BIOINFORMATICS

Vol. 18 no. 2 2002
Pages 244–250

Identification of characteristic oligonucleotides in the bacterial 16S ribosomal RNA sequence dataset

Zhengdong Zhang [1], Richard C. Willson [2] and George E. Fox [1,]*

[1]Department of Biology and Biochemistry, University of Houston, Houston, TX 77204-5001, USA and [2]Department of Chemical Engineering, University of Houston, Houston, TX 77204-4004, USA

Received on June 22, 2001; revised on August 12, 2001; accepted on October 5, 2001

ABSTRACT

Motivation: The phylogenetic structure of the bacterial world has been intensively studied by comparing sequences of 16S ribosomal RNA (16S rRNA). This database of sequences is now widely used to design probes for the detection of specific bacteria or groups of bacteria one at a time. The success of such methods reflects the fact that there are local sequence segments that are highly characteristic of particular organisms or groups of organisms. It is not clear, however, the extent to which such signature sequences exist in the 16S rRNA dataset. A better understanding of the numbers and distribution of highly informative oligonucleotide sequences may facilitate the design of hybridization arrays that can characterize the phylogenetic position of an unknown organism or serve as the basis for the development of novel approaches for use in bacterial identification.

Results: A computer-based algorithm that characterizes the extent to which any individual oligonucleotide sequence in 16S rRNA is characteristic of any particular bacterial grouping was developed. A measure of signature quality, Q_s, was formulated and subsequently calculated for every individual oligonucleotide sequence in the size range of 5–11 nucleotides and for 15mers with reference to each cluster and subcluster in a 929 organism representative phylogenetic tree. Subsequently, the perfect signature sequences were compared to the full set of 7322 sequences to see how common false positives were. The work completed here establishes beyond any doubt that highly characteristic oligonucleotides exist in the bacterial 16S rRNA sequence dataset in large numbers. Over 16 000 15mers were identified that might be useful as signatures. Signature oligonucleotides are available for over 80% of the nodes in the representative tree.

Availability: The programs described herein are available at http://prion.bchs.uh.edu/16S_signatures/programs/. A preliminary database of signature sequences identified in this paper is available at: http://prion.bchs.uh.edu/16S_signatures/.

Contact: zzhang@bayou.uh.edu; fox@uh.edu

*To whom correspondence should be addressed.

INTRODUCTION

Molecular characterization of 16S rRNA made it possible in the 1970s and early 1980s to unravel evolutionary relationships among bacteria for the first time (Fox *et al.*, 1980). It is now common to identify the phylogenetic position of an unknown bacterium by amplifying and sequencing its 16S rDNA. The resulting sequence is aligned with other 16S rRNA sequences and an appropriate method, e.g. maximum likelihood, is used to construct a phylogenetic tree. This process is reasonably fast, very accurate and facilitated by programs and data available at several sites on the Internet including the Ribosomal Database Project (RDP) web site (http://rdp.cme.msu.edu/html/; Maidak *et al.*, 2000), and the rRNA www server (http://rrna.uia.ac.be/index.html; Van de Peer *et al.*, 2000). The ARB project (http://www.mpi-bremen.de/molecol/arb/main.html) also provides similar databases as well as a graphically oriented package of software tools for establishing, handling and using hierarchical database of sequences and associated information. Many thousands of 16S rRNA/rDNA sequences, representing essentially all known genera of bacteria, are now available in these databases.

Considerable interest exists in using this data to develop rapid systems to identify specific bacteria or groups of bacteria. Thus, it is common to search the database for specific sequences that are unique to a particular organism or group of organisms that might be used as specific hybridization probes or perhaps incorporated into a large array of probes that can characterize many organisms in one experiment. There are a variety of software tools, such as Probe Match at the RDP site and the ARB PT server (http://www.arb-home.de) available to assist in this

activity. Alternatively, primers might be designed that can be used to amplify specific 16S rDNAs to the exclusion of others. However, little effort has been made to understand the extent to which the informative subsequences needed for probe or primer design actually exist.

