APPLIED AND ENVIRONMENTAL MICROBIOLOGY, June 1997, p. 2338–2346 Vol. 63, No. 6
0099-2240/97/$04.00+0

A New Computational Method for Detection of Chimeric 16S rRNA Artifacts Generated by PCR Amplification from Mixed Bacterial Populations

GEORGE A. KOMATSOULIS[1]† AND MICHAEL S. WATERMAN[1,2]*

Departments of Mathematics[1] and Biological Sciences,[2] University of Southern California, Los Angeles, California 90089-1113

Received 16 January 1997/Accepted 26 March 1997

A new computational method (chimeric alignment) has been developed to detect chimeric 16S rRNA artifacts generated during PCR amplification from mixed bacterial populations. In contrast to other nearest-neighbor methods (e.g., CHECK_CHIMERA) that define sequence similarity by *k*-tuple matching, the chimeric alignment method uses the score from dynamic programming alignments. Further, the chimeric alignments are displayed to the user to assist in sequence classification. The distribution of improvement scores for 500 authentic, nonchimeric sequences and 300 artificial chimeras (constructed from authentic sequences) was used to study the sensitivity and accuracy of both chimeric alignment and CHECK_CHIMERA. At a constant rate of authentic sequence misclassification (5%), chimeric alignment incorrectly classified 13% of the artificial chimeras versus 14% for CHECK_CHIMERA. Interestingly, only 1% of nonchimeras and 10% of chimeras were misclassified by both programs, suggesting that optimum performance is obtained by using the two methods to assign sequences to three classes: high-probability nonchimeras, high-probability chimeras, and sequences that need further study by other means. This study suggests that *k*-tuple-based matching methods are more sensitive than alignment-based methods when there is significant parental sequence similarity, while the opposite becomes true as the sequences become more distantly related. The software and a World Wide Web-based server are available at http://www-hto.usc.edu/software/mglobalCHI.

The use of 16S rRNA in the classification of bacterial species has been well established, and its effect on biology has been profound. It was 16S rRNA data that provided convincing evidence that chloroplasts and mitochondria most likely arose from free-living bacteria and that prokaryotic organisms represented not one line of evolutionary descent but two, *Bacteria* and *Archaea* (previously eubacteria and archaebacteria), that diverged from each other and from the *Eucarya* at approximately the same time (5, 26). The Ribosomal Database Project (RDP) (13) at the University of Illinois at Urbana-Champaign maintains an extensive, publicly accessible, database of 16S rRNA sequences with the long-term goal of developing complete phylogenies of all bacterial, archaebacterial, mitochondrial, and chloroplast species.

Since most bacterial species have not been cultured or are uncultivable without significant effort, methods to obtain rDNA sequences directly from bacterial biomass have been developed (8, 22, 25; see also reference 21 for a digest of comments on the inadequacy of using culture techniques to describe the members of a natural community). Some of these use PCR to directly amplify rDNA sequences from environmental samples with primers that can be targeted to regions of the 16S sequence with different degrees of cross-species conservation. Since their introduction, these methods have permitted the study of a wide range of bacterial habitats, including hot springs in Yellowstone National Park (3), oceans (6, 8), the human oral cavity (4), and the nuclei of ciliates (2) to name just a few. The power, ease, and flexibility provided by these PCR methods would seem to argue that environmentally derived sequences will come to dominate the 16S rRNA databases (and hence bacterial phylogeny) in the relatively near future. These PCR-based methods have a significant drawback, however; in some fraction of cases (estimated as 4.1 to 20% in reference 18) a recovered clone is a chimera-containing sequence derived from two microorganisms (for early observations of this problem see references 11 and 12). Obviously, inclusion of such sequences in phylogenies could cause significant errors, and the number of such occurrences must be kept to a minimum.

Despite the fact that there are many potential ways to detect chimeric sequences, including covariation analysis and analysis of predicted secondary structure (i.e., searching for mismatches in conserved helices), most detection is done by nearest-neighbor methods. In nearest-neighbor analysis, a newly recovered sequence (henceforth the query sequence or query) is split into two subsequences that are then compared with a database of similar sequences. If the sequence can be split in such a way that the phylogenetic affiliation of the parts is inconsistent with the affiliation of the sequence as a whole, a chimera is suspected. This condition for a chimera is rather vague from the standpoint of a practical method, so most nearest-neighbor chimera detection programs restate this condition as an improvement score or IS. An IS is usually defined as the sum of the similarities between the two partial sequences and their nearest neighbors minus the score for the complete sequence compared with its nearest neighbor. Since there are many ways to measure sequence similarity, there are also a wide range of possible nearest-neighbor methods, probably with different levels of sensitivity (i.e., how well they detect chimeras) and discrimination (from a practical standpoint, how many nonchimeric sequences get marked as chimeras).

There are two currently available nearest-neighbor methods,

* Corresponding author. Mailing address: Department of Biological Sciences, University of Southern California, DRB155, 1042 W. 36th Place, Los Angeles, CA 90089-1113. Phone: (213) 740-2408. Fax: (213) 740-2437. E-mail: gkoma@hto.usc.edu and msw@hto.usc.edu.

† Present address: Human Genome Sciences, Rockville, MD 20850.

the CHECK_CHIMERA method of Larsen et al. (13) and the aligned similarity method of Robison-Cox et al. (18). The former method defines similarity by the number of common oligonucleotides of length k (k-tuples) shared by a sequence. The sequence is broken at 10-base intervals, and the maximum value of the IS over all possible breakpoints is determined. The latter method computes similarity by counting the number of aligned, matched bases in two disjoint sequence domains by using the RDP universal multiple sequence alignment. Both of these methods have potential flaws. CHECK_CHIMERA does not use any alignment information at all, and the aligned similarity method suffers from problems associated with using a multiple-sequence alignment. These include the inability to properly penalize indels (insertions and deletions); plus, there are technical reasons why pairwise comparisons derived from a multiple sequence alignment are likely to be suboptimal (23). Furthermore, any fundamentally statistical method will make errors with certain types of sequences. For this reason, we developed a new method to complement these existing methods.

Our new method, called chimeric alignment, scores sequence comparisons by dynamic programming alignment (14, 19, 23), which is the method used by the Genetics Computer Group programs BestFit and Gap (7). The method calculates two improvement scores, IS_{C2S1} and IS_{C2C1}, based on the three alignments shown in Fig. 1. The first alignment (Fig. 1a) is the best alignment between the query and its closest neighbor in a database of similar sequences, which we will call the best single sequence alignment, and we will label its score $S1$. Next, two chimeric alignments are determined. We define a chimeric alignment as one in which the query sequence is broken into two parts, and then each part is aligned to a 5′ or 3′ fragment of a sequence in the database. Figure 1b shows a chimeric alignment where one database entry provides both of the nearest neighbors, and Fig. 1c shows a chimeric alignment where each query fragment has a nearest neighbor from a different sequence. For convenience we call the alignment in Fig. 1b the best chimeric alignment with a single sequence and abbreviate the score $C1$, while the alignment in Fig. 1c is simply called the best chimeric alignment, and its score is $C2$. To obtain the improvement scores mentioned above we compute raw difference scores: $C2S1 = C2 - S1$ and $C2C1 = C2 - C1$. The scores are adjusted to remove certain artifactual effects (see the program documentation at http://www-hto.usc.edu/software/mglobalCHI), and then they are labeled IS_{C2S1} and IS_{C2C1}. These scores and the alignments from which they are derived can then be used to help determine if a sequence is chimeric.

