MOLECULAR PHYLOGENY: APPLICATIONS AND IMPLICATIONS FOR MARINE MICROBIOLOGY

Craig L. Moyer
Department of Biology, MS#9160
Western Washington University
Bellingham, Washington 98225
USA

CONTENTS

INTRODUCTION TO THE APPLICATION OF MOLECULAR PHYLOGENY TO MARINE MICROBIOLOGY

The field of marine microbiology has from its inception been a methods-limited proposition, whether microbial communities are characterized through an autecology or synecology perspective. When the focus has been towards autecology, or the characterization of microbial populations through the study of cultured isolates and their physiology, the approach encompasses microbial growth procedures, such as dilution to extinction methods or enrichment culture. The primary limitation continues to be the frequent dependence upon nutrient-laden media to satisfy the nutritional requirements of every population of microorganisms which exists within the community. "The most one can hope for is a medium in which many microorganisms will grow and with which the results may be duplicated" (Zobell, 1946). The overall goal is to understand how microbial populations are able to adapt to a range of environmental parameters (or limitations) and yet influence marine microbiological processes. For a review of autecological studies emphasizing the predominant forcing functions (e.g., salinity, temperature, hydrostatic pressure, and nutrient availability) of the marine environment and their impact on microorganisms, see Morita (1986). Synecology or a systems level "black box" approach towards studying an entire community employs the central tenant that emergent properties result from the organization of the whole community which would otherwise be unobserved (i.e., the whole is greater than the sum of the parts). This general approach uses methods that estimate the *in situ* microbial biomass, viability, metabolism and growth through deterministic assays of environmental samples. For example, the most common strategy used to enumerate the total

number of microorganisms present (i.e., biomass) in a marine sample relies on direct microscopic counts which, lacks any capability for differentiation beyond simple morphology. For a detailed review of the marine microbiological methodology used in predominantly synecological studies, see Karl (1986).

A suite of molecular biological methods revolving around the idea that cellular component analyses provide a culture-independent means of investigating microorganisms as they occur in nature was developed in the mid-1980s (Olsen *et al.*, 1986; Pace *et al.*, 1986). This methodological approach targets a microbial community's primary members through molecular (i.e., cell component) means and characterizes their respective phylogeny or evolutionary history. Over the last decade, numerous studies using these molecular biological approaches have significantly changed our understanding of marine microbiology, fueling new avenues of research. Three noted examples, in chronological order, are (1) the initial dissections of bacterioplankton communities in the Atlantic (Giovannoni *et al.*, 1990) and Pacific (Schmidt *et al.*, 1991) Oceans, (2) the discovery of archaeoplankton (DeLong 1992; DeLong *et al.*, 1994), and (3) the discovery of dominant populations of iron- and sulfur-oxidizing bacteria at hydrothermal vents (Moyer *et al.*, 1994; Moyer *et al.*, 1995).

This approach has now become widespread and is used in marine microbiology to apply phylogenetic analysis to establish evolutionary relationships among organisms and to use this information as a framework for making inferences about community structure, genetic and thereby inferred organismal diversity, and (to a lesser degree) to infer physiological adaptation when applicable. This approach is possible due to the detailed theory of evolutionary relationships among the domains *Bacteria*, *Archaea* and *Eucarya* that

has emerged from comparisons of ribosomal RNA "signature" sequences (Olsen *et al.*, 1994b; Woese, 1994). Cell component analyses provide a culture-independent means of investigating microorganisms as they occur in nature, thereby eliminating the necessity for individual taxon cultivation (Amann *et al.*, 1995; Ward *et al.*, 1992). While several types of cell components are informative, SSU rDNAs (genes coding for small subunit ribosomal RNA) offer a quality and quantity of information which make them one of the most useful macromolecular descriptors of microorganisms (Ward *et al.*, 1992). Each SSU rDNA contains both highly conserved regions which are found among all living organisms, as well as diagnostic variable regions unique to a particular population or a closely related group. SSU rDNAs are widely used as informative biomarkers for the following reasons: (1) they are essential components of the protein synthesis machinery and therefore, are ubiquitously distributed and functionally conserved in all organisms, (2) they lack the interspecies horizontal gene transfer found with many prokaryotic genes, (3) they are readily isolated and identified, and (4) they contain diagnostic variable regions interspersed among highly conserved regions of primary and secondary structure, permitting phylogenetic comparisons to be inferred over a broad range of evolutionary distance (Moyer *et al.*, 1998). As a result of these studies, we are now beginning to recognize the incredible extent of diversity within the microbial world (Amann *et al.*, 1995; Head *et al.*, 1998; Hugenholtz *et al.*, 1998; Ward *et al.*, 1998). These features make SSU rDNAs particularly useful for studies of microbial ecology, where a potentially broad and unknown level of diversity of microorganisms is likely to exist. Currently, over 16,000 aligned and 30,000 unaligned SSU rRNA prokaryotic sequences have been made available for comparison by the Ribosomal Database Project II,

release 8.0 (Maidak *et al.*, 2000), which provides these data in a phylogenetically organized format. This type of approach allows for the autecology study (i.e., individual taxa) of microorganisms to be studied whether or not they can be been cultivated. In addition, the phylogenetically described taxa or "phylotypes" can be placed in a synecology context (i.e., whole community or group level) through the examination of SSU rRNA clone libraries generated from a microbial community. Depending upon the specific hypotheses to be tested, the experimental design based on molecular biological techniques can yield information regarding both autecology and synecology, in terms of community structure and phylogenetic diversity and is analogous to taking a census of a community and estimating a roadmap of evolutionary relationships for individual populations contained within. Figure 1 shows the dependence of environmental sample analysis with a sequence database (e.g., the Ribosomal Database Project or RDP).

METHODOLOGY FOR THE GENERATION AND ANALYSIS OF SSU rDNA CLONE LIBRARIES

Genomic DNA Extraction and Isolation

The first and foremost consideration is which type of nucleic acids will be efficiently extracted from environmental samples, DNA or RNA. Once group-specific oligonucleotide probes have been constructed and the goal is to assess to most physiologically robust components within a microbial community, then rRNA can be efficiently extracted using hydroxyapatite columns as described by Buckley *et al.* (1998). However, more often the generation of a clone library is needed when novel microbial communities are to be analyzed

with the goal of examining microbial community structure. This requires the direct extraction of genomic DNA (gDNA) from an environmental sample. We currently use the UltraClean "Soil" DNA Isolation kit from MoBio Laboratories, which when extracting ~0.25 to 0.5 gram microbial mat samples yields approximately 5.0 to 50 μg gDNA per gram sample (wet weight). This gDNA is consistently ≥10 kilobases in length when gently vortexed or by using a bead beater at the lowest possible speed. This method is logistically simple and consistently produces purified gDNA that is able to function as substrate in restriction digests as well as template for PCR. For every sample that is processed, the concentration, purity and size are checked by spectrophotometry (i.e., 260/280 nm ratios) and by 1% gel electrophoresis against a λ-*Hind*III DNA standard. The residual sample debris (post-extracted) is stored at -20°C and later examined by acridine orange staining with epifluorescence microscopy to confirm cellular lysis efficiency.

Amplification of SSU rDNA: Pitfalls and Perks

The success of any PCR depends largely upon the stringency of primers binding to their target template DNA during the hybridization phase. This stringency is impacted by two major factors, (1) the temperature of annealing, and (2) the concentration of free Mg^{++} ions. *Taq* polymerase is inactive in the absence of Mg^{++} and, with an excess, the polymerase has a greatly reduced fidelity that may increase the level of nonspecific amplification. Another consideration involving a successful "community" SSU rDNA PCR is the complexity of the template gDNA. Because multitemplate PCR is used to generate SSU rDNA clone libraries, the possibility for bias can arise, skewing the template-to-

amplicon ratio. Two classes of processes have been proposed based on the theoretical modeling of PCR: (1) PCR selection and (2) PCR drift (Wagner *et al.*, 1994). Considerable reduction in these biases has been demonstrated for SSU rDNA by using high template concentrations, performing fewer cycles, and mixing replicate reaction preparations as recommended by Polz and Cavanaugh (1998). An additional consideration is that template gDNA must be free of any RNA, otherwise single-stranded rRNA will duplex with coding strand rDNA templates thereby causing additional multitemplate bias (pers. comm., Thomas Schmidt). Finally, in order to reduce the possibility for preferential hybridization of degenerate primers, we design and synthesize our oligonucleotides with purine and pyrimidine analogs, dK and dP, respectively (Glen Research) and with inosine where appropriate so as to minimize primer degeneracy. Primers are also synthesized with a 5' phosphalink amidite (Applied Biosystems) to facilitate ligation reactions.

