BIOINFORMATICS

Vol. 19 no. 12 2003, pages 1461–1468
DOI: 10.1093/bioinformatics/btg200

Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA

*T. Z. DeSantis, I. Dubosarskiy, S. R. Murray and G. L. Andersen**

Center for Environmental Biotechnology, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail Stop 70A-3317, Berkeley, CA 94720, USA

Received on October 4, 2002; revised on January 27, 2003; accepted on February 19, 2003

ABSTRACT

Motivation: Prokaryotic organisms have been identified utilizing the sequence variation of the 16S rRNA gene. Variations steer the design of DNA probes for the detection of taxonomic groups or specific organisms. The long-term goal of our project is to create probe arrays capable of identifying 16S rDNA sequences in unknown samples. This necessitated the authentication, categorization and alignment of the >75 000 publicly available '16S' sequences. Preferably, the entire process should be computationally administrated so the aligned collection could periodically absorb 16S rDNA sequences from the public records. A complete multiple sequence alignment would provide a foundation for computational probe selection and facilitates microbial taxonomy and phylogeny.

Results: Here we report the alignment and similarity clustering of 62 662 16S rDNA sequences and an approach for designing effective probes for each cluster. A novel alignment compression algorithm, NAST (Nearest Alignment Space Termination), was designed to produce the uniform multiple sequence alignment referred to as the prokMSA. From the prokMSA, 9020 Operational Taxonomic Units (OTUs) were found based on transitive sequence similarities. An automated approach to probe design was straightforward using the prokMSA clustered into OTUs. As a test case, multiple probes were computationally picked for each of the 27 OTUs that were identified within the Staphylococcus Group. The probes were incorporated into a customized microarray and were able to correctly categorize *Staphylococcus aureus* and *Bacillus anthracis* into their correct OTUs. Although a successful probe picking strategy is outlined, the main focus of creating the prokMSA was to provide a comprehensive, categorized, updateable 16S rDNA collection useful as a foundation for any probe selection algorithm.

Availability: http://greengenes.llnl.gov/16S/

Contact: GLAndersen@lbl.gov

*To whom correspondence should be addressed.

Supplementary Information: Complete prokMSA in aligned fasta format: http://greengenes.llnl.gov/16S_cgi/download/update_28Dec02/prokMSA_aligned.fasta.Z Sequences within each Operational Taxonomic Unit: http://greengenes.llnl.gov/16S_cgi/download/update_15Mar02/SeqDescByOTU.txt DNA probe sequences: http://greengenes.llnl.gov/16S_cgi/download/update_15Mar02/15Mar02_StaphGrpPrbs.xls

INTRODUCTION

DNA sequence information from the small subunit 16S rRNA gene (16S rDNA or '16S') has been used to successfully identify and phylogenetically classify the biodiversity of the microbial world (Woese *et al.*, 1975; Fox *et al.*, 1980). Within communities of microorganisms, characterization of the taxa is possible using DNA probes directed against the 16S rDNA. The CASCADE-P (Comprehensive Aligned Sequence Construction for Automated Design of Effective Probes) project seeks probes intended to detect a specific product created by PCR primers capable of amplifying 16S genes from various prokaryotes in a complex sample (Weisburg *et al.*, 1991; Wilson *et al.*, 1990). High density rDNA microarrays currently offer the greatest quantity of simultaneous probe tests. With the capacity to interrogate a DNA sample using thousands of probes, microarrays can display numerous oligonucleotides directed to the rDNA of a single type of organism. Probe selection strategy has changed from reliance upon one single, completely unique target per taxa to the use of multiple, distinctive but imperfect targets. The cumulative observation allows inference of a taxon's presence (Wilson *et al.*, 2002).

Given that amplicons from universal PCR primers are to be exposed to probes, the initial step in probe design is to define the sequence spans to consider as the potential amplicons using the putative primer annealing locations to mark the boundaries. This is a critical step and requires a comprehensive multiple sequence alignment (MSA). Without an MSA, generation of possible amplicon sub-

