previously reported. TRF drift was determined for 26 bacterial species on an ABI 310 Genetic Analyzer. TRF drift increased with increasing TRF length and was significantly correlated with DNA sequence composition and TRF length.

As environmental microbiology has evolved, so have the techniques employed. The use of molecular methods to describe microorganisms and the communities they comprise have become commonplace. A recently developed tool in environmental microbiology is Terminal Restriction Fragment (TRF, a.k.a. Terminal Restriction Fragment Length Polymorphism or T-RFLP) pattern analysis. TRF patterns are produced by amplifying DNA from a bacterial community using one fluorescently labeled PCR primer and cutting the amplicons with a restriction endonuclease. Organisms in a pattern are thus differentiated based on sequence variation that results in TRFs of different lengths, which in turn create a pattern unique to that community. The resulting patterns can be used to make inferences about environmental effects on community structure or evaluate community dynamics. Several comprehensive reviews of the TRF method exist which illustrate the utility of this tool (Kitts, 2001; Marsh, 1999). An increasingly popular trend in TRF analysis studies has been to associate TRF peaks with clones or predicted matches from extensive databases of existing sequences (Braker et al., 2001; Kaplan et al., 2001; Moeseneder et al., 2001; Sakano et al., 2002). Associating sequenced clones or database matches with a TRF peak is problematic since related organisms commonly produce TRFs of the same length, requiring several enzyme digests to resolve community members. To accurately make matches requires that TRFs in a pattern migrate in such a way that their reported length accurately represents their true length. Discrepancies between sequence-determined TRF length and observed TRF length have been reported previously with estimates ranging from as little as one basepair to as much as seven basepairs (Kitts, 2001; Kaplan et al., 2001; Liu et al., 1997; Clement et al., 1998; Osborn et al., 1998). In this paper we evaluated the discrepancy between true and observed TRF lengths from the 16S rDNA region of 26 bacterial strains in an effort to quantify sources of variation and achieve more accurate database matches.

The organisms used in this study were picked from cultures available in our lab based on true TRF length and GC content of sequence (Table A1). All organisms were streaked on Tripticase Soy Agar and incubated at optimum temperature and time to provide sufficient growth for DNA extraction. Cells were then scraped from plates and transferred to MoBio® bead lysis tubes (Solano Beach, CA). The protocol given in the Mo Bio® kit was followed for the extraction process with the following exception: cells were lysed in the Bio 101 FP-120 FastPrep machine (Carlsbad, CA) running at 4.5 m/s for 25 seconds. The isolated DNA was visualized by agarose gel electrophoresis and quantified by UV spectrophotometry. Amplification of template DNA was performed by using primers 6-FAM labeled 46f (5'-GCYTAACACATGCAAGTCGA), and unlabeled 536r (5'-GTATTACCGCGGCTGCTGG). Reactions were carried out in duplicate with the following reagents in 50 µl reactions: template DNA, 10 ng; 1X Buffer (Applied Biosystems Inc., Foster City, CA); dNTPs, $3x10^{-5}$ mmols; bovine serum albumin, $4x10^{-2}$ µg; $MgCl_2$, $1.75x10^{-4}$ mmols; 46f, $1x10^{-5}$ mmols; 536r, $1x10^{-5}$ mmols; *Taq*Gold DNA polymerase (Applied Biosystems), 1.5U. Reaction temperatures and cycling for samples were as follows: 94°C for 10 min, 35 cycles of 94°C for 2 min, 46.5°C for 1 min, 72°C for 1 min, followed by 72°C for 10 min. The products were visualized on a 1.5% agarose gel and any inconsistent or unsuccessful reactions were discarded. To remove primers and concentrate amplicons, the Mo Bio® PCR Clean-Up kit was utilized according to the protocol included with the kit. The combined amplicons were then quantified by UV spectrophotometry. Restriction enzyme reactions contained 10 ng of labeled DNA,

Table A1. TRF drift bacteria predicted TRF length, drift and %R for each restriction enzyme used.

				*Dpn*II			*Rsa*I			*Msp*I			*Hha*I	
Organism name:	**Accession #**	**Group**	**TRF**	**Drift**	**%R**	**TRF**	**Drift**	**%R**	**TRF**	**Drift**	**%R**	**TRF**	**Drift**	**%R**
Bacillus licheniformis		1	266	-3.50	0.56	419	-7.76	0.57	107	-4.15	0.56	*203*	-4.36	0.55
Bacillus megaterium		1	265	-2.69	0.57	*448*	-5.91	0.58	107	-3.66	0.57	*202*	-3.62	0.56
Bacillus pumilis		1	265	-3.30	0.57	418	-7.74	0.58	107	-3.65	0.57	*202*	-4.29	0.56
Bacillus subtilis		1	266	-3.56	0.56	419	-7.77	0.57	107	-4.15	0.56	*203*	-3.94	0.55
Staphylococcus warneri		1	192	-1.36	0.57	*448*	-5.80	0.57	117	-3.57	0.58	*200*	-1.56	0.57
Citrobacter freundii		2	*235*	-1.23	0.59	*389*	-4.76	0.59	*458*	-6.53	0.59	*335*	-1.55	0.59
Enerobacter aerogenes		2	*233*	-1.20	0.59	*387*	-3.66	0.59	*456*	-6.06	0.59	*333*	-1.70	0.59
Enterobacter aerogenes		2	*233*	-1.26	0.59	*387*	-4.30	0.59	*456*	-6.59	0.59	*333*	-1.84	0.59
Enterobacter cloacae		2	*233*	-0.86	0.59	*387*	-4.08	0.59	*456*	-6.91	0.59	*333*	-1.32	0.59
Escherichia coli		2	*235*	-1.34	0.59	*389*	-4.58	0.58	*458*	-7.01	0.58	*335*	-1.80	0.59
Escherichia coli		2	*235*	-0.83	0.58	*389*	-4.93	0.58	*458*	-7.19	0.58	*335*	-1.21	0.59
Klebsiella pneumoniae		2	*233*	-1.36	0.59	**UC**	**NA**	0.58	*456*	-7.47	0.59	*333*	-1.73	0.59
Salmonella enteritidis		2	*235*	-2.00	0.58	*389*	-5.45	0.58	*458*	-7.32	0.58	*335*	-2.77	0.59
Salmonella typhimurium		2	*235*	-1.98	0.59	*389*	-4.73	0.58	*458*	-7.49	0.59	*335*	-3.09	0.59
Salmonella typhimurium		2	*235*	-2.13	0.58	*389*	-4.61	0.58	*458*	-6.84	0.58	*335*	-2.95	0.59
Serratia marcescens		2	*235*	-1.69	0.59	**UC**	**NA**	0.58	*458*	-7.32	0.59	*335*	-2.21	0.59
Shigella sonnei		2	*235*	-1.00	0.59	*389*	-4.90	0.58	*458*	-7.08	0.58	*335*	-1.45	0.59
Enterococcus hirae		3	277	-1.25	0.56	**UC**	**NA**	0.57	**ND**	**NA**	0.39	*180*	-1.59	0.57
Lactobacillus acidophilus		3	157	-1.56	0.60	**UC**	**NA**	0.58	*143*	-2.78	0.58	**UC**	**NA**	0.58
Lactobacillus casei		3	*284*	-1.91	0.56	**UC**	**NA**	0.56	**UC**	**NA**	0.56	**UC**	**NA**	0.56
Lactobacillus johnsonii		3	*288*	-1.30	0.57	**UC**	**NA**	0.58	*151*	-2.73	0.58	**UC**	**NA**	0.58
Lactobacillus murinus		3	281	-3.02	0.57	**UC**	**NA**	0.57	**UC**	**NA**	0.57	*218*	-3.40	0.56
Lactobacillus paracasei		3	**ND**	**ND**	0.50	**UC**	**NA**	0.56	**UC**	**NA**	0.56	**UC**	**NA**	0.56
Proteus mirabilis		4	83	-2.07	0.60	*389*	-2.39	0.58	*458*	-5.16	0.58	*335*	0.15	0.59
Proteus mirabilis		4	83	-2.18	0.60	*389*	-2.96	0.58	*458*	-5.45	0.58	*335*	0.13	0.59
Proteus vulgaris		4	83	-2.08	0.60	*389*	-2.79	0.58	*458*	-5.73	0.58	175	-0.36	0.58