We have developed a program (mglobalCHI) that performs the chimeric alignments described above. To determine its effectiveness at detecting chimeric sequences, we have tested both our program and the CHECK_CHIMERA program with 500 nonchimeric sequences and 300 artificial chimeras. Sequences were labeled as chimeric or nonchimeric based on IS values. Although this ignored the contribution of the user's expertise, it provided an objective assessment of sensitivity and discrimination. The chimeric alignment method was marginally better at detecting chimeric sequences, misclassifying 13% of the chimeras versus 14% for the CHECK_CHIMERA method (both programs were set to a 5%-false-positive rate). The best results were obtained when both programs were used in concert. If both programs had to agree to classify a sequence, then only 1% of nonchimeric sequences were classified as probable chimeras, 8% were classified as possible chimeras, and 91% were classified as probable nonchimeras. For the chimeras 10% were classified as probable nonchimeras, 7% as possible chimeras, and 83% as probable chimeras.

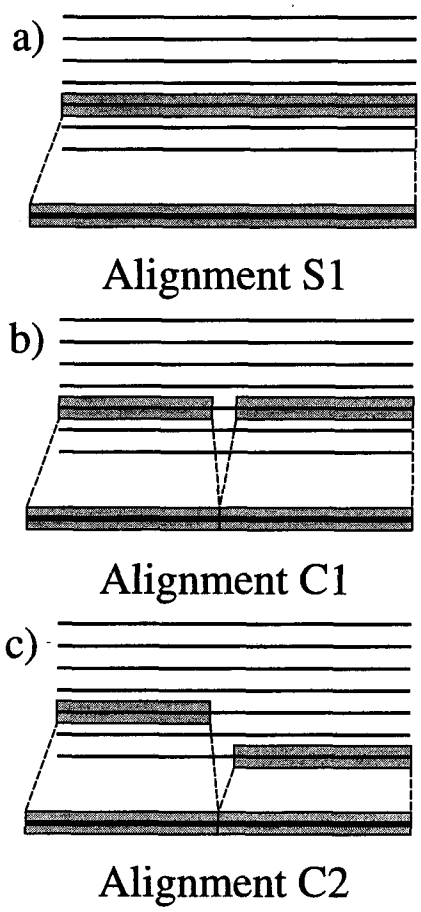

FIG. 1. Schematic representations of alignment types. Bold lines represent query sequences, fine lines represent database sequences, and shaded boxes represent aligned regions. (a) global alignment ($S1$); (b) chimeric alignment with one sequence providing both partial molecules ($C1$); (c) chimeric alignment with two sequences ($C2$).

MATERIALS AND METHODS

Program. The program works by determining the best global sequence alignment between the first i nucleotides of the query and its closest relative in a database and that between the remaining segment of the query sequence and its nearest neighbor. This is determined by using every nucleotide more than 100 bases from an end of the query as the point to divide the sequences. The scores for the two partial alignments are then summed to yield the best chimeric alignment. The best chimeric alignment with a single sequence is generated in a similar manner, except both nearest neighbors must come from the same sequence. The best single sequence alignment is determined by a straightforward extension of standard dynamic programming methods (14, 19, 23). Note that this method implicitly imposes a 200-base minimum length on query sequences. Those interested in the details of the algorithm are encouraged to review the online documentation and reference 9.

The score S for a sequence comparison is derived from the alignments by using the following relationship: $S = (\text{matches}) \times \nu - (\text{mismatches}) \times \mu - \Sigma\,(\text{indels}) \times (\text{gap penalty})$, where ν is the match score, and μ is the mismatch penalty. Gap penalties are assessed depending on length. There is a penalty α for starting a gap and a lower penalty β for each additional base in the gap. In addition, 5′ leading and 3′ trailing indels are penalized less than internal gaps, because the former are more likely related to sequencing completeness than to sequence relatedness per se. After the raw improvement scores $C2S1 = C2 - S1$ and $C2C1 = C2 - C1$ are computed, they are adjusted to remove certain artifactual effects (see the online documentation at http://www-hto.usc.edu/software/mglobalCHI) and are then labeled IS_{C2S1} and IS_{C2C1}. Readers interested in the details of the adjustments are encouraged to review the online documentation for the program and

a

```
hto.usc.edu(1): mglobalCHI-0.95 non151.seq noncheck15-16.std chimera.mat
                1 12 1 0 6 0 -m -u 151.un

Summary of alignment scores:

Query:
                          Sequence: non Bdr.IRBG2 260 bp
                            Length: 260
Database:
                          Filename: noncheck15-16.std
                      # of Sequences: 35

Best non-chimeric alignment:
                          Sequence: Bdr.ORS322 260 bp
                             Score: 780

Best chimeric alignment with a single sequence
                      Sequence: Bdr.ORS322 260 bp
                         Score: 780
               Query breakpoint: 100
                      From[1]: 1 to 100
                      From[2]: 101 to 260

Best chimeric alignment:
                    Query breakpoint: 100
                               Score: 780

                           Sequence1: Bdr.ORS322 260 bp
                                From: 1 to 100

                           Sequence2: Bdr.IRB231 264 bp
                                From: 105 to 264

Improvement scores:
                             C2S1: 0
                             C2C1: 0

                          IS(C2S1): 0
                          IS(C2C1): 0

Best single sequence alignment

   1  uuaacacaugcaagucgagcgggcauagcaauaugucagcggcagacgggugaguaacgc
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   1  uuaacacaugcaagucgagcgggcauagcaauaugucagcggcagacgggugaguaacgc

  61  gugggaacguaccuuuugguucggaacaacacagggaaacuugugcuaauaccggauaag
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  61  gugggaacguaccuuuugguucggaacaacacagggaaacuugugcuaauaccggauaag

 121  cccuuacggggaaagauuuaucgccgaaagaucggcccgcgucugauuagcuaguuggug
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 121  cccuuacggggaaagauuuaucgccgaaagaucggcccgcgucugauuagcuaguuggug

 181  ggguaauggcccaccaaggcgacgaucaguagcuggucugagaggaugaucagccacauu
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 181  ggguaauggcccaccaaggcgacgaucaguagcuggucugagaggaugaucagccacauu

 241  gggacugagacacggcccaa
      ||||||||||||||||||||
 241  gggacugagacacggcccaa

Best chimeric alignment with a single sequence:

   1  uuaacacaugcaagucgagcgggcauagcaauaugucagcggcagacgggugaguaacgc
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   1  uuaacacaugcaagucgagcgggcauagcaauaugucagcggcagacgggugaguaacgc

  61  gugggaacguaccuuuugguucggaacaacacagggaaac*uugugcuaauaccggauaa
      ||||||||||||||||||||||||||||||||||||||||||*|||||||||||||||||
  61  gugggaacguaccuuuugguucggaacaacacagggaaac*uugugcuaauaccggauaa

 120  gcccuuacggggaaagauuuaucgccgaaagaucggcccgcgucugauuagcuaguuggu
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 120  gcccuuacggggaaagauuuaucgccgaaagaucggcccgcgucugauuagcuaguuggu

 180  gggguaauggcccaccaaggcgacgaucaguagcuggucugagaggaugaucagccacau
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 180  gggguaauggcccaccaaggcgacgaucaguagcuggucugagaggaugaucagccacau

 240  ugggacugagacacggcccaa
      ||||||||||||||||||||
 240  ugggacugagacacggcccaa

Best chimeric alignment:

   1  uuaacacaugcaagucgagcgggcauagcaauaugucagcggcagacgggugaguaacgc
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
   1  uuaacacaugcaagucgagcgggcauagcaauaugucagcggcagacgggugaguaacgc

  61  gugggaacguaccuuuugguucggaacaacacagggaaac*uugugcuaauaccggauaa
      ||||||||||||||||||||||||||||||||||||||||||*|||||||||||||||||
  61  gugggaacguaccuuuugguucggaacaacacagggaaac*uugugcuaauaccggauaa

 120  gcccuuacggggaaagauuuaucgccgaaagaucggcccgcgucugauuagcuaguuggu
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 124  gcccuuacggggaaagauuuaucgccgaaagaucggcccgcgucugauuagcuaguuggu

 180  gggguaauggcccaccaaggcgacgaucaguagcuggucugagaggaugaucagccacau
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 184  gggguaauggcccaccaaggcgacgaucaguagcuggucugagaggaugaucagccacau

 240  ugggacugagacacggcccaa
      ||||||||||||||||||||
 244  ugggacugagacacggcccaa

hto.usc.edu(2):
```