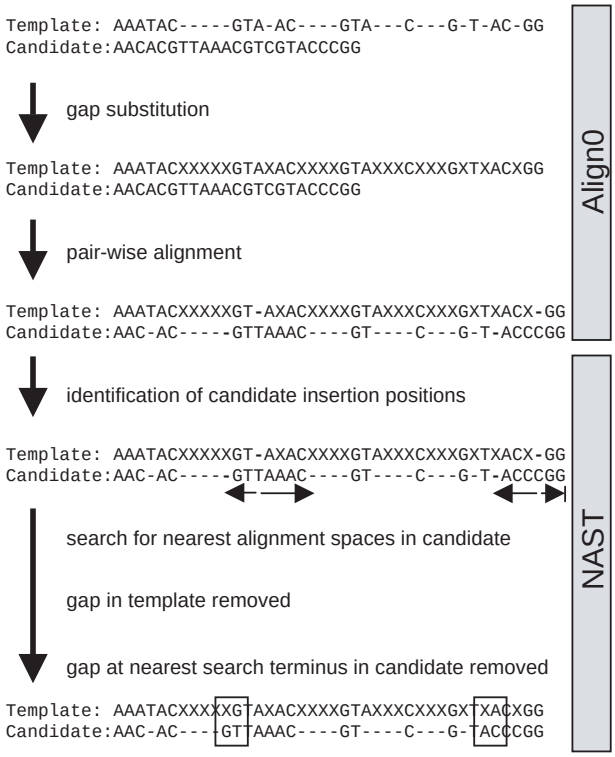

Fig. 1. Example of NAST (Nearest Alignment Space Termination) compression of an Align0 expansion using a 40 character aligned template. Template and candidate is stretched to 42 characters by Smith–Waterman algorithm used by Align0. NAST removes two characters from both sequences allowing local misalignments (boxed) while preserving the 40 character format of the global multiple sequence alignment.

strings is based on elusive textual searching, often failing because 16S sequence records may not span the region that one wishes to amplify. Thousands of 16S sequences have been situated into multiple sequence alignments (Cannone *et al.*, 2002; Maidak *et al.*, 2001; Wuyts *et al.*, 2002) allowing for extraction of sub-alignments as a prediction of the potential amplicons.

At the inception of the CASCADE-P endeavor, the Ribosomal Database Project (RDP) version 8.1 (Maidak *et al.*, 2001) housed the largest 16S MSA of 16 277 sequences. The 4218-character sequences were each placed within a hierarchical phylogenetic tree. Each node on the tree is described by a numerical designation or 'phylocode' (see Fig. 1 of Wilson *et al.*, 2002, for details).

The existing 16S MSAs represented only a fraction of the known sequences. A comprehensive, aligned collection of the >75 000 complete and incomplete 16S sequences would enable: (i) classification of each sequence into a phylogenetic category; (ii) excision of alignment slices to represent sequences internal to any given primer pair; and (iii) weighted taxon-specific consensus sequences incorporating the maximum sequence information for each locus.

The scope of this work was to computationally curate and align 16S sequences that could be phylogenetically placed, with the assumption that sequence similarities have greater importance than structural homologies when designing oligonucleotide probes. A scriptable, scalable, alignment method was conceived enabling a global 16S MSA (currently comprising 62 662 sequences) entitled the 'prokMSA' which is updated weekly. From the prokMSA, 9020 Operational Taxonomic Units (OTUs) were identified by sequence identity clustering. Finally, sets of probes were selected for each of the 27 OTUs within the Staphylococcus Group. The probe sets were synthesized on a custom Affymetrix GeneChip® array and were able to correctly categorize *Staphylococcus aureus* and *Bacillus anthracis* into their correct OTUs.

METHODS

Data sources

Fasta formatted unaligned sequences were obtained from the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov/) on 12 December, 2002. The search using '16S NOT 1.16S NOT mitochondri' across all fields in the nucleotide database returned 77 363 records (hereafter called 'candidates'). From RDP version 8.1, three aligned files were downloaded (http://rdp.cme.msu.edu/download/SSU_rRNA/). SSU_Prok.gb, SSU_Euk.gb, and SSU_Mito.gb contained 16 277, 5201, and 1503 aligned Small Sub-Unit rRNA sequences respectively. The RDP hierarchical phylocode of each sequence was attained from files of the same name with the '.phylo' extension. All 22 981 records were formatted into one BLAST (Basic Local Alignment Search Tool; Altschul *et al.*, 1990) database named 'RDP_aligned'.

