Applied and Environmental Microbiology

# Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy

**Qiong Wang, George M. Garrity, James M. Tiedje and James R. Cole**

***Appl. Environ. Microbiol.* 2007, 73(16):5261. DOI: 10.1128/AEM.00062-07.**
**Published Ahead of Print 22 June 2007.**

APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Aug. 2007, p. 5261–5267 Vol. 73, No. 16
0099-2240/07/$08.00+0 doi:10.1128/AEM.00062-07


# Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy▽†

Qiong Wang,[1] George M. Garrity,[1,2] James M. Tiedje,[1,2] and James R. Cole[1]*

*Center for Microbial Ecology[1] and Department of Microbiology and Molecular Genetics,[2] Michigan State University, East Lansing, Michigan 48824*

Received 10 January 2007/Accepted 18 June 2007

**The Ribosomal Database Project (RDP) Classifier, a naïve Bayesian classifier, can rapidly and accurately classify bacterial 16S rRNA sequences into the new higher-order taxonomy proposed in Bergey's *Taxonomic Outline of the Prokaryotes* (2nd ed., release 5.0, Springer-Verlag, New York, NY, 2004). It provides taxonomic assignments from domain to genus, with confidence estimates for each assignment. The majority of classifications (98%) were of high estimated confidence (≥95%) and high accuracy (98%). In addition to being tested with the corpus of 5,014 type strain sequences from Bergey's outline, the RDP Classifier was tested with a corpus of 23,095 rRNA sequences as assigned by the NCBI into their alternative higher-order taxonomy. The results from leave-one-out testing on both corpora show that the overall accuracies at all levels of confidence for near-full-length and 400-base segments were 89% or above down to the genus level, and the majority of the classification errors appear to be due to anomalies in the current taxonomies. For shorter rRNA segments, such as those that might be generated by pyrosequencing, the error rate varied greatly over the length of the 16S rRNA gene, with segments around the V2 and V4 variable regions giving the lowest error rates. The RDP Classifier is suitable both for the analysis of single rRNA sequences and for the analysis of libraries of thousands of sequences. Another related tool, RDP Library Compare, was developed to facilitate microbial-community comparison based on 16S rRNA gene sequence libraries. It combines the RDP Classifier with a statistical test to flag taxa differentially represented between samples. The RDP Classifier and RDP Library Compare are available online at http://rdp.cme.msu.edu/.**

Starting in the mid-1980s, Carl Woese revolutionized the field of microbiology with his rRNA-based phylogenetic comparisons delineating the three main branches of life (28). Today, rRNA-based analysis remains a central method in microbiology, used not only to explore microbial diversity but also as a day-to-day method for bacterial identification. Identification methods are conceptually easier to interpret than molecular phylogenetic analyses and are often preferred when the groups are well understood. Most rRNA identification (classification) methods, as opposed to phylogenetic (clustering) methods, have been nearest-neighbor-based classification schemes (10, 18; however, see reference 4). In some part, this was due to the lack of a consistent, higher-level bacterial classification structure (taxonomy). Several recent events have helped change this situation. In 2002, an ad hoc committee for the reevaluation of species definition in bacteriology (24) advised that all new bacterial species descriptions include an rRNA sequence from the type strain, and in 2001, Bergey's Trust published a revised higher-order taxonomy attempting to reconcile bacterial taxonomy with rRNA-based phylogeny (12, 13).

The naïve Bayesian classification method is simple yet can be extremely efficient. "Naïve" refers to the (naïve) assumption that data attributes are independent. Domingos and Pazzani (11) showed that the Bayesian method can still be optimal even when this attribute independency is violated. The method has also been reported to perform well on problems similar to the classification of sequence data, such as the classification of text documents, that have a high-dimensional feature space and sparse data (16).

The Ribosomal Database Project II (RDP) provides data, tools, and services related to rRNA sequences to the research community. As of January 2007, the RDP maintains over 300,000 bacterial sequences and averages over 5,000 new sequences each month. To handle this volume of sequences, we have developed a naïve Bayesian classifier for classifying bacterial rRNA sequences into the new Bergey bacterial taxonomy. This classifier is fast, does not require sequence alignment, and works well with partial sequences. (The vast majority of rRNA sequences in the public databases are partial.) It is capable of classifying to the genus level near-full-length and 400-base segments with an overall accuracy above 88.7%.

Microbial-community comparison based on 16S rRNA gene sequence libraries has become commonplace in microbial ecology. However, most comparison methods, whether from traditional macroecology, such as Sorensen's and the Jaccard indices, or designed specifically for sequence data, such as LIBSHUFF (22), Martin's P and F tests (19), and UniFrac (17), provide only summary information about the degree of difference between communities. These methods fail to put differences in a taxonomic context. Library Compare uses the RDP Classifier to provide rapid classification of sequences from two sample libraries. The assignments for the two samples are compared to estimate the probability of observing by chance the difference in representation for each taxon using a statistical test.

* Corresponding author. Mailing address: Center for Microbial Ecology, Michigan State University, East Lansing, MI 48824. Phone: (517) 432-4998. Fax: (517) 353-8957. E-mail: colej@msu.edu.

† Supplemental material for this article may be found at http://aem.asm.org/.

▽ Published ahead of print on 22 June 2007.

