APPLIED AND ENVIRONMENTAL MICROBIOLOGY, Apr. 1995, p. 1240–1245 Vol. 61, No. 4
0099-2240/95/$04.00+0

Evaluation of Nearest-Neighbor Methods for Detection of Chimeric Small-Subunit rRNA Sequences

J. F. ROBISON-COX,[1]* M. M. BATESON,[2] AND D. M. WARD[2]

Department of Mathematical Sciences[1] and Department of Microbiology,[2] Montana State University, Bozeman, Montana 59717

Received 3 October 1994/Accepted 13 January 1995

Detection of chimeric artifacts formed when PCR is used to retrieve naturally occurring small-subunit (SSU) rRNA sequences may rely on demonstrating that different sequence domains have different phylogenetic affiliations. We evaluated the CHECK_CHIMERA method of the Ribosomal Database Project and another method which we developed, both based on determining nearest neighbors of different sequence domains, for their ability to discern artificially generated SSU rRNA chimeras from authentic Ribosomal Database Project sequences. The reliability of both methods decreases when the parental sequences which contribute to chimera formation are more than 82 to 84% similar. Detection is also complicated by the occurrence of authentic SSU rRNA sequences that behave like chimeras. We developed a naive statistical test based on CHECK_CHIMERA output and used it to evaluate previously reported SSU rRNA chimeras. Application of this test also suggests that chimeras might be formed by retrieving SSU rRNAs as cDNA. The amount of uncertainty associated with nearest-neighbor analyses indicates that such tests alone are insufficient and that better methods are needed.

We and others have detected chimeric small-subunit (SSU) rRNA sequences when applying PCR methods to retrieve SSU rRNA sequences from natural microbial communities or mixed cultures (1, 2, 8, 12). The frequency of occurrence of these artifacts in clone libraries has been reported to range from 4.1 to 20% (1, 2, 8). It is obviously important both to minimize chimera formation (7) and to be able to detect chimeric sequences in order to ensure that the diversity of SSU rRNAs we detect in nature is real and to ensure the quality of growing SSU rRNA sequence databases like the Ribosomal Database Project (RDP) (10). This is especially important given the popularity of the PCR method in retrieving SSU rRNA sequences from nature and the likelihood that such databases will soon become dominated by environmentally derived SSU rRNA sequences, as opposed to SSU rRNA sequences of cultivated species.

Detection of chimeric SSU rRNA sequences is sometimes possible by observing base pair mismatches in secondary structures (6), but this method is not fail-safe given that some chimeras do not exhibit such abnormalities (8). Chimeras can also be detected by demonstrating that separate domains of an unknown SSU rRNA sequence are identical to different known sequences (8, 12). However, natural habitats have been shown to contain mainly uncultivated microorganisms (1, 2, 4–6, 11, 13, 16), whose SSU rRNA sequences are only now being determined. The chance that databases contain the sequences that are identical to specific domains of chimeric sequences and thus useful for the detection of chimeras depends on the degree to which the species of the habitat have been characterized by SSU rRNA analysis. Thus, methods based on demonstrating that different domains of a query sequence have different phylogenetic affiliations have been developed to detect chimeras (1, 5, 10). This difference can be demonstrated by comparing phylogenetic trees for different sequence domains or by pairwise similarity analysis of the domains. An uncertainty assessment has not been published for either approach to comparing phylogenetic affiliation. In this study, we evaluated two methods for detecting chimeras based on pairwise similarity analysis. Our main objective was to evaluate the CHECK_CHIMERA method of the RDP, which is based on determining nearest neighbors in variable unaligned sequence domains. We also developed and evaluated a method based on determining nearest neighbors to a query sequence over two defined sequence domains to observe whether the additional information provided by alignment aided in chimera detection.

* Corresponding author. Mailing address: Department of Mathematical Sciences, 2-214 Wilson Hall, Montana State University, Bozeman, MT 59717. Phone: (406) 994-5340. Fax: (406) 994-6879. Electronic mail address: jimrc@math.montana.edu.

MATERIALS AND METHODS

Though we evaluated mainly the CHECK_CHIMERA method, we describe our method first to facilitate the explanation of common features of the two approaches.

Aligned similarity method. We designed the aligned similarity method to take advantage of the alignment of homologous regions of sequences in the RDP database, since the alignment is an additional source of information beyond sequence alone. The concept was to evaluate whether distantly separated sequence domains have different nearest neighbors. Comparison of remote domains would maximize the probability of detecting fragments with different phylogenetic affiliations. The effort expended in acquiring sequence data would also be reduced if sequences could be shown to be chimeric with limited amounts of sequence data. As shown in Fig. 1a, domain 1 is a region of interest in the 5′ end of the molecule, and domain 2 is a region of interest in the 3′ end. The aligned similarity score (matching bases/total bases) was used to find the most similar nonidentical sequences in the database. S_1 was the sequence producing the highest degree of similarity to the query sequence over domain 1, S_2 showed the highest degree of similarity over domain 2, and S_{both} was most similar over the combined domains. An improvement score (IS) was calculated as follows to quantify the increase in proportion of matches obtained by permitting the program to pick two most-similar sequences rather than one overall most-similar sequence.

$$IS = \frac{\text{base matches between } S_1 \text{ and domain 1} + \text{base matches between } S_2 \text{ and domain 2}}{\text{number of bases compared}} - \frac{\text{base matches between } S_{both} \text{ and both domains}}{\text{number of bases compared}}$$

If S_1, S_2, and S_{both} are all the same sequence, then the IS will be zero. If the two domains have different most-similar sequences, then the IS will be positive since permitting the program to pick two nearest sequences instead of just one can only increase the proportion of similar bases. Large values of IS should occur if the two domains have different phylogenetic affiliations.