Sequence pre-processing

Candidate sequences <300 nt were rejected due to insufficient information for intra-species comparison. Blastn 2.1.2 was employed with default parameters to query RDP_aligned with each remaining candidate. It was assumed that the 16 277 prokaryotic sequence standards in RDP_aligned were free of non-16S data and were all sense strand oriented. Candidates were rejected from further analysis if: (1) the longest match length was <300 base pairs; (2) the highest scoring BLAST subject was not derived from a prokaryotic organism; or (3) candidate matched sequences in two or more RDP terminal tree branches equally well. Next, the phylocode of the RDP_aligned sequence producing the top BLAST HSP (High-Scoring query-subject Pair) was designated as the

phylocode of the candidate. The 'template' was the top HSP from a second BLAST process with parameter '–G 1 –E 1'. To eliminate extra-16S sequence, the candidate sequence was trimmed to that which was bound by the beginning and end points of the template HSP. Lastly, the candidate was reverse complemented whenever the top HSP paired opposite strands from the subject and query.

Multiple sequence alignment

Essentially, the prokMSA was a merger built serially by aligning each candidate to its closest relative in the RDP tree. Initial pairwise alignment between a candidate and its pre-aligned 4182-character template was achieved using Align0 (parameters –f –1 –g –1) from the FASTAv2.0u66 package (Pearson and Lipman, 1988). Align0 pre-substituted hyphen gap characters ('-') in template sequences with the character '*X*' then added new hyphens to the template whenever the candidate contains additional internal bases (insertions) compared to the template. This expansion resulted in candidates occupying over 4182 characters in its aligned format. If Align0 inserted 10 consecutive hyphens into the template, then the candidate was rejected from further analysis.

Non-rejected alignments were compressed to 4182 characters with NAST (Nearest Alignment Space Termination), an algorithm produced for this study. NAST (Fig. 1) compensated for an insertion base in the candidate by deleting the alignment gap character closest to the insertion position according to the following instructions:

DEFINE

S_t = post-Align0 template sequence.

S_c = post-Align0 candidate sequence.

H_t = alignment space (hyphen) inserted into S_t by Align0.

H_c = alignment space (hyphen) inserted into S_c by Align0.

WHILE (S_t contains one or more H_t) **DO**

L_{Ht} = character index of distal 5′ H_t within S_t

$L_{5'}$ = character index of H_c within S_c which is 5′ proximal to H_t

$L_{3'}$ = character index of H_c within S_c which is 3′ proximal to H_t

IF $((L_{Ht} - L_{5'}) > (L_{3'} - L_{Ht}))$ Delete H_c found at $L_{3'}$

ELSE Delete H_c found at $L_{5'}$

Delete H_t found at L_{Ht}

END WHILE

Lastly, aligned candidates and RDP_aligned 16S templates were stored in a relational database and collectively referred to as the 'prokMSA'.

Sequence clustering

An alignment slice from the entire 4182 character prokMSA was extracted between positions 68 and 3689 (*E.coli* positions 47 and 1473), a span targeted by broad prokaryotic PCR primers (Dojka *et al.*, 1998). The slice was taken from the preliminary prokMSA data set released 15 March 2002, when probe design (below) was initiated. Extracted sequences were required to contain at least 600 bases, of which less than 0.5% could be ambiguous. Sequences were clustered to form OTUs. All N sequences within a phylocode were pairwise compared with blastn 2.1.2 using default parameters. An $N \times N$ identity score (I_s) matrix was created, where I_s was defined as the ratio of the number of matched bases to the maximum possible matches and was calculated as:

$$I_s = I_B \times L_m / \min(L_a, L_b)$$

where I_B is the percent identities in the BLAST alignment, L_m is the sequence length over which BLAST created an alignment, and L_a and L_b are the lengths of the two sequences as trimmed by the pre-processing step. Using the matrix, sequences were clustered based on the transitive principle as applied to sequence relationships (Burke *et al.*, 1999; Gerstein, 1998). If x, y and z are sequences and R implies a relationship determined by an I_s value exceeding a given threshold then:

$$\text{if } x\ R\ y\ \&\ y\ R\ z \Rightarrow x\ R\ z$$

In this manner, each sequence assigned to a given terminal phylobranch seeds its own cluster. The final OTUs are constructed through a series of agglomerations. This procedure was repeated independently for each of the 970 archaeal and bacterial terminal phylobranches. Sequences were clustered when $I_s$was greater than 95% excepting terminal phylobranches within Flexibacter-Cytophaga-Bacteroides (2.15), Planctomyces_and_Relatives (2.20), Proteobacteria (2.28), Fusobacteria_and_Relatives (2.29), or Gram_Positive_Bacteria (2.30) where the threshold was 98%. These divisions contain the majority of organisms of current medical and environmental interest.

