Analysis of Mixed Sequencing Chromatograms and Its Application in Direct 16S rRNA Gene Sequencing of Polymicrobial Samples

Øyvind Kommedal,1,2* Bjarte Karlsen,3 and Øystein Sæbø3

Department of Microbiology and Immunology, Haukeland University Hospital, Bergen, Norway; and Section for Microbiology and Immunology, the Gade Institute, University of Bergen, Bergen, Norway; and iSentio Ltd., Thornsholmsgate 51, Bergen, Norway

Received 2 February 2008/Returned for modification 4 April 2008/Accepted 23 August 2008

Investigation of clinical samples by direct 16S rRNA gene sequencing provides the possibility to detect nonviable bacteria and bacteria with special growth requirements. This approach has been particularly valuable for the diagnosis of patients who have received antibiotics prior to sample collection. In specimens containing more than one bacterium, direct sequencing gives mixed chromatograms that complicate further interpretation. We designed an algorithm able to analyze these ambiguous chromatograms and implemented it as a Web-based service. The algorithm contains both a new base-calling procedure and a new database search procedure. 16S rRNA gene sequencing was performed on polybacterial suspensions prepared in the laboratory. The computer program identified all bacteria correctly to the species level in 23 out of 23 samples containing two different bacteria. For samples containing three different bacteria, correct identification to the species level was achieved for three out of five and to the genus level for five out of five.

DNA sequencing has become a standard method in many microbiology laboratories. It is cheap, and the capacity for generating sequences is ever increasing. Identification of bacterial species from pure culture based on 16S rRNA gene sequencing is well established, and there is increasing interest in doing 16S rRNA gene sequencing directly on clinical samples (5, 6, 11, 12). This provides the opportunity to identify bacteria that die during transportation or as a consequence of antibiotic treatment. The latest advances in PCR and sequencing technology also offer faster identification than standard phenotypical methods that depend on bacterial growth. This is of particular importance for slow-growing bacteria, bacteria with special growth requirements, and unusual bacteria for which reliable, standard phenotypic test batteries are not defined.

One of the remaining problems for this approach is samples containing more than one bacterial species. For these samples, direct sequencing results in mixed chromatograms containing two or more fluorescent signals in positions where the 16S rRNA genes differ for the bacteria present in the sample (Fig. 1A). Correct reading of these chromatograms is complicated by two factors. (i) In a chromatogram, there will always be some noise, i.e., low-intensity signals originating from the baseline or from nonspecific primer binding. Including these in the base calling can decrease the specificity of subsequent analysis. To avoid them, we used a cutoff value on the y axis (see Results). (ii) The migration of a DNA fragment through the capillary gel in the sequencing reaction is dependent on its number of bases but also, to some degree, on its base composition. Therefore, a fragment of n bases from bacterium A will migrate with a different speed than a fragment of n bases from bacterium B. This results in a relative displacement of the fluorescence peaks in the corresponding sequence position (Fig. 1B). The reading algorithm had to be able to distinguish between a base displacement and a new position. Also, with large displacements, it can be difficult to decide whether a signal from bacterium A corresponds to the signal in position Y or in position Y + 1 from bacterium B. The degrees and relative directions of displacements fluctuate through the sequence. In monobacterial chromatograms, although the distance between any two successive peaks varies, the average peak distance, D (length of chromatogram/number of bases) is very stable (this is valid for chromatograms from the ABI 3730 and 3100 sequencers). We used this average peak distance in our mixed chromatograms and divided them into N (length of trimmed chromatogram/D) blocks with size D. All fluorescence peaks within a block were determined to belong to the same sequence position. In a mixed chromatogram, in positions where the different bacteria have identical bases, the peak position of the resulting fused fluorescent signal represents a compromise between the contributing signals. These peaks are referred to as anchor peaks (Fig. 1B). To optimize block positioning, the center of the first block was defined by the first anchor peak in the trimmed sequence, and base calling started at this point. Every time a new anchor peak was detected, the block positioning was readjusted accordingly (Fig. 1B).