The sequences available in the RDP release 2.0 (January 1993) were used to find the background level IS for nonchimeric sequences (environmentally re-

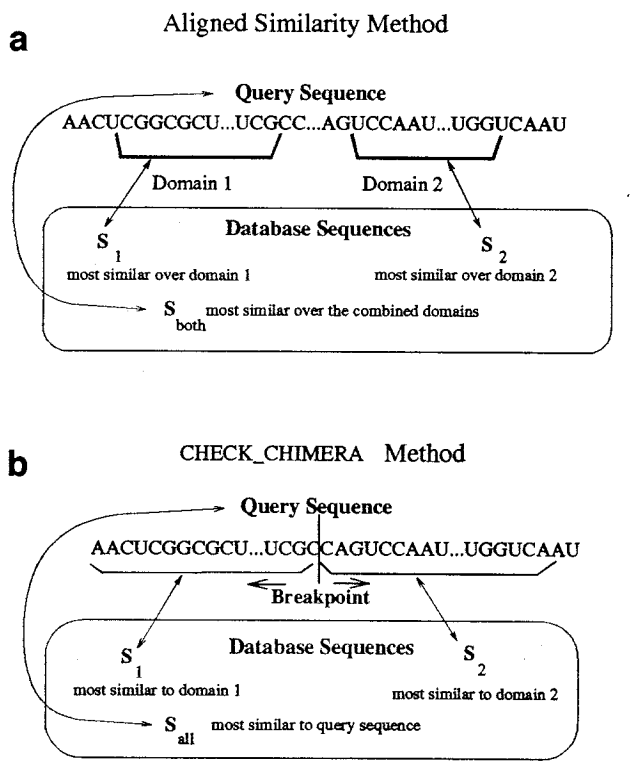

FIG. 1. Calculation of IS. (a) Aligned similarity method. Given a query sequence and defined domains 1 and 2, the most similar sequence in the database over domain 1 is labeled S_1, the nearest over domain 2 is labeled S_2, and the nearest over both domains is S_{both}. The IS is the proportion of matching bases over the two domains with S_1 and S_2 minus the proportion of matches with S_{both}. (b) CHECK_CHIMERA method. Given a query sequence and a breakpoint, S_1 is the sequence in the database most similar to the query over domain 1, S_2 is most similar over domain 2, and S_{all} is most similar overall. The IS is the number of oligomers shared by domain 1 of the query and S_1 plus those shared by domain 2 of the query and S_2 minus the number of oligomers shared by the query sequence and S_{all}. These calculations are repeated for breakpoints along the entire sequence.

trieved sequences were not used, since the chimeras observed to this point have been derived during retrieval of SSU rRNA from the environment). We first defined domains that were 200 bases long beginning in highly conserved regions which might serve as primer sites. A mask was used to remove sites which were conserved in 90% or more of the sequences (the noninformative regions) and sites which were conserved in less than 40% of the sequences (where misalignment is more probable). Artificial chimeras were created by picking two parent sequences from the release 2.0 database and splicing fragments from the appropriate domains together on the computer. Parent sequence 1 was chosen at random from all the full-length sequences, and parent sequence 2 was chosen from the 50 sequences nearest parent 1 in the RDP release 2.0 phylogenetically ordered list. The restricted randomization for parent sequence 2 was used to obtain a full range of similarities between parent sequences, since unrestricted random choice produced too few chimeras from closely related parents.

CHECK_CHIMERA method. Larsen et al. (10) of RDP have developed a method to screen for chimeric SSU rRNA sequences. The program and documentation are continuously being modified and improved; the description below was current at the time of this writing, but users should consult the RDP help facility for updated explanations and watch for further publications from Larsen. The similarities between his method and the above-described aligned similarity method stem from our intentionally mimicking his use of an IS based on the nearest neighbors in the database. CHECK_CHIMERA produces an IS by comparing two domains (Fig. 1b), but the two domains are split by a moving breakpoint and an IS is given for breakpoints at each 10th position of the sequence. CHECK_CHIMERA uses oligomer matches (currently seven bases long) as a measure of phylogenetic similarity (Sab), thus avoiding the need to align sequences. The Sab value is the number of shared oligomers (each unique oligomer is counted only once) divided by the number of unique oligomers in the shorter sequence. At a given breakpoint, the IS is calculated as

IS = oligomers shared by S_1 and domain 1 + oligomers shared by S_2 and domain 2 − oligomers shared by S_{all} and query,

where S_1 and S_2 are the database sequences most similar to the query sequence over domains 1 and 2, respectively, and S_{all} is the database sequence most similar to the entire query sequence. Like the IS for the aligned similarity method, this score will be zero if S_1, S_2, and S_{all} are all the same sequence, and scores cannot be smaller than zero. If a sequence is chimeric, the IS should increase to a maximum IS (MIS) as the breakpoint is moved closer to the point of chimera formation and should then decrease as the breakpoint moves away. The output, a plot of IS versus sequence position, is intended to permit users to make their own judgments about the possibly chimeric nature of a query sequence.