Automated design of effective probes

As a test case, the Staphylococcus Group (2.30.7.12) was selected for automated probe design. Sequences were sliced and clustered from the 15 March 2002 release of the prokMSA as described above. The slice represented the potential amplicons that could be generated from a complex environmental sample using a pair of universal PCR primers (Dojka *et al.*, 1998). The goal was to

obtain an effective set of probe pairs capable of correctly categorizing mixed amplicons into their proper OTU. For each of the 27 Staphylococcus Group OTUs (Table 1) a set of 28 specific 25mers (targets) that were prevalent in members of a given OTU but were dissimilar from sequences outside the given OTU were sought. To identify targets whose complementary probes would likely cross-hybridize with amplicons outside the OTU of interest, a simplified cross-hybridization test was employed. If the central 17mer of a 25mer potential target was found in a sequence outside the set, then the phylogenetic code of the potentially cross-hybridizing sequence was recorded. With this more conservative measure of distinctiveness, we eliminated probes that were unique solely due to a mismatch in one of the outer four bases. We have noticed that the thermodynamic instability of a mismatch towards either end of the probe was not always sufficient to distinguish it from a perfectly matching sequence (unpublished data). The alignment of the sequences allowed for discrete measurement of group size at each locus. For example, if an OTU containing seven sequences possessed a locus where one member was missing data, then the locus-specific group size was six. In ranking the possible targets, those found in all members of the locus-specific group were preferred over those found only in a fraction. Also, targets with cross-hybridization potential to sequences having a common tree node near the root were favored over those with a common node near the terminal branch. The 28 top-ranked probes (perfectly matching; PM) were synthesized by Affymetrix (Santa Clara, CA) upon a solid surface adjacent to their mismatch control (MM), an identical oligonucleotide except that the 13th nucleotide was substituted to create an internal 17mer not found in any other sequence of the potential amplicons. A 'probe pair' refers to an individual PM probe and its adjacent MM probe and a 'probe set' consists of the 28 probe pairs directed at identifying one OTU.

Probe array testing

The 16S rDNA was amplified separately from genomic DNA of *S.aureus* (ATCC 12600) and *B.anthracis* (Ames) according to an established protocol (Dojka *et al.*, 1998). The amplicons were independently quantified then combined. Three aliquots of the amplicon mix were fragmented, biotinylated, and hybridized to the Affymetrix arrays according to the array manufacturer's instructions. The final concentrations of the amplicons in the 200 μl hybridization cocktail were 5 pM for *B.anthracis* and 40 pM for *S.aureus*. The array was subsequently washed, stained and scanned. Reagents, conditions, and equipment are detailed elsewhere (Masuda and Church, 2002). The scan was recorded as a pixel image and analyzed using standard Affymetrix software (GeneChip Analysis Suite, version 3.3) that reduced the data to an individual signal value for each probe. Signal noise (Q) was calculated as previously described (Wilson *et al.*, 2002). For an individual probe pair to be scored as positive, two criteria had to be satisfied. The intensity of fluorescence from the perfectly matched probe (PM) must exceed 1.3 times the intensity from the mismatched control (MM), and the difference in intensity, PM minus MM, must exceed 150Q. The percent of positive pairs for each OTUs probe set (each consisting of 28 probe pairs) was calculated. An OTU was considered present only if it's probe set responded with 100% of its probe pairs as positive.

Table 1. Percent of probe-pairs scored positive for each probe set in the Staphylococcus Group

OTU[1]	Positive pairs (%)[2]
2.30.7.12.1.013[3]	100
2.30.7.12.1.014	46–57
2.30.7.12.1.015	54–61
2.30.7.12.1.016	39–54
2.30.7.12.1.017	18
2.30.7.12.2.002	11
2.30.7.12.2.003	14
2.30.7.12.2.005	14–32
2.30.7.12.2.006	18–32
2.30.7.12.2.007	21–25
2.30.7.12.2.008	14–29
2.30.7.12.3.001	7–25
2.30.7.12.3.002	8
2.30.7.12.3.003	4
2.30.7.12.3.004	7–11
2.30.7.12.3.005	4–14
2.30.7.12.3.006	11
2.30.7.12.3.007	14–29
2.30.7.12.3.008	7
2.30.7.12.3.009	4–11
2.30.7.12.3.010	0–4
2.30.7.12.4.001	21–36
2.30.7.12.4.004[4]	100
2.30.7.12.4.005	0–11
2.30.7.12.4.006	29–54
2.30.7.12.4.007	11–14
2.30.7.12.4.008	11

[1]Operational Taxonomic Units contain sequences with 98% transitive sequence identity.
[2]Ranges encompass total variation among the three replicates.
[3]OTU 2.30.7.12.1.013 contains *S.aureus*.
[4]OTU 2.30.7.12.4.004 contains *B.anthracis*.