Multitemplate gDNA PCR: Mixtures and conditions

First Master Mix:	10X PCR buffer (1X final)
	25 mM MgCl (2.5 mM final)
	50 µM oligo primers (1 µM final for each)
	2.5 mM dNTPs (200 µM of each dNTP final)
	Best sterile water to 50 µl per reaction
Second Master Mix:	10 mg/ml BSA (200 ng/µl final)
	5 Units Ampli-Taq Gold per reaction (Applied Biosystems)

Combine the following master mix components for a minimum of 5 PCR reactions and a negative control for each SSU rDNA library to be constructed. Final volume for each reaction is 50 µl. Aliquant first master mix to each reaction tube inside a laminar flow hood using aerosol resistant pipette tips. UV irradiate for 5 to 10 minutes. Then add second master mix and finally add 100 to 500 ng gDNA per reaction. No template gDNA is placed into negative control. Reaction mixtures are sealed and incubated in a thermal cycler (e.g., GeneAmp 9700; Applied Biosystems) as follows: "hot start" at 95°C for 8 min, 25 to 30 cycles of 94°C for 1 min, annealing at 55 to 60°C for 1.5 min, with extension at 72°C for 3 min, then a final 7 min extension at 72°C, followed by a 4°C hold. Amplification products are assayed for size by 1% gel electrophoresis against a 1kb-ladder DNA standard. Only reactions yielding no amplification of negative controls are used. Ensuing ligation step must be completed within 24 hrs to insure "A" overhangs are not degraded.

Ligation, transformation and screening of SSU rDNA clones

For the construction of SSU rDNA clone libraries, five independent amplification reactions from each initial sample are pooled and then quantified by spectrophotometry. This mixture is then ligated into the pTA cloning vector and transformed using the manufacturer's protocol (Clontech). Clones are screened by α-complementation using X-gal and IPTG (~1 mg/plate each) as the substrate on LB agar plates containing 100 mg/ml ampicillin. Each putative positive clone is then selected and additionally screened by PCR using primers binding near the pTA cloning site (i.e., M13F and M13R) to determine the relative size of the insert sequence.

Putative positive screening PCR

Master Mix: 10X PCR buffer containing NP-40 and/or TritonX-100 (1X final)
25 mM MgCl (2.5 mM final)
50 µM oligo primers (0.5 µM final of both M13F and M13R)
2.5 mM dNTPs (250 µM of each dNTP final)
Best sterile water to 20 µl per reaction
10 mg/ml BSA (200 ng/µl final)
2 Units *Taq* polymerase

Combine these master mix components and aliquant to each reaction tube to a final volume of 20 µl inside a laminar flow hood using aerosol resistant pipette tips. A small amount of cloned cells from each white colony is then added to corresponding reactions with a sterile toothpick. The mixtures are then incubated using the previous protocol described for amplification of SSU rDNA from gDNA, except that one preincubation for 10 min at 94°C (to lyse the cells and inactivate any nucleases) is substituted for the 8 min "hot start" step. Negative controls exhibiting no amplification products are required for each series of screening reactions. Amplification products are then separated and visualized on a 1% agarose gel against a 1kb-ladder DNA standard. Clones containing correctly sized inserts are grown overnight at 37°C in ~10 ml LB broth with ampicillin (100 mg/ml) and are vigorously shaken. A 1 ml subsample of each overnight broth is aseptically transferred to a cryovial containing 0.5 ml of sterile 80% glycerol and then quick frozen and stored at -80°C. The remaining broth is used to isolate and purify plasmids using a Qiaprep spin plasmid kit

according to the manufacturers protocol (Qiagen), with the final plasmid elution in 100 µl of 0.1X Tris buffer (1.0 mM Tris-HCl, 0.1 mM EDTA, pH 8.0) and stored at -20°C.

Amplified ribosomal DNA restriction analysis or ARDRA

The ARDRA approach allows for the cataloging (based on restriction data) of SSU rDNA sequences or operational taxonomic units (OTUs) contained within a clone library thereby estimating the dominant microbial taxa contained within the sampled microbial community. The level of discrimination using four tetrameric restriction enzymes (i.e., the double-double digest) has been shown to differentiate among known SSU rDNA sequences (i.e., phylotypes) that have >98% sequence similarity (Moyer *et al.*, 1995) and has also been found to distinguish among >99% of the bacterial taxa present within a modeled dataset of maximized diversity (Moyer *et al.*, 1996).

As ARDRA is potentially sensitive to the orientation of the cloned insert, SSU rDNA sequences are amplified from plasmid templates using oligonucleotide primers specific to proximal flanking vector sequences of the pTA plasmid. The following primers have been designed to hybridize adjacent to the pTA cloning site and are used to generate templates for the restriction digest: (5'-ACGGCCGCCAGTGTGCTG) in the forward orientation and (5'-GTGTGATGGATATCTGCA) in the reverse.

ARDRA template PCR

Master Mix: 10X PCR buffer (1X final)

25 mM MgCl (2.5 mM final)

50 µM oligo primers (0.5 µM final for both)

2.5 mM dNTPs (200 µM of each dNTP final)

Best sterile water to 50 µl per reaction

10 mg/ml BSA (200 ng/µl final)

5 Units *Taq* polymerase

Combine these master mix components and aliquant to each reaction tube to a final volume of 50 µl inside a laminar flow hood using aerosol resistant pipette tips, include ~50 ng of purified plasmid to each reaction separately. Reactions are incubated for 1 min at 95°C followed by 30 cycles of denaturation, annealing and extension at 94°C for 1 min, 50°C for 1.5 min, and 72°C for 3 min respectively. This is followed by an additional extension at 72°C for 7 min, and a 4°C hold. A 5 µl subsample of each amplification is assayed for size and purity on a 1% agarose gel against a 1kb-ladder DNA standard.

Restriction digests of amplification products are performed in a microtiter dish format. Each of the two treatments (i.e., the double-double digest) consists of a well containing 15 µl of each amplification reaction and 15 µl of a restriction cocktail. Each restriction cocktail contains 3 µl of 10X restriction digest buffer (e.g., NEBuffer 2) and either 10 units of both *Hha*I and *Hae*III or 10 units of both *Rsa*I and *Msp*I (New England Biolabs) per 15 µl. Restriction digest components are mixed in microtiter wells to a total volume of

30 µl, sealed with a mylar sheet and incubated for 16 hrs at 37°C. After incubation, 6 µl of Orange G loading buffer [15% (w/v) Ficoll Type 400 and 0.25% (w/v) Orange G dye] is added to each digestion reaction. DNA standards are prepared by mixing 20 µl of DNA Marker V (0.25 µg/ml; Roche) and 4 µl Orange G loading buffer. Separation of restriction fragments and DNA standards are performed by electrophoresis in a cold room at 4°C with 3.5% MetaPhor agarose (BioWhittaker Molecular Applications) gels run at 5 volts/cm for ~4 hrs. Gels are stained with 0.5% (w/v) ethidium bromide solution for 20 min, destained in tap water for 20 min, and visualized by UV excitation. Gel images are captured using a digital gel documentation system (Figure 2).

The cluster analysis of digitized restriction fragment patterns is carried out using the GelCompare software (version 4.0; Applied Maths). All gel images are digitally optimized and then normalized to a single DNA Marker V standard to reduce gel-to-gel restriction pattern variability. Cluster analysis is performed on the ARDRA patterns from all clones obtained from SSU rDNA libraries using unweighted pair group analysis of Pearson product-moment correlations. Restriction pattern clusters with correlation values between 70 and 80% are defined as discrete OTUs. As Pearson correlation coefficients are sensitive to band intensity as well as size, threshold levels must be empirically determined depending upon the type of gel documentation system used and by subjective visual examination of corresponding to restriction patterns for each OTU (Figure 3). This process allows for an estimate of the number of representative SSU rDNA clones per OTU contained within a clone library (Heyndrickx *et al.*, 1996).