TABLE 1. Number of taxa at different ranks

| Taxonomy | No. of sequences in corpus | No. of: | | | | | |
|---|---|---|---|---|---|---|---|
| | | Domains | Phyla | Classes | Orders | Families | Genera |
| Bergey's | 5,014 | 1 | 24 | 33 | 79 | 211 | 988 |
| NCBI | 23,095 | 1 | 24 | 31 | 82 | 209 | 1,187 |

## MATERIALS AND METHODS

**Type sequences with Bergey's taxonomy.** Small-subunit rRNA sequences consisting of 5,014 bacterial species type strain sequences in 988 genera (Table 1), along with associated taxonomic assignment information, were obtained from Bergey's *Taxonomic Outline of the Prokaryotes* (release 5.0 [2004]) (13). The sequences averaged 1,460 bases in length and had a range of 1,200 to 1,833 bases. These sequences formed the Bergey corpus. Each sequence was labeled with a set of taxa from domain to genus. From the highest rank to the lowest, the major formal taxonomic ranks are domain, phylum, class, order, family, genus, and species. There are occasional intermediate ranks such as "subclass" and "suborder".

**Complete rRNA database sequences with NCBI's taxonomy.** All available (23,095) near-full-length (≥1,200 bases) 16S rRNA sequences were obtained from the January 2004 release of the RDP. Taxonomic information for these sequences was obtained from the January 2004 release of GenBank. These sequences are classified into 1,187 genera in the NCBI taxonomy (3, 26). Initial placement into the NCBI taxonomy is based on information provided by the sequence submitter, with modifications by the NCBI staff. The sequences averaged 1,454 bases in length. These sequences formed the NCBI corpus.

**Algorithm.** The Classifier uses a feature space consisting of all possible 8-base subsequences (words). Word sizes between 6 and 9 bases were tested in preliminary experiments. Sizes of 8 and 9 bases gave nearly identical results, while sizes of 6 and 7 bases were less accurate, especially with shorter test sequences (not shown). As there are fewer possible words of size 8 than size 9, size 8 was chosen for all further work to reduce memory requirements. The position of a word in a sequence is ignored. As with text-based Bayesian classifiers, only those words occurring in the query contribute to the score (16). A similar word-based classification scheme has been used to search for horizontal gene transfer events in whole-genome sequences (21).

**Word-specific priors.** Let $W = \{w_1, w_2, \ldots, w_d\}$ be the set of all possible eight-character subsequences (words). From the corpus consisting of $N$ sequences, let $n(w_i)$ be the number of sequences containing subsequence $w_i$. The expected-likelihood estimate (determined with the Jeffreys-Perks law of succession) calculated for each word over the entire corpus with the formula $P_i = [n(w_i) + 0.5]/(N + 1)$ was used as a word-specific prior estimate of the likelihood of observing word $w_i$ in an rRNA sequence. The values 0.5 in the numerator and 1 in the denominator keep the probabilities in the range $0 < P_i < 1$.

**Genus-specific conditional probabilities.** For genus $G$ with a training set consisting of $M$ sequences, let $m(w_i)$ be the number of these sequences containing word $w_i$. The conditional probability that a member of $G$ contains $w_i$ was estimated with the equation $P(w_i \mid G) = [m(w_i) + P_i]/(M + 1)$. Ignoring the dependency between words in an individual sequence, the joint probability of observing from genus $G$ a (partial) sequence, $S$, containing a set of words, $V = \{v_1, v_2, \ldots, v_f\}$ ($V \sqsubseteq W$), was estimated as $P(S \mid G) = \Pi\, P(v_i \mid G)$.

**Naïve Bayesian assignment.** By Bayes' theorem, the probability that an unknown query sequence, $S$, is a member of genus $G$ is $P(G \mid S) = P(S \mid G) \times P(G)/P(S)$, where $P(G)$ is the prior probability of a sequence being a member of $G$ and $P(S)$ the overall probability of observing sequence $S$ (from any genus). Assuming all genera are equally probable (equal priors), the constant terms $P(G)$ and $P(S)$ can be ignored. We classify the sequence as a member of the genus giving the highest probability score, but we ignore the actual numerical probability estimate.

**Bootstrap confidence estimation.** For each query sequence, the collection of all eight-character subsequences (words) in the query was first calculated. Normally, when data consist of independent features, a bootstrap sample size equal to the number of features in the original sample is chosen. In this case, the number of completely independent features equals the number of nonoverlapping words. So for each bootstrap trial, a subset of one-eighth of the words was randomly chosen (with replacement) and the words in this subset were then used to calculate the joint probability. The number of times a genus was selected out of 100 bootstrap trials was used as an estimate of confidence in the assignment to that genus. For higher-rank assignments, we sum the results for all genera under each taxon.

**Library comparison.** For each of the greater than 1,000 bacterial taxa, the two samples are compared to estimate the probability that the observed membership could be drawn by chance from a single underlying distribution.

For taxa with greater than five sequences assigned, the standard two-population proportions test is used to estimate the probability of the observed differences (7). The $P$ value is estimated from the $Z$ critical value, where

$$Z = \frac{\dfrac{x}{N_1} - \dfrac{y}{N_2}}{\sqrt{\mu(1 - \mu)\left(\dfrac{1}{N_1} + \dfrac{1}{N_2}\right)}}$$

and where $N_1$ and $N_2$ are the total number of sequences for library 1 and 2, respectively, $x$ and $y$ are the number of sequences assigned to taxon $T$ from library 1 and 2, respectively, and $\mu$ equals $(x + y)/(N_1 + N_2)$.