Matching against a reference database. The search was performed by dividing the query sequence into overlapping pieces of w bases, which were then compared successively to the sequences in the database (the subject sequences). When a piece contained ambiguous bases, all possible combinations (words) were constructed and included in the comparison. Only words that had a similarity over a given threshold were recognized as a match (hit). Pieces with a high proportion of ambiguous bases could generate several thousand words, so a traditional unrestricted use of this approach would have given a lot of irrelevant hits all along the subject sequences (Fig. 2A). By utilizing the primer sequences used in the sequencing reaction to detect the hypothetical primer binding sites on all the sequences in the database, it was

* Corresponding author. Mailing address: Department of Microbiology and Immunology, Haukeland University Hospital, 5021 Bergen, Norway. Phone: 47 48 19 26 19. Fax: 47 94 76 59 36. E-mail: oyvind.kommedal@isentio.com.

7 Published ahead of print on 3 September 2008.

MATERIALS AND METHODS

Algorithm. The interpretation of mixed chromatograms is dependent upon both correct reading of the chromatograms (base calling) and the subsequent matching procedure (search).

Reading the chromatogram. Direct 16S rRNA gene sequencing of polymicrobial samples results in mixed chromatograms containing two or more fluorescent signals in positions where the 16S rRNA genes differ for the bacteria present in the sample (Fig. 1A). Correct reading of these chromatograms is complicated by two factors. (i) In a chromatogram, there will always be some noise, i.e., low-intensity signals originating from the baseline or from nonspecific primer binding. Including these in the base calling can decrease the specificity of subsequent analysis. To avoid them, we used a cutoff value on the y axis (see Results). (ii) The migration of a DNA fragment through the capillary gel in the sequencing reaction is dependent on its number of bases but also, to some degree, on its base composition. Therefore, a fragment of n bases from bacterium A will migrate with a different speed than a fragment of n bases from bacterium B. This results in a relative displacement of the fluorescence peaks in the corresponding sequence position (Fig. 1B). The reading algorithm had to be able to distinguish between a base displacement and a new position. Also, with large displacements, it can be difficult to decide whether a signal from bacterium A corresponds to the signal in position Y or in position Y + 1 from bacterium B. The degrees and relative directions of displacements fluctuate through the sequence. In monobacterial chromatograms, although the distance between any two successive peaks varies, the average peak distance, D (length of chromatogram/number of bases) is very stable (this is valid for chromatograms from the ABI 3730 and 3100 sequencers). We used this average peak distance in our mixed chromatograms and divided them into N (length of trimmed chromatogram/D) blocks with size D. All fluorescence peaks within a block were determined to belong to the same sequence position. In a mixed chromatogram, in positions where the different bacteria have identical bases, the peak position of the resulting fused fluorescent signal represents a compromise between the contributing signals. These peaks are referred to as anchor peaks (Fig. 1B). To optimize block positioning, the center of the first block was defined by the first anchor peak in the trimmed sequence, and base calling started at this point. Every time a new anchor peak was detected, the block positioning was readjusted accordingly (Fig. 1B).

Matching against a reference database. The search was performed by dividing the query sequence into overlapping pieces of w bases, which were then compared successively to the sequences in the database (the subject sequences). When a piece contained ambiguous bases, all possible combinations (words) were constructed and included in the comparison. Only words that had a similarity over a given threshold were recognized as a match (hit). Pieces with a high proportion of ambiguous bases could generate several thousand words, so a traditional unrestricted use of this approach would have given a lot of irrelevant hits all along the subject sequences (Fig. 2A). By utilizing the primer sequences used in the sequencing reaction to detect the hypothetical primer binding sites on all the sequences in the database, it was
possible to estimate for every piece of the query sequence the corresponding area on the subject sequences. Matching of the words derived from a specific piece was then restricted to its corresponding area (window), and accidental hits in other parts of the subject sequences were avoided (Fig. 2B). To secure maximum sensitivity, the areas, the chromatograms were divided into equal-size blocks (×) based on the average peak distance in monobacterial chromatograms. All bases inside a block were defined to the same position. Block positioning was adjusted correspondingly. In this example, it is adjusted to the left (red circle). The 16S rRNA gene sequences for the respective genes are given under the chromatogram. (B) Examples of anchor peaks and sequence blocks. The chromatogram was obtained by direct 16S rRNA gene sequencing of a bacterial mixture containing 10×1000 μl of acid-washed glass beads (size, ±100 μm; Sigma Aldrich), and lysis was performed by a 45-s run at a speed of 6.5 m/s in a FastPrep instrument (Qiagen). This gave homogeneous signal intensities for all bacteria except Staphylococcus aureus, which still tended to give low peak signals.