ND – TRF outside detection limit of 36 – 600 basepairs.
UC – Uncut TRF, not included in dataset.

and restriction endonuclease enzyme (*Hha*I, 0.1 Units; or *Msp*I, 0.1 Units; *Rsa*I, 0.2 Units; or *Dpn*II, 0.2 Units; or *Hae*III, 0.2 Units (New England Biolabs, Beverly, MA. USA) in the manufacturer's recommended reaction buffers. Reactions were digested for 2 hours at 37°C. Samples were ethanol precipitated then dissolved in 9 µl of Hi-DI formamide (Applied Biosystems), with 0.5 µl each of Genescan Rox 500 (Applied Biosystems) and Rox 600 (BioVentures, Murfreesboro, TN) size standards. The DNA was denatured at 95°C for 5 minutes and snap-cooled in an ice slurry for 10 minutes. Samples were run on an ABI Prism™ 310 Genetic Analyzer at 15 kV and 60°C. TRF sizing was performed on electropherogram output from Genescan™ 3.1.2 software using Local Southern method with heavy smoothing. For DNA sequencing, extracted DNA samples were amplified by PCR as described above except that the forward and reverse primers were replaced with 8df (5'-AGAGTTTGTTCMTGGCTCAG) and 803r (5'-CTACCAGGGTATCTAATCC). Sequencing reactions (10µl) contained: DNA, 4ng; primer, 1.6e10^{-5} mmol; ABI Big Dye (Perkin Elmer), 4µl; PCR water, 0.4µl. Samples were run on an ABI 377 DNA sequencer and the resulting sequences analyzed in SeqMan™ (DNAStar, Madison, WI). Sequences were analyzed for TRF cut sites of each enzyme used in this study for comparison with TRF pattern data.

TRF data were analyzed using five different analysis methods (2nd order least square, 3rd order least square, local southern, global southern, cubic spline) available with Genescan 3.1.2 software. Different analysis methods produced different standard curves for the internal ladder, thus creating differences in observed TRF length. As a previous report has shown (Osborn et al., 1998), the local southern method produced a standard curve with the least variability between true sequence determined and observed TRF lengths (data not shown). To facilitate a statistical analysis we defined "TRF drift" as the observed TRF length minus the true TRF length. Amplicons that did not contain an enzyme cut site, resulting in an uncut TRF, were not included in this dataset. This is because *Taq* polymerase can add a 3' adenine residue to a PCR product resulting in a longer fragment than predicted from the sequence.

The average TRF drift was approximately –3 basepairs over the lengths analyzed, with a standard deviation of 1.26 basepairs. Longer TRFs had larger TRF drift associated with them (Figure A1). The trend in TRF drift was similar among related bacteria suggesting that sequence composition may affect TRF drift. *Proteus* spp. had the least TRF drift at any length (~2 basepairs), while *Bacillus* spp. had the most TRF drift (~4 basepairs). Purine content was negatively correlated with TRF drift ($p < 0.001$). Purine content across the entire dataset was 58% (+/– 2%). TRFs from *Proteus* spp. had an average purine content of 59% (+/– 1%) while *Bacillus* spp. had an average purine content of 57% (+/– 1%).

Analysis of electropherogram data suggested that TRF drift resulted from two sources of variation: differential migration of ladder and sample DNA and sequence composition. Differential migration is the variation between the internal ROX-labeled ladder and the 6-FAM-labeled sample DNA. The effect of this dissimilar migration manifested itself as progressively shorter observed TRFs as retention time in the capillary increased. In fact, fragment analysis software from some manufacturers automatically compensates for differential dye migration (Beckman Coulter). Alternatively, this source of variation can also be corrected by using the equations below. In this dataset, TRF drift was most accurately predicted using equation 1.