RESULTS

Sequence pre-processing

Of the 77 363 '16S' records obtained from NCBI, 20% had identical accession numbers to RDP_aligned sequences and were deemed duplicates and 19% were rejected for other reasons (Fig. 2). Many sequences were

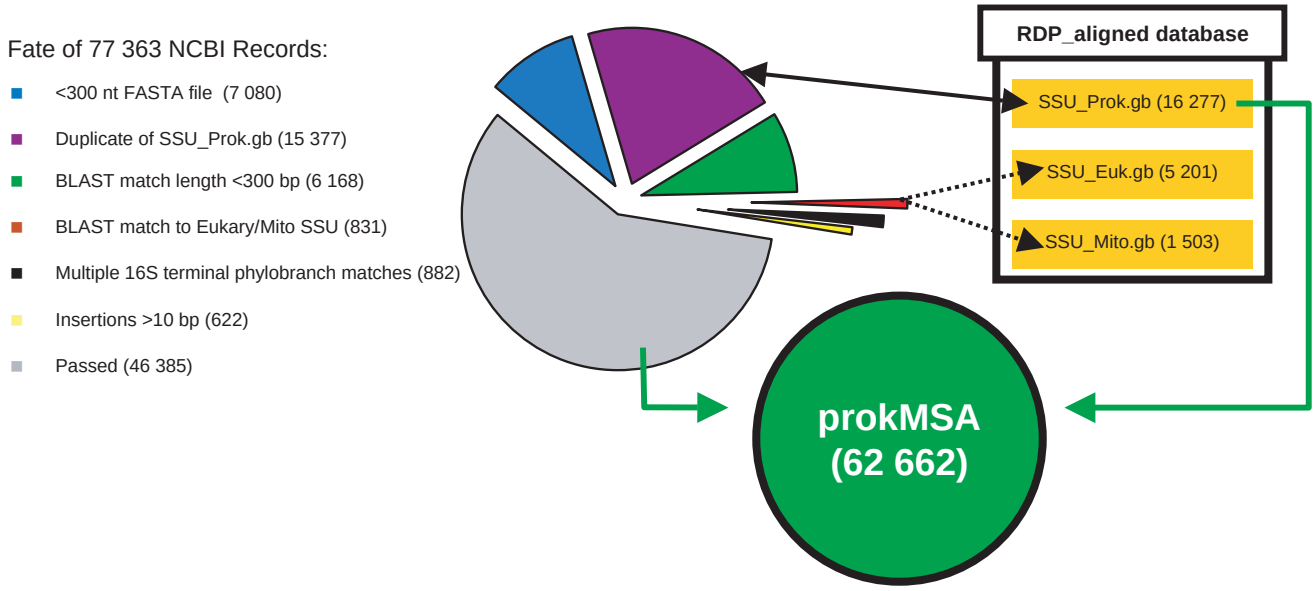

Fig. 2. Summary of pre-processing and evaluation of '16S' NCBI sequences and construction of prokMSA database. Number of sequences fulfilling each condition is shown in parentheses.

abandoned when they failed to produce BLAST HSPs to known 16S rDNA standards from RDP. Interestingly, 831 '16S' NCBI records more closely matched sequences from eukaryotic origin. Of the 47 007 sequences which were phylogenetically placed into a single terminal branch, 32 159 were partial sequences under 1300 bases in length.

Multiple sequence alignment

Alignments were performed on the 47 007 candidate sequences that were not eliminated by the pre-processing filters. Align0 sequence output was successfully compressed to 4182 characters in all but 622 attempts. In these sequences, at least one stretch of 11–25 consecutive gap characters was necessary in the template to accommodate the candidate. These candidates were not NAST condensed and were not included in the final prokMSA.

Sequence clustering

From the 15 March 2002 prokMSA of 49 310 sequences, a quality set of 26 637 contained $\geqslant$600 bases of which less than 0.5% were ambiguous between positions 68 and 3689. Application of the transitive clustering procedure assembled this set into 9020 OTUs. Of the collection's 5562 organisms described at the genus-species level, 399 were represented by sequences placed in two or more separate OTUs. For example, 16S sequences derived from various isolates of *Buchnera aphidicola* were placed into 25 different OTUs as expected considering this taxon's unusual 16S variability (Fukatsu, 2001).