256 **Rarefaction analysis**

257 In order to estimate the OTU richness as a function of diversity, the rarefaction
258 technique is used. This is a deterministic transform of OTU abundance data. Rarefaction has
259 the feature that it allows for the comparison of diversity from clone libraries of unequal
260 sample size and estimates the number of phylotypes (E_s) in a random sample of n clones
261 samples without replacement from a finite parent collection of N clones, where n_i is the
262 number of clones of the ith phylotype (Tipper, 1979). Rarefaction is described by the
263 following equation:

$$E_s = \sum_{i=1}^{s} \left\{ 1 - \binom{N - n_i}{n} \binom{N}{n}^{-1} \right\}$$

264 Rarefaction analysis with corresponding standard deviations are performed for each clone
265 library with Matlab software (Mathworks; Moyer *et al*., 1998) using the algorithm developed
266 by Simberloff (1978). A comparative example of rarified data from samples of various
267 habitats is demonstrated in Figure 4.

268 **METHODOLOGY FOR THE GENERATION AND PHYLOGENETIC ANALYSIS**
269 **OF SSU rDNA SEQUENCES**

270 **SSU rDNA sequencing**

271 Representative SSU rDNA clones from OTUs containing three or more clones are
272 generally the primary targets for sequencing. The most common approach currently available
273 is to use a BigDye Terminator Cycle Sequencing Kit, which uses fluorescently labeled

dideoxy-terminators via cycle sequencing (Applied Biosystems) in conjunction with an automated DNA Sequencer (e.g., Model 310 or 377). SSU rDNA templates used for sequencing can be generated from purified plasmids using M13F and M13R primers and PCR conditions identical to those for ARDRA analysis. Amplification products from sequencing PCR reactions are pooled and purified by size exclusion using Microcon 50 filters (Millipore) prior to sequencing. Oligonucleotides used as primers internal to the archaeal and bacterial SSU rDNA are as previously described (Lane, 1991; Moyer *et al.*, 1998).

The process of transforming raw sequence data files output by automated sequencing to contiguous SSU rDNA sequences for phylogenetic analysis is performed using the software program GeneTool with the assembly editor function (BioTools). Many programs are available that perform a similar task, however GeneTool has been found to be extremely efficient and easy for novices to use for the purpose of "contig" file generation and data quality control. All data should optimally be sequenced in both directions to minimize the possibility for the introduction of errors into the database.

Phylogenetic analysis: Preliminary steps

The first step in a successful and descriptive phylogenetic analysis is the proper alignment of SSU rDNA sequences with a collection of similar and perhaps not so similar aligned sequences from an existing database so that a hierarchical context based on molecular evolution may be inferred. This is where the Ribosomal Database Project II (RDP) functions as an invaluable resource and starting point. The RDP is an internet accessed

295 database (www.cme.msu.edu/RDP) that supplies phylogenetically ordered sequence
296 alignments (their major contribution), previously constructed phylogenetic trees, ribosomal
297 secondary structures, and distributes various software programs for constructing, analyzing,
298 and viewing alignments and trees (Maidak *et al.*, 2000).

299 The usual strategy begins with a similarity search using a newly generated SSU rDNA
300 sequence to query the database for sequences that are the most similar. This can be
301 accomplished directly through the RDP using the SEQUENCE_MATCH utility and also by
302 using a basic BLAST search for the latest Genbank accessions (www.ncbi.nlm.nih.gov). This
303 approach achieves two tasks, first to find if any identical or closely related sequences exist
304 in the database and second to ascertain the level of dissimilarity between a potentially novel
305 sequence and any previously recorded phylogenetic groups. Both of these searching functions
306 are based on estimating S_{ab} values and cannot be used to infer in-depth phylogenetic
307 relationships.

308 Another consideration regarding multitemplate PCR of SSU rDNAs is the potential
309 generation of nonextant chimeras and thus artefactual sequences leading to the erroneous
310 description of nonexistent microorganisms. At this point, sequences should be submitted to
311 detect possible chimeric artefacts using the nearest-neighbor based CHECK_CHIMERA
312 function online at RDP (Robinson-Cox *et al.*, 1995) and/or the *k*-tuple matching method of
313 mglobalCHI available at www-hto.usc.edu/software/mglobalCHI (Komatsoulis and
314 Waterman, 1997). Chimeras are certainly not a rarity and every sequence must be thoroughly
315 tested, including a complete secondary structure analysis looking for non-compensatory base
316 changes. Chimeras have been found to occur at ~5% in multitemplate clone libraries even

under the most stringent of PCR conditions. However, an advantage of the ARDRA approach is that no chimera sequence has occurred more than once within any OTU detected from any single clone library. Once this stage has been completed, then the initial choices for comparative microbial sequences used in the phylogenetic analysis can be made.

The next phase is by far the most critical step in an accurate phylogenetic analysis regardless of the algorithm used to model evolutionary distance. Phylogenetic analysis is restricted to the comparison of highly to moderately conserved nucleotide positions that are unambiguously alignable in all sequences to be examined. The basic assumption is that these data then represent homologous positions of common ancestry. This step involves the alignment of novel sequences to previously aligned sequences, which again can be obtained from the RDP. One must realize that although the alignment of sequences is relatively simple among closely related taxa, it can be very difficult as the sequences become more divergent. Multiple sequence alignments can be constructed with programs such as the Genetic Data Environment (GDE) distributed by RDP or with the graphically oriented "ARB: a software environment for sequence data" (www.biol.chemie.tu-muenchen.de) which links sequence data files to a dendrogram hierarchy (Strunk *et al.*, 1998). The ARB package has the added advantage of an automated aligner function. However, in either case, this process weighs heavily upon secondary structure considerations and alignments must be checked against known secondary structures, as all rRNA molecules regardless of ancestry share a common core of secondary structure. Generally, this process is achieved by the construction of a "mask" or row of 1's and 0's allowing the phylogenetic algorithm to process specific columns of data from the alignment file. Since data removal means information loss, it is

advantageous to analyze each dataset with multiple mask variations. This potentially shows the robustness of a given tree topology and gives an estimate as to whether there is a substantial influence from the more highly variable positions. Both ARB and GDE have the capacity to use weighted masks with multiple sequence alignments.

Phylogenetic analysis: Which algorithm should I use?

There are basically three approaches used for the reconstruction of phylogenetic trees: distance matrix, maximum parsimony, and maximum likelihood methods. These algorithms are based on evolutionary models with different criteria for estimating evolutionary distance and maximizing the congruency of tree topologies (Ludwig *et al.*, 1998). Assumptions common among each of these approaches are: (1) each character is evolving independently, (2) nucleotide changes are primarily neutral, (3) comparisons are among orthologous genes, and (4) positional homology has been inferred correctly.

Distance matrix methods revolve around a two-step approach where first a matrix of pairwise distance values is calculated based on various nucleotide substitution formulas (i.e., the Jukes and Cantor one-parameter or Kimura two-parameter models). Then after the distance matrix is calculated, binary sequence differences are transformed into a tree using a clustering algorithm such as the neighbor-joining or DeSoete methods. This approach is advantageous when many taxa are compared and high-throughput tree building is necessary as it is computationally the least expensive. The disadvantages are that sequence data is converted into distance values, thereby reducing some phylogenetic information. Overall, distance matrix methods represent a compromise, but are especially useful for initial

phylogenetic screening or when taxa for diverse and yet established lineages are compared (Figure 5). Both the ARB and GDE (with the inclusive PHYLIP software) packages are able to produce distance matrices and generate trees from distance data.