At the end of the search, when all possible words from the ambiguous query sequence had been matched against their corresponding areas on the sequences in the database, a similarity percentage was calculated for each of the subject sequences. The similarity percentage points out the most similar sequence construct of the mixed sequence in question.

**Interpretation of results.** The results were presented as a list sorted with the highest similarity percentage on top. To decide which ones and how many to include in the final answer, a cutoff value had to be established. In addition, the risk that bacteria similar to those actually present in the sample could reach the relative base displacements in corresponding positions differ. In addition, for most samples, unequal 16S rRNA gene lengths and locations of evolutionary insertions and deletions give dissimilar mixing in the forward and reverse directions. The results are presented in Table 3. On the basis of these results, a piece size of 13 together with a window size of 13 + (2 × 3) (combination C) was considered the most favorable.

**Trimming and definition of signal cutoff values.** were done on November 1, 2008.

**RESULTS.** In order to determine how well the algorithm described above works on mixed chromatograms, a number of suspensions containing two or three different bacterial species were sequenced as shown in Table 1 and Table 2, respectively. Each mixed chromatogram, 1 to 23, forward plus reverse, was then run with eight different combinations of piece and window sizes in order to identify the optimal parameter settings. The forward and reverse chromatograms for one sample in this context can be regarded as unique chromatograms because the relative base displacements in corresponding positions differ. In addition, for most samples, unequal 16S rRNA gene lengths and locations of evolutionary insertions and deletions give dissimilar mixing in the forward and reverse directions. The results are presented in Table 3. On the basis of these results, a piece size of 13 together with a window size of 13 + (2 × 3) (combination C) was considered the most favorable.

**Trimming and definition of signal cutoff values.** were done on November 1, 2008.

**RESULTS.** In order to determine how well the algorithm described above works on mixed chromatograms, a number of suspensions containing two or three different bacterial species were sequenced as shown in Table 1 and Table 2, respectively. Each mixed chromatogram, 1 to 23, forward plus reverse, was then run with eight different combinations of piece and window sizes in order to identify the optimal parameter settings. The forward and reverse chromatograms for one sample in this context can be regarded as unique chromatograms because the relative base displacements in corresponding positions differ. In addition, for most samples, unequal 16S rRNA gene lengths and locations of evolutionary insertions and deletions give dissimilar mixing in the forward and reverse directions. The results are presented in Table 3. On the basis of these results, a piece size of 13 together with a window size of 13 + (2 × 3) (combination C) was considered the most favorable.

**Trimming and definition of signal cutoff values.** were done on November 1, 2008.
set well above this. A cutoff higher than 80 seldom contributed to specificity.

The results were presented as a list with the highest-scoring subjects on top. The challenge was to decide which ones and how many to include in the answer. Close inspection of the chromatograms usually gave an indication of how many but could also be misleading. We therefore adopted an empirically set cutoff at 99.3% similarity. All subjects reaching this cutoff were included in the final answer (rule 1). If the answer included two bacteria from the same genus, both above the cutoff, only the higher-scoring species was chosen, although we could not exclude the possibility that both were present (rule 2). This rule was also applied to bacteria from different genera known to have very similar 16S rRNA genes, e.g., some *Citrobacter/Enterobacter/Klebsiella* spp. or *E. corrodens/Kingella denitrificans*. Based upon our experience with the method, the finding of two bacteria with similarity in the sequenced area of 95% should be interpreted with caution. Typically, they could be easily distinguishable if they were the only bacteria present in the mixture, but if they were together with a third bacterium, only the one with the higher score should be accepted.

