JOURNAL OF CLINICAL MICROBIOLOGY, May 2011, p. 1799–1809 Vol. 49, No. 5
0095-1137/11/$12.00 doi:10.1128/JCM.02350-10

Automated Identification of Medically Important Bacteria by 16S rRNA Gene Sequencing Using a Novel Comprehensive Database, 16SpathDB▽

Patrick C. Y. Woo,[1,2,3,4]†* Jade L. L. Teng,[3]† Juilian M. Y. Yeung,[3]† Herman Tse,[1,2,3,4] Susanna K. P. Lau,[1,2,3,4]* and Kwok-Yung Yuen[1,2,3,4]

State Key Laboratory of Emerging Infectious Diseases,[1] Research Centre of Infection and Immunology,[2] Department of Microbiology,[3] and Carol Yu Centre for Infection,[4] The University of Hong Kong, Hong Kong

Received 22 November 2010/Returned for modification 3 January 2011/Accepted 27 February 2011

Despite the increasing use of 16S rRNA gene sequencing, interpretation of 16S rRNA gene sequence results is one of the most difficult problems faced by clinical microbiologists and technicians. To overcome the problems we encountered in the existing databases during 16S rRNA gene sequence interpretation, we built a comprehensive database, 16SpathDB (http://147.8.74.24/16SpathDB) based on the 16S rRNA gene sequences of all medically important bacteria listed in the *Manual of Clinical Microbiology* and evaluated its use for automated identification of these bacteria. Among 91 nonduplicated bacterial isolates collected in our clinical microbiology laboratory, 71 (78%) were reported by 16SpathDB as a single bacterial species having >98.0% nucleotide identity with the query sequence, 19 (20.9%) were reported as more than one bacterial species having >98.0% nucleotide identity with the query sequence, and 1 (1.1%) was reported as no match. For the 71 bacterial isolates reported as a single bacterial species, all results were identical to their true identities as determined by a polyphasic approach. For the 19 bacterial isolates reported as more than one bacterial species, all results contained their true identities as determined by a polyphasic approach and all of them had their true identities as the "best match in 16SpathDB." For the isolate (*Gordonibacter pamelaeae*) reported as no match, the bacterium has never been reported to be associated with human disease and was not included in the *Manual of Clinical Microbiology*. 16SpathDB is an automated, user-friendly, efficient, accurate, and regularly updated database for 16S rRNA gene sequence interpretation in clinical microbiology laboratories.

In the last decade, as a result of the widespread use of PCR and DNA sequencing, 16S rRNA gene sequencing has played a pivotal role in accurate identification of medically important bacteria in both clinical microbiology and research laboratories (26). For the management of individual patients, accurate and objective identification of clinical isolates has assisted clinicians in the choice and duration of antibiotic therapy, as well as infection control measures. On the population scale, accurate identification has greatly improved our understanding of the epidemiology and, hence, the empirical treatment of infectious disease syndromes. In the past 10 years, our group and others have used this technology for the identification of a large number of medically important bacteria, resulting in major impacts on infectious diseases.

Interpretation of 16S rRNA gene sequence results is one of the most difficult problems faced by inexperienced clinical microbiologists and technical staff, despite the wide range of software and databases available. The best-known software and databases include GenBank (1), the Ribosomal Database Project (RDP-II) (2, 3, 13), MicroSeq (15, 17), Ribosomal Differentiation of Medical Microorganisms (RIDOM) (4–6), and the SmartGene Integrated Database Network System (SmartGene IDNS) (16). The databases of RDP-II and SmartGene IDNS contain sequences downloaded from GenBank, whereas all sequences in the databases of RIDOM and MicroSeq were obtained by sequencing the 16S rRNA genes of bacterial strains from culture collections. Due to the large number of unvalidated 16S rRNA gene sequences in GenBank, it is often not easy for inexperienced users to decide whether the "first hit" or the "closest match" is the real identity of a bacterial isolate. As for the other software and databases, the usefulness is further limited by the choice of bacterial species in the database. If a bacterial species is not included in the database, it would never be the identity of an isolate. If the database includes bacterial species with minimal differences in their 16S rRNA gene sequences, and hence, that cannot be identified with confidence by 16S rRNA gene sequencing, they may also give rise to wrong identification if the software reports only that the "first hit" or "closest match" is the identity of the bacterium.