Prior to the development of modern sequencing methods, 16S rRNA sequences were characterized by determining their ribonuclease T_1 catalogs. These catalogs were lists of oligonucleotide subsequences of the 16S rRNA that consisted of only those fragments which ended in G and contained no internal Gs. Analysis of this data revealed that there were in fact many characteristic 'signature' oligonucleotides (Woese *et al.*, 1980; Zablen, 1976) that were uniquely found in and hence highly characteristic of specific groups of bacteria. In order to quantify signature information, a signature quality index, that ranged from 0 (no meaningful signature) to 1 (perfect signature) was developed for use with the ribonuclease T_1 oligonucleotides (McGill *et al.*, 1986). This index allowed the quantitative characterization of the utility of any T_1 oligonucleotide in determining if an unknown organism belonged to any particular genetic grouping in a phylogenetic tree based on 16S rRNA catalog data.

In the work described here, the idea of a signature quality index is refined to quantify the extent to which any particular oligonucleotide sub-sequence found in a 16S rRNA/rDNA is a useful indicator, e.g. a 'signature,' that the larger sequence belongs to a particular phylogenetic grouping. An algorithm is given that defines Q_s, the signature quality index. Systems and methods have been developed that allow the calculation of Q_s for any 16S rRNA/rDNA sub-sequences of length N with respect to all nodes in a predefined phylogenetic tree. In the implementation presented here, Q_s is used to determine which nodes in the phylogenetic tree are potentially identified by each of the individual oligonucleotide sub-sequences of either 9, 11 or 15 nucleotides.

ALGORITHM

Signature quality index

A signature quality index that could be used in conjunction with a database of complete 16S rRNA sequences and a phylogenetic tree constructed from these sequences was defined as follows:

$$Q_s = ({}^{I}f_s) \times (1 - {}^{O}f_s) \tag{1}$$

where Q_s is a measure of the signature quality of a specific oligonucleotide sequence for a particular phylogenetic cluster. ${}^{I}f_s$ is the frequency of the oligonucleotide sequence under consideration within a specific phylogenetic group, and ${}^{O}f_s$ is the frequency of the total number of occurrences of the oligonucleotide sequence under consideration that occur outside the cluster of interest. ${}^{I}f_s$ and ${}^{O}f_s$ can be formally defined as:

$${}^{I}f_s = N_{IN}/N_C \tag{2}$$

$${}^{O}f_s = (N_T - N_{IN})/N_T \tag{3}$$

where N_T is the total number of sequences in which the target oligonucleotide occurs in the entire dataset, N_{IN} is the number of sequences in the cluster being analyzed that carry the putative signature oligonucleotide, and N_C is the total number of sequences in the target cluster. By combining (1) with (2) and (3), we have:

$$\begin{aligned} Q_s &= (N_{IN}/Nc) \times (1 - (N_T - N_{IN})/N_T) \\ &= (N_{IN}^2)/(Nc \times N_T). \end{aligned} \tag{4}$$

Currently the system uses (4) to calculate the signature quality index Q_s and in order to do so during run time it keeps tracking N_{IN}, Nc, and N_T of every oligonucleotide of a specific length at every internal tree node. Many alternative definitions of signature quality might be devised.

Data structure

All types of data structures that are built-in to the Perl programming language, namely scalar, array, and hash, were used. In addition, because of the complexity of the data presentations, more sophisticated data structures such as a bi-directional binary tree and a composite hash were also required.

Given the characteristic structure of phylogenetic trees, it was natural to represent them as binary trees in the programs. However, in the implementation developed here the tree structure is unusual in that it is bi-directional. The parent tree node has a pointer to each of its two child tree nodes and the child tree node also has a pointer back to its parent tree node. This structure is required to facilitate the calculation of Q_s at each branch tree node (excluding the tree root and all the leaf nodes).

Each leaf tree node has five data fields: 'shortName,' 'fullName,' 'nodeNumber,' 'isValid,' and 'isMatched.' The first two fields hold the abbreviated name and the full name of the organism. leafNumber records the sequentially assigned number of the leaf node in the tree. The last two fields are Boolean variables used mainly for calculation purposes. Each branch tree node has four data fields: 'nodeNumber,' 'numLeaves,' 'numValidLeaves,' and 'numMatchedLeaves.' The first field records the sequentially assigned number of the branch tree node. The rest of the fields record the numbers of leaves, 'valid' leaves, and 'matched' leaves descended from each particular branch tree node.