To better illustrate the principles for interpretation, two examples are given. The interpretation of chromatogram 15 forward was simple. Only *S. agalactiae* and *A. aphrophilus* scored above the cutoff (according to rule 1), and they were not assumed to have similar 16S rRNA genes. The interpretation of chromatogram 13 forward was more complex. *S. pneumoniae* scored 100% and was included in the answer. Other streptococci also scored higher than 99.3% but lower than 100% and were therefore excluded (according to rule 2). *K. pneumoniae, Klebsiella oxytoca,* and *Enterobacter aerogenes* all scored above the cutoff. *K. oxytoca* scored lower than *K. pneumoniae* and was excluded (according to rule 2). A pairwise alignment was performed between *K. pneumoniae* and *E. aerogenes*. This showed 98.2% similarity, and consequently, only *K. pneumoniae* was included in the final answer (according to rule 2).

The detailed results for the mixed samples containing two bacteria using parameter combination C are presented in Table 1. An overview of the parameter combinations is given in Table 3. All 23 samples (combined interpretation of forward and reverse chromatograms) were correctly identified to the species level. If we look at the individual chromatograms, correct identification to the species level was achieved in 40 out of 46. Correct identification to the genus level was achieved in 46 out of 46. It was not possible to differentiate between *E. coli* and *Shigella* spp. This was not considered an error, as their sequences are identical with the primers used here. The results for the triple chromatograms (24-F to 28-F plus 24-R to 28-R) are presented in Table 2. These chromatograms were run with a piece size of 20, together with a window size of $20 + (2 \times 2)$ (combination B from Table 3). If we look at each chromatogram individually, 5 out of 10 were correctly identified to the species level, whereas 9 out of 10 were correctly

---

**FIG. 2.** Search principle. The mixed query sequence was divided into pieces of size $w$. For each piece, all possible base combinations (words) were constructed and compared to the subject sequences. In this example, from the random piece $n$, it is possible to create eight different words. (A) An unrestricted matching procedure for these eight words against one of the sequences in the database. Three of them have a similarity above the cutoff with different parts of the subject sequence and are recognized as hits. Only word 5 (red circle) represents a relevant hit. Words 1 and 4 have by chance significant similarity to other parts of the subject sequence and contribute to reduced search specificity. (B) A restricted matching procedure. The area ($w + 2x$) on the subject sequence that corresponds to the area of piece $n$ on the query sequence is estimated based on the hypothetical binding site for the forward primer (F-PBS) and the distance $d$ from the start of the query sequence to the beginning of piece $n$. It is made slightly larger than the piece size ($w$) to secure maximum sensitivity. Matching of the words from piece $n$ is now restricted to this specific area (window), and the nonspecific hits from words 1 and 4 are avoided.
identified to the genus level. On the sample level, three out of five were identified to the species level and five out of five to the genus level.

To achieve the above-mentioned results, a single manual correction of the base calling was performed in chromatograms 24 reverse and 25 reverse. For the remaining 54 chromatograms, no corrections were necessary.

**DISCUSSION**

Direct 16S rRNA gene sequencing has several advantages compared to culture (2, 9). The possibility to analyze chromatograms from samples containing up to three different bacteria could increase its usefulness and make it relevant for a broader range of infections and samples. Also, the potential for reducing the time to identification is even larger for mixed samples, as a cultivation-based approach requires isolation and subcultivation of individual colonies.

In our opinion, the most important feature of direct 16S rRNA gene sequencing is the possibility to analyze samples collected after the administration of antibiotics. Foci in internal organs are often difficult to discover in the first place, and once they are discovered, sample collection frequently requires complicated invasive procedures. Consequently, initiation of antibiotic treatment cannot await sample collection, and cultivation often yields a negative answer. In polymicrobial infections, cultivation can actually yield misleading answers if the antibiotics administered have affected the involved bacteria unequally, permitting some to grow and not others.

