Development of an automated 16S rRNA gene sequence database, 16SpathDB, for identification of medically important bacteria

JLL Teng *, PCY Woo, SKP Lau

KEY MESSAGES

1. We developed 16SpathDB, a database for identification of medically important bacteria.
2. 16SpathDB offers efficient and accurate analysis of 16S rRNA gene sequences of medically important bacteria.
3. The updated 16SpathDB version 2.0 is available at http://www.microbiology.hku.hk/16SpathDB and is updated periodically for every new edition of the *Manual of Clinical Microbiology*.

Hong Kong Med J 2018;24(Suppl 6):S40-2
RFCID project number: 11101102

JLL Teng, PCY Woo, SKP Lau

Department of Microbiology, The University of Hong Kong

* Principal applicant and corresponding author: llteng@hku.hk

Introduction

Clinical microbiology and research laboratories regularly receive clinical isolates that require accurate bacterial identification and differentiation, particularly for medically important bacteria. Traditional identification via phenotypic characteristics is time-consuming, and bacteria isolates often cannot be fully characterised owing to poorly defined phenotypes and subjective bias. In the past 10 years, laboratories have relied on 16S rRNA gene sequencing as an identification tool. High-throughput analysis of 16S rRNA sequencing can be performed using a wide range of databases, including GenBank,[1] Ribosomal Database Project,[2] MicroSeq,[3] Ribosomal Differentiation of Medical Microorganisms,[4] and SmartGene Integrated Database Network System.[5] The database sequences of Ribosomal Database Project and SmartGene Integrated Database Network System are derived from GenBank, whereas the database sequences of Ribosomal Differentiation of Medical Microorganisms and MicroSeq are based on the sequencing of 16S rRNA genes of bacterial culture strains. Each database produces a list of top hits according to its own algorithm, and the 'first hit' or 'closest match' can be interpreted as the identity of the bacterial isolate. Nonetheless, accurate interpretation of 16S rRNA gene sequencing results is difficult for inexperienced users. For example, if the database does not contain the bacterial species of interest, the real identity of the isolate cannot be determined. If there is minimal difference between the 16S rRNA gene sequences of multiple bacterial species, it may be difficult to differentiate the identity of the isolate based on the list of top hits. Hence, we aimed to develop a database that allows determination of the most likely identity of medically important bacteria using 16S rRNA gene sequencing. An automated user-friendly platform was generated, and the accuracy of the database in identification of bacteria was evaluated.

Methods

16SpathDB is a web-based database that uses 16S rRNA gene sequences for identification of medically important bacteria. The most representative 16S rRNA gene sequence of each medically important bacterial species listed in the 9th edition of *Manual of Clinical Microbiology* was retrieved from GenBank. Strain sequences with the following criteria were preferred: good phenotypic and/or genotypic characterization, isolated from humans, few undetermined bases, and longer sequence length. More than one 16S rRNA gene sequence was included for bacterial species that had >2% intragenomic difference in their 16S rRNA gene sequences or intervening sequences in their 16S rRNA genes.

To evaluate the usefulness of 16SpathDB, 16S rRNA gene sequences of 250 medically important bacterial isolates, including those of 82 aerobic Gram-positive bacteria (staphylococci, streptococci, enterococci, mycobacteria, corynebacteria, nocardia, and members of *Bacillus*), 82 aerobic Gram-negative bacteria (*Bartonella*, *Bordetella*, *Burkholderia*, *Neisseria*, *Desulfovibrio*, *Campylobacter* and *Helicobacter* species and members of *Aeromonadaceae*, *Enterobacteriaceae*, *Legionellaceae*, *Pasteurellaceae*, *Moraxellaceae*, *Pseudomonadaceae*, and *Vibrionaceae*), 85 anaerobic bacteria (*Actinomyces*, *Clostridium*, *Bacteroides*, *Porphyromonas*, and *Prevotella* species), and one

Mycoplasma hominis that had been archived in our clinical microbiology laboratory in the past 10 years were input to the database for analysis. The exact identities of these isolates were determined by a polyphasic approach using a combination of phenotypic tests and 16S rRNA gene sequencing.

Results

16SpathDB includes 1014 16S rRNA gene sequences from 1010 unique bacterial species. The database interface is clear and simple to use. One can enter the 'query page' by clicking the 'Identify bacteria by 16S rRNA gene sequence' hyperlink. Users can then submit their query by inputting one or more queries for 16S rRNA gene sequences in the textbox or by uploading a file that contains the sequences via the 'Browse' button. Next, after clicking the 'Begin identification' button, the percentage nucleotide identity calculated from the alignment between the query sequence and each of the sequences in 16SpathDB is then displayed on the 'query results' page and can be used to determine the identity of the query sequence. The length of the input sequence affects the sequence percentage identity for species identification. In general, the 5' end of the 16S rRNA gene is more variable than the other parts of the gene and is thus preferred. An example data file is also provided to allow users to familiarise themselves with the system.