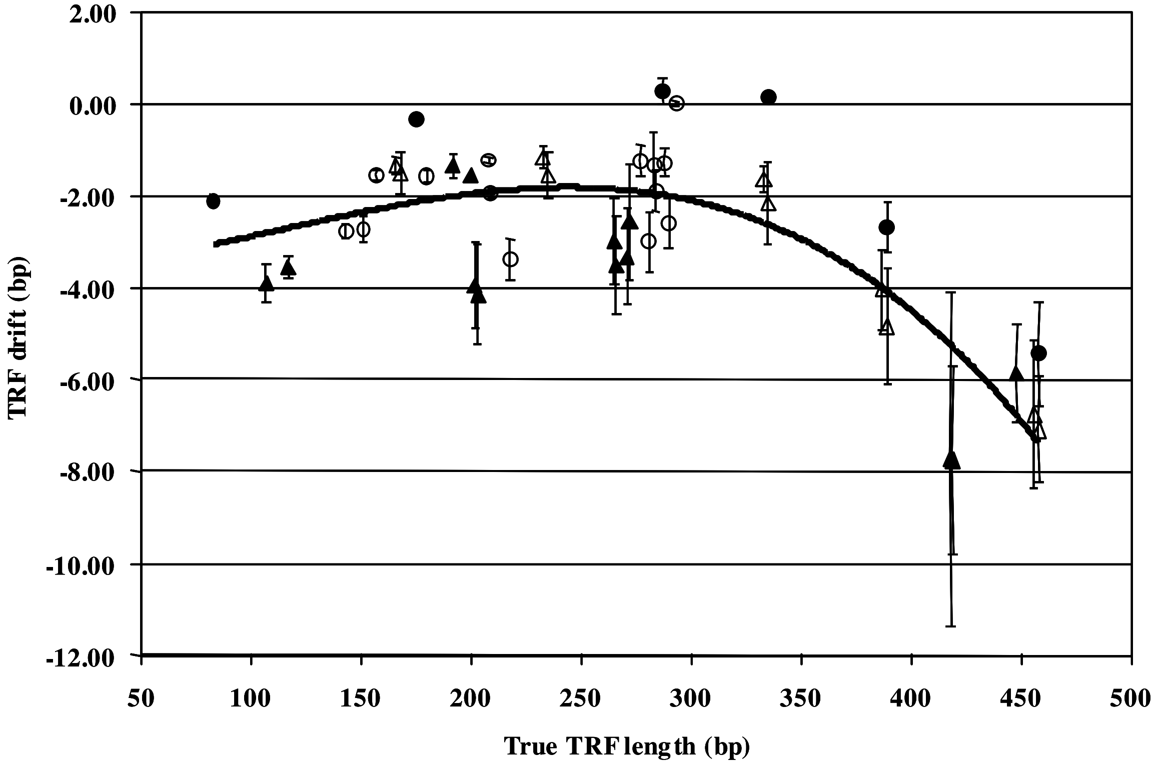


Figure A1. TRF drift for different phylogenetic groups used in the study. *Proteus* spp. had the least drift of any group at any length. Enterics and *Lactobacillus* spp. had similar TRF drift, while *Bacillus* spp. species had the most drift of any species used in this study. A 3rd order fit is indicated by solid line. *Proteus* spp., closed circle; Enterics, open circle, *Lactobacillus* spp., open triangle; *Bacillus* spp., closed triangle.

1) TRF Drift = $-2.24\text{x}10^{-7}$(Observed TRF length)^3 + $8.15\text{x}10^{-5}$(Observed TRF length)^2 + $1.39\text{x}10^{-3}$(Observed TRF length) - 3.48

2) Adjusted TRF length = Observed TRF length – TRF Drift

Retention time alone only accounted for 65% of the variation in TRF drift. An additional 6% of the variation could be account for by incorporating purine content of the TRFs into the analysis. This source of variation was most obvious among organisms with the same true TRF length, but different observed TRF lengths (Figure A1). Each group of organisms had a trend in TRF drift that was displaced from the average overall TRF drift by a constant amount. Secondary structure was not a likely candidate since fragment analysis was performed at 60°C in a denaturing gel matrix. However, purine content clearly affected TRF drift. A 1% difference in average purine content resulted in a 1 basepair shift in average TRF drift for both *Proteus* spp. and *Bacillus* spp. This implies that subtle differences in molecular weight can significantly alter the observed TRF length.

Machine variation was observed in this dataset and manifested itself in the form of variation between observed TRF lengths in replicate runs. The primary cause of this variation was attributed to fluctuations in ambient temperature during runs. While not strictly a source of TRF drift, machine variation resulted in an alarming fluctuation in observed TRF length of up to 5 basepairs (Figure A1) suggesting that this source of variation could have unpredictable effects if proper care is not taken to maintain a constant lab temperature. The amount of TRF drift may differ on other machines.

Using the equation and recommendations presented here it is possible to minimize the effects of TRF drift. However, a certain amount of drift between true and observed TRF lengths will remain. This means that matching observed TRF peaks to database predicted TRFs should include a window of +/– 1 bp. A more liberal window of +/– 2 bp could be used but since this would increase the number of matches and include more inaccurate matches caution should be taken when interpreting matches with a large window. Multiple enzyme digests could be used to sufficiently narrow the number of matches and facilitate more accurate database matching.