A composite hash was used to store all the oligonucleotides of a specific length derived from a dataset of prokaryotic 16S rRNA/rDNA sequences and their related information. The 'infrastructure' of this composite hash

was implemented with Perl's built-in hash, which is composed of unique keys and their corresponding values. The keys of the outmost layer of the composite hash were the sequences of the oligonucleotides and the value of each key is an anonymous hash that has three sub-keys—'matchingTimes,' 'matchingOrg,' and 'treeNodeValues.' The value of 'matchingTimes' counts how many times the oligonucleotide occurs in the sequence dataset. The value of 'matchedOrg' is the set of the names of the organisms whose 16S rRNA/rDNA sequences are matched by this oligonucleotide. Because of the special nature of the hash (its keys must be unique), the set is also implemented with an anonymous hash, whose keys are the names of the matched organisms and the corresponding values are set to 'undef.' The value of 'treeNodeValues' records the five highest quality index values at the branch nodes. This is implemented with an anonymous hash whose keys are the branch tree node numbers and the corresponding values are the quality index values.

Finally, it is inherent to the algorithm that only those oligonucleotide sequences that actually occur in the dataset are possible signature sequences. Sequences that never occur need not be considered. This is important because as the target sequence length is increased the number of possible sequences of that length increases exponentially with the result that the required CPU utilization can increase dramatically. In order to prevent this problem, the system described here first determines which oligonucleotides of the target length actually occur in the dataset multiple times and then only calculates values of Q_s for those sequences. For example, in the case of decamers, only 133 599 of the 1 048 570 (4^{10}) decamers actually occur in the dataset more than once.

SYSTEM AND METHODS

Except for the program readseq (Gilbert, 1990), all the other programs needed to find and evaluate signature sequences were written in Perl (version 5.6 on a Linux Intel 486 personal computer). As implemented, the system consists of eight principal programs. It can be roughly divided into four functionally different subsystems that carry out sequence file format conversion, internal data structure preparation, function value calculation, and finally presentation and evaluation of results.

Subsystem I

The sequence file preparation subsystem is comprised of three programs, readseq, fasta2flat and select_seq. The 7322 unaligned prokaryotic 16S rRNA sequences in RDP Release 7.1 were downloaded from the RDP in GenBank format. Readseq was used to change the 16S rRNA/rDNA sequence file from GenBank format to FASTA format. In this step, only the names of the organisms and the 16S rRNA/rDNA sequences were retained while all other information was discarded. Subsequently, the program fasta2flat was used to convert the FASTA format sequence data to a 'flat' format, in which every line is a data entry beginning with the organism name, followed by a tab character ('/t') as the separator, followed by a string of letters (A, U/T, G, C), which is the 16S rRNA/rDNA sequence. The next program, select_seq, was used to identify and exclude sequences that were either insufficiently complete or contained ambiguous or unsequenced positions. A total of 1921 sequences that contained at least 1400 contiguous nucleotides (nt) of complete sequence were extracted from the initial dataset.

Subsystem II

The purpose of this subsystem was to build a representative binary prokaryotic phylogenetic tree and the composite oligonucleotide hash. The comprehensive 16S rRNA/rDNA phylogenetic tree, SSU_Proc.Newick, which contains 7322 leaf nodes, was obtained from the RDP web site. The tree file, which is in Newick format, was parsed in a stepwise and bottom-up manner. The program tree_parser scans the tree file and adds one leaf node a time to a nascent internal tree facilitated by a stack of references. The tree structure reconstructed from SSU_Proc.Newick could have been used to calculate values of Q_s, but the process would have been very inefficient because nearly 74% of the leaf nodes are based on incomplete sequences. Greater efficiency was obtained by using a representative tree based on only the highest quality sequences. To construct this representative tree, 929 phylogenetically diverse sequences were manually selected from the 1921 complete sequences. The list of the leaf node numbers of these 929 bacterial species was kept in the text file called selected_leaf_node_list.

In order to keep the topology of the representative tree in accordance with that of the comprehensive tree, the program tree_parser used the list of selected leaf nodes in the file selected_leaf_node_list as the reference to 'trim away' the unselected nodes in the tree (Figure 1). This trimmed tree structure was later used in calculating the signature quality, Q_s.