For taxa with fewer than five sequences assigned, a statistical test first developed for comparing transcript levels in "digital Northern" analysis (2) is used. The probability of the observed difference in assignment to taxon $T$ is estimated as follows:

$$P(y \mid x) = \left(\frac{N_2}{N_1}\right)^y \frac{(x + y)!}{x!y!\left(1 + \dfrac{N_2}{N_1}\right)^{(x+y+1)}}$$

Since each taxon is tested (multiple tests), reported significance values must be interpreted with caution. The taxa (and hence the tests) are nested, and no attempt is made to correct for multiple tests.

**SeqMatch k-NN classifier.** The RDP SeqMatch tool is a k-nearest-neighbor (k-NN) classifier (8). It uses a word-matching strategy not requiring alignment to determine the percentage of shared seven-character words between a query and members of a database of sequences. This tool assigns queries to the lowest taxon shared by the highest-scoring k database sequences. For the tests described here, the Bergey corpus was used as the database, and a query was assigned to the same genus as the closest-matching sequence from the Bergey corpus (effectively, $k = 1$). Setting $k$ to 1 ensured that all queries were classified at all taxonomic ranks for comparison with the RDP Classifier.

**Phylogenetic analysis.** Aligned 16S rRNA sequences were selected at the RDP website and analyzed using the RDP Tree Builder tool. This tool utilizes Weighbor (5), which uses the weighted neighbor-joining algorithm of phylogenetic reconstruction. Weighbor parameters were set to alphabet size 4 and effective sequence length 1,000, and the Jukes-Cantor distance correction was employed. Tree Builder incorporates branch-order confidence estimation using 100 bootstrap samples. The resultant tree was downloaded in PostScript format and embellished with Adobe Illustrator.

**Implementation and availability.** The Classifier engine and related software were written in Java (API v1.4.1) and have been tested on the Solaris (2.8), Linux (2.4.23), Macintosh (OS 10.4), and Microsoft Windows XP operating systems using Java virtual machines from Sun and Apple. The online version of Classifier and Library Compare are built on Java technology, including Java Server Page and Java Servlet Technology (Sun Microsystems).

An online version of the Classifier is available at http://rdp.cme.msu.edu/classifier. A user can submit one sequence or a group of sequences for classification. The sequences can be saved in a file for file upload or inserted into the text area on the start page. The online Classifier requires a sequence of at least 200 bases in length. Up to 10,000 query sequences may be submitted in Fasta, GenBank, and EMBL formats. Single sequences may also be submitted in raw-text format. The Classifier automatically checks both the forward and reverse orientations of the sequence and returns only the results for the correct orientation. Taxonomic assignments above a user-specified confidence threshold are presented in an interactive taxonomic hierarchy. Each line in the hierarchy contains summary information, including the taxon rank and name and the number of query sequences assigned to that taxon. Users can navigate through the hierarchy by clicking on the various taxa. Clicking on the "show assignment detail" link will display the detailed classification results for all or a specified subset of the user's queries. Each result contains the name of a user's sequence, a list of assigned taxa, and the corresponding confidence estimates. In the detail view, a "−" after the sequence name indicates that the minus strand of the sequence was submitted by the user. Results can be downloaded in a text format and imported into a spreadsheet program for further analysis.

The online version of Library Compare is available at http://rdp.cme.msu.edu/comparison. A user can upload two files containing sequences from two libraries. After the sequences are classified, the statistical comparison analysis is

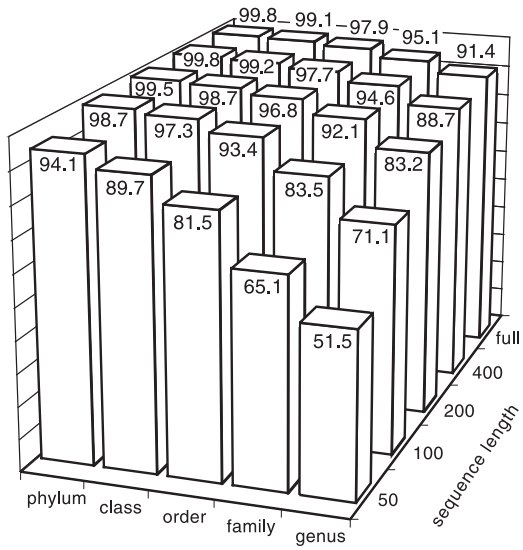


FIG. 1. Overall classification accuracy by query size (exhaustive leave-one-out testing using the Bergey corpus). Numbers are percentages of tests correctly classified.

performed based on the default confidence threshold of 80%. The comparison results are presented in a taxonomic hierarchy for easy navigation. The hierarchy view shows the summary of the assignments and the significance values from the comparison. Taxa with significant differences in representation are marked. The comparison results can also be displayed in a table sorted by significance. The user can choose to change the classification confidence threshold from these two pages. The number of sequences assigned to each taxon will then be recalculated based on the user-specified confidence and compared to obtain a new set of statistical results. Both comparison results and classification assignments can be downloaded in text format and imported into a spreadsheet program for further analysis.

The Classifier command line program trained on Bergey's release 5.0 data, along with the source code, javadoc, example taxonomy and sequence files, and help files, is freely available from http://sourceforge.net/projects/rdp-classifier/ and is released under the terms of the GNU General Public License (http://www.gnu.org/copyleft/gpl.html).