For analysis of this method, the 1,698 full-length nonenvironmental sequences from RDP release 4.0 (June 1994) were used as a nonchimeric background. The background sequences and the artificial chimeras described below were processed by the RDP server between 27 July and 17 August 1994, using the 1,744 sequences in the RDP database with at least 1,200 bases as the reference database. The graphical output from the server was converted to numerical scores by using the axis provided with each output graph. Artificial chimeras (900 sequences) were created by choosing parent sequence 1 at random and choosing parent sequence 2 at random from the 100 sequences closest to the first in the RDP release 3.0 phylogenetically ordered list. Points of chimera formation were chosen randomly at least 200 bases from the ends of the sequences.

Estimated similarity and regression. For the aligned similarity method, we computed the estimated similarity between two possibly parental domains of a query sequence by computing the similarity (based on the RDP release 2.0 alignment) of the entire sequences of the nearest neighbors of each domain, i.e., between S_1 and S_2. When CHECK_CHIMERA was used, the estimated similarity was computed in the same way by the use of the RDP release 4.0 alignment, with the two domains defined as the 5′ and 3′ ends of the query sequence split by the breakpoint which produced the MIS.

Weighted regression was used to describe the relationship between IS or MIS and estimated similarity. Ordinary least-squares regression was not appropriate for these data, since the standard deviation of the response IS or MIS was not constant but changed with the estimated similarity (see Results). As the standard deviation of IS or MIS decreased linearly with estimated similarity, it was possible to correct for the change in spread by using weighted regression. Using this method, we found the equation of the best-fitting line by minimizing

$$\sum \frac{(y_i - \hat{y}_i)^2}{(1 - x_i)^2}$$

where y_i is the observed IS or MIS value, x_i is the estimated similarity, and $\hat{y}_i = a + bx_i$ is the predicted IS or MIS value on the line of best fit for a given x_i. The numerator terms are the squared residuals or deviations from the line, and the denominators weight each observation, with more weight given to observations of higher estimated similarity values which have smaller variance. The computation of lower prediction bounds for IS or MIS at a given estimated similarity was done as in an ordinary regression except that the distribution of residuals ($y_i - \hat{y}_i$) was assumed to be normal, with standard deviations proportional to $1 - x_i$.

RESULTS

Aligned similarity method. We initially used domains of 200 bases beginning at conserved primer sites, but tests often failed to detect artificial chimeras. Masking out highly variable and highly conserved regions improved discrimination only slightly. To examine the limitations of this method, the results reported here focus on the case which uses the entire sequence in order to find the greatest power of the methodology. The two domains were taken to be the 3′ and 5′ halves of the SSU rRNA molecule, splitting at position 1100 in the RDP release 2.0 alignment, which is position 745 of the *Escherichia coli* SSU rRNA sequence. Examination of the limiting case does not provide a workable test for new, possibly chimeric sequences, since it is very unlikely that the point of chimera formation will be in the middle of the molecule. Rather, we were testing the feasibility of using this method at all. The IS was computed for each of 639 sequences in the database, and the distribution of IS was observed (Fig. 2a). Note that 95% of the sequences have an IS below 0.009. If new sequences yield IS with the same distribution as that observed for these 639 sequences, then calling a query sequence chimeric when the IS is over 0.009 will result in misclassifying a new authentic sequence 5% of the time. If a chimera were to have an IS below 0.009, it would be misclassified as authentic. Of the 500 artificial chi-

meras tested, 14% had ISs below 0.009 (Fig. 2b) and would not be detected as chimeras by the use of this fixed cutoff.

To estimate the nondetection probabilities more accurately, the effect of similarity of parents was included in the analysis by calculating the estimated similarity of the query sequence and relating it to the IS by using weighted regression. The relationship between IS and estimated similarity is shown in Fig. 3 for 500 artificial chimeras. The detection limit of 0.009 is the horizontal line in this plot, so data below the line represent chimeras which were not detected. The broken lines in Fig. 3 show 80, 90, and 95% lower prediction limits for a new observed chimera. Since the 95% lower bound crosses the detection limit line when estimated similarity is 82%, we can say that we are 95% confident that a new chimera will be detected if the estimated parental similarity is no more than 82%. If we are willing to take greater risk, the probability of detection is 90% when estimated similarity is 89%, 80% when estimated similarity is 93%, and 50% when estimated similarity is 96%. These results are based on the very optimistic scenario in which the point of chimera formation is the middle of the molecule. Chimeras formed at points other than the middle of the molecule would not be as easy to detect as those created for this demonstration, but the results discussed here were computed to examine the limits of the aligned similarity method under optimal conditions.