FIG. 2. Sample output from the mglobalCHI program. (a) Authentic sequence Bdr.IRBG2; (b) artificial chimera chi105. Note that the best single sequence alignment and the best chimeric alignment with a single sequence in panel b are truncated for improved visibility. Both of these alignments have long 3′ trailing indels. The name of the query sequences and their lengths are listed at the top, along with the names of the files containing the 16S database and the numbers of sequences in those files.

to run the program on test cases to observe the relationships between the unadjusted and adjusted scores.

Sequences. The database of sequences used in these comparisons was obtained from the RDP (13). Sequences were obtained from the aligned database (RDP version 5.0) with the RDP World Wide Web server. The database contained all prokaryotic sequences except those denoted env., which are derived from environmental samples. Most of the sequences denoted sym. (endosymbionts) were also deleted. The sequences were edited to remove extraneous characters and converted into our standard file format. The total number of sequences was 2,520. Nonchimeric test sequences were obtained by randomly selecting sequences from the database described above. In some cases database entries were missing various amounts of internal sequences. The RDP designates areas where sequence is missing by using a period rather than a dash in the universal alignment (periods are generally found where data are unavailable or where there were more than six Ns in a row in the original sequence). In such cases the longest continuous piece of sequence was used in the test. In extreme cases, an entry was simply bypassed.

Artificial chimeras were obtained from authentic sequences by randomly selecting a sequence from the database and then randomly selecting a sequence from the 100 nearest sequences on either side of the first in the phylogenetically ordered list. This restriction was designed to give a wide range of similarity scores for the parental sequences of chimeras. The breakpoint was chosen at random in the RDP-aligned sequence, and the two sequences were joined together to form the chimera. If the chimera had a breakpoint less than 200 bases from either end, a new breakpoint was chosen. Occasionally it was necessary to reverse the order of the two partial sequences in the chimera in order to create a continuous sequence from database entries that were missing internal data (see above).

The nonchimeric sequences varied in length from 225 to 1,740 bases; chimeras ranged from 438 to 1,609 bases. Estimated similarity of the parents of the partial sequences that were assembled into a chimera (i.e., the sequences used to construct the chimera) was determined by aligning the two sequences by using the three gap function method described above and the same parameters as those for the chimeric-alignment program. Percent similarity was then measured as the number of identically matched bases divided by the total number of positions aligned (excluding leading and trailing indels). If data were missing from a parental sequence, it was broken into continuous pieces and the fragments were aligned separately. Overall percent similarity was then calculated as the length-weighted mean of the similarity of the fragments. Estimated percent similarity of the parental sequences of the chimeras ranged from <67 to 99%.

Program evaluation. To explore the effectiveness of the program, 500 nonchimeric sequences were used to obtain an estimate of the distribution of ISs for authentic sequences with this database and parameter set. For each sequence, IS_{C2C1} and IS_{C2S1} were determined (the query sequence was removed from the database in all cases) and the scores were ordered. A cutoff value for a chimeric sequence was determined by finding a score that would give a 5% false-positive rate (this is the same method used by Robison-Cox et al. as described in reference 18). Three hundred artificial chimeras were then tested by using the above database (minus both source sequences) to determine the false-negative rate.

The chimeric alignment program used the following parameters. The match and mismatch values are from matrix chimera.mat (available at the University of Southern California (USC) computational biology World Wide Web server, http://www-hto.usc.edu/). Generally, matches were scored +3 and mismatches were scored −6. In the case of mismatches with N (any nucleotide) the score was penalized −4. R-A, R-G, R-R, Y-T, Y-U, Y-C, and Y-Y mismatches were penalized −2. The gap penalties used the following parameters: $\alpha_{5'} = \alpha_{3'} = -1$, $\alpha_{\text{internal}} = -12$, $\beta_{5'} = \beta_{3'} = 0$, and $\beta_{\text{internal}} = -6$.

For comparison purposes, the 800 sequences described above were also tested with the CHECK_CHIMERA method at the University of Illinois, Urbana-Champaign. The studies were conducted between 19 March 1996 and 6 May 1996. The database used excluded the same env. and sym. sequences as those excluded by the one used for the chimeric alignment program. In addition the RDP recommends that sequences shorter than 1,300 bases be excluded. A previous study (18) of CHECK_CHIMERA effectiveness removed all sequences shorter than 1,200 bases and so this was also done for consistency. The maximum IS was calculated with the graph axes provided by the RDP, and similar background and false-negative values were calculated.