## RESULTS

We tested the Classifier by exhaustive leave-one-out testing on the entire Bergey corpus. For each test, we reserved a single sequence from the corpus as the test sequence and retrained the Classifier on the remaining sequences. The process was repeated for all sequences in the corpus. The corpus contained 453 genera represented by single sequences. For these sequences, the (obviously) incorrect genus assignment was not considered in the summary statistics, but the results were included for higher ranks containing other genera with valid training sets (Fig. 1). In addition to testing the near-full-length sequence, we repeated the leave-one-out testing of the RDP Classifier on small contiguous regions of 400, 200, 100, and 50 bases chosen at random from the test sequence. The overall accuracy for near-full-length and 400-base segments was above 88.7% down to the genus level. For 200-base segments, the accuracy was still above 92.1% down to the family level, while 83.2% were classified accurately at the genus level. The accuracy for the 50-base segments was only 94.1% at the phylum level and decreased dramatically to 51.5% at the genus level.

Bootstrap analysis was used to estimate the confidence of each assignment. Overall, for near-full-length sequences, 97.5% of taxon assignments matched in 95 or more of the 100 bootstrap trials, and these assignments were correct 98% of the time (Tables 2 and 3).

TABLE 2. Classifier accuracy versus bootstrap confidence for the Bergey corpus

| Length of segment (bases) | % of correct classifier assignments within a bootstrap confidence range of[a]: | | | | | |
|---|---|---|---|---|---|---|
| | 100–95% | 94–90% | 89–80% | 79–70% | 69–60% | 59–50% |
| Full | 98.0 | 66.4 | 69.2 | 41.8 | 46.2 | 34.7 |
| 400 | 98.3 | 86.1 | 75.9 | 65.4 | 61.1 | 49.2 |
| 200 | 98.2 | 90.1 | 83.0 | 75.6 | 64.6 | 55.7 |
| 100 | 97.4 | 89.8 | 82.5 | 75.6 | 64.7 | 55.6 |
| 50 | 94.9 | 83.9 | 76.8 | 67.9 | 59.5 | 49.7 |

[a] Bootstrap confidence reflects the frequency of most common assignments out of 100 bootstrap samplings. Percentages of correct assignments at all ranks and within this bootstrap confidence range are shown.

We also conducted a similar series of exhaustive leave-one-out testing using the RDP SeqMatch tool. The overall accuracies for near-full-length test sequences were similar to those of the results for the RDP Classifier: 99.5%, 98.8%, 97.8%, 95.1%, and 91.9% for the ranks phylum, class, order, family, and genus, respectively.

Using the RDP 9 alignment (8), we extracted segments corresponding to 100-base regions of the *Escherichia coli* reference sequence J01695 at 25-base intervals and used each region independently for exhaustive leave-one-out testing, removing the corresponding full-length sequence from training for each test. In general, accuracy was greater for regions mapping to the 16S hypervariable regions (Fig. 2A). The average bootstrap confidence estimate was similarly higher in the hypervariable regions (Fig. 2B).

For each of the 393 near-full-length sequences misclassified at the genus level, we determined the distance to other members of the Bergey corpus by calculating the pairwise identity between sequences using the corresponding aligned sequences from the RDP website. For 250 of the misclassified sequences, 1 or more sequences in different genera were closer than all other sequences in the genus to which the sequence was assigned in the taxonomy (Table 4).

To understand the nature of three of these misclassifications, we explored sequences from the *Alicyclobacillaceae* family. We chose this family because it contains three sequences misclassified by the RDP Classifier, only one of which was also

TABLE 3. Percentage of matching bootstrap assignments for various query lengths for the Bergey corpus

| Length of segment (bases) | % of classification tests within a bootstrap confidence range of[a]: | | | | | |
|---|---|---|---|---|---|---|
| | 100–95% | 94–90% | 89–80% | 79–70% | 69–60% | 59–50% |
| Full | 97.5 | 0.5 | 0.5 | 0.4 | 0.4 | 0.4 |
| 400 | 93.4 | 1.5 | 1.4 | 1.0 | 0.9 | 0.9 |
| 200 | 86.3 | 3.2 | 2.8 | 2.0 | 1.9 | 1.6 |
| 100 | 70.5 | 6.5 | 6.6 | 4.2 | 3.7 | 3.1 |
| 50 | 46.2 | 10.1 | 10.6 | 7.4 | 6.8 | 6.2 |

[a] Frequency of most common assignments out of 100 bootstrap samplings. Percentages of classification tests that were within the specified range of most common bootstrap assignment are indicated.