The remaining two approaches are both character-based methods where the aligned sequence data (i.e., individual nucleotide positions) are used directly by the respective algorithm. Maximum parsimony is popular due to its logically simple and truly cladistic model known as Ocham's Razor, where the simplest solution is decidedly the best solution assuming that homoplasy (i.e., parallelism or convergence) is minimal. This is where the selected tree(s) has/have the shortest overall tree length and is supported by the largest number of synapomorphies (i.e., shared and derived character sites). The disadvantages are that maximum parsimony relies heavily upon synapomorphies (i.e., much information is lost) and a single best-fit tree may not necessarily be found. Also, it requires a greater computational capacity than any of the distance matrix methods. ARB and the new PAUP* (Sinauer) are examples of software packages which allow both the estimation of branch lengths as well as the generation of trees according to maximum parsimony.

The maximum likelihood approach for tree reconstruction is the most sophisticated and robust of the three methods, and allows for the inequality of transition and transversion rates. This statistically motivated approach calculates the tree for which the observed data are most probable, using a given nucleotide substitution model (e.g., Kimura 2-parameter). The algorithm itself functions as a two-step process where first it defines the tree topology and then optimizes the branch lengths on that particular topology (Felsenstein, 1981). The big advantage is that this method uses all of the character data and as such looks at every

382 possible scenario of evolutionary change at each nucleotide position. The primary
383 disadvantage is that due to the tremendous number of calculations it is by far the most
384 computationally intensive. However, using the enhanced version (i.e., fastDNAml) which
385 significantly improves computational performance (Olsen *et al.*, 1994a) and with the advent
386 of modern computer technology, this has become much less of a burden and enabled
387 phylogenetic tree reconstruction with ≥25 taxa with a Sun workstation (Figure 6). Trees are
388 constructed using jumbled orders for the addition of taxa and allowing for the global
389 swapping of branches. Using these parameters, the search for an optimal tree is repeated until
390 the best log likelihood score is reached in at least three independent searches. The
391 fastDNAml program is also distributed by the RDP.

392 In order to further test the confidence of branching orders, resampling techniques
393 such as bootstrapping can be used in conjunction with any of the phylogenetic approaches
394 so that node reproducibility and robustness can be determined (Felsenstein, 1985). Bootstrap
395 values are assigned to each internal node of a tree, indicating the percentage of the time that
396 a subtree defined by that respective branch appears as monophyletic. When used with
397 fastDNAml, generally a threshold of ≥50% is used and bootstrapping occurs ≥100 times
398 again with a jumbled addition of taxa and the search for each optimal tree is repeated until
399 the best log likelihood score is reached in at least two independent searches (Figure 6). The
400 collection of bootstrapped trees is compiled using the consensus tree function in either the
401 GDE (with the inclusive PHYLIP software) or PAUP* software packages in order to
402 calculate bootstrap values. For a comprehensive review of the methods used in phylogenetic
403 analysis, including an in-depth description of the mathematical modeling and theory, see

Swofford *et al.* (1996).

CONCLUDING REMARKS

This paper describes an avenue for the application of modern molecular biological techniques to marine microbiology. Many promising molecular-based applications are also viable alternatives such as fluorescent *in situ* hydridization (FISH) of group specific oligonuclotide probes (Amann *et al.*, 1995) or the high-throughput method of terminal restriction fragment length polymorphism (T-RFLP) used to the track of specific populations through space and time (Marsh *et al*., 2000). However, as shown in Figure 1, environmental sample analysis remains dependent upon the available database of known (and aligned) sequences. This, coupled with the observation that >>1% of physiologically defined microorganisms found in culture collections have been detected in environmental samples, points to the efficacy of the clone library approach coupled with the phylogenetic analysis of SSU rDNA sequences when attempting to understand the microbial community structure and diversity from marine habitats.

418 **References**

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

421 Buckley, D. H., J. R. Graber, and T. M. Schmidt. (1998). Phylogenetic analysis of
422 nonthermophilic members of the kingdom *Crenarchaeota* and their diversity and abundance
423 in soils. *Appl. Environ. Microbiol.* **64,** 4333-4339.

424 DeLong, E. F. (1992). Archaea in coastal marine environments. *Proc. Natl. Acad. Sci.* USA
425 **89,** 5685-5689.

426 DeLong, E. F., K. Y. Wu, B. B. Prezelin, and R. V. M. Jovine. (1994). High abundance of
427 *Archaea* in Antarctic marine picoplankton. *Nature* **371,** 695-697.

428 Emerson, D., and C. L. Moyer. (1997). Isolation and characterization of novel iron-oxidizing
429 bacteria that grow at circumneutral pH. *Appl. Environ. Microbiol.* **63,** 4784-4792.

430 Felsenstein, J. (1981). Evolutionary trees from DNA sequences: a maximum likelihood
431 approach. *J. Mol. Evol.* **17,** 368-376.

432 Felsenstein, J. (1985). Confidence limits on phylogenies: an approach using bootstrap.
433 *Evolution* **39,** 783-791.

434 Giovannoni, S. J., T. B. Britschgi, C. L. Moyer, and K. G. Field. (1990). Genetic diversity
435 in Sargasso Sea bacterioplankton. *Nature* **345,** 60-63.

436 Head, I. M., J. R. Saunders, and R. W. Pickup. (1998). Microbial evolution, diversity, and
437 ecology: A decade of ribosomal RNA analysis of uncultivated microorganisms. *Microb.*
438 *Ecol.* **35,** 1-21.

439 Heyndrickx, M., L. Vauterin, P. Vondamme, K. Kersters, and P. De Vos. (1996).
440 Applicability of combined amplified ribosomal DNA restriction analysis (ARDRA) patterns
441 in bacterial phylogeny and taxonomy. *J. Microbiol. Meth.* **26,** 247-259.

442 Hugenholtz, P., B. M. Goebel, and N. R. Pace. (1998). Impact of culture-independent studies
443 on the emerging phylogenetic view of bacterial diversity *J. Bacteriol.* **180,** 4765-4774.

444 Karl, D. M. (1986). Determination of *in situ* microbial biomass, viability, metabolism, and
445 growth. In: *Bacteria in nature*, (J. S. Poindexter and E. R. Leadbetter, eds.), Vol. 2, pp. 85-
446 176. Plenum Press, New York.

447 Komatsoulis, G. A., and M. S. Waterman. (1997). A new computaional method for detection
448 of chimeric 16S rRNA artifacts generated by PCR amplification from mixed bacterial
449 populations. *Appl. Environ. Microbiol.* **63,** 2338-2346.

450 Lane, D. J. (1991). 16S/23S rRNA sequencing. In: *Nucleic acid techniques in bacterial*
451 *systematics* (E. Stackebrandt and M. Goodfellow, eds.), pp. 115-175. John Wiley & Sons,
452 Ltd., London, England.

453 Ludwig, W, O. Strunk, S. Klugbauer, N. Klugbauer, M. Weizenegger, J. Neumaier, M.
454 Bachleitner, and K. H. Schleifer. (1998). Bacterial phylogeny based on comparative sequence
455 analysis. *Electrophoresis* **19,** 554-568.

456 Maidak, B. L., J. R. Cole, G. Lilburn, C. T. Parker, P. R. Saxman, J. M. Stredwick, G. M.
457 Garrity, B. Li, G. J. Olsen, S. Pramanik, T. M. Schmidt, and J. M. Tiedje. (2000). The RDP
458 (Ribosomal Database Project) continues. *Nucl. Acids Res.* **28,** 173-174.

459 Marsh, T. L., P. Saxman, J. Cole, and J. Tiedje. (2000). Terminal restriction fragment length
460 polymorphism analysis program, a web-based research tool for microbial analysis. *Appl.*
461 *Environ. Microbiol.* **66,** 3616-3620.

462 Morita, R. Y. (1986). Autecological studies and marine ecosystems. In: *Microbial*
463 *autecology: A method for environmental studies* (R. L. Tate, ed.), pp. 147-181. John Wiley
464 & Sons, Ltd., London, England.

465 Morita, R. Y., and C. L. Moyer. (2000). Biodiversity of psychrophiles. In: *Encyclopedia of*
466 *biodiversity,* (S. A. Levin, R. Colwell, G. Daily, J. Lubchenco, H. A. Mooney, E.-D. Schulze,
467 G. D. Tilman, eds.), In Press. Academic Press, San Diego.