Probe design and validation

Probes intended to categorize sequences from the Staphylococcus Group were selected using the automated process described above. For each OTU, a 'probe set' containing 28 'probe pairs' (PM, MM) was defined and can be inspected in file 15Mar02_StaphGrpPrbs.xls. Probes were synthesized upon an Affymetrix GeneChip® and tested with two spikes of amplicons from two of the 27 Staphylococcus Group OTUs. Hybridization results of each probe set were summarized in Table 1.

DISCUSSION

This work has established the largest dedicated collection of 16S sequences to date. The NCBI string search strategy was effective for nominating prokMSA candidates from the public repositories. The possibility exists that some true 16S rDNA sequences were overlooked if '16S' was not incorporated in the record. These sequences will have to be uncovered by other methods. Interestingly, 882 candidates had equal similarity to more than one phylogenetic taxon. Further assessment of these sequences may reveal chimeras or partial sequences spanning only the region conserved among taxa.

Alignment of the collection was critical in establishing the potential amplicons generated from universal primers. Without an MSA, potential amplicons would have to be located within each sequence using textual-based primer searches. If a search failed, it would not be clear whether the primer site was present but altered enough to evade

the pattern match, or whether the primer locus was outside the available sequence data. Alternately, an MSA arranged as horizontal rows of characters allows vertical slices to be easily extracted between columns of conserved primer annealing positions.

The prokMSA data collection represents significant advantages over accumulating data from other sources. It is a thorough set derived from the National Center for Biotechnology Information. Each record has been verified to contain only 16S sequence; proximal sequences (tRNA genes, intergenic spacer regions, and 23S rDNA) have been truncated from the 16S span. All sequences have been oriented so that the sense strand is reported. Every sequence is in a common alignment format composed of 4182 characters, allowing data availability either as continuous sequence or with necessary gaps to correspond with the popular RDP format. The user can obtain a slice of the alignment by indicating a phylogenetic group of interest and base position end-points (using *E.coli* base numbering). For each sequence, the phylocode of the closest matching RDP v8.1 sequence is associated. In addition to taxon identification, these designations also allow: (1) the average rRNA operon copy number to be estimated using the compatible Ribosomal RNA Operon Copy Number Database (http://rrndb.cme.msu.edu/; Klappenbach *et al.*, 2001); and (2) identification of taxon specific probes using the compatible PRIMROSE package (Ashelford *et al.*, 2002).

Phylogenetic placements were established for 47 007 records using BLAST with default parameters (Fig. 2). Many sequences less than 1300 bases were placed with this method. Speculation on evolutionary relationships is especially unreliable when considering only fragments of the 16S gene (Hugenholtz *et al.*, 1998). Thus, a set of prokMSA sequences that share a common phylocode should be considered as associated only by their primary structure similarity over the revealed span.

Selection of an alignment template and truncation of extra-16S sequence was accomplished by repeating the BLAST search with reduced penalties for gap introduction and elongation. It was observed that a candidate's best default-BLAST match was, at times, to a sequence shorter than the candidate. This could have resulted in over truncation since the boundaries of the 16S gene were defined by the span included by the top HSP. By diminishing gap penalties, HSPs were returned with lower identity percentages but with longer match lengths compared to the default-BLAST. This method reduced trimming of valuable 16S sequence data.

Since BLAST HSPs do not preserve the original alignment spaces, Align0 was chosen for expanding the candidate because it does not remove characters from sequence inputs. The default Align0 scoring matrix heavily penalized the introduction of alignment space characters in the candidate (-16 for first space, -4 for additional consecutive spaces). This penalty was unreasonable due to the necessity of extensive gap generation within the candidate for alignment to the heavily gapped template. In this study, the penalty for any gap was reduced to -1 which facilitated the candidate expansion in alignment to the template.

A novel procedure for sequence compression, NAST (Nearest Alignment Space Termination), was employed whenever the template sequence was expanded by Align0 to accommodate the candidate. Candidates that caused template expansion were those displaying 'insertions' relative to the template. NAST compressed the over-expanded candidate by deletion of place-holders proximal to the loci of insertions. The result was a local misalignment from the insertion base to the deleted alignment space. The minimum span of the misalignment is equal to the number of consecutive insertions (extra bases in candidate relative to the template) plus one (Fig. 1). In theory, NAST's utility is not limited to 16S data. Sequences can be merged into any existing MSA providing the trade-off between fixed total alignment string length and the extent of local misalignment is acceptable.