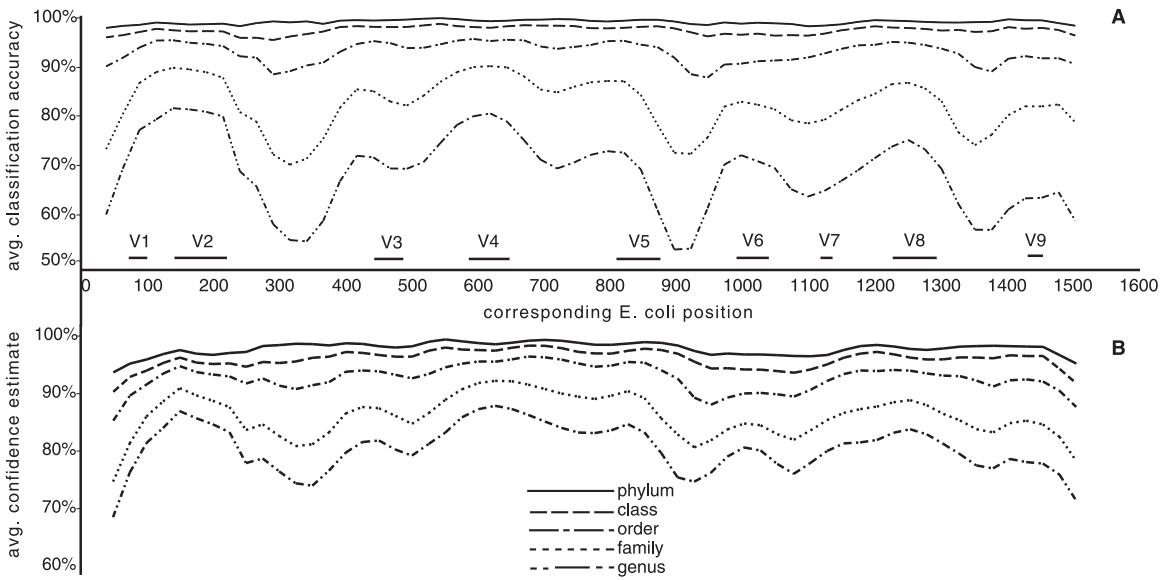


FIG. 2. (A) Classification accuracy rate for the Bergey corpus with sequence segments of 100 bases, moving 25 bases a time. The gray bars on the *x* axis define the hypervariable regions. The average classification accuracy rate at the genus level was 70% over all 100-base regions. (B) Average bootstrap confidence estimate for each segment.

misclassified by SeqMatch. Bergey's taxonomic placement of these three sequences appeared not to be congruent with the derived phylogeny (Fig. 3).

We tested the performance of the Classifier on a more diverse corpus consisting of 23,095 16S rRNA sequences as classified by the NCBI in its taxonomy database. We randomly chose a test set of 20% (4,619) of these for leave-one-out testing, as described above. The accuracy using the NCBI taxonomy database and various query lengths was very similar to the previous results (Table 5).

## DISCUSSION

The RDP Classifier was developed to provide rapid taxonomic placement based on rRNA sequence data. As technical improvements have made it easier to obtain rRNA data, the use of these data has spread beyond the dedicated molecular phylogenist. For many such users, a complete and thorough phylogenetic analysis may not be an option. Instead, these users may simply want a rapid method that provides a taxonomic placement for their unknowns. In addition, high-throughput environmental rRNA projects routinely produce hundreds to hundreds of thousands of sequences per sample. It is unrealistic to imagine that all of these sequences will be examined in detail. For these high-throughput experiments, the RDP Classifier can provide rapid taxonomic placement and summary data, including the number of input sequences belonging to each taxon.

For the near-full-length and 400-base partial rRNA sequences, the Classifier was accurate down to the genus level, while even with 200-base partial sequences, the Classifier was accurate down to the family level. The Classifier did not perform well for partial sequences with a length of 50 bases, likely due to insufficient features provided by such short partial sequences.

For shorter sequences, classification confidence was greatly improved by using the bootstrap to estimate classification reliability (Table 2). In our tests, most classification assignments were made with high confidence. For near-full-length sequences, most misclassifications were probably due to errors in the underlying taxonomy (see below), but for shorter sequences, misclassifications may reflect the lack of data. The bootstrap helps determine whether available data are sufficient for a robust classification. The majority of available rRNA sequences are short (under 1,200 bases in length), and short partial sequences are commonly used for environmental survey sequencing.

Naïve Bayesian classifiers are one example of supervised classification methods. The k-NN method is another supervised classification method that is arguably simpler and often works well in practice (16). We compared the RDP naïve Bayesian Classifier with the RDP SeqMatch k-NN classifier.

TABLE 4. Number of genus misclassifications for near-full-length sequences by phylum (the Bergey corpus)

| Phylum[a] | Total[b] | No. (%) misclassified | No. (%) of taxonomic anomalies[c] |
|---|---|---|---|
| *Firmicutes* | 1,295 | 165 (12.7) | 80 (6.2) |
| *Proteobacteria* | 1,641 | 154 (9.4) | 115 (7.0) |
| *Actinobacteria* | 1,220 | 31 (2.5) | 23 (1.9) |
| *Bacteroidetes* | 192 | 27 (14.1) | 20 (10.4) |
| *Cyanobacteria* | 14 | 6 (42.9) | 6 (42.9) |
| *Fusobacteria* | 33 | 4 (12.1) | 2 (6.1) |
| *Aquificae* | 12 | 2 (16.7) | 1 (8.3) |
| *Spirochaetes* | 50 | 2 (4.0) | 2 (4.0) |
| *Chlorobi* | 12 | 2 (16.7) | 1 (8.3) |
| Other phyla | 92 | 0 (0) | 0 (0) |
| Overall | 4,561 | 393 (8.6) | 250 (5.5) |

[a] Only phyla with at least one misclassification are listed.
[b] Number of test sequences excluding singletons at the genus level.
[c] Numbers of misclassified sequences with their nearest neighbors in a different genus, indicating that the genus is not monophyletic.