b

```
Summary of alignment scores:

Query:
                        Sequence: chi Hb.marism2 Hc.spBr3 1 to 226 1 212 brkpt:530
                          Length: 438
Database:
                        Filename: /tmp/13248.db
                  # of Sequences: 101

Best non-chimeric alignment:
                        Sequence: > Hb.marism1 1472 bp
                           Score: 1036

Best chimeric alignment with a single sequence
                      Sequence: > Hb.marism1 1472 bp
                         Score: 1039
            Query breakpoint: 297
                     From[1]: 1 to 297
                     From[2]: 299 to 1472

Best chimeric alignment:
                   Query breakpoint: 226
                         Score: 1235

                      Sequence1: > Hb.marism1 1472 bp
                           From: 1 to 226

                      Sequence2: > Hc.Blp 212 bp
                           From: 1 to 212

Improvement scores:
                       C2S1: 199
                       C2C1: 196

                    IS(C2S1): 198
                    IS(C2C1): 195


Best single sequence alignment

   1 auuccgguugauccugccggaggccauugcuaucggagucccgauuuagccaugcuaguug
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |
   1 auuccgguugauccugccggaggccauugcuaucggagucccgauuuagccaugcuagucg

  61 cacgaguuuagacucguagcauauagcucaguaacacguggccaaacuacccuacagacc
     |||| | |||||| ||| ||||||||||||||||||||||||||||||||||||||||||
  61 cacgggcuuagacccguggccauauagcucaguaacacguggccaaacuacccuacagacc

 121 gcaauaaccucgggaaacugaggccaauagcggauauaacucucaugcuggagugcagag
     || |||||||||||||||||||||||||||||||||||||||||||||| |||||||| |||
 121 gcgauaaccucgggaaacugaggccaauagcggauauaacucucaugcuggagugccgag

 181 aguuagaaacguuccggcgcuguaggauguggcugcggccgauuagguagaucggugggu
     |||||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||
 181 aguuagaaacguuccggcgcuguaggauguggcugcggccgauuagguagauggugggu

 241 aacggcccaccgugccgauaaucgguacggguuugugagagcaagaacccggagacggacu
     |||||||||| |||||||||||||||||||||||||||||||||||||||| |||||||| |
 241 aacggcccaccaugccgauaaucgguacgggguugugagagcaagaaccuggagacgguau

 301 cugagacaagaguccgggccnuacggggcgcagcagacgcgaaaccuuuacacugcacga
     |||||||||| ||||||| |||||||||||||||| ||||||||||||||||||||||||
 301 cugagacaagauaccgggcccuacggggcgcagcaggcgcgaaaccuuuacacugcacga

 361 cagugcgauaaggggaucccgagugcgagggcauacagucccucgcuuuucgugaccguaa
     |||||||||| |||||  ||||||| ||||||||| || ||||||||||   |||||||
 361 cagugcgauaggggggacuccgagugugagggcauauagcccucgcuuuucuguaccguaa

 421 gguggucucagaauaagg----------------------------------
     ||||||    ||| ||||
 421 ggugguacaggaacaaggacugggcaagaccggugccagccgccgcggguaauac......
```

```
Best chimeric alignment with a single sequence:

   1 auuccgguugauccugccggaggccauugcuaucggagucccgauuuagccaugcuaguug
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |
   1 auuccgguugauccugccggaggccauugcuaucggagucccgauuuagccaugcuagucg

  61 cacgaguuuagacucguagcauauagcucaguaacacguggccaaacuacccuacagacc
     |||| | |||||| ||| ||||||||||||||||||||||||||||||||||||||||||
  61 cacgggcuuagacccguggccauauagcucaguaacacguggccaaacuacccuacagacc

 121 gcaauaaccucgggaaacugaggccaauagcggauauaacucucaugcuggagugcagag
     || |||||||||||||||||||||||||||||||||||||||||||||| |||||||| |||
 121 gcgauaaccucgggaaacugaggccaauagcggauauaacucucaugcuggagugccgag

 181 aguuagaaacguuccggcgcuguaggauguggcugcggccgauuagguagaucggugggu
     |||||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||
 181 aguuagaaacguuccggcgcuguaggauguggcugcggccgauuagguagauggugggu

 241 aacggcccaccgugccgauaaucgguacgggguugugagagcaagaacccggagacggac
     |||||||||| |||||||||||||||||||||||||||||||||||||||| |||||||| |
 241 aacggcccaccaugccgauaaucgguacgggguugugagagcaagaaccuggagacgguau

 301 cugagacaagaguccgggccnuacggggcgcagcagacgcgaaaccuuuacacugcacga
     |||||||||| ||||||| |||||||||||||||| ||||||||||||||||||||||||
 301 cugagacaagauaccgggcccuacggggcgcagcaggcgcgaaaccuuuacacugcacga

 361 cagugcgauaaggggaucccgagugcgagggcauacagucccucgcuuuucgugaccguaa
     |||||||||| |||||  ||||||| ||||||||| || ||||||||||   |||||||
 361 cagugcgauaggggggacuccgagugugagggcauauagcccucgcuuuucuguaccguaa

 421 gguggucucagaauaagg----------------------------------
     ||||||    ||| ||||
 421 gguggucucagaauaagg----------------------------------
     |||||||   ||| ||||
 420 gguggucucagaauaagg----------------------------------
     |||||||   ||| ||||
 420 aggugguacaggaacaaggacugggcaagaccggugccagccgccgcggguaaua......

Best chimeric alignment:

   1 auuccgguugauccugccggaggccauugcuaucggagucccgauuuagccaugcuaguug
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||| |
   1 auuccgguugauccugccggaggccauugcuaucggagucccgauuuagccaugcuagucg

  61 cacgaguuuagacucguagcauauagcucaguaacacguggccaaacuacccuacagacc
     |||| | |||||| ||| ||||||||||||||||||||||||||||||||||||||||||
  61 cacgggcuuagacccguggccauauagcucaguaacacguggccaaacuacccuacagacc

 121 gcaauaaccucgggaaacugaggccaauagcggauauaacucucaugcuggagugcagag
     || |||||||||||||||||||||||||||||||||||||||||||||| |||||||| |||
 121 gcgauaaccucgggaaacugaggccaauagcggauauaacucucaugcuggagugccgag

 181 aguuagaaacguuccggcgcuguaggauguggcugcggccgauuag*guagacgguggggg
     |||||||||||||||||||||||||||||||||||||||||||||||||||||*||||||||
 181 aguuagaaacguuccggcgcuguaggauguggcugcggccgauuag*guagacgguggggg

 240 uaacggcccaccgugccgauaaucgguacggguuugugagagcaagaaccccggagacggac
     |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
  14 uaacggcccaccgugccgauaaucgguacggguuugugagagcaagaaccccggagacggac

 300 ucugagacaagaguccgggccnuacggggcgcagcagacgcgaaaccuuuacacugcacg
     |||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||
  74 ucugagacaagaguccgggccnuacggggcgcagcagacgcgaaaccuuuacacugcacg

 360 acagugcgauaaggggauccccgagugcgagggcauacagucccucgcuuuucgugaccgua
     ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
 134 acagugcgauaaggggauccccgagugcgagggcauacagucccucgcuuuucgugaccgua

 420 agguggucucagaauaagg
     ||||||||||||||||||||
 194 agguggucucagaauaagg
```

FIG. 2—*Continued.*

RESULTS

Program. Figure 2 shows typical outputs from the mglobalCHI program. Figure 2a shows an authentic sequence, while Fig. 2b shows the sequence of an artificial chimera. Note that in Fig. 2b the single-sequence alignment and chimeric alignment with a single sequence are truncated to improve readability (a long 3′ trailing indel was removed from each alignment). The program computes the score for the optimal alignment of the query and a single sequence in the database (*S*1; Fig. 1a) and those for two chimeric alignments, one where both partial sequences come from a single sequence (*C*1; Fig. 1b) and one where the two partial sequences are derived from two database entries (*C*2; Fig. 1c). In each case the aligned sequence(s), its start and stop points, and the point in the query where the sequence was split to make the chimeric alignments (which we label a "chimeric breakpoint") are also given. Two raw statistics are given, *C*2*S*1 and *C*2*C*1; the former is the score for alignment *S*1 subtracted from the score for the best chimeric alignment (*C*2) and the latter is the score for alignment *C*1 subtracted from the score for alignment *C*2. The next lines contain the final ISs, IS_{C2S1} and IS_{C2C1}, which have been adjusted by the software. The software adjusts for two types of effects. First, it compensates for different amounts of aligned (paired) sequence in the three alignments. Second, it corrects for two cases where obviously nonbiological alignments have occurred (5′ end of one sequence aligned with the 3′ end of another and substantial internal deletions). Full details are available in the online documentation. Finally, the three alignments are printed out, allowing the researcher to study the sources of the scores. In addition, the program can be set to generate a graph of *C*2*C*1 for each possible breakpoint along the query. Figure 3 shows two examples of such a graph, one of an authentic sequence and the other of an artificial chimera. IS_{C2C1} was the most satisfactory statistic, so in the subsequent discussions IS will refer to IS_{C2C1} unless otherwise specified.