**TABLE 1. Description of bacterial mixtures containing two bacteria and results from the RipSeq software analysis**

<table>
<thead>
<tr>
<th>ID</th>
<th>Bacterium 1</th>
<th>%</th>
<th>Bacterium 2</th>
<th>%</th>
<th>Forward chromatogram</th>
<th>Bacterium 1</th>
<th>%</th>
<th>Bacterium 2</th>
<th>%</th>
<th>Reverse chromatogram</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>E. coli</td>
<td>100</td>
<td>B. fragilis</td>
<td>99.7</td>
<td>E. coli</td>
<td>100</td>
<td>B. fragilis</td>
<td>99.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>E. coli</td>
<td>100</td>
<td>C. gingivalis</td>
<td>100</td>
<td>E. coli</td>
<td>99.5</td>
<td>C. gingivalis</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>E. coli</td>
<td>100</td>
<td>E. faecalis</td>
<td>100</td>
<td>E. coli</td>
<td>100</td>
<td>E. faecalis</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>E. coli</td>
<td>100</td>
<td>K. pneumoniae</td>
<td>100</td>
<td>E. coli</td>
<td>100</td>
<td>K. pneumoniae</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>E. coli</td>
<td>100</td>
<td>P. vulgaris</td>
<td>100</td>
<td>E. coli</td>
<td>100</td>
<td>P. vulgaris</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>E. coli</td>
<td>100</td>
<td>P. vulgaris</td>
<td>100</td>
<td>E. coli</td>
<td>100</td>
<td>P. vulgaris</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>E. coli</td>
<td>100</td>
<td>P. vulgaris</td>
<td>100</td>
<td>E. coli</td>
<td>100</td>
<td>P. vulgaris</td>
<td>99.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>H. influenzae</td>
<td>100</td>
<td>E. corrodens</td>
<td>100</td>
<td>H. influenzae</td>
<td>100</td>
<td>E. corrodens</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>H. influenzae</td>
<td>100</td>
<td>E. faecalis</td>
<td>100</td>
<td>H. influenzae</td>
<td>100</td>
<td>E. faecalis</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>H. influenzae</td>
<td>100</td>
<td>M. catarrhalis</td>
<td>100</td>
<td>H. influenzae</td>
<td>99.7</td>
<td>M. catarrhalis</td>
<td>100</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>H. influenzae</td>
<td>100</td>
<td>S. pneumoniae/mitis/infantis/oralis</td>
<td>99.7</td>
<td>H. influenzae</td>
<td>100</td>
<td>S. pneumoniae</td>
<td>99.7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**TABLE 2. Description of bacterial mixtures containing three bacteria and results from the RipSeq software analysis**

<table>
<thead>
<tr>
<th>ID</th>
<th>Bacterium 1</th>
<th>%</th>
<th>Bacterium 2</th>
<th>%</th>
<th>Bacterium 3</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>24-F</td>
<td>E. coli</td>
<td>100</td>
<td>B. fragilis</td>
<td>99.5</td>
<td>S. aureus/hominis/lugdunensis/warneri</td>
<td>99.7</td>
</tr>
<tr>
<td>24-R</td>
<td>E. coli</td>
<td>100</td>
<td>B. fragilis</td>
<td>100</td>
<td>S. aureus/capitis/caprae/epidermidis/warneri</td>
<td>99.5</td>
</tr>
<tr>
<td>25-F</td>
<td>E. coli</td>
<td>100</td>
<td>B. fragilis</td>
<td>100</td>
<td>S. maltophilia</td>
<td>99.5</td>
</tr>
<tr>
<td>25-R</td>
<td>E. coli</td>
<td>100</td>
<td>B. fragilis</td>
<td>99.7</td>
<td>S. maltophilia</td>
<td>99.4</td>
</tr>
<tr>
<td>26-F</td>
<td>E. coli</td>
<td>100</td>
<td>K. pneumoniae</td>
<td>99.7</td>
<td>P. vulgaris</td>
<td>99.7</td>
</tr>
<tr>
<td>26-R</td>
<td>E. coli</td>
<td>99.7</td>
<td>K. pneumoniae</td>
<td>99.7</td>
<td>P. vulgaris</td>
<td>99.7</td>
</tr>
<tr>
<td>27-F</td>
<td>S. agalactiae</td>
<td>100</td>
<td>A. aphrophilus</td>
<td>100</td>
<td>E. faecalis</td>
<td>100</td>
</tr>
<tr>
<td>27-R</td>
<td>S. agalactiae</td>
<td>100</td>
<td>A. aphrophilus</td>
<td>100</td>
<td>E. faecalis/faecium/casseliflavus/durans</td>
<td>100</td>
</tr>
<tr>
<td>28-F</td>
<td>S. pneumoniae/mitis/infantis/infantis/oralis</td>
<td>99.7</td>
<td>K. pneumoniae</td>
<td>100</td>
<td>M. catarrhalis</td>
<td>100</td>
</tr>
<tr>
<td>28-R</td>
<td>S. pneumoniae</td>
<td>ND</td>
<td>K. pneumoniae</td>
<td>100</td>
<td>M. catarrhalis</td>
<td>100</td>
</tr>
</tbody>
</table>