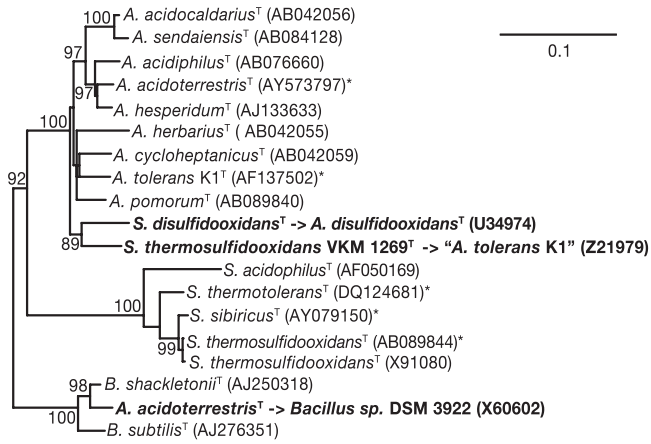


FIG. 3. Phylogenetic analysis of the *Alicyclobacillaceae*, including the genera *Sulfobacillus* and *Alicyclobacillus*. Sequences for each of the 11 species type strains of *Alicyclobacillaceae* available in release 5.0 of Bergey's taxonomic outline, along with additional sequences for four *Alicyclobacillaceae* species type strains that became available after the release of the outline (marked with an asterisk), and two *Bacillus* species type strains were analyzed using the weighted neighbor-joining method (5). The tree is rooted using *Escherichia coli* sequence J01695 as the outgroup. Bootstrap confidence estimates above 85% are shown. Three misclassifications made by the RDP Classifier are highlighted, with the original (release 5.0) description appended with a corrected description. *S. disulfidooxidans*$^T$ became *A. disulfidooxidans*$^T$. In 2005, *S. disulfidooxidans* was formally reclassified as a new combination, *A. disulfidooxidans* (14). *S. thermosulfidooxidans* VKM 1269$^T$ became "*A. tolerans* K1." In release 5, sequence Z21979 was listed as coming from the type strain (VKM 1269) of *S. thermosulfidooxidans*. This agrees with the original publication for the sequence (25). The same group later reported that the sequence was probably from *S. thermosulfidooxidans* strain K1, not VKM 1269 (15). Two independent sequences (X91080 and AB089844) for the type strain of *S. thermosulfidooxidans* are available. They are nearly identical to each other (0.2% difference) and 19% different from Z21979. In 2005, *S. thermosulfidooxidans* strain K1 was reclassified as the type strain of a new species, *A. tolerans* (14). In our analysis, however, the sequence for K1 given in the naming paper (accession number AF137502) is 8% different from that for Z21979. Although still listed in the GenBank record as from *A. tolerans* strain K1, Z21979 is most probably not from strain K1 and not even from a member of the species *A. tolerans*. *A. acidoterrestris*$^T$ became *Bacillus* sp. Sequence X60602 was published in 1991 as from the type strain (DSM 3922) for *A. acidoterrestris* (1). The GenBank record also lists this as from strain DSM 3922, but a strain mix-up in the culture supplied by the DSMZ was reported in 1992 (27). The type strain (ATCC 49025, listed by the DSMZ as equivalent to DSM 3922) was resequenced in 2005 (AY573797 [9]). These two sequences are 14% different, and X60602 is now described by GenBank as from a *Bacillus* sp.

We chose to use SeqMatch for the comparison, because it, like the RDP Classifier, is offered through the RDP. Also, SeqMatch has been shown to be more accurate than BLAST in finding the most similar rRNA sequence as defined by pairwise aligned distance (8). The RDP Classifier and RDP SeqMatch had nearly identical overall error rates. However, only about two-thirds of these errors were in common (not shown). For both classifiers, many of these misclassifications appeared to result from errors in the underlying taxonomy (not shown); however, the two classifiers often responded differently to specific taxonomic anomalies.

Many, if not most, of the 393 misclassifications made by the RDP Classifier with the Bergey corpus are likely due to discrepancies between the underlying (16S rRNA) phylogeny and Bergey's taxonomy. Over 60% of misclassified sequences were actually more similar to sequences in other genera than to sequences in their own genera, strong evidence that these genera may not be monophyletic. These misclassifications are not evenly distributed among the taxa. Of the three most represented phyla (*Firmicutes*, *Proteobacteria*, and *Actinobacteria*), the *Actinobacteria* had the fewest errors (3%, or 31 in 1,220 trials), while the misclassification rate for the *Firmicutes* was about five times as high (13%, or 165 in 1,295 trials). Within the *Firmicutes*, more than two-thirds of the errors occurred in the class *Clostridia* (108 of 449 trials), and within that class, the genus *Eubacterium* stood out, with 30 misclassifications in 41 trials (see the supplemental material).

To identify an example of the sources of errors in the taxonomy, we conducted a phylogenetic analysis of classification errors in the family *Alicyclobacillaceae* (Fig. 3). The RDP Classifier misclassified three sequences from the *Alicyclobacillaceae*: *Sulfobacillus disulfidooxidans* and *Sulfobacillus thermosulfidooxidans*, which were both misclassified in the genus *Alicyclobacillus*, and *Alicyclobacillus acidoterrestris*, which was misclassified in the genus *Bacillus*. SeqMatch misclassified only *A. acidoterrestris*. Interestingly, these three "errors" made by the RDP Classifier appear to be in agreement with phylogenetic analysis. Literature searches confirmed that, subsequent to the publication of version 5.0 of Bergey's taxonomic outline, all three had been reevaluated in line with the classification given by the RDP Classifier. The strains originally attributed to *Sulfobacillus* were both reevaluated in 2005 and reassigned to the genus *Alicyclobacillus*, matching the classification of the RDP Classifier but no longer matching the classification from SeqMatch. The sequence originally attributed to *A. acidoterrestris* was the result of a culture mix-up and belongs in the genus *Bacillus*, as assigned by both classifiers. Unlike the SeqMatch nearest-neighbor classifier, the RDP naïve Bayesian Classifier uses information averaged over the entire genus and hence is less influenced by individual misplaced training sequences.