CHECK_CHIMERA method. Typical outputs of CHECK_CHIMERA are plotted in Fig. 4a, where chimeras (solid lines) are readily distinguishable from nonchimeric sequences (dashed lines). Figure 4b illustrates the detection problem. Some nonchimeric background sequences (dashed lines) exhibit high MISs (up to 194), and some chimeric sequences (solid lines) exhibit relatively low MISs. The actual similarities between the full-length parental sequences of the chimeras 1 through 6 plotted in Fig. 4 are 75, 81, 92, 93, 95, and 97%, respectively.

To summarize these plots, we concentrated on the MIS, because in an analysis of numerical summary measures of the plots (MIS, peakedness, mean, median, and standard deviation) the MIS was by far the most informative. The other statistics did not improve discrimination of artificial chimeras from nonchimeras. The distribution of MISs for 1,698 authentic sequences from release 4.0 of RDP showed that 95% of the MIS were below 55. The use of 55 as the detection cutoff resulted in misclassification of 20% of the artificial chimeras. Again, detection was easier when parent sequences were more distantly related, so MIS was viewed as a function of estimated similarity (Fig. 5) and was modeled by the same methods used for IS and estimated similarity in the aligned similarity method described above. Since the 95% lower bound crosses the detection limit line at an estimated similarity of 84%, we can be 95% confident that a newly observed chimera will be detected when the estimated similarity of the parents is no more than 84%. When estimated similarity is 89, 92, and 95%, we can be 90, 80, and 50% confident, respectively, that the chimera would be detected. This use of an MIS of ≥55 as an indication of a chimera is a naive use of the CHECK_CHIMERA output, since expert opinion is ignored. By stating a specific rule for chimera indication, we are able to assess the sensitivity of the procedure in detection of artificial chimeras.

To further investigate properties of this CHECK_CHIMERA test, we submitted five SSU rRNA chimeras which had been reported previously in the literature for testing. Table 1 shows that of the three full-length chimeras identified by Barns et al. (1), two were detected by our use of CHECK_CHIMERA and the third was not detected. The two chimeras identified by Kopczynski et al. (8) were both detected. We also tested all long sequences

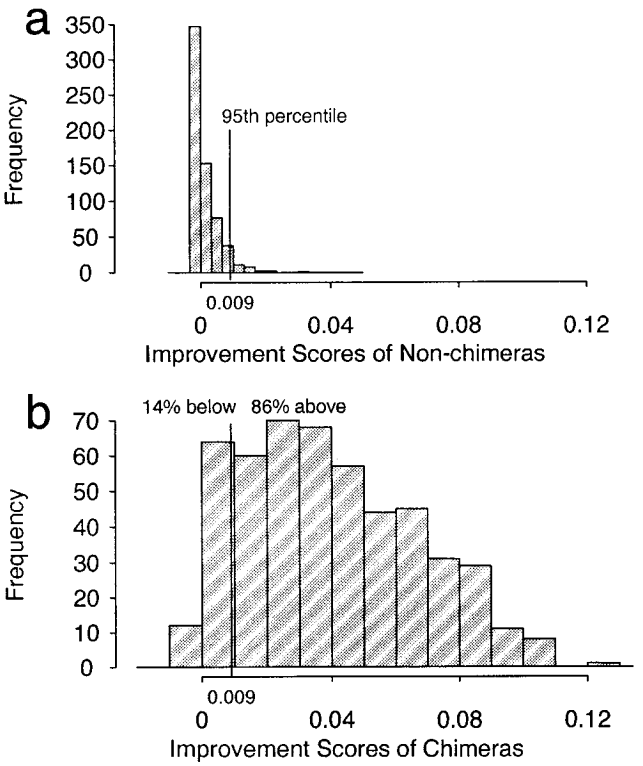

FIG. 2. Distribution of aligned similarity IS for 639 nonchimeric sequences (a) and for 500 artificially created chimeras (b). The 95th percentile of IS for authentic sequences is indicated by the vertical line in both panels. Frequency is the number of sequences.

obtained thus far from the Octopus Spring cyanobacterial mat community by using CHECK_CHIMERA (Table 2). Two sequences obtained as cDNA clones were indicated as possible chimeras.