In 622 candidate sequences, evidence was found for a large insertion exceeding 10 bases. The local misalignment that would have resulted from the NAST compression would span at least 11 positions. Dismissal of these sequences was based on the interpretation that they contained poor data from previously described taxa or acceptable data from a yet-to-be described phylogenetic classification. The scope of the prokMSA collection was limited to sequences that could be phylogenetically placed in the RDP tree.

A similarity clustering of all the quality sequences was achieved to enable automated probe selection. Restricting the sequences for cluster analysis to those $\geqslant$600 bases barred sequence data obtained from single observation sequence reactions while still including a majority of partial 16S data likely derived from both a forward and a reverse strand. Also, short sequences would have increased transitive grouping effects. Because any two sequences whose identity score (I_s) surpassed the threshold were placed in the same cluster, seq**A** and seq**B** would be clustered even if they share below-threshold similarity when there exists a seq**C** with above-threshold similarity to both seq**A** and seq**B**. Use of sequences $<$600 bases would have caused the agglomeration of many otherwise distinct clusters.

To demonstrate that the design of effective probes could be automated using the prokMSA, probe selection rules were applied to the 27 OTUs that were identified within the Staphylococcus Group. Contemporary 16S probe design approaches (Ashelford *et al.*, 2002; Maidak *et al.*, 2001; Zhang *et al.*, 2002) all assist in evaluating the phylogenetic scope and cross-hybridization of potential probes.

Unfortunately, they did not allow: (i) restriction of probe selection between two primer positions; (ii) consideration of phylogenetic distance between the sequences which putatively cross-hybridize to the same probe; (iii) consideration of locus-specific group sizes; and (iv) generation and evaluation of mismatch control probes. Since these features were desired, special probe ranking rules were established.

The advantages of probe pair comparison were utilized when screening potential probes for cross-hybridization problems. In the probe pair test, the fluorescence intensity from a PM probe must be sufficiently greater than that emitted from the MM probe for the pair to be positive. Specifically, the intensity from the PM probe must exceed 1.3 times the intensity from the MM probe, and the difference in intensity, PM minus MM, must be >150-fold greater than the noise. To reduce the possibility of spurious positives, each potential probe target (25mer) was tested for putative cross-hybridization by searching for its internal 17mer in sequences outside the OTU of interest. By applying this test, a 25mer probe would not be considered as cross-reactive even with up to 20 contiguously matched bases to a sequence outside the OTU. In other words, it was insufficient if a 25mer's uniqueness was reliant upon the outer four positions. We have observed in prototype arrays that mismatched amplicon sequences near the center of a PM hybrid correlate with low PM intensities (unpublished data) thus reducing the PM minus MM difference below the threshold required for a pair to be deemed as positive.

Because unanimous agreement among multiple probe pairs was required to confirm an OTUs presence, the liberal cross-hybridization tolerance for any one PM probe did not diminish the array's ability to correctly categorize the mixed amplicons. Of the 27 probe sets, the only two that responded with 100% of the probe pairs as positive were the two expected from the spike-ins.

We conclude the prokMSA is a comprehensive collection of quality 16S rDNA data and the OTU clusters represent taxa for which probes can be computationally generated. Sets of probes, each with a central mismatch control probe, were useful for categorizing the 16S amplicons of a mixed sample. Although this work does briefly describe a successful probe picking strategy, the main focus of creating the prokMSA was to provide a comprehensive, categorized, updateable 16S rDNA collection useful as a foundation for any probe design algorithm. The clusters of the 15 March 2002 version of the prokMSA have since been used for the design of a 500 000 feature Affymetrix array intended for the identification of mixed environmental 16S amplicons using the probe picking strategy described in this work. These probes are being compared against collections of probes generated by a variety of approaches (Loy *et al.*, 2003) to identify where commonality exists.

ACKNOWLEDGEMENTS

We thank the RDP for providing the phylogenetic framework that made this study possible, Tom Kuczmarski for early conceptual work on assembling large 16S rDNA collections, Lisa Corsetti for Unix System Administration, Tim Harsh for database schema consultations and Art Kobayashi, Peter Agron and Sadhana Chauhan for helpful advise in preparing the manuscript. This work was performed under the auspices of the US Department of Energy by the University of California, Lawrence Berkeley National Laboratory under Contract No. DE-AC03-76SF00098. This work was funded by the Chemical and Biological Non-Proliferation program NN-20 for the Department of Energy.

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **15**, 403–410.

Ashelford,K.E., Weightman,A.J. and Fry,J.C. (2002) PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. *Nucleic Acids Res.*, **30**, 3481–3489.