**Boldface indicates that correct differentiation to species level was not possible; the correct species are underlined. ND, not determined. All bacteria had a final concentration of 1.0 McFarland standard in their respective mixtures.**
Because of the need to perform susceptibility testing, nucleic acid-based identification cannot replace cultivation. Cultivation is also more sensitive than PCR-based detection when samples contain viable bacteria. Large volumes of sample material can be cultivated in liquid media, with a theoretical sensitivity of 1 CFU/sample. In sequencing, the sensitivity is limited both by the method of DNA extraction and by the number of cycles that can be run in the first PCR before contaminant DNA in the reagents gives a false-positive result (4).

Some publications have expressed that primer cross-reactivity to eukaryotic DNA is a concern when bacterial 16S rRNA genes are amplified directly from a human clinical sample (7). This results in mixed chromatograms. The RipSeq algorithm can in part solve this problem with its ability to read mixed chromatograms and extract the information from the bacterial DNA. However, interference from human DNA decreases the differentiating power of the 16S rRNA gene and sometimes makes it impossible to distinguish between bacterial species with very similar 16S rRNA genes. Primers should therefore be selected carefully, and this was the main argument when we chose our primers. Although they do not target the most variable parts of the 16S rRNA gene, in our experience, they do not cross-react with human DNA.

When direct sequencing is used on polymicrobial samples, additional issues have to be addressed. The main challenge of bacterial identification on the basis of a mixed chromatogram from the 16S rRNA gene is the enormous number of possible combinations in relation to the relatively short variable segments upon which differentiation is dependent (1). This applies in particular to species with very similar 16S rRNA genes. If, for example, the 16S rRNA genes of bacteria A and B are different only in position P and bacterium A is mixed with bacterium C, which is identical to bacterium B in position P, the algorithm will fail to decide whether the mixture contains bacteria A and C or B and C (or A, B, and C). However, the chance for this to happen in both the forward and reverse chromatogram is small, because the alignments of the two sequences normally will be different. This is exemplified in Table 1 for samples 11, 13, 17, 18, 19, and 21. When three bacteria are mixed, the chance for this to happen is substantially higher, as exemplified by sample 24 in Table 2.

The difference in the sequenced area between S. pneumoniae and Streptococcus mitis/Streptococcus oralis/Streptococcus intermedius is only a single base. Reliable differentiation between these species based on the 16S rRNA gene alone is probably not possible, even from pure culture (9). The difference between S. pneumoniae and Streptococcus sanguinis is 6 bases. Staphylococcus capitis/Staphylococcus caprae/Staphylococcus epidermidis are identical in the sequenced area and differ from S. aureus in three positions. Staphylococcus warneri, Staphylococcus lugdunensis, and Staphylococcus intermedius differ from S. aureus in three, four, and six positions, respectively. The difference between B. subtilis and Bacillus licheniformis is 5 bases.