The program probes_hash_table_generator is responsible for generating a composite hash, which records the needed information for each of the oligonucleotides of a specific length that occurs in at least one sequence in the 939 sequences dataset. It reads in the selected set of 16S rRNA/rDNA sequences and for each sequence it excises oligonucleotides of the specified length from the 5′ end, shifting one nucleotide at a time, to the 3′ end. Since an oligonucleotide can occur in many 16S rRNAs/rDNAs and/or several times in one particular sequence, the total number of occurrences of an oligonucleotide in the hash must be greater than or be equal to the number of the organisms whose 16S rRNAs/rDNAs it occurs in.

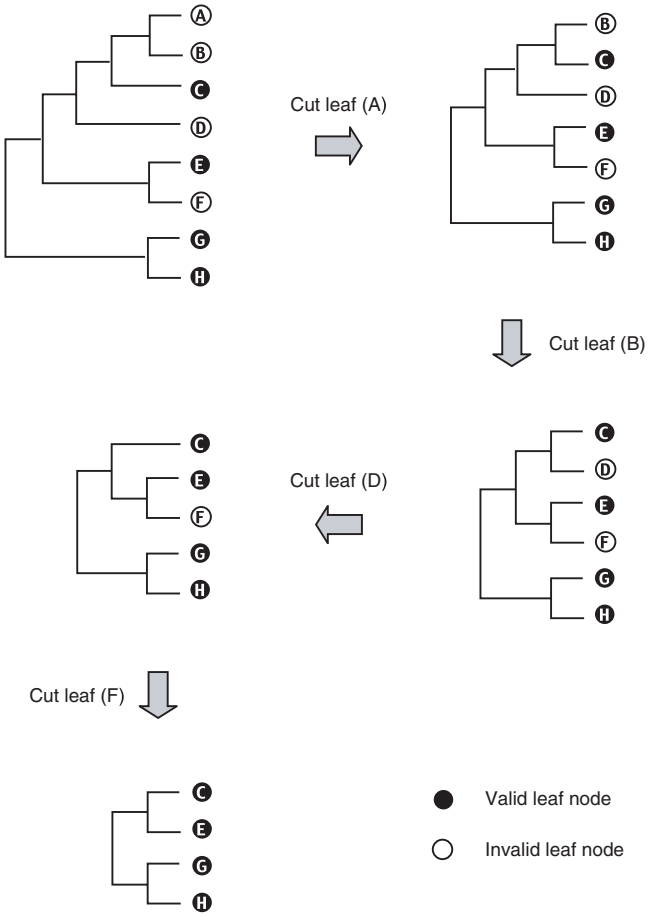

Fig. 1. The trimming process is stepwise and topology conserving.

Subsystem III

At this point, with the necessary sequence file format conversions and the data structure constructions complete, the system is now ready to calculate the value of Q_S at each branch tree node. This is accomplished with one program—calc_node_value. For each oligonucleotide of the specified size that occurs multiple times in the hash reconstructed from the binary hash file, leaf nodes in the phylogenetic tree are marked if this sequence occurs in the 16S rRNAs/rDNAs of the organisms at these leaf nodes. At each branch node, the number of descendant marked leaf nodes is counted by backward pointers in the tree structure. The Q_S index values are calculated at all the branch nodes and then sorted in descending order. The top five highest values and their corresponding branch node numbers are kept as the value/key pairs in the tree NodeValues anonymous hash field for each oligonucleotide sequence in the composite hash.

Subsystem IV

The result presentation subsystem reconstructs the composite probe hash and retrieves the calculation results from the binary data structure output by subsystem III. This is the open end of the system: the calculation result can be analyzed and presented in a variety of ways because any program, as long as it can reconstruct the composite hash from the binary file, can 'plug into' the system via subsystem IV and interpret the calculation results in its own way. A simple search tool, probe_set_analyzer is used to count the number of signatures that can identify phylogenetic groups with three or more members at quality level $= 1.0$, $\geqslant 0.8$, and >0.6 and determine the coverage of phylogenetic groupings in the tree for signature of each length. This program also evaluates the extent to which 'perfect' signatures match sequences in version 8.1 of the RDP database that were not included in the 929 sequence representative tree.

RESULTS

The potential utility of sequences in the size range from 5 to 11 and 15mers as signatures of phylogenetic relationship in a predetermined tree was examined in detail. Although the approach will work for any underlying tree, the signature sequences found will depend on the specific tree used. In the present case, we used the tree provided by the RDP because it covers the entire dataset and was constructed by well-established procedures. It should be noted that an alternative tree that might have been used is available as part of the ARB project (http://www.arb-home.de/). In the event that a particular sequence was not a perfect signature of some phylogenetic grouping in the tree (Q_S less than 1), it was not clear what value would be sufficient to deem it useful, only that a higher value is better than a lower one. Therefore, in what is presented here oligonucleotide sequences were considered in three groupings based on values of Q_S: perfect, greater than 0.8 but less than 1.0 and finally values greater than 0.6 up to 0.8.