In view of these problems, in 2005 we started developing our own database, which includes the most representative 16S rRNA gene sequences of all medically important bacteria listed in the most current edition of the *Manual of Clinical Microbiology* (14), for identification of medically

* Corresponding author. Mailing address: Department of Microbiology, The University of Hong Kong, University Pathology Building, Queen Mary Hospital, Hong Kong. Phone: (852) 22554892. Fax: (852) 28551241. E-mail for P. C. Y. Woo: pcywoo@hkucc.hku.hk. E-mail for S. K. P. Lau: skplau@hkucc.hku.hk.
† P.C.Y.W., J.L.L.T., and J.M.Y.Y. contributed equally to the study.
▽ Published ahead of print on 9 March 2011.

TABLE 1. Identification of clinical bacterial isolates using 16SpathDB

Strain no.	Identification by polyphasic approach	Reference	Identification results reported by 16SpathDB		Best match in 16SpathDB	Nucleotide identity (%)	Second-best match in 16SpathDB	Nucleotide identity (%)	Third-best match in 16SpathDB	Nucleotide identity (%)
			Bacterial species	Nucleotide identity (%)						
1	*Abiotrophia defectiva*	21	*Abiotrophia defectiva*	99.54	*Abiotrophia defectiva*	99.54	*Granulicatella adiacens*	92.77	*Granulicatella elegans*	92.67
2	*Actinobaculum schaalii*		*Actinobaculum schaalii*	99.35	*Actinobaculum schaalii*	99.35	*Actinobaculum massiliense*	95.69	*Actinobaculum suis*	94.52
3	*Actinobaculum urinale*		*Actinobaculum urinale*	100.00	*Actinobaculum urinale*	100.00	*Actinobaculum massiliense*	91.57	*Actinobaculum schaalii*	91.01
4	*Actinomyces meyeri*		*Actinomyces meyeri*	100.00	*Actinomyces meyeri*	100.00	*Actinomyces odontolyticus*	97.59	*Actinomyces georgiae*	97.16
5	*Actinomyces odontolyticus*	22	*Actinomyces odontolyticus*	98.92	*Actinomyces odontolyticus*	98.92	*Actinomyces meyeri*	97.78	*Actinomyces turicensis*	96.07
6	*Actinomyces urogenitalis*		*Actinomyces urogenitalis*	100.00	*Actinomyces urogenitalis*	100.00	*Actinomyces slackii*	97.37	*Actinomyces radicidentis*	97.24
7	*Actionmyces turicensis*		*Actinomyces turicensis*	99.44	*Actinomyces turicensis*	99.44	*Actinomyces odontolyticus*	97.04	*Actinomyces meyeri*	96.65
8	*Aggregatibacter actinomycetemcomitans*		*Aggregatibacter actinomycetemcomitans*	98.62	*Aggregatibacter actinomycetemcomitans*	98.62	*Aggregatibacter aphrophilus*	94.12	*Haemophilus aegyptius*	93.73
9	*Aggregatibacter aphrophilus*	10	*Aggregatibacter aphrophilus*	99.59	*Aggregatibacter aphrophilus*	99.59	*Aggregatibacter segnis*	97.29	*Haemophilus parahaemolyticus*	95.31
10	*Aggregatibacter segnis*	10	*Aggregatibacter segnis*	99.59	*Aggregatibacter segnis*	99.59	*Aggregatibacter aphrophilus*	97.54	*Haemophilus influenzae*	95.73
11	*Agrobacterium tumefaciens*		*Agrobacterium tumefaciens*	98.86	*Agrobacterium tumefaciens*	98.86	*Ochrobactrum anthropi*	94.25	*Brucella melitensis*	93.95
12	*Arcanobacterium haemolyticum*		*Arcanobacterium haemolyticum*	99.80	*Arcanobacterium haemolyticum*	99.80	*Arcanobacterium phocae*	97.27	*Arcanobacterium pluranimalium*	94.39