DISCUSSION

Our intent is to make those who retrieve SSU rRNA sequences from natural habitats aware of the limitations of the methods available for detecting chimeric artifacts. Such tools

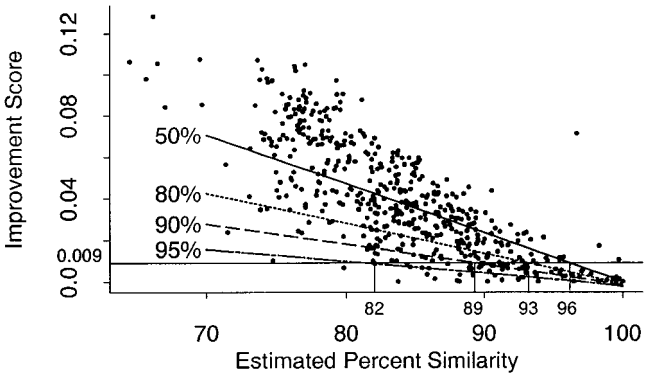

FIG. 3. The relationship between IS and estimated similarity of parental sequences (similarity of S_1 to S_2 [Fig. 1a]) of artificial chimeric SSU rRNAs, as determined by the aligned similarity method. The weighted regression line is labeled 50% (for a test of zero slope; $P = 0.0000$); lower prediction limits for a single new sequence are indicated by the dashed lines. Estimated similarities are indicated where the prediction limit lines cross the detection limit of 0.009.

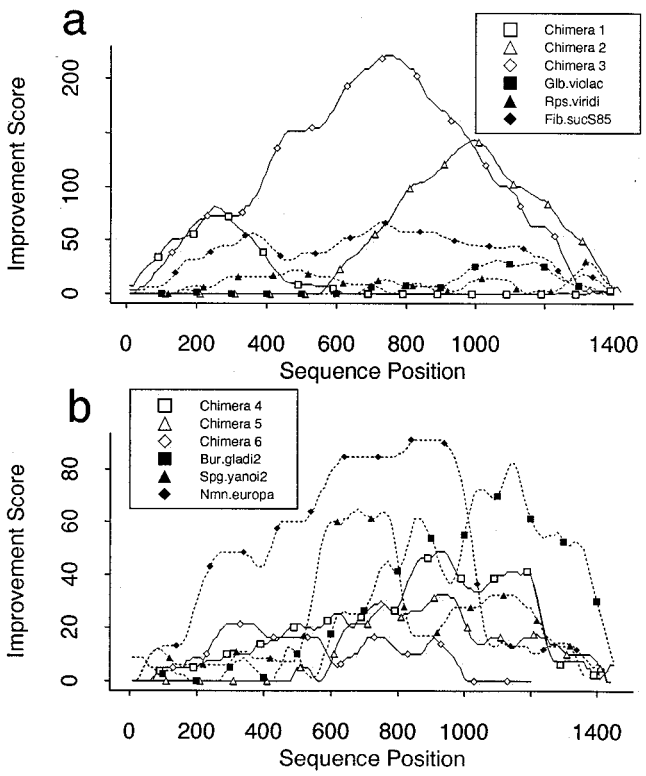

FIG. 4. CHECK_CHIMERA output from typical authentic sequences and artificial chimeras (a) and low-scoring chimeras and high-scoring authentic sequences (b). Chimeras were formed from *Runella slithyformis* and *Azospirillum lipoferum* combined at base 258 (chimera 1), *Bacillus laterosporus* and *Eubacterium biforme* combined at base 1084 (chimera 2), *Eubacterium dolichum* and *Bacillus mycoides* combined at base 802 (chimera 3), *Haemophilus somnus* and *Haemophilus influenzae* combined at base 1216 (chimera 4), *Actinobacillus capsulatus* P243 and *Haemophilus parasuis* combined at base 760 (chimera 5), and *Mycobacterium senegalense* and *Mycobacterium chubuense* combined at base 333 (chimera 6). The authentic sequences used are *Gloeobacter violaceus* (Glb. violac), *Rhodopseudomonas viridis* (Rps. viridi), *Fibrobacter succinogenes* S85 (Fib. sucS85), *Burkholderia gladioli* pathovar gladioli (Bur. gladi2), *Sphingomonas yanoikuyae* (Spg. yanoi2), and *Nitrosomonas europaea* (Nmn. europa).

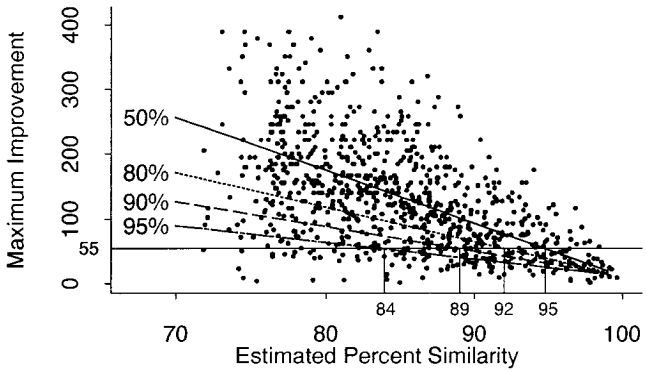

FIG. 5. The relationship between MIS and estimated similarity of parental sequences (similarity of S_1 and S_2 [Fig. 1b]) of artificial chimeric SSU rRNAs as determined by the CHECK_CHIMERA method. Weighted regression was used as for Fig. 3 (for a test of zero slope; $P = 0.0000$). Estimated similarities are indicated where the prediction limit lines cross the detection limit of 55.

must be used with an appreciation of their capabilities, since classification errors are unfortunate but unavoidable. We have tried to quantify the probabilities of making classification errors when these tests are used.