Table 1 lists the numbers of signature sequences in each of the three categories of Q_S for each oligonucleotide sequence length that detect groupings with three or more organisms. A searchable database of signature sequences can be found at http://prion.bchs.uh.edu/16S_signatures/. Table 1 also indicates the portion of the internal nodes in the tree for which there is at least one signature sequence. The table immediately reveals that the number of useful sequences and the extent to which the tree is covered increases dramatically when oligonucleotides of length 8 and larger are used. Potential signature sequences of different lengths are distributed quite differently in the phylogenetic tree. The general observation is that long and short sequences have polar distributions in the tree: the long oligonucleotides tend to identify the branch nodes

Table 1. Numbers of potentially useful signature sequences in the pentamer to 15mer size range

Sequence length	Number of signature oligonucleotides at various quality levels as measured by Q_s[a]			Coverage of phylogenetic groupings (%)[b]
	=1.0	⩾0.8	>0.6	
5	35	482	674	1.99
6	0	371	680	4.29
7	4	372	1 151	24.81
8	450	1 715	5 396	69.52
9	2 512	5 559	14 342	81.32
10	4 995	9 191	22 381	83.92
11	6 774	11 593	27 866	84.84
15	10 487	16 629	39 502	86.37

[a]Only oligonucleotide sequences that can identify phylogenetic groups with three or more members are counted.
[b]The coverage is calculated by a computer program. Any branch nodes other than those that have two leaf nodes as their two child nodes in the representative tree are regarded as phylogenetic groups (635 in total).

near the tree leaves while the short ones are more likely to pick out those near the tree root. Of the perfect ($Q_s = 1.0$) pentameric signatures, 35 out of 35 (100%) identify the root node while 11 958 out of the 18 746 undecameric signatures (64%) and 18 824 of the 29 311 15mers (64%) identify terminal clusters of just two organisms.

In order to illustrate what happens at a more detailed level, we describe here (Table 2) the results with reference to a local region of the comprehensive tree, which originally contained 16S rRNAs representing 38 organisms. A total of 23 of these sequences were of high quality but many of them were very similar so a total of 12 sequences were selected for final inclusion in the representative tree (Figure 2). The numbers of nonameric, undecameric and 15mer signature sequences at each of the 11 branch tree nodes in this 12 organism sub-tree in different ranges of quality levels are summarized in Table 2. Tree node 5547 does not have any signatures at the Q_s 1.0 level whereas its parent branch, node 5549, has 14 perfect nonameric/undecameric/15mer signatures. Several of these are the same sequences, which serve as signatures for node 5547 at values of Q_s at the 0.8 level. This result draws attention to the fact that many individual oligonucleotides are signatures of several branch nodes at differing levels of Q_s. The signatures identifying the taxonomical group represented by the local root node 5577 of the representative tree illustrate another common feature. Of the 17 perfect signatures for node 5577, 5 are nonameric, 6 undecameric and 6 are 15mers. However, every one of these 5 nonameric signatures appears as a part of one of the 6 undecameric signatures. This inclusion of shorter signature sequences is part of a longer one and is frequently seen regardless of the signature length, the signature quality level and the position of interest in the phylogenetic tree.