Evaluation of the mglobalCHI program. To determine a baseline value for IS_{C2C1}, 500 authentic, nonchimeric (length, 225 to 1,740 bases) sequences were obtained from the RDP-derived database as described in Materials and Methods. After

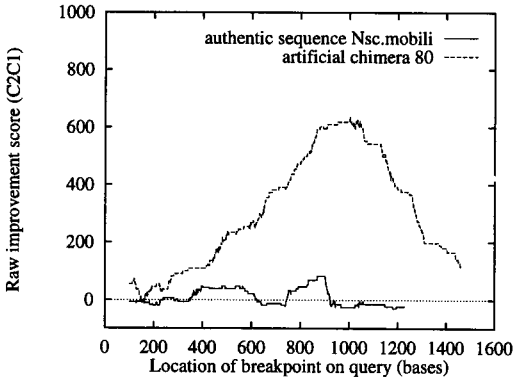

FIG. 3. Sample graphs of *C2C1* as a function of chimeric breakpoint from the mglobalCHI program for authentic sequence Nsc.mobili and chimeric sequence chi80.seq which consists of the sequence Ace.longum from bases 1 to 1001 concatenated with sequence Dfm.austra from bases 1001 to 1560.

the sequences were edited to remove discontinuities, they were tested by using the mglobalCHI program. Figure 4a shows the distribution of the improvement scores (IS_{C2C1}) for these sequences. An IS of 145 was taken to be the cutoff value for a chimeric sequence, as this gave a false positive rate of 5%. Next, 300 artificial chimeras were tested by using the mglobalCHI program to estimate the fraction of such sequences that the program can detect. With a cutoff value of 145, this resulted in the misclassification of 39 chimeras, or 13% of the total. Figure 4b shows the distribution of IS_{C2C1} for the artificial chimeras. Not surprisingly, as the estimated percent similarities of the partial sequences increased, the fraction of detected chimeras decreased, as can be seen in Fig. 5a. However, as the large degree of dispersion indicates, this is not the only factor that is acting to influence detection efficiency. The distance between the nearest sequence end and chimeric breakpoint also affected detection (although the correlation is less striking), as can be seen in Fig. 5b.

Evaluation of the CHECK_CHIMERA method. For comparison purposes the same set of sequences was tested by using the CHECK_CHIMERA program on the RDP e-mail server (13) as described in Materials and Methods. The output from CHECK_CHIMERA is a graph of IS, which we label $IS_{7\text{-tuple}}$ (because scores are based on counts of 7-tuples), plotted against chimeric breakpoint along the query sequence. The maximum value of this function was extracted and used as the score for a given sequence. This score was used even though it was of necessity a rather naive use of the output. By using only the numerical values of both IS_{C2C1} (also a naive use of a program) and $IS_{7\text{-tuple}}$ an objective measure of the two programs relative efficacy could be obtained.

An $IS_{7\text{-tuple}}$ of 50 gives a cutoff value which yields a 5% false-positive rate for the CHECK_CHIMERA program on the test set of 500 sequences. Figure 6a shows the distribution of scores for the authentic sequences. When the 300 chimeric sequences were tested (Fig. 6b), 42 sequences (14%) were misclassified as nonchimeric. As with the mglobalCHI program, chimera detection is generally more likely with decreasing parental similarity (Fig. 7a) and increasing distance between breakpoint and the nearest sequence end (Fig. 7b).

Comparison of mglobalCHI and CHECK_CHIMERA errors. Although the sensitivity and discrimination of both programs are approximately similar, the sets of sequences that were misclassified are only partially overlapping, particularly in the category of false positives (i.e., authentic sequences labeled as chimeric). Only 5 of 500 nonchimeric sequences (1%) were misclassified as chimeric by both programs, leaving 455 sequences that were classified by both as nonchimeric and 40 that were classified by only one program as chimeras. Similarly, only 30 of 300 artificial chimeras were classified by both programs as authentic sequences. Many of the misclassified chimeras are derived from sequences with high levels of local similarity and/or are sequences with chimeric breakpoints near one of the ends, that is, chimeras with a low probability of detection under any nearest-neighbor method. This leaves 249 sequences that were labeled chimeric by both methods and 21 sequences labeled as chimeric by one program only. Thus, if agreement between the programs is required for assignment as a chimera or as an authentic sequence, the nonchimeras would be partitioned into three groups: probable chimeras (1%), possible chimeras (8%), and probable nonchimeras (91%). For the chimeras these values would be 83, 7, and 10%, respectively.

As indicated earlier, using the two programs in the manner described above is a naive use of the data that they generate. In addition to the single value of the IS, the CHECK_CHIMERA program provides a graph of the value of $IS_{7\text{-tuple}}$ plotted with respect to the breakpoint in the query sequence, and the shape of this curve also contains information. A genuine chimera should have a consistently rising score until the

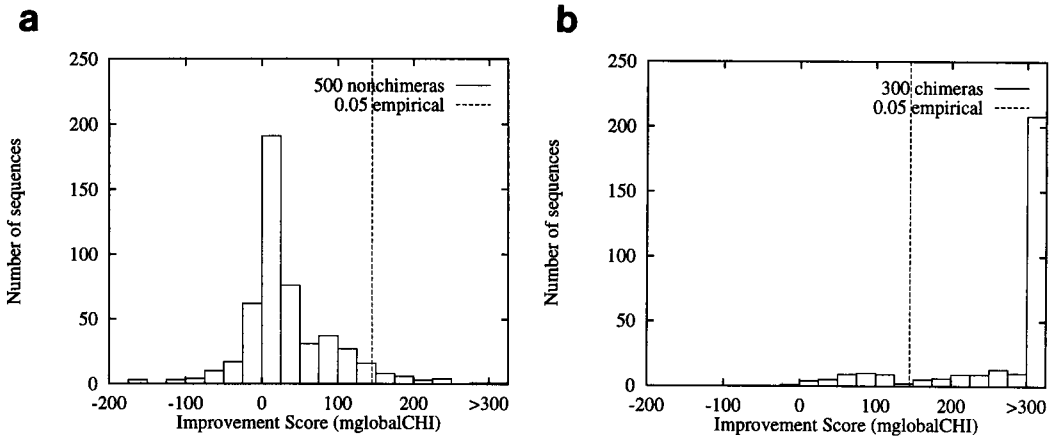

FIG. 4. Histograms of IS_{C2C1} from the mglobalCHI program for test sequences. The vertical line represents the cutoff value (145) that gives a false-positive rate of 0.05. (a) Histogram for 500 nonchimeric sequences; (b) histogram for 300 chimeric sequences.

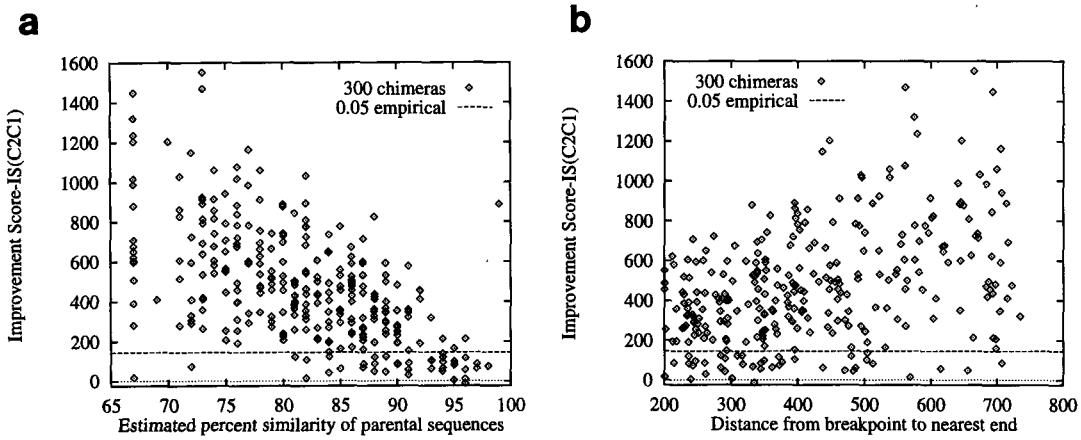

FIG. 5. Effects of similarity of parental sequences (a) and distance between chimeric breakpoint and nearest sequence end (b) on IS_{C2C1} (mglobalCHI) for 300 artificial chimeras. The dashed horizontal line is the cutoff value for a chimeric sequence (145).

breakpoint is reached, and then the score should begin a consistent fall. The mglobalCHI program also provides such a curve (although it is of the unadjusted value *C2C1* rather than of IS_{C2C1} per se). Generally, the mglobalCHI graph of *C2C1* from a high-probability chimera (i.e., one with a large IS_{C2C1}) is of the same form as the graph of a chimera detected by CHECK_CHIMERA (a consistent rise until the breakpoint is reached followed by a consistent fall); however, our observations suggest that this becomes much less clear as the maximum value of IS_{C2C1} decreases.