13	*Arcobacter butzleri*		*Arcobacter butzleri*	99.64	*Arcobacter butzleri*	99.64	*Arcobacter cryaerophilus*	96.76	*Arcobacter skirrowii*	96.61
14	*Arcobacter cryaerophilus*	7	*Arcobacter cryaerophilus*	100.00	*Arcobacter cryaerophilus*	100.00	*Arcobacter skirrowii*	98.42	*Arcobacter butzleri*	97.18
15	*Averyella dalhousiensis*		*Averyella dalhousiensis* *Enterobacter cancerogenus* *Enterobacter aerogenes*	99.18 99.02 99.02	*Averyella dalhousiensis*	99.18	*Enterobacter cancerogenus*	99.02	*Enterobacter aerogenes*	99.02
16	*Bifidobacterium breve*		*Bifidobacterium breve*	99.68	*Bifidobacterium breve*	99.68	*Bifidobacterium longum*	97.77	*Bifidobacterium bifidum*	95.96
17	*Burkholderia pseudomallei*	35	*Burkholderia pseudomallei* *Burkholderia mallei*	100.00 99.93	*Burkholderia pseudomallei*	100.00	*Burkholderia mallei*	99.93	*Burkholderia thailandensis*	98.99
18	*Campylobacter coli*	7	*Campylobacter coli* *Campylobacter jejuni*	99.80 99.79	*Campylobacter coli*	99.80	*Campylobacter jejuni*	99.79	*Campylobacter lari*	98.33
19	*Campylobacter fetus*	7	*Campylobacter fetus*	99.85	*Campylobacter fetus*	99.85	*Campylobacter hyointestinalis*	97.98	*Campylobacter mucosalis*	96.40
20	*Campylobacter hyointestinalis*		*Campylobacter hyointestinalis* *Campylobacter fetus*	100.00 99.01	*Campylobacter hyointestinalis*	100.00	*Campylobacter fetus*	99.01	*Campylobacter mucosalis*	96.06
21	*Campylobacter rectus*		*Campylobacter rectus*	99.34	*Campylobacter rectus*	99.34	*Campylobacter showae*	97.94	*Campylobacter concisus*	96.11
22	*Capnocytophaga sputigena*		*Capnocytophaga sputigena*	99.24	*Capnocytophaga sputigena*	99.24	*Capnocytophaga ochracea*	95.67	*Capnocytophaga cynodegmi*	92.24
23	*Citrobacter koseri*		*Citrobacter koseri* *Citrobacter farmeri* *Salmonella enterica*	99.92 98.74 98.74	*Citrobacter koseri*	99.92	*Citrobacter farmeri*	98.74	*Salmonella enterica*	98.74
24	*Clostridium barati*	24	*Clostridium baratii* *Eubacterium budayi* *Eubacterium nitritogenes*	99.83 99.41 99.16	*Clostridium baratii*	99.83	*Eubacterium budayi*	99.41	*Eubacterium nitritogenes*	99.16
25	*Clostridium difficile*	24	*Clostridium difficile*	99.51	*Clostridium difficile*	99.51	*Clostridium sordellii*	95.90	*Eubacterium tenue*	95.80
26	*Clostridium disporicum*	24	*Clostridium disporicum*	98.99	*Clostridium disporicum*	98.99	*Clostridium celatum*	97.57	*Clostridium carnis*	96.91
27	*Clostridium hathewayi*	28	*Clostridium hathewayi*	98.28	*Clostridium hathewayi*	98.28	*Clostridium indolis*	94.17	*Clostridium sphenoides*	93.86
28	*Clostridium Innocuum*	24	*Clostridium innocuum*	99.19	*Clostridium innocuum*	99.19	*Holdemania filiformis*	85.33	*Erysipelothrix inopinata*	84.82
29	*Clostridium orbiscindens*	24	*Clostridium orbiscindens*	100.00	*Clostridium orbiscindens*	100.00	*Bacteroides capillosus*	97.44	*Clostridium sporosphaeroides*	86.62
30	*Clostridium paraputrificum*	24	*Clostridium paraputrificum*	99.42	*Clostridium paraputrificum*	99.42	*Clostridium carnis*	96.81	*Clostridium disporicum*	96.53