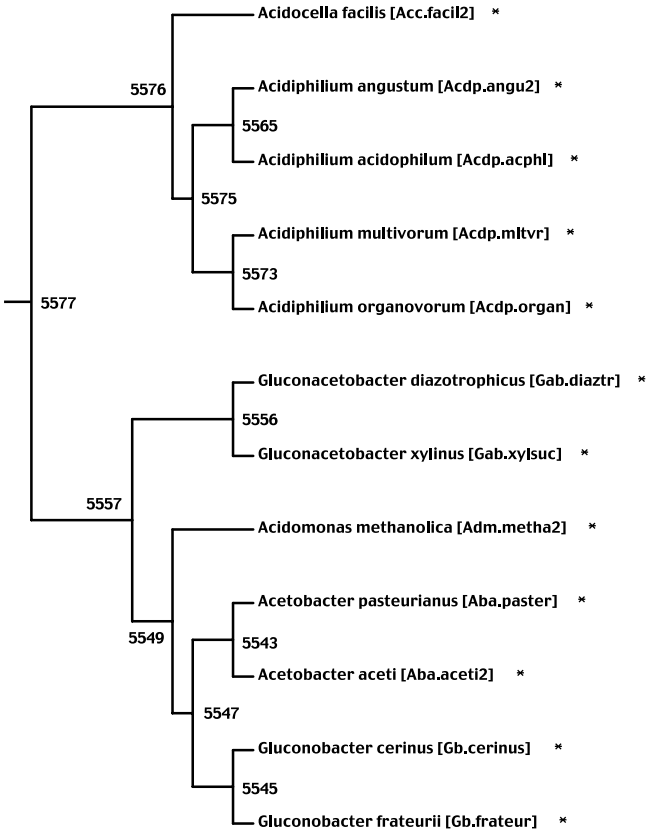

Fig. 2. A representative section of the tree following trimming from 38 to 12 representative sequences. The branch numbers in the representative tree are labeled in the picture and can be correlated with the results given in Table 2.

A remaining issue is the effect of using a 929 organism representative tree in lieu of the entire 7322-sequence dataset of RDP Release 7.1. In order to evaluate this, we examined the extent to which the 29 311 perfect 15mers signature sequences actually occur outside their target group in the whole 7322-sequence dataset. It was found that 13 066 sequences continued to be uniquely found in their target group. An additional 5236 were encountered in one organism outside the group, 2842 in two organisms and 2098 in three. Thus, approximately 80% of the predicted signature sequences are extremely effective. It is expected that further improvement can be obtained by better selection of organisms for inclusions in the representative tree. This result verifies that there are a very large number of signature sequences available and that the approached described here is an effective tool for finding them.

Table 2. The numbers of nonameric, undecameric, and 15meric signatures at different branch tree nodes in different ranges of signature quality level

Branch node number	Number of nonameric, undecameric, and 15meric oligonucleotides sequences in various Q_S ranges											
	1.0				[0.8, 1.0)				[0.6, 0.8)			
	Σ	9	11	15	Σ	9	11	15	Σ	9	11	15
5543	77	12	26	39	0	0	0	0	36	5	12	19
5545	176	26	58	92	0	0	0	0	163	25	51	87
5547	0	0	0	0	27	4	10	13	98	10	36	52
5549	14	4	5	5	33	5	14	14	183	24	62	97
5556	47	6	13	28	0	0	0	0	29	2	8	19
5557	19	1	8	10	61	9	28	24	118	27	36	55
5565	298	42	99	157	0	0	0	0	108	24	46	38
5573	419	42	136	241	0	0	0	0	139	32	50	57
5575	90	12	30	48	165	24	48	93	102	23	39	40
5576	93	11	28	54	134	22	49	63	154	27	47	80
5577	17	5	6	6	61	15	21	25	109	26	41	42

DISCUSSION

At the time this system was developed, 7322 bacterial 16S rRNA/rDNA sequences were available in RDP Release 7.1. The recent release of version 8.1 has increased the number to 16 277. These sequences have multifarious qualities—some are fully determined in terms of both the length and every position of the sequence while many others are either partially sequenced (sometimes as little as a few hundred positions) and/or contain one or more undetermined positions. This problem is compounded by the fact that the same primers are used in many sequencing studies with the result that the same regions are unsequenced in many organisms. Inclusion of low quality sequences is problematic. When the distribution of an oligonucleotide is being examined it may appear to be missing solely because the sequence being searched is incomplete or contains un-sequenced positions at the site where the target sequence would occur. This would then produce a false negative. A related problem is that in some cases large numbers of essentially identical sequences are in the dataset. Inclusion of such sequences would result in substantial decrease in the apparent Q_S of any sequence that had a spurious occurrence in a sequence that was essentially entered in the dataset multiple times. In the initial implementation presented here the simplest solution to these problems was used, mainly elimination of the numerous problematic sequences from consideration by using a representative tree. Subsequently, the reasonableness of the results was evaluated by comparing the signature sequences that were identified to the newest version of the entire sequence database.