Both methods we evaluated are able to detect artificially chimeric SSU rRNA sequences when the parental sequences are quite different. However, as the similarity of the parental molecules increases, the reliability of the methods becomes limited and the two approaches perform equally poorly (Fig. 3 and 5). Surprisingly, use of the extra information available in the alignment did not improve the sensitivity of the method in detecting chimeras. Our discussion focuses on common aspects of the two approaches, since in our opinion limitations are largely due to the nearest-neighbor approach and properties of the sequences in the database and not to the specific program.

Confidence in detection of chimeras by both methods decreases from 95 to 50% as the estimated similarity between parental sequences increases from 82 to 96% (with the probability of labeling an authentic sequence chimeric set at 5%). This limitation is based in part on the fact that authentic SSU rRNA sequences from cultivated species exhibit separate sequence domains which more closely resemble different database sequences. This phenomenon skews the IS (Fig. 2a) and MIS distributions among background sequences, raising the 95th percentiles which we established as detection limits. It is interesting to speculate that such sequences might be natural chimeras, as has been previously proposed (14). The limitation on detection is also based on the fact that some chimeric sequences exhibit relatively low ISs and MISs. These are mainly chimeras formed from parental sequences which are closely related. The mechanism of chimera formation is thought to be based on cross-hybridization of different templates during PCR (12). If this is true, it is possible that chimeras form more frequently from parent molecules with more similar sequences, which would have an increased propensity for cross-hybridization. In other words, the most difficult chimeras to detect may be those which are most readily formed.

The testing of known and suspected chimeras reveals several interesting points. Barns et al. (1) used CHECK_CHIMERA as one means of testing sequence pJP 75 and as a result of that test and other evidence (secondary structure and separate phylogenetic trees for the 5′ and 3′ ends separated by the point at which the CHECK_CHIMERA MIS occurs) reported that the sequence is chimeric. The use of the naive test based on CHECK_CHIMERA MIS output does not lead to the same conclusion. Since estimated similarity is only 82% for pJP 75, this should not have been a difficult chimera to detect. In the set of artificial chimeras we created, 340 had estimated similarity below 82% and 6% of those sequences were not detected. This illustrates the difficulty of using nearest-neighbor methods to confidently detect all chimeras.

Chimeras reported by Choi et al. (2) were not used as another test of the CHECK_CHIMERA method, since the

TABLE 1. CHECK_CHIMERA results for reported SSU rRNA chimeras of 1,150 bases or more

Sequence	Estimated similarity (%)[a]	MIS	Chimera indicated[b]
pJP 101[c]	82	69	Yes
pJP 102[c]	87	108	Yes
pJP 75[c]	82	43	No
OP-I 6[d]	77	140	Yes
OP-I 8[d]	76	194	Yes

[a] Between S_1 and S_2 (see the text).
[b] The presence of a chimera is indicated for an MIS of at least 55.
[c] Data from reference 1.
[d] Data from reference 8.

TABLE 2. CHECK_CHIMERA results for long SSU rRNA sequences of cDNA and PCR clones obtained from the Octopus Spring cyanobacterial mat community

Clone and phylogenetic type	Sequence[a]	Estimated similarity (%)[b]	MIS	Chimera indicated[c]	No. of bases compared
cDNA clones					
Cyanobacteria	OS type B	89	19	No[d]	1,110
	OS type I	90	71	Yes	1,280
	OS type J	91	33	No[d]	1,060
	OS type P[e]	91	32	No	1,250
Green nonsulfur	OS type C	75	15	No[d]	1,070
	OS-V-L-20	79	31	No[d]	850
Planctomyces	OS type L	83	53	No[d]	1,060
Spirochete	OS type K	83	38	No	1,330
Proteobacteria	OS-V-L-28	90	24	No	1,320
	OS type O	79	29	No[d]	680
Unstable	OS-VI-L-4	79	64	Yes[d]	810
PCR clones					
Uncertain	OP-I 2	80	18	No	1,390
	OP-I 4	83	38	No[d]	1,000
	OP-I 7	80	37	No[d]	1,020

[a] Described in references 8 and 9 and 15 to 17.
[b] Between S_1 and S_2 (see the text).
[c] The presence of a chimera is indicated for an MIS of at least 55.
[d] Approximate because of short sequence length.
[e] Identical to cultivated *Synechococcus* sp. strain B10 (15).

length of these clones is about 500 bases. The CHECK_CHIMERA program does yield output for sequences of any length, but the 95th percentile of MIS that we have used as a cutoff may not be correct for testing these partial sequences. One would have to reexamine the background, using the domain of these fragments to find a new 95th percentile of the authentic MIS.