468 Moyer, C. L., F. C. Dobbs, and D. M. Karl. (1994). Estimation of diversity and community
469 structure through restriction fragment length polymorphism distribution analysis of bacterial
470 16S rRNA genes from a microbial mat at an active, hydrothermal vent system, Loihi
471 Seamount, Hawaii. *Appl. Environ. Microbiol.* **60,** 871-879.

472 Moyer, C. L., F. C. Dobbs, and D. M. Karl. (1995). Phylogenetic diversity of the bacterial
473 community from a microbial mat at an active, hydrothermal vent system, Loihi Seamount,
474 Hawaii. *Appl. Environ. Microbiol.* **61,** 1555-1562.

475 Moyer, C. L., J. M. Tiedje, F. C. Dobbs, and D. M. Karl. (1996). A computer-simulated
476 restriction fragment length polymorphism analysis of bacterial small subunit rRNA genes:
477 Efficacy of selected tetrameric restriction enzymes for studies of microbial diversity in
478 nature. *Appl. Environ. Microbiol.* **62,** 2501-2507.

Moyer, C. L., J. M. Tiedje, F. C. Dobbs, and D. M. Karl. (1998). Diversity of deep-sea hydrothermal vent *Archaea*. *Deep-Sea Res.* II. **45,** 303-317.

Olsen, G. J., D. J. Lane, S. J. Giovannoni, and N. R. Pace. (1986). Microbial ecology and evolution: a ribosomal RNA approach. *Ann. Rev. Microbiol.* **40,** 337-365.

Olsen, G. J., H. Matsuda, R. Hagstrom, and R. Overbeek. (1994a). fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. *CABIOS* **10,** 41-48.

Olsen, G. J., C. R. Woese, and R. Overbeek. (1994b). The winds of (evolutionary) change: breathing new life into microbiology. *Microbiol. Rev.* **176,** 1-6.

Pace, N. R., D. A. Stahl, D. J. Lane, and G. J. Olsen. (1986). The analysis of natural microbial populations by ribosomal RNA sequences. *Adv. Microb. Ecol.* **9,** 1-55.

Polz, M. F., and C. M. Cavanaugh. (1998). Bias in template-to-product ratios in multitemplate PCR. *Appl. Environ. Microbiol.* **64,** 3724-3730.

Robison-Cox, J. F., M. M. Bateson, and D. M. Ward. (1995). Evaluation of nearest-neighbor methods for detection of chimeric small-subunit rRNA sequences. *Appl. Environ. Microbiol.* **61,** 1240-1245.

Schmidt, T. M., E. F. DeLong, and N. R. Pace. (1991). Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. *J. Bacteriol.* **173,** 4371-4378.

Simberloff, D. (1978). The use of rarefaction and related methods in ecology. In: *Biological data in water pollution assessment: Quantitative and statistical analyses* (K. L. Dickson, J. Cairns, Jr., and R. J. Livingston, eds.), pp. 150-165. ASTM STP 652.

Strunk, O., W. Ludwig, O. Gross, B. Reichel, N. Stuckmann, M. May, B. Nonhoff, M. Lenke, T. Ginhart, A. Vilbig, and R. Westram. (1998). *ARB: a software environment for sequence data.* Department of Microbiology, Technical University of Munich, Munich, Germany.

Swofford, D. L., G. J. Olsen, P. J. Waddell, D. M. Hillis (1996). Phylogenetic inference. In: *Molecular systematics* 2nd ed. (D. M. Hillis, C. Moritz, B. K. Mabel, eds.), pp. 407-514. Sinauer, Sunderland, MA.

Tiedje, J. M., J. -Z. Zhou, K. Nüsslein, C. L. Moyer, and R. R. Fulthorpe. (1997). Extent and patterns of soil microbial diversity. In: *Progress in microbial ecology: Proceedings of the 7th International Symposium on Microbial Ecology*, (M. T. Martins, M. I. Z. Sato, J. M. Tiedje, L. C. N. Hagler, J. Döbereiner, and P. S. Sanchez, eds.), pp. 35-41. Brazilian Society for Microbiology, São Paulo, Brazil.

Tipper, J. C. (1979). Rarefaction and rarefiction – the use and abuse of a method in paleoecology. *Paleobiology* **5,** 423-434.

Wagner, A., N. Blackstone, P. Cartwright, M. Dick, B. Misof, P. Snow, G. P. Wagner, J. Bartels, M. Murtha, and J. Pendleton. (1994). Surveys of gene families using polymerase chain reaction: PCR selection and PCR drift. *Syst. Biol.* **43,** 250-261.

Ward, D. M., M. M. Bateson, R. Weller, and A. L. Ruff-Roberts. (1992). Ribosomal RNA analysis of microorganisms as they occur in nature. *Adv. Microb. Ecol.* **12,** 219-286.

Ward, D. M., M. J. Ferris, S. C. Nold, and M. M. Bateson. (1998). A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. *Microbiol. Mol. Biol. Rev.* **62,** 1353-1370.

Woese, C. R. (1994). There must be a prokaryote somewhere: microbiology's search for itself. *Microbiol. Rev.* **58,** 1-9.

Zobell, C. E. (1946). *Marine microbiology: A monograph on hydrobacteriology*. Chronica Botanica Co. Waltham, Mass.

List of Suppliers

Applied Biosystems
850 Lincoln Centre Drive
Foster City, CA 94404

Tel.: 650-570-6667
1-877-477-3675
Fax: 650-572-2743
Web: www.appliedbiosystems.com
BigDye Terminator Cycle Sequencing Kit.

Applied Maths BVBA
Risquons-Toutstraat 38
8511 Kortrijk, Belgium

Tel.: 32-56-424144
Fax: 32-56-402145
Web: www.applied-maths.com
GelCompar Software Program.

BioTools Incorporated
420 Sun Life Place
10123 99 Street
Edmonton, Alberta, Canada T5J 3H1

Tel.: 1-780-423-1133
Fax: 1-780-423-1333
Web: www.biotools.com
GeneTool Software Program.

BioWhittaker Molecular Applications
191 Thomaston Street
Rockland, MD 04841

Tel.: 207-594-3400
1-800-341-1574
Fax: 207-594-3426
Web: www.bmaproducts.com
MetaPhor agarose for high resolution separation of small DNA fragments.

Clontech Laboratories, Inc.
1020 East Meadow Circle
Palo Alto, CA 94303

Tel.: 650-424-8222
1-800-662-2566
Fax: 650-424-1064
Web: www.clontech.com
AdvanTAge PCR Cloning Kit.

Glen Research
22825 Davis Drive
Sterling, VA 20164

Tel.: 703-437-6191
1-800-327-4536
Fax: 703-435-9774
Web: www.glenres.com
dK and dP nucleotide analogs.

Millipore Corp.
80 Ashby Road
Bedford, MA 01730

Tel.: 1-800-645-5476
Fax: 1-781-533-3110
Web: www.millipore.com
Microcon 50 ultrafiltration units.

Mo Bio Laboratories, Inc.
P.O. Box 606
Solana Beach, CA 92075

Tel.: 760-929-9911
1-800-606-6246
Fax: 760-929-0109
Web: www.mobio.com
“Soil” DNA Isolation Kit.

New England Biolabs
32 Tozer Road
Beverly, MA 01915

Tel.: 1-800-632-5227
Fax: 1-800-632-7440
Web: www.neb.com
Source of Tetrameric Endonucleases.

Qiagen Inc. - USA
28159 Avenue Stanford
Valencia, CA 91355

Tel.: 1-800-426-8157
Fax: 1-800-718-2056
Web: www.qiagen.com
QIAprep Spin Plasmid Minprep Kit.

Roche Molecular Biochemicals
9115 Hague Road
P.O. Box 50414
Indianapolis, IN 46250

Tel.: 1-800-428-5433
Fax: 1-800-428-2883
Web: biochem.roche.com
Supplier of low molecular weight DNA standard Marker V.

Sinauer Associates, Inc.
P.O. Box 407
23 Plumtree Road
Sunderland, MA 01375

Tel.: 413-549-4300
Fax: 413-549-1118
Web: www.sinauer.com
PAUP* 4.0 (beta version) Software Programs.