Prior to the beginning of this project, it was known that when a small set of 16S rRNA/rDNA sequences were analyzed at least some signature sequences existed that were highly representative of the phylogenetic groups identified by tree constructions based on the complete 16S rRNA sequences. However, it was not known what was the consequence of having thousands of such sequences in the dataset. Would noise build up to the extent that useful signatures are obscured? Even if such sequences continued to exist in the larger dataset it was not clear that their numbers would be useful nor was it clear that they could be readily identified. Finally, it was not obvious that a sufficient number of interesting clusters could be identified. In the work described herein, these questions were examined with the assistance of suitable computer programs. The implementation described establishes beyond any doubt that characteristic oligonucleotides in the bacterial 16S rRNA/rDNA sequence dataset do in fact exist in huge numbers.

The existence of large numbers of oligonucleotide signature sequences is a direct demonstration and an innate characteristic of the evolution of bacterial 16S rRNAs that can be utilized to identify an unknown prokaryotic agent by elucidating its immediate phylogenetic neighborhood. It is anticipated that these characteristic oligonucleotides

can be used in the future as the basis for developing experimental methodology, e.g. possibly array hybridization, that can establish the phylogenetic position of an unknown organism without any preconceived notion as to what it was. This will not necessarily be straightforward, however, as current technology typically utilizes probes that are fifteen nucleotides or larger and some regions of the RNA are more or less accessible than others (Fuchs *et al.*, 1998). For some nodes, a suitable signature sequence of sufficient length may not exist. In such instances, it will be necessary to devise alternative strategies taking advantage of the existence of highly characteristic shorter sequences as a starting point. This might include the incorporation of inosine, which pairs with T, C or A at ambiguous positions to connect two or more short signature sequences together or simply including a family of related probe sequences on the array such that a hit with any one of them is indicative of the same grouping. In summary, the results presented here demonstrate that a very large number of potentially useful signature sequences do in fact exist in the 16S rRNA dataset and provide an effective approach for finding them. In the future it may be useful to extend this analysis to other RNA sequence sets. There remains, however, much to be done with the 16S rRNA dataset itself to better understand the distribution of the signature sequences with respect to different phylogenetic levels and their locations in the 16S rRNA molecule.

ACKNOWLEDGEMENTS

This work is dedicated to Carl Woese. Financial support was provided by a grant from the National Space Biomedical Research Institute (NASA Cooperative Agreement NCC-9-58) to G.E.F. and R.C.W.

REFERENCES

Fox,G.E., Stackebrandt,E., Hespel,R.B., Gibson,J., Maniloff,J., Dyer,T.A., Wolfe,R.S., Balch,W.E., Tanner,R.S., Magrum,L.J., Zablen,L.B., Blakemore,R., Gupta,R., Bonen,L., Lewis,B.J., Stahl,D.A., Luehrsen,K.R., Chen,K.N. and Woese,C.R. (1980) The phylogeny of prokaryotes. *Science*, **209**, 457–463.

Fuchs,B.M., Wallner,G., Beisker,W., Schwippl,I., Ludwig,W. and Amann,R. (1998) Flow cytometric analysis of the *in situ* accessibility of *Escherichia coli* 16S rRNA for fluorescently labeled oligonucleotide probes. *Appl. Environ. Microbiol.*, **64**, 4973–4982.

Gilbert,D.G. (1990) ReadSeq. On-line documentation available on the World Wide Web at http://iubio.bio.indiana.edu/soft/molbio/readseq/classic/Readme.

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

McGill,T.R., Jurka,J., Sobieski,J.M., Pickett,M.H., Woese,C.R. and Fox,G.E. (1986) Characteristic archaebacterial 16S rRNA oligonucleotides. *Syst. Appl. Microbiol.*, **7**, 194–197.

Van de Peer,Y., De Rijk,P., Wuyts,J., Winkelmans,T. and De Wachter,R. (2000) The European small subunit ribosomal RNA database. *Nucleic Acids Res.*, **28**, 175–176.

Woese,C.R., Maniloff,J. and Zablen,L.B. (1980) Phylogenetic analysis of the mycoplasmas. *Proc. Natl Acad. Sci. USA*, **77**, 494–498.

Zablen,L.B. (1976) *Prokaryotic Phylogeny by Ribosomal Ribonucleic Acid Sequence Homology*, PhD Dissertation, University of Illinois, Urbana, Illinois.