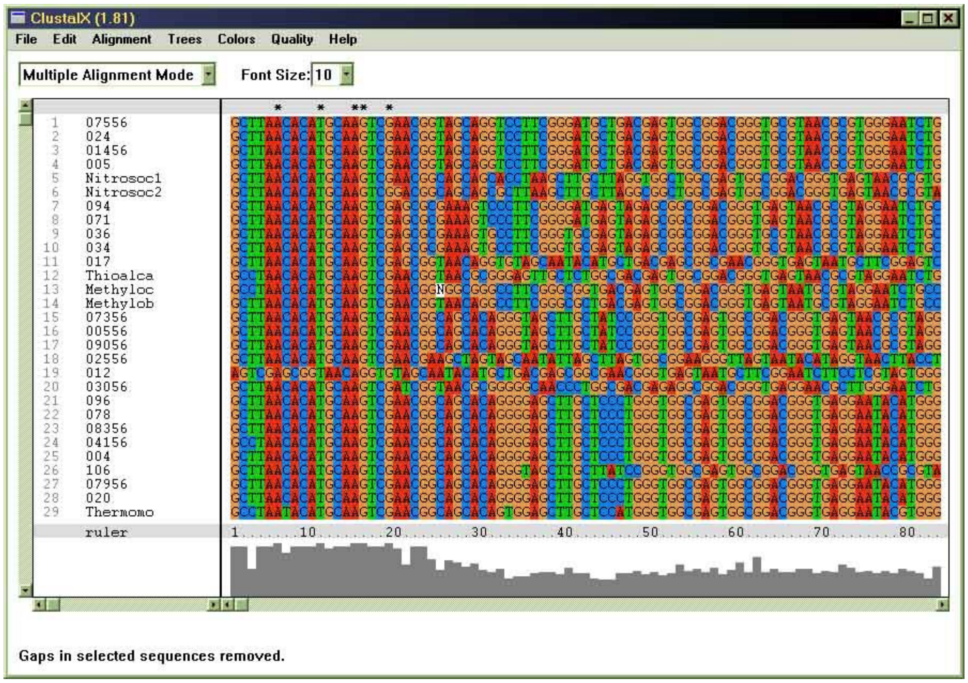
APPENDIX B

Phylogenetic analysis using Phylip v3.6a2.1

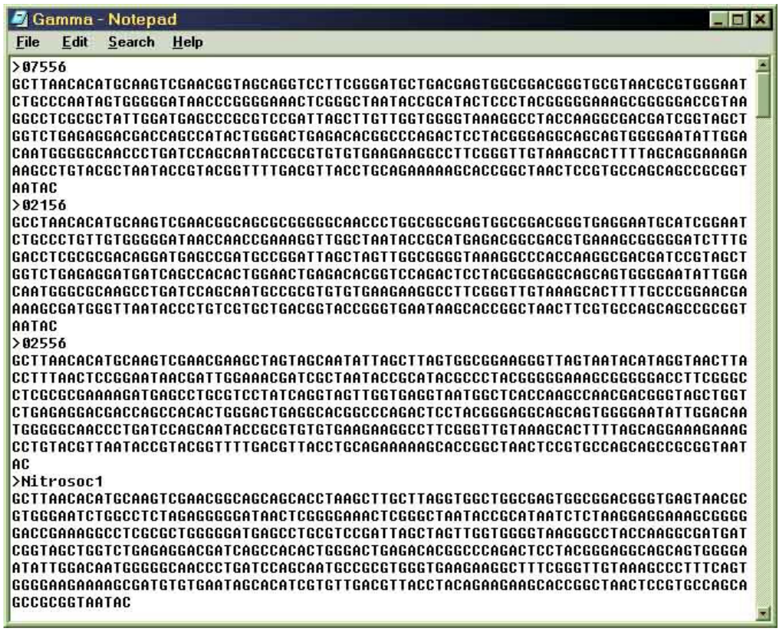
by Chris Kaplan

Disclaimer – This is not meant to be a thorough explanation of the theory and methods of tree building, rather a guide. Building an accurate tree, and understanding and interpreting it correctly require knowledge of systemactics theories and methods. Trees resulting from inappropriate analysis will be misleading and incorrect. Further information on all programs and settings can be found in documentation files supplied with the various programs.

1. Load sequences from a text document into ClustalX from the **File → Load Sequences** menu. Sequences should be in FASTA format as shown below. Loaded sequence will be unaligned when they are entered into ClustalX.

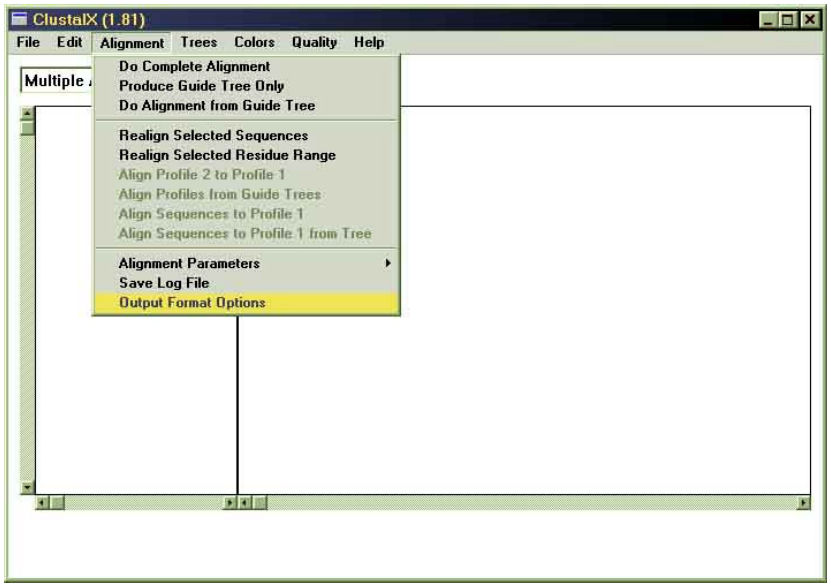


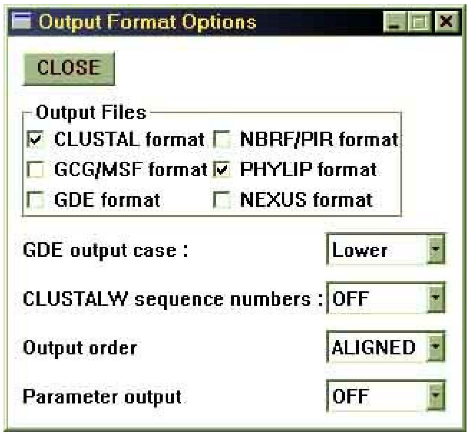
Gaps in selected sequences removed.



```
>07556
GCTTAACACATGCAAGTCGAACGGTAGCAGGTCCTTCGGGATGCTGACGAGTGGCGGACGGGTGCGTAACGCGTGGGAAT
CTGCCCAATAGTGGGGGATAACCCGGGGAAACTCGGGCTAATACCGCATACTCCCTACGGGGGAAAGCGGGGGACCGTAA
GGCCTCGCGCTATTGGATGAGCCCGCGTCCGATTAGCTTGTTGGTGGGGTAAAGGCCTACCAAGGCGACGATCGGTAGCT
GGTCTGAGAGGACGACCAGCCATACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGGA
CAATGGGGGCAACCCTGATCCAGCAATACCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTTAGCAGGAAAGA
AAGCCTGTACGCTAATACCGTACGGTTTTGACGTTACCTGCAGAAAAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGT
AATAC
>02156
GCCTAACACATGCAAGTCGAACGGCAGCGCGGGGGCAACCCTGGCGGCGAGTGGCGGACGGGTGAGGAATGCATCGGAAT
CTGCCCTGTTGTGGGGGATAACCAACCGAAAGGTTGGCTAATACCGCATGAGACGGCGACGTGAAAGCGGGGGATCTTTG
GACCTCGCGCGACAGGATGAGCCGATGCCGGATTAGCTAGTTGGCGGGGTAAAGGCCCACCAAGGCGACGATCCGTAGCT
GGTCTGAGAGGATGATCAGCCACACTGGAACTGAGACACGGTCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGGA
CAATGGGCGCAAGCCTGATCCAGCAATGCCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTTGCCCGGAACGA
AAAGCGATGGGTTAATACCCTGTCGTGCTGACGGTACCGGGTGAATAAGCACCGGCTAACTTCGTGCCAGCAGCCGCGGT
AATAC
>02556
GCTTAACACATGCAAGTCGAACGAAGCTAGTAGCAATATTAGCTTAGTGGCGGAAGGGTTAGTAATACATAGGTAACTTA
CCTTTAACTCCGGAATAACGATTGGAAACGATCGCTAATACCGCATACGCCCTACGGGGGAAAGCGGGGGACCTTCGGGC
CTCGCGCGAAAAGATGAGCCTGCGTCCTATCAGGTAGTTGGTGAGGTAATGGCTCACCAAGCCAACGACGGGTAGCTGGT
CTGAGAGGACGACCAGCCACACTGGGACTGAGGCACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGAATATTGGACAA
TGGGGGCAACCCTGATCCAGCAATACCGCGTGTGTGAAGAAGGCCTTCGGGTTGTAAAGCACTTTTAGCAGGAAAGAAAG
CCTGTACGTTAATACCGTACGGTTTTGACGTTACCTGCAGAAAAAGCACCGGCTAACTCCGTGCCAGCAGCCGCGGTAAT
AC
>Nitrosoc1
GCTTAACACATGCAAGTCGAACGGCAGCAGCACCTAAGCTTGCTTAGGTGGCTGGCGAGTGGCGGACGGGTGAGTAACGC
GTGGGAATCTGGCCTCTAGAGGGGGATAACTCGGGGAAACTCGGGCTAATACCGCATAATCTCTAAGGAGGAAAGCGGGG
GACCGAAAGGCCTCGCGCTGGGGGATGAGCCTGCGTCCGATTAGCTAGTTGGTGGGGTAAGGGCCTACCAAGGCGATGAT
CGGTAGCTGGTCTGAGAGGACGATCAGCCACACTGGGACTGAGACACGGCCCAGACTCCTACGGGAGGCAGCAGTGGGGA
ATATTGGACAATGGGGGCAACCCTGATCCAGCAATGCCGCGTGGGTGAAGAAGGCTTTCGGGTTGTAAAGCCCTTTCAGT
GGGGAAGAAAAGCGATGTGTGAATAGCACATCGTGTTGACGTTACCTACAGAAGAAGCACCGGCTAACTCCGTGCCAGCA
GCCGCGGTAATAC
```