As expected, the smallest differentiating margins were seen between some species within the genus Staphylococcus, between some species in the genus Streptococcus, and between some species in the genus Enterococcus. In some situations, the inability to distinguish between different species within these genera may not be clinically problematic because bacteria with similar 16S rRNA genes frequently have similar susceptibilities to antibiotics and similar clinical importance. It becomes a problem, however, when it is not possible to distinguish between, e.g., S. aureus and a coagulase-negative staphylococcus or between S. pneumoniae and other members of the S. mitis group. In these situations, other genes can be sequenced in addition, such as sodA or rpoB (8–10). The limitation that lies in the variability of the 16S rRNA gene is the main reason for us to say that in chromatograms containing more than three different species, specificity will not be sufficient to obtain reliable answers on a routine basis. This needs to be taken into consideration when deciding which samples to test. We expect suitable samples to include aspirates from abscesses in internal organs, except those related directly to the intestine, as well as aspirates from deep cutaneous, subcutaneous, muscular, and retroperitoneal abscesses. Body fluids, like cerebrospinal fluid, synovial fluid, pleural fluid, and bile, may also be relevant. The often serious nature of infections in these foci, in combination with the resources put into collecting the specimens in the first place, can justify the extra cost of the sequencing procedure. Samples from open wounds or other areas that communicate directly with mucous membranes or skin are not likely to be suitable. Swabs, small-needle biopsy specimens, or very small amounts of aspirate are probably not suitable, since the dilution rate in the DNA extraction procedure will become very high. If a sample that contains more than three species is sequenced, two situations might occur. (i) The 16S rRNA genes of all bacteria present are successfully sequenced and expressed in the chromatogram. The resulting answer from the RipSeq analysis contains more than three species and should be rejected. (ii) Because of unequal concentrations of the different species, only three or fewer are detectable in the chromatogram, giving an acceptable but incomplete answer. We have found that this occurs if the difference in concentrations exceeds 1:10, with some variations dependent on the affinity of the primers for the different targets and the number of copies of the 16S rRNA gene in the respective bacteria (3). Differences in concentration are in theory not a problem in cultivation, but competition for nutrients, different growth speeds,
similar-looking colonies, and swarming may camouflage the presence of some species.

The triple chromatograms were more complex than those containing only two bacteria, and consequently, there was a higher risk for erroneous base calling. To compensate for this, more relaxed search parameters could have been beneficial. However, the higher number of words that could possibly be derived from each piece increased the risk for random hits against nonrelevant sequences in the database. Consequently, the samples had to be run with increased stringency to achieve sufficient specificity. In a clinical sample, the number of bacteria is a priori unknown. We therefore suggest an approach in which the samples are first run with parameter combination C. If the resulting answer yields one or two bacteria, the answer is accepted. If the answer yields three or more bacteria, the chromatogram will have to be reanalyzed using the more stringent parameter combination B (Table 3) and only this answer accepted.

If the respective bacteria were present in unequal concentrations, or unevenly lysed, correspondingly large differences in signal intensities were seen, and the cutoff value had to be set lower to detect all the relevant peaks. A cutoff lower than 30 generally led to the inclusion of a high number of nonspecific signals, which was the situation for chromatograms 11 forward and 19 reverse. In addition, it made base calling more vulnerable to relative base displacements in corresponding positions by reducing the number of anchor peaks.

We have shown in this study that it is possible to analyze a mixed chromatogram containing up to three different bacteria. How well this will work on clinical samples and to what degree it will provide valuable information remain to be established in a study of patient samples. Preliminary results from our laboratory indicate that the method is a valuable supplement to culture, in particular when the patient has received antibiotics prior to specimen collection.

ACKNOWLEDGMENTS

This work was supported by The Research Council of Norway, Innovation Norway, and Bergen Technology Transfer.

We gratefully acknowledge Harald G. Wiker and Lars Haar for critical comments and suggestions on the manuscript.

A patent application (B. Karlsen, Ø. Sæbø, and Ø. Kommedal, patent PCT/NO2007/00314) has been filed for several aspects of the algorithm described in this article. The RipSeq program will be accessible through a commercial Web service partly owned by the authors.

REFERENCES