Comparison of two approaches for the classification of 16S rRNA gene sequences

Sonia Chatellier,1 Nathalie Mugnier,2 Françoise Allard,1 Bertrand Bonnaud,2 Valérie Collin,1 Alex van Belkum,1 Jean-Baptiste Veyrieras2 and Stefan Emler3

1bioMérieux SA, Unit Microbiology, R&D Microbiology, La Balme Les Grottes, France
2bioMérieux SA, Unit Innovation and Systems, Marcy l'Etoile, France
3SmartGene, Zug, Switzerland

The use of 16S rRNA gene sequences for microbial identification in clinical microbiology is accepted widely, and requires databases and algorithms. We compared a new research database containing curated 16S rRNA gene sequences in combination with the LCA (lowest common ancestor) algorithm (RDB-LCA) to a commercially available 16S rDNA Centroid approach. We used 1025 bacterial isolates characterized by biochemistry, matrix-assisted laser desorption/ionization time-of-flight MS and 16S rDNA sequencing. Nearly 80 % of isolates were identified unambiguously at the species level by both classification platforms used. The remaining isolates were mostly identified correctly at the genus level due to the limited resolution of 16S rDNA sequencing. Discrepancies between both 16S rDNA platforms were due to differences in database content and the algorithm used, and could amount to up to 10.5 %. Up to 1.4 % of the analyses were found to be inconclusive. It is important to realize that despite the overall good performance of the pipelines for analysis, some inconclusive results remain that require additional in-depth analysis performed using supplementary methods.

INTRODUCTION

Identification of bacterial species on the basis of 16S rDNA sequencing continues to be used widely (Claridge, 2004; Woo et al., 2008; Schlaberg et al., 2012). GenBank, the largest public nucleic acid sequence database, contains >4730000 16S rRNA gene sequences (June 2013). To identify bacterial isolates of unknown identity, the basic strategy is to compare their 16S rRNA sequence to those in public databases. A 1–3 % difference between rRNA sequences reflects the range where closely related species can still be differentiated, although this is sometimes confounded by intra-species genetic diversity (Stackebrandt & Goebel, 1994).

A prerequisite for reliable 16S rDNA comparison is the availability of high-quality sequencing data so that only true differences are documented. Several initiatives have led to the creation of public but ‘curated’ 16S rDNA databases, such as RDP (Ribosomal Database Project; http://rdp.cme.msu.edu) (Cole et al., 2009), SILVA (www.arb-silva.de) (Quast et al., 2013) and greengenes (http://greengenes.lbl.gov/cgi-bin/nph-index.cgi) (DeSantis et al., 2006). Commercial libraries such as the integrated database network system (IDNS) Eubacterial Database (SmartGene) (Simmon et al., 2006; Conville et al., 2010; Helal et al., 2011) and the MicroSeq database (Life Technologies) (Woo et al., 2003) are also available. All databases need to be updated continuously, generating a constant need for bio-informatics expertise, secure servers and associated costs. This is why many microbiologists prefer to create proprietary rRNA databases and usually rely on freely available alignment methods such as BLAST (Basic Local Alignment Search Tool) (Altschul et al., 1990).

Here, we compared a new research database containing curated 16S rRNA gene sequences in combination with the LCA (lowest common ancestor) algorithm (RDB-LCA) and the SmartGene Centroid approach for assessing the accuracy of identification for a large set of de novo 16S rDNA sequences.

METHODS

Bacterial isolates. A random collection of 1025 bacterial strains provided a dataset of 1025 sequences covering 377 bacterial species
(between one and 11 strains per species) from 118 genera. Strains were diverse, and were recovered from clinical, veterinary and food specimens. Geographical origins of the samples included Europe and North America. The 16S rDNA sequence length ranged from 456 to 1563 bp with only 32 sequences being <1000 bp (see below).