2. A multiple sequence alignment can be performed in ClustalX. From the **Alignment → Output Format Options** menu, check **CLUSTAL format** and **Phylip format** selected as the output file types. The CLUSTAL format can be reopened in ClustalX and sequences realigned if necessary. Phylip format will be used in phylogenetic analysis using Phylip.

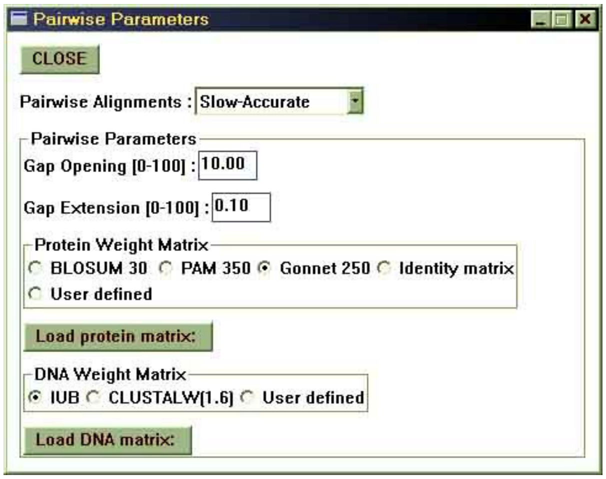

ClustalX (1.81)
File Edit Alignment Trees Colors Quality Help
Multiple
Do Complete Alignment
Produce Guide Tree Only
Do Alignment from Guide Tree
Realign Selected Sequences
Realign Selected Residue Range
Align Profile 2 to Profile 1
Align Profiles from Guide Trees
Align Sequences to Profile 1
Align Sequences to Profile 1 from Tree
Alignment Parameters
Save Log File
Output Format Options



Output Files
☑ CLUSTAL format ☐ NBRF/PIR format
☐ GCG/MSF format ☑ PHYLIP format
☐ GDE format ☐ NEXUS format

GDE output case : Lower
CLUSTALW sequence numbers : OFF
Output order ALIGNED
Parameter output OFF

3. From the Alignment → Alignment Parameters → Pairwise Alignment Parameters menu, change the Gap Opening and Gap Extension values. Make sure the setting for Pairwise Alignments is Slow-Accurate, which is the default setting. Also, from the Alignment → Alignment Parameters → Multiple Alignment Parameters menu, change the Gap Opening and Gap Extension values. Pairwise and multiple sequence alignment parameters should be changed together and be the same values. Reasonable alignment parameters vary with different sequence data, but reasonable starting values for Gap Opening and Gap Extension are 10 and 6 respectively. Vary these parameters by 1 or two units up or down to achieve an optimal alignment. Large Gap Opening values will result in fewer gaps, while small values will result more gaps. Large Gap Extension values will result in smaller gaps, while small values will result larger gaps. An optimal alignment is one that: follows current theories of systematics (i.e. closely related species align together better than with less closely related species), and changing alignment parameters may results in no change to alignment (i.e. same gaps and gaps lengths). The basic idea is to make an alignment in which related nucleotide are aligned. Related nucleotides are nucleotide