Use of alignments from mglobalCHI. The most important difference between the mglobalCHI and CHECK_CHIMERA programs is that mglobalCHI displays the alignments from which the inferred classification is drawn. This can be useful in correctly classifying marginal sequences (although it is important to note that none of the information in the discussion below was used to alter the cutoff values used in the program evaluation). For example, based on IS_{C2C1} only, the mglobalCHI program incorrectly classifies the sequence Stc.oralis (from *Streptococcus oralis*) as a chimera (IS_{C2C1} = 160), but a study of the alignments reveals that the sequence is in fact nonchimeric. The *S1* and *C1* alignments were made with sequence Stc.pneumo (*Streptococcus pneumoniae*), which, while complete, contained a number of ambiguous bases (N). The chimeric alignment of Stc.oralis was made with a concatenation of Stc.pneumo with Stc.pneum2, another database entry from the same species. Sequence Stc.pneum2 was incomplete, but there were no ambiguous bases. The alignments showed that all but 18 points of IS_{C2C1} were the result of replacing matches with ambiguous nucleotides in alignment *C1* with matches with defined bases in alignment *C2*. In reality therefore, the true IS was 18 and the sequence should be classified as nonchimeric (in addition, the fact that both partial sequences were from the same species suggests that the sequence is authentic). Similar observations would probably reclassify the "possibly chimeric" sequence Par.halden (*Paracoccus halodenitrificans*) as "probable nonchimera."

Other detectable misclassifications can occur when the nearest neighbor to a query is a sequence that is missing some internal data (approximately 10% of the sequences in the database lack some internal sequence data, see above). This can yield a global sequence alignment with one or more gaps. Misclassification can occur when there is a shorter, but still relatively closely related sequence without the missing internal data. The program aligns a concatenation of the sequence of the original nearest neighbor and this sequence to the query, the gap(s) gets eliminated, and IS_{C2C1} reaches a large value. The sequence Lcc.lacti2 (*Lactococcus lactis* IL1403; IS_{C2C1} =

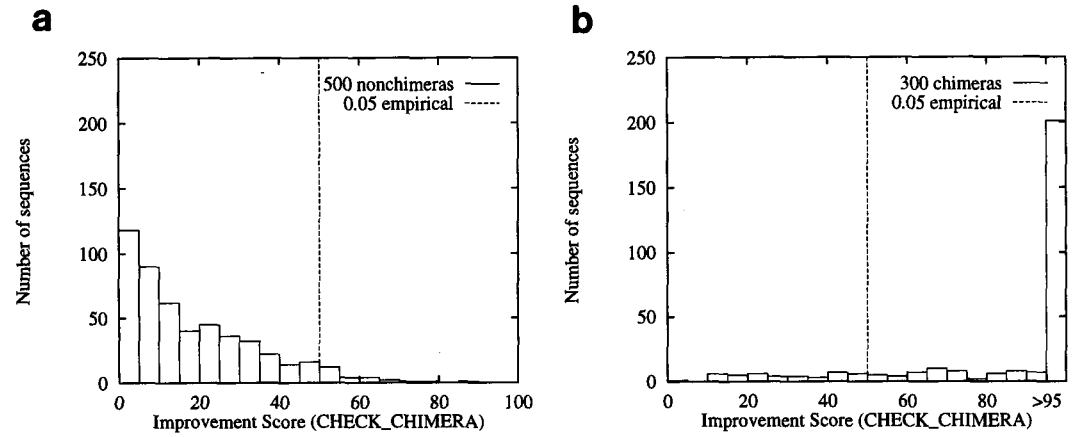

FIG. 6. Histograms of $IS_{7\text{-tuple}}$ from the CHECK_CHIMERA program for test sequences. The vertical line represents the cutoff value (50) that gives a false-positive rate of 0.05. (a) Five hundred nonchimeric sequences; (b) 300 chimeric sequences.

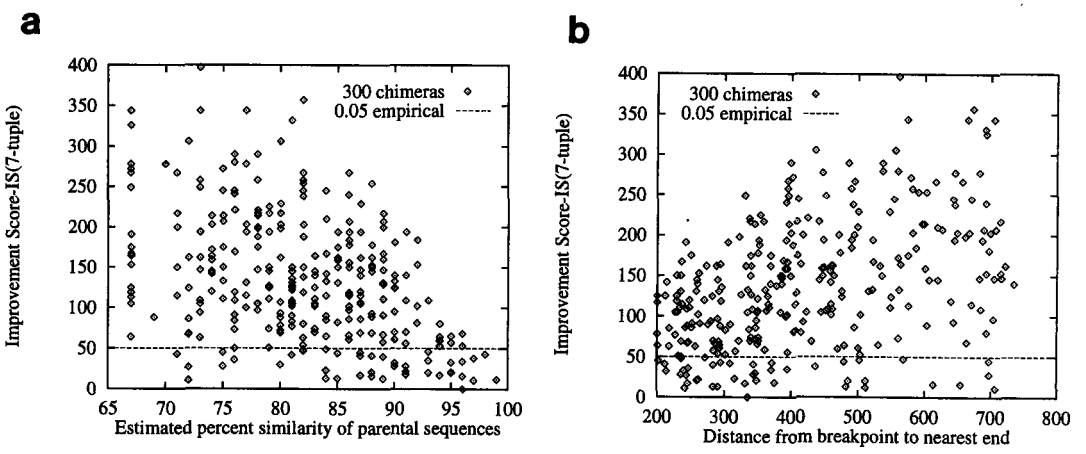

FIG. 7. Effects of similarity of parental sequences (a) and distance between chimeric breakpoint and nearest sequence end (b) on $IS_{7\text{-tuple}}$ (CHECK_CHIMERA) for 300 artificial chimeras. The dashed horizontal line is the cutoff value for a chimeric sequence (50).

155) is an example of this type of error and would most likely be reclassified as a nonchimera after study of the alignments.

DISCUSSION

Sequence similarity, scoring systems, and chimera detection. Clearly, the single most important factor in determining the success or failure of chimera detection (i.e., sensitivity) by both mglobalCHI or CHECK_CHIMERA is the similarity of the two parental molecules that form the chimera. For example, of the 14 artificial chimeras with 95% or greater estimated parental sequence similarity, only 3 were correctly classified by mglobalCHI (a dismal 21%); on the other hand, for chimeras where parental similarity was less than or equal to 80%, 121 of 123 were properly classified (98%). Table 1 lists the fraction of chimeras detected for several ranges of parental sequence percent similarity for both mglobalCHI and CHECK_CHIMERA. Interestingly, this table and Fig. 5a and 7a indicate that the distribution of misclassified chimeras with regard to parental sequence similarity varies between the two methods. The CHECK_CHIMERA method works slightly better at higher percent similarities and makes more errors at moderate and low percent similarities.

This difference may arise from the fact that as two sequences diverge their *k*-tuple match score decreases in a much less linear fashion than their dynamic programming alignment score (10). The nonlinear decrease in *k*-tuple score is caused by multiple mismatches within the same *k*-tuple. While the first mismatch in any region of a sequence may affect up to *k* *k*-tuples, the second mismatch might only affect $(k-1)$ *k*-tuples, and so on. In contrast to *k*-tuple match methods, dynamic programming alignments score each mismatch (and indel for that matter) alike. The first, second, and *n*th mismatches are treated exactly the same, so that as two sequences diverge, the score decreases in a more linear fashion (under the assumption that the optimal alignment does not change dramatically).