Pyrosequencing can now produce up to 100 Mb in less than 8 h (Roche product literature). Sogin and colleagues recently used the first version of pyrosequencing, which provided up-to-100-bp reads, to rapidly analyze large numbers of 16S rRNA V6 variable regions from environmental samples (23). In that study, taxonomic diversity was assessed through alignment to members of a reference database of vetted V6 sequences. The classification accuracy of the RDP Classifier for 100-base fragments varied greatly by region along the 16S rRNA molecule (Fig. 2). Variable regions were, in general, more accurately

TABLE 5. Classifier accuracy at various query lengths (NCBI's taxonomy)

| Length of segment (bases) | % of segments accurately identified in: | | | | |
|---|---|---|---|---|---|
| | Phylum | Class | Order | Family | Genus |
| Full | 99.8 | 99.3 | 98.6 | 97.1 | 92.1 |
| 400 | 99.7 | 99.3 | 98.5 | 97.0 | 90.4 |
| 200 | 99.7 | 99.2 | 98.1 | 95.7 | 86.6 |
| 100 | 99.2 | 98.4 | 95.7 | 88.9 | 74.9 |
| 50 | 94.6 | 90.9 | 81.6 | 69.2 | 52.8 |

classified than more-conserved regions. Variable regions V2 and V4 (20) were classified correctly at rates as high as 82% and 90% at the genus and family levels, respectively (at all levels of confidence), while accuracy for fragments containing the V6 region were less reliably classified (73% at the genus level). The V6 region is rather small (46 bases) and flanked by highly conserved regions (6). In contrast, the V2 region is flanked on both sides and the V4 region is flanked on the 3′ side by semiconserved regions. These regions likely account for both the increased accuracy for and the broadened peaks around these variable regions. These regions may make good targets as the read length of pyrosequencing technology increases.

In bacteriology, although nomenclature is governed by an official code, there is no definitive or official taxonomy. The RDP Classifier is not limited to using the bacterial taxonomy proposed in Bergey's *Taxonomic Outline of the Prokaryotes* (13). We chose Bergey's taxonomy for use because it is readily available and *Bergey's Manual of Systematic Bacteriology* is widely respected in the microbiological community. The RDP Classifier worked equally well when trained on the NCBI taxonomy. There is significant congruence between the two taxonomies, and both are based, at least in part, on phylogenetic principals. The RDP Classifier likely can be adapted to additional phylogenetically coherent bacterial taxonomies that may gain acceptance in the future.

For query sequences from regions of bacterial diversity with less-defined taxonomy, the RDP Classifier tends to provide classification results with low confidence estimates. For example, in one library of *Acidobacteria* environmental clone sequences, 72 of 77 sequences were classified with less than 80% confidence even at the phylum level (not shown). Such low confidence classification results may help identify sequences for which a more thorough analysis is likely to give the highest payoff.

The RDP Classifier is fast enough to handle large sample volumes. For example, on a 2.66-GHz Apple Intel Xeon processor, the program classifies approximately nine sequences per second (with 100 bootstrap samples of each). The online version of the Classifier returns results for submissions of up to 10,000 user sequences in a few minutes and provides both summary and detailed classification assignments. The Library Compare tool enables the analysis of microbial community composition based on the taxonomic group assignments. This tool leverages the Classifier to provide rapid comparison of two samples, each containing up to 5,000 sequences. The results can be explored interactively and can also be downloaded in a form suitable for importing into common spreadsheet programs.

The RDP has been using the RDP Classifier with Bergey's taxonomy internally for over 4 years to give order to its collection of over 300,000 bacterial rRNA sequences. During that time, the RDP has updated the Classifier to use data from three successive versions of Bergey's taxonomic outline. This taxonomy is still evolving as species are reevaluated and discrepancies are resolved. As these updates have been released, it has proved relatively simple to retrain the Classifier and update the assignments for sequences in the RDP library. As bacterial taxonomy continues to evolve, we expect the Classifier's performance to continue to improve.

## ACKNOWLEDGMENTS

This research was supported by the Office of Science (BER), U.S. Department of Energy, grant DE-FG02-99ER62848, and the National Science Foundation, grant DBI-0328255.

We thank Sue Barnes for allowing us access to her unpublished sequence data and Phillip Neal for constructive suggestions.