The two Octopus Spring mat sequences retrieved by the cDNA method and detected as possible chimeras by our use of CHECK_CHIMERA output (Table 2) may be examples of the 5% of high-scoring authentic sequences (possibly natural chimeras). Alternatively, the cDNA method, like the PCR method, may lead to chimera formation. For PCR chimeras (OP-I 6 and OP-I 8 in Table 1), Kopczynski et al. reported that three of the four possible parental sequences matched database sequences (8). For cDNA sequence OS type I, we found no matching sequences for domains identified by CHECK_CHIMERA as possibly originating from different parent sequences. Furthermore, there were no secondary structure abnormalities. At this point, the only evidence we have that this sequence is chimeric is the high MIS from CHECK_CHIMERA. In contrast, the OS-VI-L-4 sequence does exhibit a match to the cyanobacterial OS type J sequence between *E. coli* positions 238 and 366 (including the site of chimera formation suggested by CHECK_CHIMERA at *E. coli* position 332). This observation must be interpreted with caution, since these 129 positions do not include a hypervariable region and are relatively well conserved. As previously reported (18), the phylogenetic character of the molecule at positions toward the 3′ end becomes very different from that of a cyanobacterium (e.g., low percent similarity and atypical secondary structure features). This sequence was also unstable in phylogenetic tree analysis (17). Furthermore, a base pair mismatch between *E. coli* positions 29 and 554 exists. Thus, several lines of evidence support the interpretation that the OS-VI-L-4 sequence is a chimeric artifact.

We advise caution in labeling any sequence as chimeric or nonchimeric on the basis of nearest-neighbor methods, especially when results are near the 95th-percentile cutoff levels. The probability that a given test result (chimeric or not) is correct can be computed by using Bayes' rule (3). This computation is commonly used in random drug testing, which is similar in that use of drugs in the target population is a rare event, as we hope is true for chimera formation in the population of SSU rRNA sequences. First, consider the case in which only 4% of the population are chimeras (true positives) and 96% are nonchimeras (lowest reported chimera frequency). Positive test results are the combination of false positives (5% of authentic sequences are allowed to be misclassified) and true positives. The probability of correctly judging a query sequence to be chimeric is the ratio of true positives over all positives: $(0.04)/(0.05 \times 0.96 + 0.04) = 0.04/0.088 = 0.45$. This is in the optimistic case which assumes that all chimeras are detected (i.e., low percent similarity of parents). In fact, this test will fail to detect some of the chimeras, which lowers this probability further. If the proportion of true chimeras in an SSU rRNA population is 20% (highest reported chimera frequency), then the probability of a true positive given a positive test result is approximately 0.80. The probability that a sequence indicated as nonchimeric actually is nonchimeric (true negative) can be computed in the same manner. For distantly related parent sequences (estimated similarity of $<82\%$), the probability of a true negative, given that a chimera is not indicated, is high (0.998 if 4% of the population is chimeric and 0.987 if 20% of the population is chimeric).

A final caveat concerns assumptions made in applying the probabilities given above to real sequences. The confidence levels and probabilities are accurate if the current database is representative of the entire population of SSU rRNA sequences in nature and if the MIS residuals are distributed close to normally. The importance of the reference database is underscored by comparing the results reported herein with preliminary results from analysis of the RDP release 3.0 database (October 1993). When 1,036 sequences from release 3.0 were used as an authentic background and the 1,368 sequences of 1,200 bases or longer were used as the reference database, the 95th percentile of MIS was 62 and 26% of the artificial chimeras were misclassified as authentic. With the current database,

the 95th percentile is 55 and the misclassification rate is only 20%. Thus, detection has improved. Unfortunately, additions to the database are not done at random. Close relatives of sequences in the database are often added, which causes a general decrease in MIS for authentic sequences. For instance, in release 3.0, *E. coli* was most similar (96%) to *Serratia marcescens* and yielded an MIS of 40. In release 4.0, several variants of the sequence have been added, each of which is 99% similar to the *E. coli* sequence, and the MIS has dropped to 0. In general, the MISs of authentic sequences of the two RDP releases are moderately correlated ($r = 0.75$). However, the correlation between the MISs for the artificial chimeras obtained by using the release 3.0 reference database and their MISs obtained by using release 4.0 is only 0.09. This weak correlation points to the sensitivity of nearest-neighbor methods to additions to the database. The probabilities stated herein apply to only the release 4.0 database, but detection probabilities should improve as the database expands; however, the nonrandom nature of the sampling process which adds sequences to the database has unknown effects on these results. The artificial chimeras created for these tests cannot be considered representative of true chimeras, since little is known about true chimeras from PCR or cDNA methods (e.g., distribution of sites of chimera formation and frequency of chimera formation relative to parent sequence similarity).