It is also interesting to note that there are no user-definable parameters that can be changed to make the relationship between similarity and score more linear in *k*-tuple-based scoring systems. In contrast, dynamic programming alignment scoring functions have several user-definable parameters that ultimately dictate the form of the alignment and its score, as well as the slope of the relationship between sequence similarity and score. The parameter values chosen for this study are most efficient at detecting chimeras whose parental sequences are approximately 70 to 85% similar; however, it should be possible to develop parameter sets that are optimized for other ranges of parental similarity. It is important to note however, that more stringent parameters optimized for chimeras with high parental sequence similarity may yield more false positives.

Alignment parameters and database sequence distribution. 16S rRNA sequence databases are nonuniform; certain clades (i.e., *Mycobacteria*) have many entries in the database, and by extension, a large fraction of the clade's members are represented in the database. The situation is the opposite for other clades, many *Archaea* for example. The former case is an example of a dense region of the database; the latter is one of a sparse region. A new sequence that aligns into a dense region of the database is likely to have a small evolutionary distance between itself and its nearest neighbor, while sequences that align into sparse regions will likely have large evolutionary distances between themselves and their nearest neighbors. Under a given set of alignment parameters this means that alignment scores are likely to be higher in dense regions of the database and lower in sparse regions.

Any given set of alignment parameters is likely to be suboptimal for some or most sequences. Although the current understanding of the distributions of global alignment scores under different parameter choices is somewhat limited (interested readers are referred to references 20 and 24 for more details), it is known that the most informative alignments are obtained when the parameters reflect evolutionary distance.

TABLE 1. Fraction of chimeras detected by the mglobalCHI and CHECK_CHIMERA programs partitioned by percent similarity of chimera parental sequences

Estimated similarity (%)	No. of chimeras	% Correct classification by:	
		Chimeric alignment (mglobalCHI)	*k*-tuple matching (CHECK_CHIMERA)
<70	18	94	100
70–74	38	97	92
75–79	54	100	94
80–84	72	93	90
85–89	71	89	86
90–94	33	61	73
95–99	14	21	29

TABLE 2. Overall rates of misclassification and undetected chimeras for several nearest-neighbor methods assuming 10% chimeras in the population[a]

Method[b]	% of total sequences misclassified	% of nonchimeras misclassified	% of chimeras misclassified
mglobalCHI	5.8	5	13
CHECK_CHIMERA	5.9	5	14
Both: Agree for chimera	2.6	1	17
Both: Either for chimera	9.1	9	10
Both: Three categories[c]	1.9	1	10

[a] See text for an explanation of methods.
[b] Both: agree for chimera, both methods must yield a chimera classification to label a sequence a chimera; both: either for chimera, both methods must yield an authentic classification to label a sequence authentic; both: three categories, both methods must agree for either label to be applied.
[c] Of the total number of sequences, 7.9% would be classified as possible chimeras, requiring study by other methods.

Generally, the more closely related the sequences, the more stringent the parameters should be; hence, the PAM250 matrix has a higher mean value than the PAM25 matrix. Given this, we would expect that more stringent alignment parameters should be used to align a sequence that is phylogenetically affiliated with a dense region of the database than to align a sequence affiliated with a sparse region. This points out another advantage of the mglobalCHI program, namely, adaptability. It should be possible to develop a set of parameters and cutoff scores tailored to sequences that align best into regions of varying database sparseness. The *k*-tuple scoring system of CHECK_CHIMERA on the other hand, is analogous to an invariant, highly stringent set of alignment parameters that cannot be customized.

Changes in the sequence database will also affect chimera detection. Clearly, as a greater fraction of the total spectrum of available sequences becomes available, the number of exact and close matches will increase, with a corresponding decrease in the number of false positives. The degree to which this improvement occurs (and the rate at which it happens) will be dependent, however, on whether the new sequences added to the databases are uniformly distributed over the entire spectrum of sequences or whether they are primarily close relatives of existing sequences. In addition, as the database becomes more complete, the mean difference in score between a perfect match and the observed alignment is likely to decrease, with the result that the expectation value of IS_{C2C1} for an authentic sequence will also likely decrease. The cutoff value(s) of IS_{C2C1} (and $IS_{7\text{-tuple}}$) for classification as a chimera will therefore need to be adjusted as well. As a result, when a new release of the database is made available, it will again be necessary to reestimate the distribution of scores generated by authentic and chimeric sequences, and it may be necessary to alter the value of some or all of the alignment parameters to reflect the new mean similarity level between a randomly selected 16S sequence and its nearest neighbor in the database.

Misclassification probability. The data presented here with regard to the performance of the chimeric alignment and CHECK_CHIMERA methods tend to argue for their concerted rather than individual use. Table 2 lists overall misclassification rates and the percentages of misclassified authentic sequences and chimeras, given a population of sequences that was 10% chimeras (an intermediate value in the reported range of 4.1 to 20% [18]). As can be seen in the table the two programs can be used together to minimize the overall misclassification rate, the false-positive rate, the false-negative rate, or some combination of the three, depending on whether the programs must (i) agree to label a sequence chimeric, (ii) agree to label a sequence authentic, or (iii) agree to label a sequence either way. This last method seems to be the best way to use the programs. In the optimistic case that careful study detects all chimeras originally classified as possible chimeras (but is not used on all sequences because it takes too much time or too many resources), the overall error rate is 1.9%, with a missed chimera rate of only 10%.

TABLE 3. Approximate probability that a classification as a chimera is correct for several methods and fractions of chimeras in the population[a]

Method[b]	Probability of correct classification for:		
	4% Chimeras	10% Chimeras	20% Chimeras
mglobalCHI	0.42	0.66	0.81
CHECK_CHIMERA	0.42	0.66	0.81
Both: agree for chimera	0.78	0.90	0.95
Both: either for chimera	0.29	0.53	0.71
Both: three categories[c]	0.78	0.90	0.95

[a] All estimates reflect false-positive and false-negative rates derived from simulation studies.
[b] Methods labeled "both" are as defined for Table 2.
[c] Probabilities are for sequences for which both programs agree only.

With this data in mind, it is instructive to compute the probability that a sequence that is labeled a chimera by these methods is in fact chimeric. Table 3 lists the probabilities that a chimera designation is correct (computed with Bayes' rule) by using various assignment methods and for three levels of chimeras in the total population. Note that the estimates for the three-category method only reflect the probabilities for sequences for which the CHECK_CHIMERA and mglobalCHI programs agree. The 7.9% of total sequences (10% chimeras in the population) over which the programs disagreed need to be studied by other methods for which no estimates of the probabilities of correct classification exist. In all of the cases, however, the calculations are based on the data derived experimentally during this study (i.e., the empirical false-positive and false-negative rates). A similar analysis can be conducted for sequences that are labeled as nonchimeric (Table 4).