**Reference identification.** All isolates were identified using phenotypic tools [motility, catalase, Gram staining, oxidase, API strips and VITEK2 (bioMérieux)]. Isolates were also identified by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) MS (VITEK MS; bioMérieux). In addition, 16S rDNA sequencing was performed for all strains. After overnight culture on blood agar plates, DNA was extracted using EasyMag (bioMérieux). The 16S rDNA sequencing was performed using the Sanger approach. Three overlapping regions of the gene were amplified and sequenced using universal primers (Greisen et al., 1994). Sequence length ranged from 456 to 1563 bp. Identification using GenBank BLAST analysis was performed prior to and independent of the two 16S rDNA classifications introduced below.

**Comparison of two 16S rRNA gene databases and algorithms.**

Two approaches were selected and used for comparative identification of the 1025 sequences introduced above.

First, for the new RDB-LCA, full 16S rRNA gene sequences were downloaded from GenBank (September 2011), identical entries were removed and sequences were grouped by species. Then, alignments were done for all sequences for each species using the NUCmer (nucleotide mummer) alignment script. We defined the alignment quality between sequences by:

\[
\min(L_1, L_2) \times 100 \\
\min(L_3, L_4)
\]

where \(\min(L_1, L_2)\) is the minimal length of the alignment (\(L_1\) and \(L_2\) are the length of aligned bases for the request and query sequences, respectively), \(\% Id\) is the identity percentage and \(\min(L_3, L_4)\) is the maximal length of alignable bases (\(L_3\) and \(L_4\) are the total lengths of the request and query sequences, respectively). For each species, we defined a symmetry matrix of similarity between the various GenBank sequences. We applied a 95th percentile median similarity rule for each species to exclude outliers with ambiguous taxonomic assignment. We removed short sequences with uncertain 16S rRNA gene annotation and deleted entries with a base ambiguity percentage >10 %. The number of sequences decreased from 290 709 to 68 038. In order to classify strain-specific 16S rDNA gene sequences, we used BLAST and retained optimal hits. We tried to identify ‘best hits’, which were defined as BLAST hits that presented the highest identity percentages. If multiple ‘best hits’ were found (several alignments with the same identity percentage), the result returned by our classifier was identified as the LCA (Harel & Tarjan, 1984). Ultimately, RDB-LCA contained 16S rRNA gene sequences with a length ranging from 1000 to 3000 bp (including upstream and downstream sequences) with a mean length of 1425 nt. RDB-LCA covered 7708 bacterial species corresponding to 1348 genera (Fig. 1); 1110 species (14 %) were characterized by a single representative sequence with a mean number of 8.8 sequences per species. Neisseria meningitidis, Bacillus cereus, Lactobacillus helveticus, Bacillus pumilus, Lactobacillus plantarum, Bacillus licheniformis, Pseudomonas aeruginosa, Pseudomonas putida, Lactococcus lactis and Haemophilus influenza were represented by >500 sequences. Highly represented genera were Bacillus spp. (5396 sequences), Lactobacillus spp. (3930 sequences), Streptomyces spp. (3847 sequences) and Pseudomonas spp. (3082 sequences).

The second approach involved the SmartGene 16S rRNA Eubacterial Database, which is governed by a proprietary automated daily ribosomal sequence extraction from GenBank/EMBL. Extraction profiles and criteria to ‘filter’ irrelevant entries are applied based upon information on overall sequence conservation in the ribosomal genes along with a species annotation process to identify potentially erroneous entries. This IDNS Centroid Database was prepared using a proprietary algorithm (SmartGene), which determined the single most representative ‘Centroid’ sequence per species. Sequence variability at the intra-species level was assessed. All sequence entries annotated with a valid species by their submitters were labelled with a ‘confidence score’ allowing them to be recognized as species-typical or not. The confidence score for each sequence carrying a valid species name takes into account the sequence similarity to the Centroid sequence and factors such as sequence length and coverage. The most representative sequences per species were used to build a non-redundant reference database called the ‘Centroid Database’. As compared with the IDNS Eubacterial Database, the Centroid Database contains only the most representative 16S rRNA gene entry per species and ‘Centroids’ are defined only for taxonomically valid species, not subspecies.