Pairwise Alignments : Slow-Accurate

Pairwise Parameters

Gap Opening [0-100] : 10.00

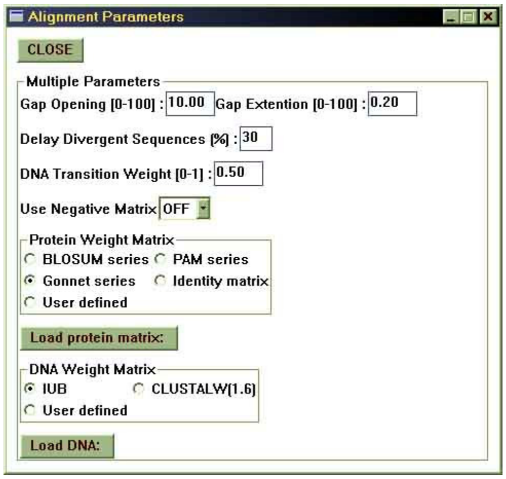
Gap Extension [0-100] : 0.10

Protein Weight Matrix

○ BLOSUM 30 ○ PAM 350 ● Gonnet 250 ○ Identity matrix
○ User defined

DNA Weight Matrix

● IUB ○ CLUSTALW(1.6) ○ User defined



Multiple Parameters

Gap Opening [0-100] : 10.00 Gap Extention [0-100] : 0.20

Delay Divergent Sequences (%) : 30

DNA Transition Weight [0-1] : 0.50

Use Negative Matrix OFF

Protein Weight Matrix

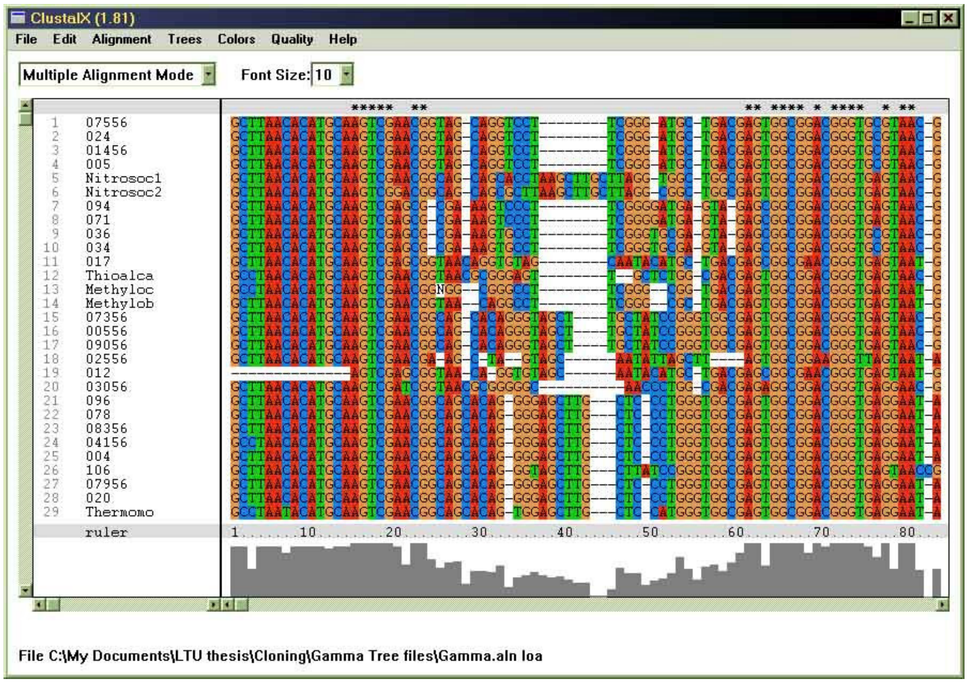
○ BLOSUM series ○ PAM series
● Gonnet series ○ Identity matrix
○ User defined

DNA Weight Matrix

● IUB ○ CLUSTALW(1.6)
○ User defined

that share the same evolutionary history, so should therefore be present in the same column of a multiple sequence alignment. Some subjective decision-making is necessary in this process, but in general, it should be relatively easy for a discerning eye to recognize correctly and incorrectly aligned sequences. In some cases it will not be possible to align every base as desired, so do not lament over a few misaligned bases if you are dealing with a few hundred bases in your alignment unless the species have very little variation between them (i.e. only a few bases are different).

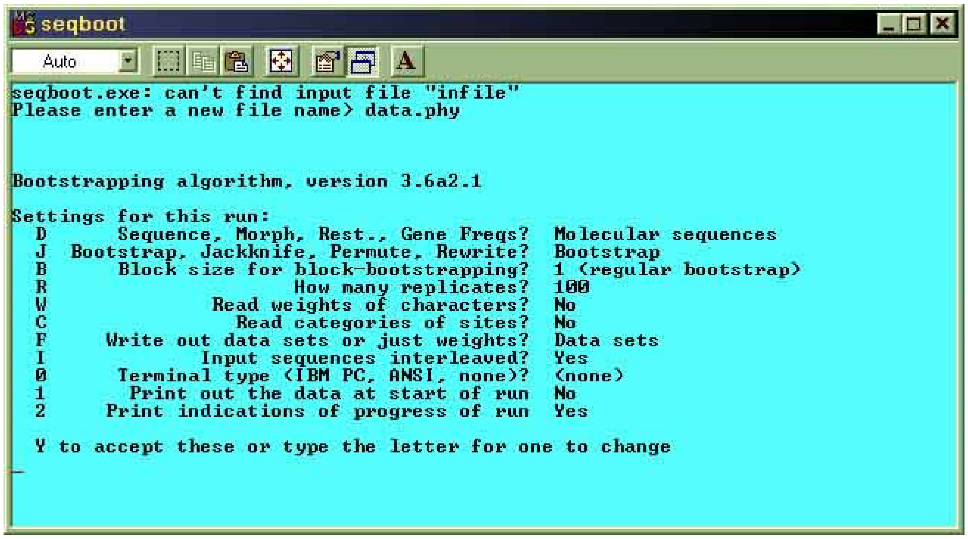
4. After alignment parameters have been selected, choose **Alignment → Do Complete Alignment**. The alignment will proceed until each sequence has been compared to every other sequence and aligned to each other. The completed alignment will appear when the alignment process is finished. Assess the alignment and adjust alignment parameters as necessary. It will be helpful to save each alignment with a different name stating the alignment parameters used (e.g. "name-10-6" for an alignment using a Gap Opening of 10 and a Gap Extension of 6). It will be necessary to try several



different alignment parameters settings to make an accurate assessment of an optimal alignment. This process can be extremely time consuming, but is the critical to accurate phylogenetic analysis since all resulting analyses produce results from this alignment. An inappropriate or inaccurate alignment will result in an incorrect and misleading tree.

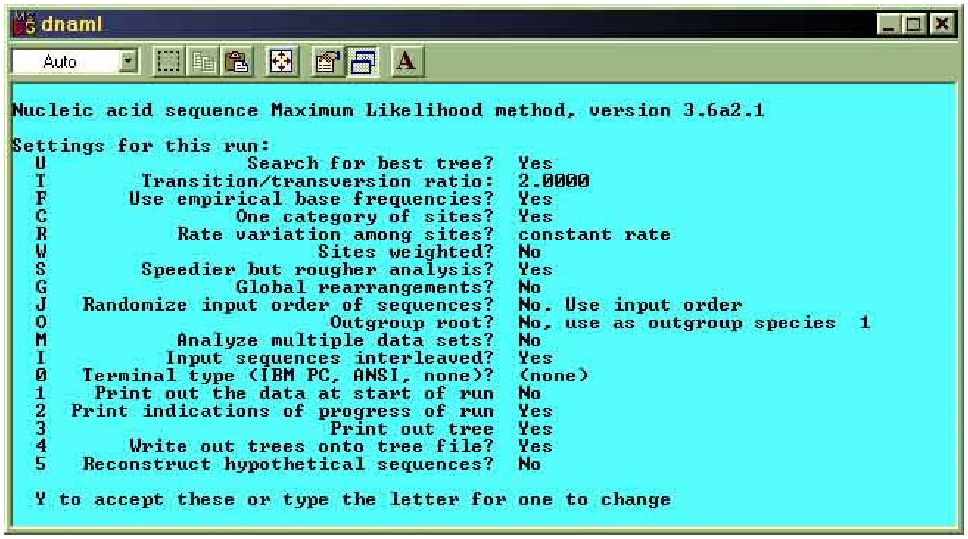
5. Copy the saved Phylip format aligned sequence file into the Phylip “exe” folder. Phylip formatted files have a “.phy” extension name.
6. Before starting with Phylip, it is important to point out that several steps are required in an analysis and that multiple programs within the Phylip package will be used, often times more than once. After each analysis is complete, files will be generated with the generic names “outfile” and “outtree”. The “outfile” contains information relevant to the analysis performed such as p-values and is generated by each program, but the information contained within the file varies. The “outtree” contains text formatted trees that can be read by a tree visualization program, such as TreeView, and is only generated for programs that construct trees. It is important to rename these files after each analysis with different, yet relevant names since each program creates output files with the same name and will overwrite previous analyses. Recommendations for file names will be given, but may be varied to suit ones preference. A key thing to remember is when entering file names pay careful attention that the complete name with file extension is used (e.g. “sequences.phy” is a file called “sequences” with a “.phy” Phylip extension). Since Phylip output files do not have extension names they are entered as the file name only.