Speed considerations. The most severe problem with a naive use of this method is its running time. On a SPARCstation20 computer work station equivalent this program takes 5 to 6 h to run for a query of 1,400 to 1,500 bases with a database of 2,520 sequences. This is a significant amount of time, particularly since every time the database changes the IS_{C2C1} cutoff needs to be recalibrated. The obvious solution to this problem

TABLE 4. Approximate probability that a classification as an authentic sequence is correct for several methods and fractions of chimeras in the population[a]

Method[b]	Probability of correct classification for:		
	4% Chimeras	10% Chimeras	20% Chimeras
mglobalCHI	0.994	0.985	0.967
CHECK_CHIMERA	0.994	0.984	0.964
Both: agree for chimera	0.993	0.981	0.959
Both: either for chimera	0.995	0.988	0.973
Both: three categories[c]	0.995	0.988	0.973

[a] All estimates reflect the false-positive and false-negative rates derived from simulation studies.
[b] Methods labeled "both" are as defined for Table 2.
[c] Probabilities are for sequences for which both programs agree only.

is to limit the number of sequences that are processed by the computationally expensive dynamic programming, either by selecting likely candidates, or by removing unlikely sequences. Several methods might be used to achieve these goals, including double filtration (16, 17) and rapid database searching tools such as BLAST (1) and FASTA (15).

We therefore implemented an interactive interface for the mglobalCHI program that prescreens the database using FASTA (15) to select likely candidates for later dynamic programming alignment. For FASTA searches, eight sequence segments are used to search the database. Two segments correspond to the 5′ and 3′ half molecules; the remaining six are 100-base segments that start or end 0, 50, or 100 bases from the 5′ or 3′ end, respectively, of the sequence. In a test set of 200 sequences (130 authentic and 70 chimeric) the preprocessing routine found all of the database entries that were used by mglobalCHI in alignments (data not shown), suggesting that there will be no loss in chimera detection efficiency by using these reduced databases. The preprocessing decreases the running time for a full-length sequence to 15 to 40 min (including the FASTA runs) on a SPARCstation20, depending on the size of the reduced database.

Testing and obtaining the mglobalCHI software. The mglobalCHI program (for UNIX workstations), as well as an interactive interface called chidetect, is available on the USC Computational Biology World Wide Web server (http://www-hto.usc.edu/). In addition to the chimera detection programs themselves, the package includes queuing software, so that no more than one of the computationally intensive mglobalCHI jobs is running at any time. Also included are Latex and online versions of a general manual, a quick reference sheet, several file format conversion routines, the database and matrix files used in these experiments, and UNIX man pages in nroff and html format, etc.

For evaluation purposes we have implemented a World Wide Web-based server for the mglobalCHI program. Users who wish to test the program before deciding to download and install the software can submit sequences by cutting and pasting them into a simple, user-friendly World Wide Web page. The mglobalCHI program will then be run, and the results will be e-mailed to the user. Also associated with the chimera server is a set of online user manuals for both the download and Web versions. The chimera detection home page is located at uniform resource locator http://www-hto.usc.edu/software/mglobalCHI/ on the USC Computational Biology Web site. The server is accessible from that page, and users are encouraged to begin there; however, the server can be directly accessed at http://www-hto.usc.edu/software/mglobalCHI/chidetect-query.html if desired.

ACKNOWLEDGMENTS

We thank Jim Robison-Cox of Montana State University for inspiring and assisting with this project. We thank Jed Fuhrman and Joel Mefford of USC for critical reading of the manuscript and for agreeing to test the software. We also thank David Ward and Mary M. Bateson of Montana State University and Bonnie L. Maidak of the RDP for helpful discussions. Finally, we thank Clyde Adley, Ann Frost and Paul Hardy of the USC Department of Mathematics for programming assistance.

This work was supported by NIH grant GM36230 to M.S.W.

REFERENCES

1. **Altshul, S. F., W. Gish, W. Miller, E. W. Meyers, and D. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
2. **Amann, R., N. Springer, W. Ludwig, H.-D. Görtz, and K.-H. Schleifer.** 1991. Identification in situ and phylogeny of uncultured bacterial endosymbionts. Nature (London) **351:**161–164.
3. **Barns, S. M., R. E. Fundyga, M. W. Jeffries, and N. R. Pace.** 1994. Remarkable phylogenetic diversity detected in a Yellowstone National Park hot spring environment. Proc. Natl. Acad. Sci. USA **91:**1609–1613.
4. **Choi, B. K., B. J. Paster, F. E. Dewhirst, and U. B. Göbel.** 1994. Diversity of cultivable and uncultivable oral spirochetes from a patient with severe destructive periodontitis. Infect. Immun. **62:**1889–1895.
5. **Fox, G. E., E. Stackbrandt, R. B. Hespell, J. Gibson, J. Maniloff, T. A. Dyer, R. S. Wolfe, W. E. Balch, R. S. Tanner, L. H. Magium, L. B. Zablen, R. Blakemore, R. Gupta, L. Bonen, B. J. Lewis, D. A. Stahl, K. R. Luehrsen, K. N. Chen, and C. R. Woese.** 1980. The phylogeny of prokaryotes. Science **209:**457–463.
6. **Fuhrman, 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.
7. **Genetics Computer Group.** 1994. Program manual for the Wisconsin Package, version 8. Genetics Computer Group, Madison, Wis.
8. **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.
9. **Komatsoulis, G. A., and M. S. Waterman.** 1997. Chimeric alignment by dynamic programming: algorithm and biological uses, p. 174–180. *In* Proceedings of the First International Meeting on Computational Molecular Biology (RECOMB97). ACM Press, New York, N.Y.
10. **Komatsoulis, G. A., and M. S. Waterman.** Statistics in molecular biology: an example from detection of chimeric 16S rRNA artifacts. *In* Proceedings of the 2nd IASC world conference on computational statistics and data analysis, in press.
11. **Kopczynski, E. D., M. M. Bateson, and D. M. Ward.** 1994. Recognition of chimeric small-subunit ribosomal DNAs composed of genes from uncultured microorganisms. Appl. Environ. Microbiol. **60:**746–748.
12. **Liesack, W., H. Weyland, and E. Stackbrandt.** 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:**192–198.
13. **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.
14. **Needleman, S. B., and C. D. Wunsch.** 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48:**443–453.
15. **Pearson, W. R., and D. J. Lipman.** 1988. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA **85:**2444–2448.
16. **Pevzner, P. A., and M. S. Waterman.** 1995. Multiple filtration and approximate pattern matching. Algorithmica **13:**135–154.
17. **Pevzner, P. A., and M. S. Waterman.** 1993. A fast filtration algorithm for the substring matching problem. Combinatorial pattern matching. Lect. Notes Comput. Sci. **684:**197–213.
18. **Robison-Cox, J. F., M. M. Bateson, and D. M. Ward.** 1995. Evaluation of nearest-neighbor methods for detection of chimeric small-subunit rRNA sequences. Appl. Environ. Microbiol. **61:**1240–1245.
19. **Smith, T. F., and M. S. Waterman.** 1981. Identification of common molecular subsequences. J. Mol. Biol. **147:**195–197.
20. **Vingron, M., and M. S. Waterman.** 1994. Sequence alignment and penalty choice: review of concepts, case studies and implications. J. Mol. Biol. **235:**1–12.
21. **Ward, D. M., M. M. Bateson, R. Weller, and A. L. Ruff-Roberts.** 1992. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Adv. Microb. Ecol. **12:**219–286.
22. **Ward, D. M., R. Weller, and M. M. Bateson.** 1990. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature (London) **345:**63–65.
23. **Waterman, M. S.** 1995. Introduction to computational biology, p. 183–232. Chapman and Hall, New York, N.Y.
24. **Waterman, M. S., M. Eggert, and E. Lander.** 1994. Parametric sequence comparisons. Proc. Natl. Acad. Sci. USA **89:**6090–6093.
25. **Weller, R., and D. M. Ward.** 1989. Selective recovery of 16S rRNA sequences from natural microbial communities in the form of cDNA. Appl. Environ. Microbiol. **55:**1818–1822.
26. **Woese, C. R., O. Kandler, and M. L. Wheelis.** 1990. Towards a natural system of organisms: proposals for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA **87:**4576–4579.