The MathWorks, Inc.
3 Apple Hill Drive
Natick, MA 01760

Tel.: 508-647-7000
Fax: 508-647-7001
Web: www.mathworks.com
Matlab Software Program used with “Rarefier” Program.

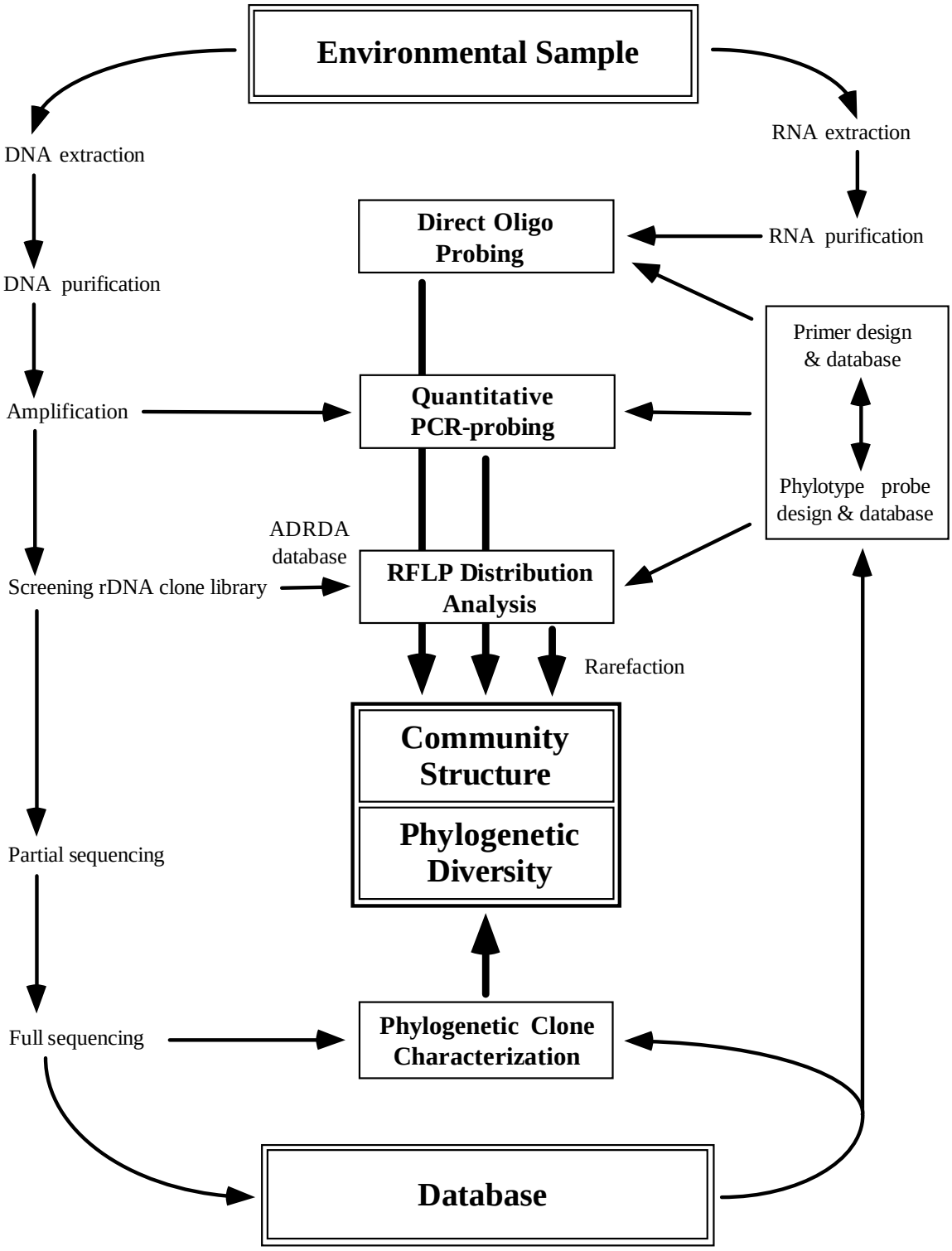

Figure 1. Flowchart describing dependency of experimental design for **Environmental Sample** analysis with sequence **Database**, while maintaining the ultimate goal of determining microbial **Community Structure** and **Phylogenetic Diversity**.

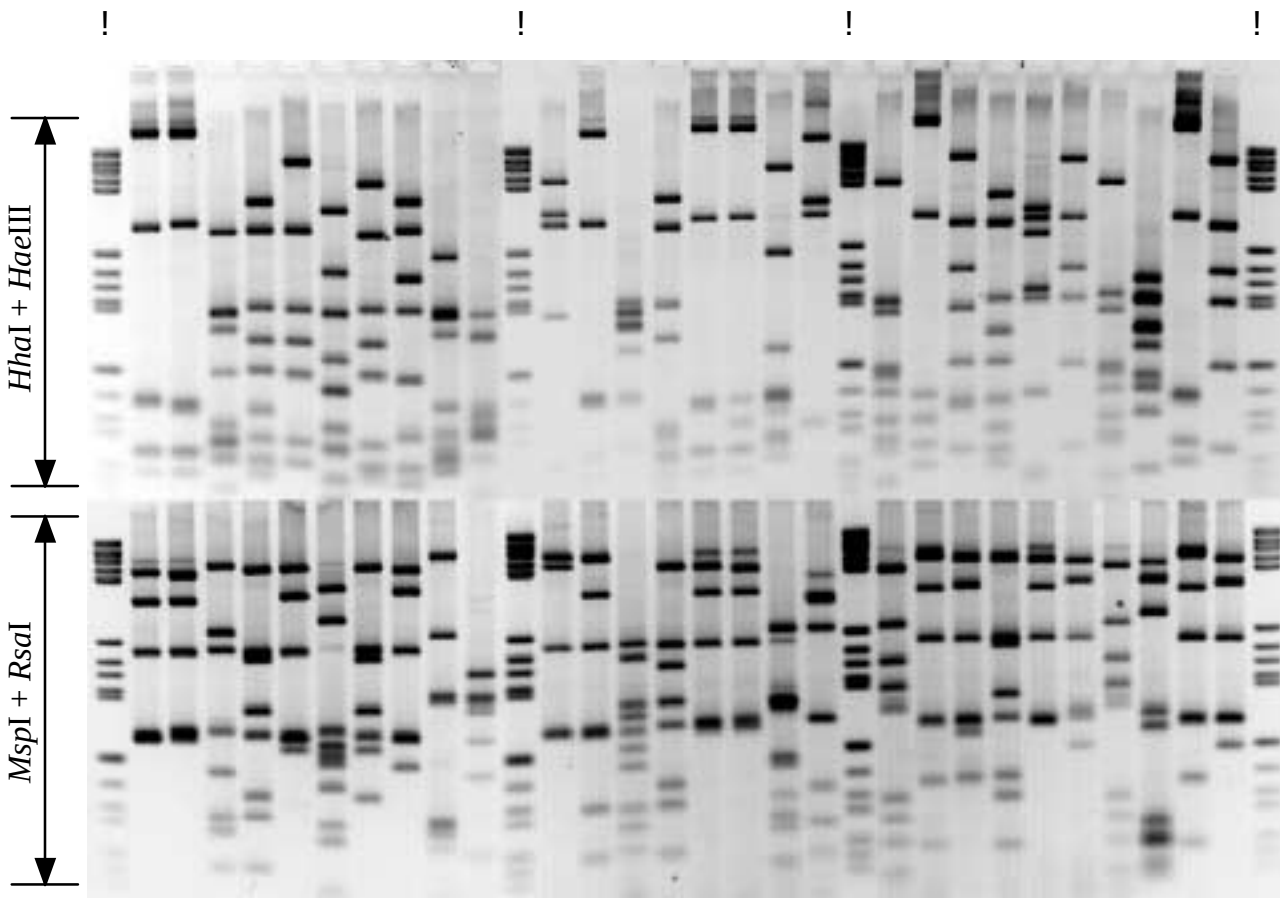

Figure 2. ARDRA gel mosaic image showing double-double digest treatments in top and bottom lanes. Lanes 1, 12, 21 and 32 (designated by !) have DNA Marker V as standard, remaining lanes represent individual SSU rDNA clones.