7. The first step in an analysis is to create a bootstrap dataset. In the **Phylip → exe** folder double click **seqboot**. The first line is an error message, which will be present in all the programs, indicating that the default file name "infile" could not be found, and can be ignored. Enter the filename with extension of the Phylip formatted aligned sequence file created in ClustalX (e.g. data.phy) and press **enter**. The default setting will produce 100 bootstrap replicates, which is sufficient for this analysis. Enter **"Y"** and then press **enter** to accept the settings. Enter an odd number less than 32000 with can be divided by 4n+1, then press **enter**. The program will then proceed with creating a bootstrapped data set. Rename the "outfile" generated in the **Phylip → exe** folder (e.g. dataSB100).



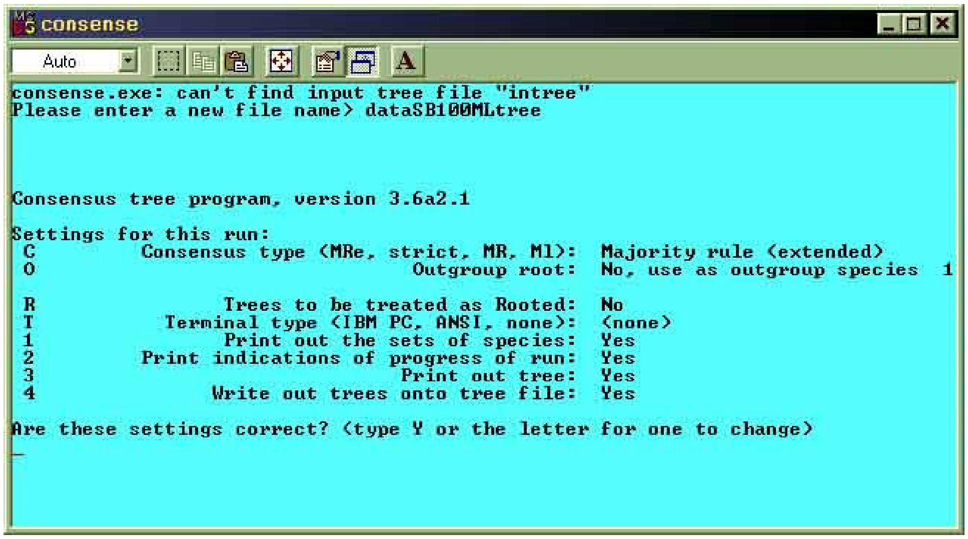
```
seqboot.exe: can't find input file "infile"
Please enter a new file name> data.phy


Bootstrapping algorithm, version 3.6a2.1

Settings for this run:
  D      Sequence, Morph, Rest., Gene Freqs?  Molecular sequences
  J  Bootstrap, Jackknife, Permute, Rewrite?  Bootstrap
  B      Block size for block-bootstrapping?  1 (regular bootstrap)
  R                     How many replicates?  100
  W              Read weights of characters?  No
  C                Read categories of sites?  No
  F     Write out data sets or just weights?  Data sets
  I             Input sequences interleaved?  Yes
  0      Terminal type (IBM PC, ANSI, none)?  (none)
  1       Print out the data at start of run  No
  2     Print indications of progress of run  Yes

  Y to accept these or type the letter for one to change
_
```

8. The next step is to analyze the bootstrapped data sets with the maximum likelihood algorithm. In the **Phylip → exe** folder double click **DNAml**. Enter the file name generated by seqboot (e.g. dataSB100) and press **enter**. In the following menu select **"M"**. The following prompt will ask if data sets or weights are being used. Enter **"D"** to signify that datasets and press **enter**. The next prompt asks how many data sets. Enter **"100"** to signify that 100 bootstrapped data sets in present in the input file, then press **enter**. The next prompt asks for a random seed. Enter a random seed as discussed above and press **enter**. The next prompt as how many times to jumble the data. Enter **"1"** and press **enter**. Enter **"Y"** and then press **enter** to accept the settings and start the analysis. When the analysis is complete (it may take a few hours depending on the number and length of the sequences in the data set) rename the output files (e.g. outtree → dataSB100MLtree and outfile → dataSB100MLout).



```
Nucleic acid sequence Maximum Likelihood method, version 3.6a2.1

Settings for this run:
  U                 Search for best tree?  Yes
  T        Transition/transversion ratio:  2.0000
  F       Use empirical base frequencies?  Yes
  C               One category of sites?  Yes
  R           Rate variation among sites?  constant rate
  W                       Sites weighted?  No
  S        Speedier but rougher analysis?  Yes
  G               Global rearrangements?  No
  J   Randomize input order of sequences?  No. Use input order
  O                        Outgroup root?  No, use as outgroup species  1
  M           Analyze multiple data sets?  No
  I          Input sequences interleaved?  Yes
  0   Terminal type (IBM PC, ANSI, none)?  (none)
  1    Print out the data at start of run  No
  2  Print indications of progress of run  Yes
  3                        Print out tree  Yes
  4       Write out trees onto tree file?  Yes
  5   Reconstruct hypothetical sequences?  No

  Y to accept these or type the letter for one to change
```