Burke,J., Davison,D. and Hide,W. (1999) d2_cluster: a validated method for clustering EST and full-length cDNA sequences. *Genome Res.*, **9**, 1135–1142.

Cannone,J.J., Subramanian,S., Schnare,M.N., Collett,J.R., D'Souza,L.M., Du,Y., Feng,B., Lin,N., Madabusi,L.V. and Müller,K.M. *et al.* (2002) The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics*, **3**, 2.

Dojka,M.A., Hugenholtz,P., Haack,S.K. and Pace,N.R. (1998) Microbial diversity in a hydrocarbon- and chlorinated-solvent-contaminated aquifer undergoing intrinsic bioremediation. *Appl. Environ. Microbiol.*, **64**, 3869–3877.

Fox,G.E., Stackebrandt,E., Hespell,R.B., Gibson,J., Maniloff,J., Dyer,T.A., Wolfe,R.S., Balch,W.E., Tanner,R.S., Magrum,L.J. *et al.* (1980) The phylogeny of prokaryotes. *Science*, **209**, 457–463.

Fukatsu,T. (2001) Secondary intracellular symbiotic bacteria in aphids of the genus Yamatocallis (Homoptera:Aphididae: Drepanosiphinae). *Appl. Environ. Microbiol.*, **67**, 5315–5320.

Gerstein,M. (1998) Measurement of the effectiveness of transitive sequence comparison, through a third 'intermediate' sequence. *Bioinformatics*, **14**, 707–714.

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

Hugenholtz,P. and Huber,T. (2002) Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *Int. J. Sys. Evol. Microbiol. Papers*. in Press, http://www.sgm.ac.uk/IJSEM/PiP/ijsem02441.pdf

Klappenbach,J.A., Saxman,P.R., Cole,J.R. and Schmidt,T.M. (2001) rrndb: the ribosomal RNA operon copy number database. *Nucleic*

Acids Res., **29**, 181–184.
Liu,W.T., Mirzabekov,A.D. and Stahl,D.A. (2001) Optimization of an oligonucleotide microchip for microbial identification studies: a non-equilibrium dissociation approach. *Environ. Microbiol.*, **3**, 619–629.
Loy,A., Horn,M. and Wagner,M. (2003) probeBase—an online resource for rRNA-targeted oligonucleotide probes. *Nucleic Acids Res.*, **31**, 514–516.
Maidak,B.L., Cole,J.R., Lilburn,T.G., Parker,Jr,C.T., Saxman,P.R., Farris,R.J., Garrity,G.M., Olsen,G.J., Schmidt,T.M. and Tiedje,J.M. (2001) The RDP-II (Ribosomal Database Project). *Nucleic Acids Res.*, **29**, 173–174.
Masuda,N. and Church,G.M. (2002) *Escherichia coli* gene expression responsive to levels of the response regulator EvgA. *J. Bacteriol.*, **184**, 6225–6234.
Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.
Weisburg,W.G., Barns,S.M., Pelletier,D.A. and Lane,D.J. (1991) 16S ribosomal DNA amplification for phylogenetic study. *J. Bacteriol.*, **173**, 697–703.
Wilson,K.H., Blitchington,R.B. and Greene,R.C. (1990) Amplificationof bacterial 16S ribosomal DNA with polymerase chain reaction [published erratum appears in *J. Clin. Microbiol.* 1991 Mar; **29**(3): 666]. *J. Clin. Microbiol.*, **28**, 1942–1946.
Wilson,K.H., Wilson,W.J., Radosevich,J.L., DeSantis,T.Z., Viswanathan,V.S., Kuczmarski,T.A. and Andersen,G.L. (2002) High-density microarray of small-subunit ribosomal DNA probes. *Appl. Environ. Microbiol.*, **68**, 2535–2541.
Woese,C.R., Fox,G.E., Zablen,L., Uchida,T., Bonen,L., Pechman,K., Lewis,B.J. and Stahl,D. (1975) Conservation of primary structure in 16S ribosomal RNA. *Nature*, **254**, 83–86.
Wuyts,J., Van de Peer,Y., Winkelmans,T. and De Wachter,R. (2002) The European database on small subunit ribosomal RNA. *Nucleic Acids Res.*, **30**, 183–185.
Zhang,Z., Willson,R.C. and Fox,G.E. (2002) Identification of characteristic oligonucleotides in the bacterial 16S ribosomal RNA sequence dataset. *Bioinformatics*, **18**, 244–250.