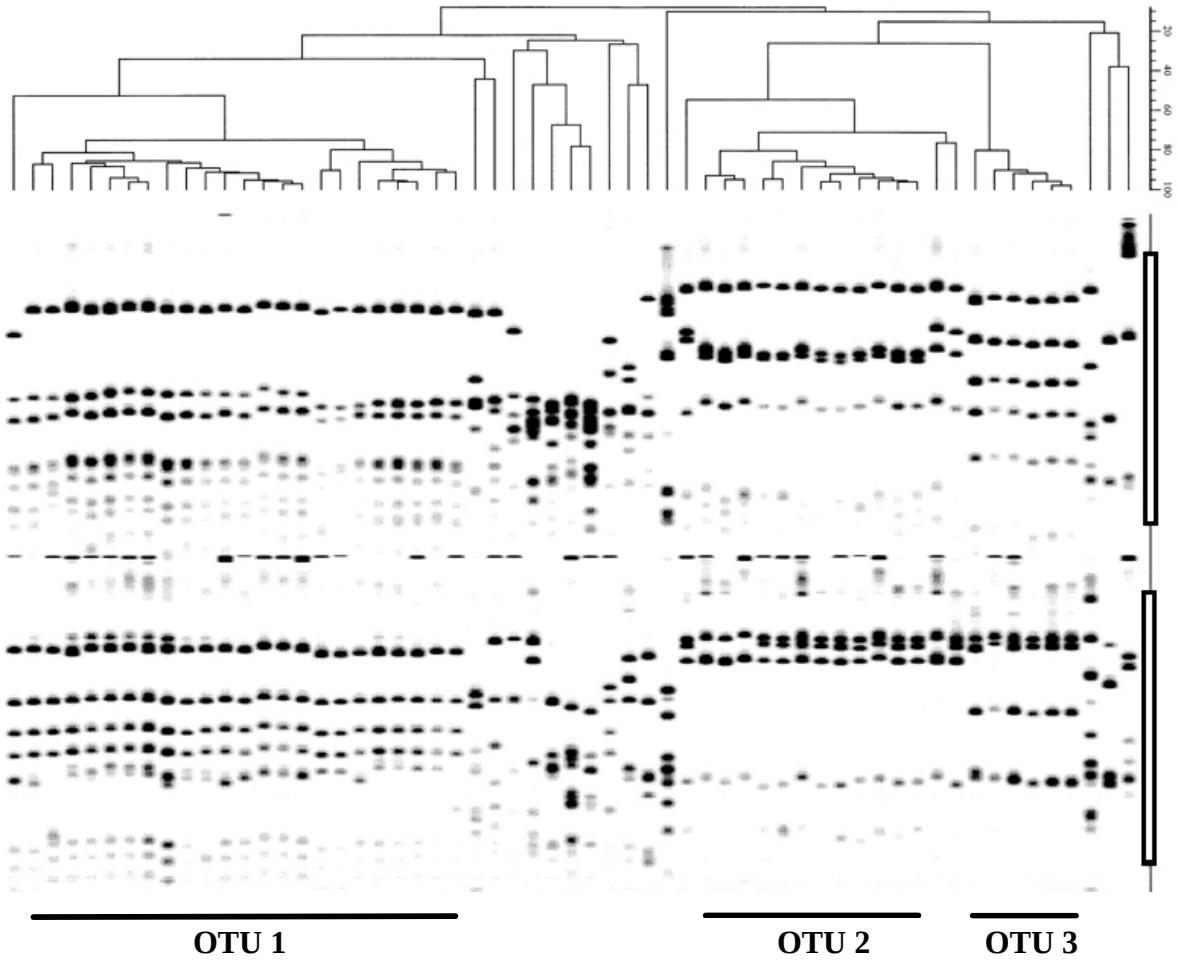

Figure 3. UPGMA cluster analysis of digitized and normalized ARDRA patterns indicating OTUs. Open bars on right indicate data region used in analysis which corresponds to size range of DNA standard for both treatment 1 and 2. OTU groupings are indicated by horizontal bars on bottom.

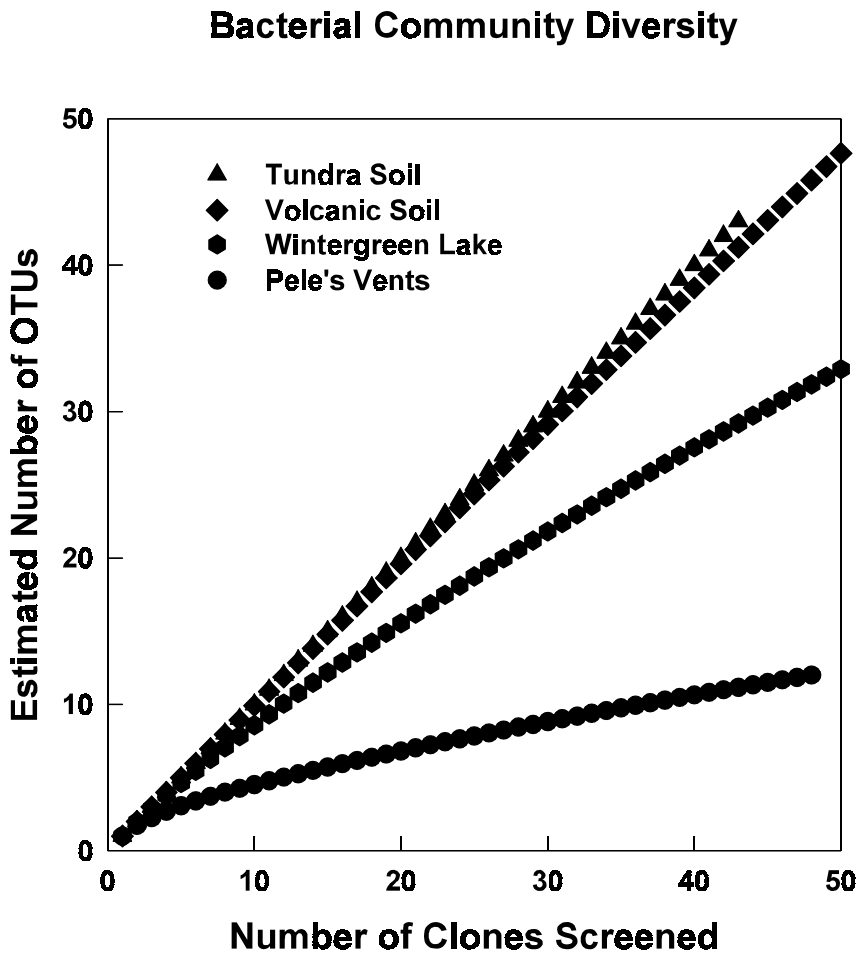

Figure 4. Rarefaction curves as indicators of bacterial community diversity from four different habitats: Soil communities are most diverse, lake bacterioplankton community is intermediate, and hydrothermal vent microbial mat community is least diverse. All four communities were analyzed using ARDRA with the double-double digest as the basis for operational taxonomic unit (OTU) classification (Tiedje *et al.*, 1997).