**RESULTS AND DISCUSSION**

The reference identification showed that the 377 bacterial species included in the dataset displayed variable biochemical backgrounds, ribosomal protein contents and 16S rDNA gene sequences (results not shown). Therefore, the agreement level between the methods used can be close to but never equal to 100 %. The 16S rDNA sequencing provided a single-choice species identification for almost 80 % of all cases (see Table 1). Correct identification to the genus level was observed for all remaining isolates. The 16S rDNA failed to distinguish between some closely related species or genera (e.g. *Bacillus* spp., *Escherichia coli* versus *Shigella*) (Dahllof et al., 2000; Clarridge, 2004; Rajendhran & Gunasekaran, 2011).
The comparison of the two classification approaches with respect to precision of identification is presented in Fig. 2. Compared with the reference identification, straightforward 16S rDNA sequencing provided the correct genus identification but failed to reach the species level for 166 out of 1025 (16.1%) samples. The number of samples identified solely at the genus level was slightly higher (192 out of the 1025 isolates) using the two alternative approaches (Fig. 2). The concordant or discrepant results obtained using the various classification approaches are illustrated in Fig. 3.

A total of 728 out of 1025 (71%) isolates gave the same identification results based on reference identification, RDB-LCA analysis and Centroid Database analysis. Among those isolates, 644 displayed reference identification at the species level and 84 strains at the subspecies level.

A small subset of strains (93 out of 1025) was concordantly identified with the reference identification and the Centroid Database analysis, but not with RDB-LCA analysis (Table S1, available in the online Supplementary Material). These differences covered a wide range of species, highlighting the need to have a comprehensive 16S rDNA sequence database to cover difficult and genetically closely related species. Identification calls matching between reference identification and results obtained with RDB-LCA analysis, but discordant with the 16S Centroid Database analysis, were also observed (82 out of the 1025 sequences) (Table S2). In order to clarify non-conclusive results, it is recommended to query the sequence against the Eubacterial Database, which is an integral part of the SmartGene Bacteria Module. This Eubacterial Database contains >500,000 entries. Thus, when minor mismatches with a ‘Centroid’ match reflect a sequence variation observed commonly for a particular species, a better match to a sequence variant in the Eubacterial Database can improve the identification (results not shown).

Our results illustrate that the two approaches we compared generated similar results but still displayed significant differences, leading to variable identification calls for

### Table 1. Distribution of samples in the different taxonomic levels according to the different microbial identification methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Above species rank (group, genus, etc.)</th>
<th>Single species</th>
<th>Subspecies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference</td>
<td>0</td>
<td>931 (90.8)</td>
<td>94 (9.2)</td>
</tr>
<tr>
<td>RDB-LCA</td>
<td>216 (21.1)</td>
<td>809 (78.9)</td>
<td>0</td>
</tr>
<tr>
<td>Centroid Database</td>
<td>206 (20.1)</td>
<td>819 (79.9)</td>
<td>0</td>
</tr>
</tbody>
</table>

![Fig. 2. Agreement for taxonomic levels of identification between the three classification approaches (reference, IDNS Centroid and RDB-LCA). Agreement of taxonomic levels between the reference identification and the IDNS Centroid Database (black bar) or RDB-LCA (white bar). Agreement between the IDNS Centroid Database and RDB-LCA (grey bar).](http://jmm.sgmjournals.org)
including biochemistry, MALDI-TOF MS and targeted sequencing might eventually lead to more accurate microbial identification. In the clinical microbiology laboratory, where MALDI-TOF MS is replacing classical identification at an unprecedented speed, polyphasic identification of bacteria will probably be captured by a combination of proteomic MALDI-TOF MS and gene or genome sequencing approaches very soon.

**ACKNOWLEDGEMENTS**

We thank Thierry Lombardot, an employee at SmartGene at the time of this study, and the bioMérieux Research and Development VITEK MS, DKL and Biomathematic teams for their technical support. S. E. is working for SmartGene GmbH, the company that markets the Centroid solution package, whilst all of the other authors are bioMérieux employees. bioMérieux is a company developing and marketing infectious disease diagnostics.

**REFERENCES**