The following algorithm was used to report results generated by 16SpathDB. If there is one species in 16SpathDB with >98.0% nucleotide identity with the query sequence, this bacterial species, as well as the percentage nucleotide identity between the query sequence and the sequence of the most likely bacterial species, is reported (category 1). If there is more than one species in 16SpathDB with >98.0% nucleotide identity with the query sequence, the species that shares the highest nucleotide identity with the query sequence ('best match in 16SpathDB') as well as those with 16S rRNA gene sequences having <1% difference with the 'best match in 16SpathDB' are reported, and the user is alerted that further tests, such as biochemical tests or sequencing additional genes, may be necessary for differentiation between the most probable identities (category 2). If there are no species in 16SpathDB with >98.0% nucleotide identity with the query sequence, but there is one or more species in 16SpathDB with >96.0% nucleotide identity with the query sequence, only the genus is reported (category 3). The user is also reminded that further tests are necessary for definite species identification. If there are no species in 16SpathDB with >96.0% nucleotide identity with the query sequence, the results page would state, "No species in 16SpathDB was found to share high nucleotide identity with your query sequence" (category 4). This indicates that the query sequence may represent a bacterial species not included in the *Manual of Clinical Microbiology* or a novel bacterial species. When this occurs, users are advised to perform a BLAST search against the GenBank nr database to differentiate between the two possibilities.

Following the analysis of the submitted 16S rRNA gene sequences, users can inspect the detailed contents of the database as well as the information of individual sequences. This can be performed by clicking the 'Browse bacterial 16S rRNA gene information' hyperlink to enter the 'sequence information' page.

In 16SpathDB, among the 250 medically important bacterial isolates tested, 140 (56%) were reported as a single bacterial species having >98.0% nucleotide identity with the query sequence (category 1), 109 (43.6%) as more than one bacterial species having >98.0% nucleotide identity with the query sequence (category 2), none as genus level matches (category 3), and one (0.4%) as "No species in 16SpathDB was found to share high nucleotide identity with your query sequence" (category 4). For the 140 bacterial isolates reported as a single bacterial species, all results were identical to the true identities of the isolates as determined by the polyphasic approach. For the 109 bacterial isolates reported as more than one bacterial species, all results contained the true identities of the isolates as determined by the polyphasic approach.

Discussion

We developed 16SpathDB for identification of medically important bacteria using 16S rRNA gene sequencing. The platform has a simple user-friendly interface. One advantage of 16SpathDB is that it includes only the 16S rRNA gene sequences of medically important bacteria listed in the *Manual of Clinical Microbiology*, which contains nearly all bacterial strains ever recovered from patients. In contrast, other databases, such as the Ribosomal Database Project and SmartGene Integrated Database Network System, include 16S rRNA gene sequences from other bacterial species that have never been isolated from patients. In the clinical setting, inclusion of non-medically important bacteria could potentially hinder accurate differentiation and identification of clinical isolates. 16SpathDB is also superior to MicroSeq, which does not include an adequate number of medically important bacteria for proper isolate identification via 16S rRNA gene sequencing. Moreover, MicroSeq generates only a single 'identity' result of the query sequence while disregarding other bacterial species with similar 16S rRNA gene sequences. In contrast, 16SpathDB reports the species that shows the highest nucleotide identity to the query sequence (ie, 'best match') as well as those with <1% difference from the 'best match' to alert the user that further tests may be required to differentiate between the probable isolate identities.

16SpathDB is accurate for identification of the 16S rRNA gene sequences of medically important bacteria. 16SpathDB successfully identified all 250 bacterial isolates archived in our clinical microbiology laboratory. Among these 250 bacterial isolates, 43.6% showed multiple possible identities, which reflected an inherent limitation of using 16S rRNA gene sequencing for bacterial identification. Phenotypic tests or sequencing of additional gene loci should be performed to differentiate among the reported bacterial species. It should be noted that this bacterial collection is from a single laboratory in Hong Kong, and thus, complete assessment of the database is not provided.

One limitation of 16SpathDB is that it includes only the sequences of bacterial species listed in the *Manual of Clinical Microbiology*, but in the clinical setting, the exclusion of bacterial species that have never been reported to cause infection is beneficial to data interpretation, as it would markedly minimise the results that have multiple possible identities. The 16S rRNA gene sequence of one isolate was reported as "No species in 16SpathDB was found to share high nucleotide identity with your query sequence". This is because *Gordonibacter pamelaeae* has never been reported to be associated with human disease and is not included in the *Manual of Clinical Microbiology*. Users should perform a BLAST search against the GenBank nr database to identify such 16S rRNA gene sequences. Users should also bear in mind that bacterial species that have never been reported to be associated with infection may still have the potential to do so.

An updated version, 16SpathDB 2.0, is available at http://www.microbiology.hku.hk/16SpathDB. It is updated periodically for every new edition of the *Manual of Clinical Microbiology*.

Acknowledgement

This study was supported by the Research Fund for the Control of Infectious Diseases, Food and Health Bureau, Hong Kong SAR Government (#11101102).

References

1. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res 2008;36:D25-30.
2. Cole JR, Chai B, Farris RJ, et al. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res 2005;33:D294-6.
3. Woo PC, Ng KH, Lau SK, et al. Usefulness of the MicroSeq 500 16S ribosomal DNA-based bacterial identification system for identification of clinically significant bacterial isolates with ambiguous biochemical profiles. J Clin Microbiol 2003;41:1996-2001.
4. Harmsen D, Rothganger J, Frosch M, Albert J. RIDOM: Ribosomal Differentiation of Medical Micro-organisms Database. Nucleic Acids Res 2002;30:416-7.
5. Simmon KE, Croft AC, Petti CA. Application of SmartGene IDNS software to partial 16S rRNA gene sequences for a diverse group of bacteria in a clinical laboratory. J Clin Microbiol 2006;44:4400-6.