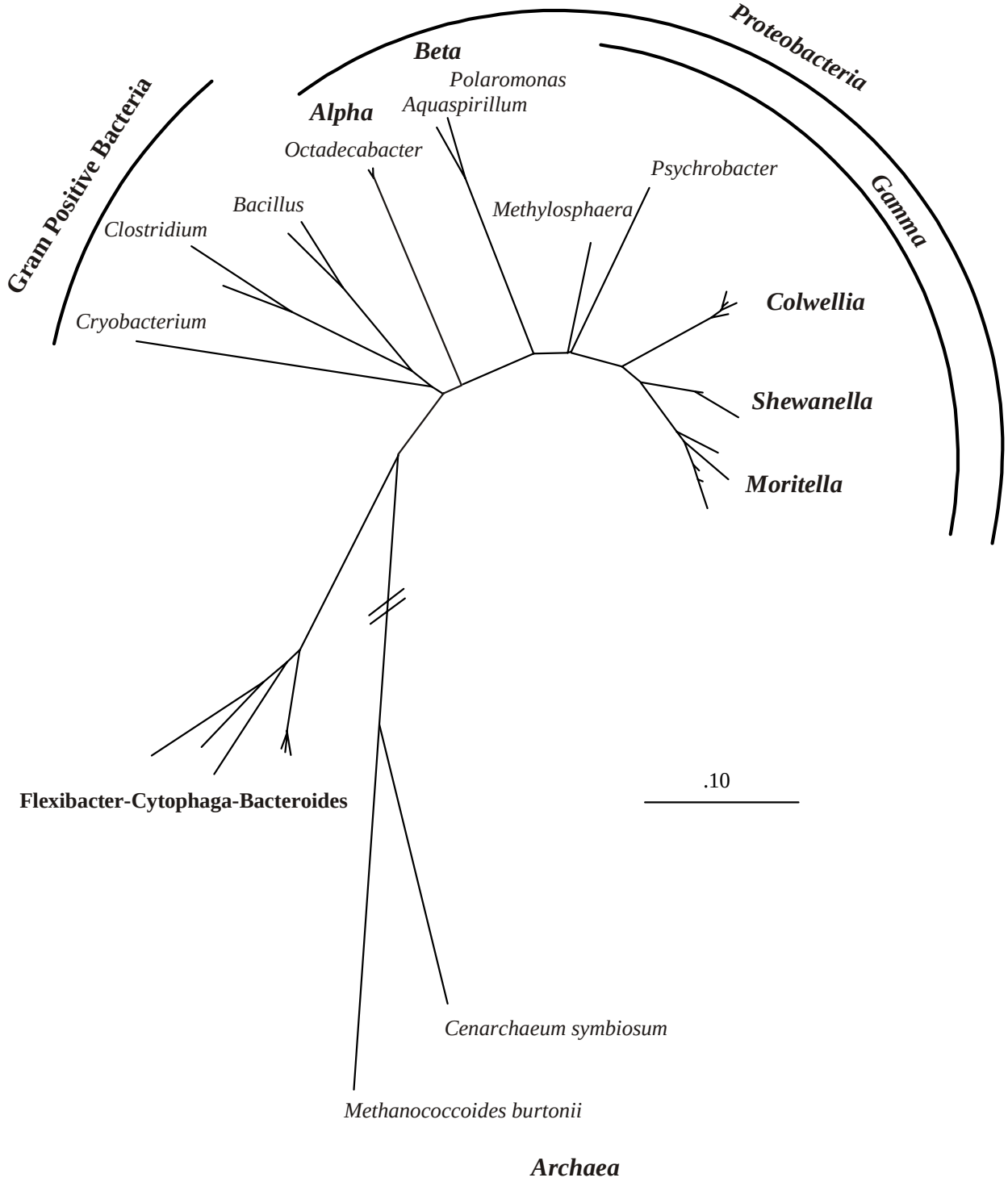

Figure 5. Radial phylogenetic tree using the neighbor-joining distance method demonstrating the evolutionary relationships among cultivated obligate psychrophiles. The tree was constructed using complete SSU rRNA sequences from the Ribosomal Database Project (RDP) with the additions of *Cenarchaeum symbiosum* and *Moritella sp.* ANT-300. The scale bar represents 0.10 fixed mutations per nucleotide position (Morita and Moyer, 2000).

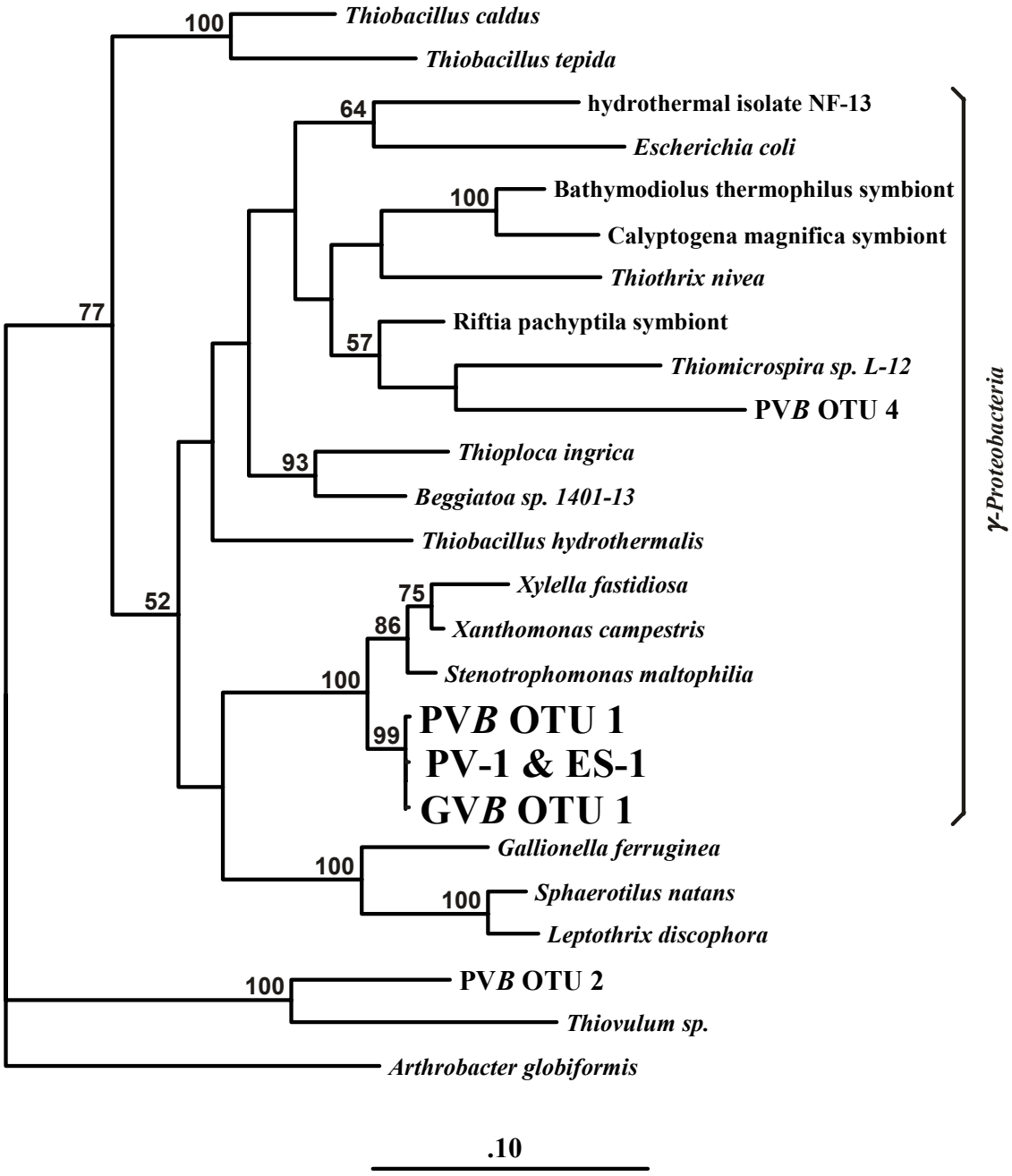

Figure 6. Phylogenetic tree demonstrating the relationships of the PV-1 & ES-1 cultured isolate phylotypes, which are included in Guaymas Vent *Bacteria* (GV*B* OTU 1) and Pele's Vents *Bacteria* (PV*B* OTU 1) lineage, with other γ-*Proteobacteria* and additional representative iron- and sulfur-oxidizers, as determined by maximum likelihood analysis of SSU rDNA sequences. Numbers at nodes represent bootstrap values (percent) for that node (based on 200 bootstrap resamplings). An outgroup is represented by *Arthrobacter globiformis*. The scale bar represents 0.10 fixed mutations per nucleotide position. Bootstrap values are shown for frequencies at or above a threshold of 50% (Emerson and Moyer, 1997; unpublished data).