In summary, detection of chimeric SSU rRNA sequences is a complex problem. We now have a large database of authentic sequences to use in making comparisons, but comparisons based on nearest neighbors do not provide a very powerful test. We recommend that researchers use CHECK_CHIMERA but not rely on it exclusively. Caution is also required when using methods based on phylogenetic trees, since the uncertainty of trees has not been quantified. It is important to use all available tests until better discrimination methods become available. At present, certainty in labeling a sequence chimeric is possible only when one can find exactly matching parent sequences. Other methods of identifying chimeras involve uncertainty because some authentic sequences have chimera-like characteristics. Stating that a sequence of uncultivated origin is nonchimeric is never absolutely certain, since all methods will break down as parental similarity increases and as fragment length decreases. A corollary of this statement is that we cannot obtain a reliable estimate of the true proportion of chimeras in the population of SSU rRNA sequences derived through PCR or cDNA methods. The 4 to 20% figures observed to date may be overly optimistic. Clearly, improved methods for detection of chimeric sequences are needed.

ACKNOWLEDGMENTS

We thank Niels Larsen of RDP for useful discussions.

We thank Frank Roberto of the Idaho National Energy Laboratory for arranging support by the U.S. Department of Energy Office of Health and Environmental Research Subsurface Science Program, under DOE Idaho Operations Office contract DE-AC07-76ID01570. J.F.R.-C. was also supported by Montanans On a New Track for Science (NSF EPSCOR program) grant 190138. Other funds for this research were provided by NSF (BSR-9209677) and NASA (NWAG-2764).

REFERENCES

1. **Barns, S. M., R. E. Fundyga, M. W. Jeffries, and N. R. Pace.** 1994. Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. Proc. Natl. Acad. Sci. USA **91:**1609–1613.
2. **Choi, B. K., B. J. Paster, F. E. Dewhirst, and U. B. Gobel.** 1994. Diversity of cultivable and uncultivable oral spirochetes from a patient with severe destructive periodonitis. Infect. Immun. **62:**1889–1895.
3. **DeGroot, M. H.** 1989. Probability and statistics. Addison-Wesley, Reading, Mass.
4. **DeLong, E. F.** 1992. Archaea in coastal marine environments. Proc. Natl. Acad. Sci. USA **89:**5685–5689.
5. **Fuhrmann, J. A., K. McCallum, and A. A. Davis.** 1993. Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific oceans. Appl. Environ. Microbiol. **59:**1294–1302.
6. **Giovannoni, S. J., T. B. Britschgi, C. L. Moyer, and K. G. Field.** 1990. Genetic diversity in Sargasso Sea bacterioplankton. Nature (London) **345:** 60–63.
7. **Kane, M. D., L. K. Poulsen, and D. A. Stahl.** 1993. Monitoring the enrichment and isolation of sulfate-reducing bacteria by using oligonucleotide hybridization probes designed from environmentally derived 16S rRNA sequences. Appl. Environ. Microbiol. **59:**682–686.
8. **Kopczynski, E. D., M. M. Bateson, and D. M. Ward.** 1994. Recognition of chimeric small-subunit ribosomal DNAs composed of genes from uncultivated microorganisms. Appl. Environ. Microbiol. **60:**746–748.
9. **Kopczynski, E. D., M. M. Bateson, and D. M. Ward.** Unpublished data.
10. **Larsen, N., G. J. Olsen, B. L. Maidak, M. J. McCaughey, R. Overbeek, T. J. Macke, T. L. Marsh, and C. R. Woese.** 1993. The Ribosomal Database Project. Nucleic Acids Res. **21**(Suppl.)**:**3021–3023.
11. **Liesack, W., and E. Stackebrandt.** 1992. Occurrence of novel groups of the domain *Bacteria* as revealed by analysis of genetic material isolated from an Australian terrestrial environment. J. Bacteriol. **174:**5072–5078.
12. **Liesack, W., H. Weyland, and E. Stackebrandt.** 1991. Potential risks of gene amplification by PCR as determined by 16S rDNA analysis of a mixed-culture of strict barophilic bacteria. Microb. Ecol. **21:**191–198.
13. **Schmidt, T. M., E. F. DeLong, and N. R. Pace.** 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J. Bacteriol. **173:**4371–4378.
14. **Sneath, P. H. A.** 1993. Evidence from *Aeromonas* for genetic crossing-over in ribosomal sequences. Int. J. Syst. Bacteriol. **43:**626–629.
15. **Ward, D. M., M. J. Ferris, S. C. Nold, M. M. Bateson, E. D. Kopczynski, and A. L. Ruff-Roberts.** 1994. Species diversity in hot spring microbial mats as revealed by both molecular and enrichment culture approaches—relationship between biodiversity and community structure. NATO ASI Ser. G Ecol. Sci. **35:**33–44.
16. **Ward, D. M., R. Weller, and M. M. Bateson.** 1990. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature (London) **344:**63–65.
17. **Weller, R., J. W. Weller, and D. M. Ward.** 1991. 16S rRNA sequences of uncultivated hot spring cyanobacterial mat inhabitants retrieved as randomly primed cDNA. Appl. Environ. Microbiol. **57:**1146–1151.
18. **Wilmotte, A.** 1994. Molecular evolution and taxonomy of the cyanobacteria, p. 1–25. *In* D. A. Bryant (ed.), The molecular biology of cyanobacteria. Kluwer Academic Publishers, Dordrecht, The Netherlands.