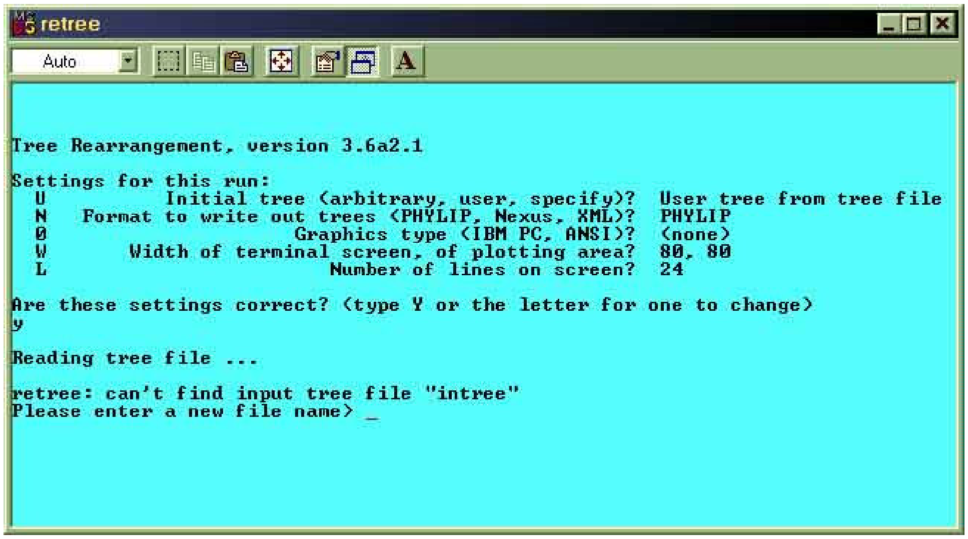
9. The next program will condense the trees generated from the maximum likelihood analyzed bootstrapped data sets in to one consensus tree. In the **Phylip → exe** folder double click **consense**. Enter the file name for the tree data generated by DNAml (e.g. dataSB100MLtree) and press **enter**. Enter **"Y"** and then press **enter** to accept the default settings. After a short pause the program will be complete. Rename the output file (e.g. outtree → dataSB100MLCtree, only a "C" was added to the name).



```
consense.exe: can't find input tree file "intree"
Please enter a new file name> dataSB100MLtree



Consensus tree program, version 3.6a2.1

Settings for this run:
 C         Consensus type (MRe, strict, MR, Ml):  Majority rule (extended)
 O                                Outgroup root:  No, use as outgroup species  1

 R                Trees to be treated as Rooted:  No
 T           Terminal type (IBM PC, ANSI, none):  (none)
 1                Print out the sets of species:  Yes
 2         Print indications of progress of run:  Yes
 3                               Print out tree:  Yes
 4               Write out trees onto tree file:  Yes

Are these settings correct? (type Y or the letter for one to change)
```

10. To get an estimation of phylogenetic distance displayed on the consensus tree a few more steps are necessary. The first step involves unrooting the consensus tree. In the **Phylip → exe** folder double click **retree**. Enter **"Y"** and then press **enter** to accept the default settings. Enter the file name for the tree data generated by consense (e.g. dataSB100MLCtree) and press **enter**. Enter **"W"** and then press **enter** to write the tree to file. At the next prompt enter **"U"** and then press **enter** to save an unrooted tree. Enter **"Q"** and then press **enter** to quit the program. Rename the output file (e.g. outtree → dataSB100MLCREtree and outfile → dataSB100MLCREout, only "RE" was added to the file name).

11. The next step involves analyzing a user-supplied tree and a data set from which to estimate phylogenetic distance. A user tree must first be created using Microsoft Word. In Word, open the unrooted tree data (e.g. dataSB100MLCREout). All distance information needs to be removed from the tree by performing a global replace. To do this in Word, press **Ctrl+H** to open the replace menu. Enter "**:*.0**" (without quotes) in the "Find what" box and nothing in the "Replace with" box. Ensure that the "**Use wildcards**" box is checked in the "More" button settings. Click the "**Replace All**" button and then close the "Find and Replace" window. Select the tree data and press **Ctrl+C** to copy the text. In the **Phylip → exe** folder open the original Phylip formatted aligned sequence file (e.g. data.phy). Scroll to the bottom of this file and paste the copied tree on a blank line below the last sequence data. Type a "1" on a blank line directly above the tree data. A truncated example of the file is shown below. Save this file with a different name using **File → Save As…** (e.g. dataWUtree). Before closing the file select all of the sequence data and the "1" and delete it so that only the tree data remains. Save this file with a different name using **File → Save As…** (e.g. dataUtree).



```
Tree Rearrangement, version 3.6a2.1

Settings for this run:
  U          Initial tree (arbitrary, user, specify)?  User tree from tree file
  N   Format to write out trees (PHYLIP, Nexus, XML)?  PHYLIP
  0                     Graphics type (IBM PC, ANSI)?  (none)
  W       Width of terminal screen, of plotting area?  80, 80
  L                        Number of lines on screen?  24

Are these settings correct? (type Y or the letter for one to change)
y

Reading tree file ...

retree: can't find input tree file "intree"
Please enter a new file name> _
```

```
GCAGCCGCGG TAATAC
GCAGCCGCGG TAATAC
GCAGCCGCGG TAATAC
1
((Rhodoc,((00156,(Microbac1,(Microbac2,02356))),
(Deino2,Deino1))),Mycobact,07856);
```

12. Now the created user tree and data set will be used to create a tree with estimated distances using DNAml. In the **Phylip → exe** folder double click **DNAml**. Enter the file name of the data set with the user tree added to the end (e.g. dataWUtree.phy) and press **enter**. Enter **"U"** and then press **enter** to tell the program that a user tree is being supplied with the data set. Enter **"Y"** and then press **enter**. Enter the name of the file containing the user tree (e.g. dataUtree.txt) and press **enter**). Rename the output files (e.g. outtree → dataSB100MLCDtree and outfile → dataSB100MLCDout).
13. If an outgroup root was used in the data set the tree can be rooted using **retree**. In the **Phylip → exe** folder double click **retree**. Enter **"Y"** and then press **enter** to accept the default settings. Enter the file name for the tree generated by DNAml (e.g. dataSB100MLCDtree) and press **enter**. Enter **"O"** and then press **enter**. Enter the number of the outgroup node and press **enter**. Enter **"W"** and then press **enter** to write the tree to file. At the next prompt enter **"R"** and then press **enter** to save an rooted tree. Enter **"Q"** and then press **enter** to quit the program. Rename the output

file (e.g. outtree → dataSB100MLCDFinaltree and outfile → dataSB100MLCDFinalout).

14. Use TreeView 1.81 to visualize the tree. A ".tre" extension can be added to the end of all tree files enabling TreeView to open when the files are double-clicked. Bootstrap values for each tree node can be obtained from the consense outtree file (e.g. dataSB100MLCtree).