## REFERENCES

1. **Ash, C., J. A. E. Farrow, S. Wallbanks, and M. D. Collins.** 1991. Phylogenetic heterogeneity of the genus bacillus revealed by comparative analysis of small subunit ribosomal RNA sequences. Lett. Appl. Microbiol. **13:**202–206.
2. **Audic, S., and J. M. Claverie.** 1997. The significance of digital gene expression profiles. Genome Res. **7:**986–995.
3. **Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler.** 2000. GenBank. Nucleic Acids Res. **28:**15–18.
4. **Brown, M. P. S.** 1999. RNA modeling using stochastic context-free grammars. Ph.D. thesis. University of California, Santa Cruz.
5. **Bruno, W. J., N. D. Socci, and A. L. Halpern.** 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. Mol. Biol. Evol. **17:**189–197.
6. **Cannone, J. J., S. Subramanian, M. N. Schnare, J. R. Collett, L. M. D'Souza, Y. Du, B. Feng, N. Lin, L. V. Madabusi, K. M. Muller, N. Pande, Z. Shang, N. Yu, and R. R. Gutell.** 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.
7. **Christensen, H. B.** 1992. Introduction to statistics: a calculus-based approach, 1st ed., p. 510–512. Harcourt Brace Jovanovich, Inc., Orlando, FL.
8. **Cole, J. R., B. Chai, R. J. Farris, Q. Wang, S. A. Kulam, D. M. McGarrell, G. M. Garrity, and J. M. Tiedje.** 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res. **33:**D294–D296.
9. **Connor, C. J., H. Luo, B. B. M. Gardener, and H. H. Wang.** 2005. Development of a real-time PCR-based system targeting the 16S rRNA gene sequence for rapid detection of *Alicyclobacillus* spp. in juice products. Int. J. Food Microbiol. **99:**229–235.
10. **DeSantis, T. Z., I. Dubosarskiy, S. R. Murray, and G. L. Andersen.** 2003. Comprehensive aligned sequence construction for automated design of effective probes (CASCADE-P) using 16S rDNA. Bioinformatics **19:**1461–1468.
11. **Domingos, P., and M. Pazzani.** 1997. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning **29:**103–130.
12. **Garrity, G. M., J. A. Bell, and D. B. Searles.** 2001. Taxonomic outline of the prokaryotes. Bergey's manual of systematic bacteriology, 2nd ed., release 1.0. Springer-Verlag, New York, NY.
13. **Garrity, G. M., J. A. Bell, and T. G. Lilburn.** 2004. Taxonomic outline of the procaryotes. Bergey's manual of systematic bacteriology, 2nd ed., release 5.0. Springer-Verlag, New York, NY.
14. **Karavaiko, G. I., T. I. Bogdanova, T. P. Tourova, T. F. Kondrat′eva, I. A. Tsaplina, M. A. Egorova, E. N. Krasil'nikova, and L. M. Zakharchuk.** 2005. Reclassification of '*Sulfobacillus thermosulfidooxidans* subsp. *thermotolerans*' strain K1 as *Alicyclobacillus tolerans* sp. nov. and *Sulfobacillus disulfidooxidans* Dufresne et al. 1996 as *Alicyclobacillus disulfidooxidans* comb. nov., and emended description of the genus *Alicyclobacillus*. Int. J. Syst. Evol. Microbiol. **55:**941–947.
15. **Karavaïko, G. I., T. P. Turova, I. A. Tsaplina, and T. I. Bogdanova.** 2000. The phylogenetic position of aerobic, moderately thermophilic bacteria of the Sulfobacillus species, oxidizing $Fe^{2+}$, $S^0$ and sulfide minerals. Mikrobiologiia **69:**857–860.
16. **Li, Y. H., and A. K. Jain.** 1998. Classification of text documents. Comput. J. **41:**537–546.
17. **Lozupone, C., and R. Knight.** 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Appl. Environ. Microbiol. **71:**8228–8235.
18. **Maidak, B. L., N. Larsen, M. J. McCaughey, R. Overbeek, G. J. Olsen, K. Fogel, J. Blandy, and C. R. Woese.** 1994. The Ribosomal Database Project. Nucleic Acids Res. **22:**3485–3487.
19. **Martin, A. P.** 2002. Phylogenetic approaches for describing and comparing the diversity of microbial communities. Appl. Environ. Microbiol. **68:**3673–3682.
20. **Neefs, J. M., Y. Van de Peer, P. De Rijk, S. Chapelle, and R. De Wachter.** 1993. Compilation of small ribosomal subunit RNA structures. Nucleic Acids Res. **21:**3025–3049.
21. **Sandberg, R., G. Winberg, C. I. Branden, A. Kaske, I. Ernberg, and J. Coster.** 2001. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. Genome Res. **11:**1404–1409.
22. **Singleton, D. R., M. A. Furlong, S. L. Rathbun, and W. B. Whitman.** 2001. Quantitative comparisons of 16S rRNA gene sequence libraries from environmental samples. Appl. Environ. Microbiol. **67:**4374–4376.
23. **Sogin, M. L., H. G. Morrison, J. A. Huber, D. M. Welch, S. M. Huse, P. R. Neal, J. M. Arrieta, and G. J. Herndl.** 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc. Natl. Acad. Sci. USA **103:**12115–12120.

24. **Stackebrandt, E., W. Frederiksen, G. M. Garrity, P. A. D. Grimont, P. Kämpfer, M. C. J. Maiden, X. Nesme, R. Rosselló-Mora, J. Swings, H. G. Trüper, L. Vauterin, A. C. Ward, and W. B. Whitman.** 2002. Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. Int. J. Syst. Evol. Microbiol. **52:**1043–1047.
25. **Turova, T. P., A. B. Poltoraus, I. A. Lebedeva, E. S. Bulygina, I. A. Tsaplina, T. I. Bogdanova, and G. I. Karavaiko.** 1995. Determination of the phylogenetic position of *Sulfobacillus thermosulfidooxidans* on the basis of analysis of the 5S and 16S ribosomal RNA. Mikrobiologiia **64:**366–374.
26. **Wheeler, D. L., C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova, and B. A. Rapp.** 2000. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. **28:**10–14.
27. **Wisotzkey, J. D., P. Jurtshuk, G. E. Fox, G. Deinhard, and K. Poralla.** 1992. Comparative sequence analyses on the 16S rRNA (rDNA) of *Bacillus acidocaldarius*, *Bacillus acidoterrestris*, and *Bacillus cycloheptanicus* and proposal for creation of a new genus, *Alicyclobacillus* gen. nov. Int. J. Syst. Bacteriol. **42:**263–269.
28. **Woese, C. R., O. Kandler, and M. L. Wheelis.** 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA **87:**4576–4579.