31	*Clostridium perfringens*		*Clostridium perfringens*	99.92	*Clostridium perfringens*	99.92	*Clostridium baratii*	94.60	*Eubacterium budayi*	94.19
32	*Clostridium ramosum*	24	*Clostridium ramosum*	99.92	*Clostridium ramosum*	99.92	*Clostridium spiroforme*	95.25	*Lactobacillus vitulinus*	86.78
33	*Clostridium sordellii*		*Clostridium sordellii* *Eubacterium tenue*	99.33 98.44	*Clostridium sordellii*	99.33	*Eubacterium tenue*	98.44	*Clostridium ghonii*	98.15
34	*Clostridium symbiosum*		*Clostridium symbiosum*	99.49	*Clostridium symbiosum*	99.49	*Clostridium bolteae*	94.05	*Clostridium clostridioforme*	93.37
35	*Clostridium tertium*	24	*Clostridium tertium*	100.00	*Clostridium tertium*	100.00	*Clostridium carnis*	98.89	*Clostridium chauvoei*	97.90
36	*Clostridium sporosphaeroides*	24	*Clostridium sporosphaeroides*	98.80	*Clostridium sporosphaeroides*	98.80	*Clostridium orbiscindens*	87.17	*Bacteroides capillosus*	85.97
37	*Desulfovibrio desulfuricans*		*Desulfovibrio desulfuricans*	99.07	*Desulfovibrio desulfuricans*	99.07	*Desulfovibrio fairfieldensis*	95.58	*Desulfovibrio piger*	95.07
38	*Desulfovibrio fairfieldensis*		*Desulfovibrio fairfieldensis*	99.92	*Desulfovibrio fairfieldensis*	99.92	*Desulfovibrio desulfuricans*	96.75	*Desulfovibrio piger*	95.83
39	*Dolosigranulum pigrum*		*Dolosigranulum pigrum*	98.97	*Dolosigranulum pigrum*	98.97	*Alloiococcus otitis*	91.40	*Aerococcus sanguinicola*	88.79
40	*Eggerthella hongkongensis*	11	*Eggerthella hongkongensis*	100.00	*Eggerthella hongkongensis*	100.00	*Eggerthella lenta*	93.08	*Slackia exigua*	90.09
41	*Eggerthella lenta*	9	*Eggerthella lenta*	99.74	*Eggerthella lenta*	99.74	*Slackia exigua*	89.48	*Slackia heliotrinireducens*	89.10
42	*Enterobacter hormaechei*		*Enterobacter hormaechei* *Enterobacter asburiae* *Enterobacter ludwigii*	98.60 98.45 98.22	*Enterobacter hormaechei*	98.60	*Enterobacter asburiae*	98.45	*Enterobacter ludwigii*	98.22
43	*Enterococcus cecorum*	29	*Enterococcus cecorum*	98.72	*Enterococcus cecorum*	98.72	*Enterococcus columbae*	96.50	*Enterococcus casseliflavus*	96.06
44	*Enterococcus raffinosus*		*Enterococcus raffinosus* *Enterococcus gilvus* *Enterococcus malodoratus*	99.93 99.39 99.25	*Enterococcus raffinosus*	99.93	*Enterococcus gilvus*	99.39	*Enterococcus malodoratus*	99.25
45	*Facklamia hominis*		*Facklamia hominis*	99.38	*Facklamia hominis*	99.38	*Facklamia languida*	96.82	*Facklamia ignava*	93.16
46	*Fusobacterium necrophorum*		*Fusobacterium necrophorum*	99.78	*Fusobacterium necrophorum*	99.78	*Fusobacterium gonidiaformans*	97.65	*Fusobacterium russii*	93.95
47	*Gemella haemolysans*	7	*Gemella haemolysans*	99.62	*Gemella haemolysans*	99.62	*Gemella sanguinis*	97.67	*Gemella morbillorum*	97.15
48	*Gemella morbillorum*		*Gemella morbillorum*	99.68	*Gemella morbillorum*	99.68	*Gemella haemolysans*	98.30	*Gemella sanguinis*	97.50
49	*Gordonia terrae*		*Gordonia terrae*	100.00	*Gordonia terrae*	100.00	*Gordonia polyisoprenivorans*	98.31	*Gordonia bronchialis*	98.30
50	*Gordonibacter pamelaeae*	30	No match	-	*Eggerthella sinensis*	94.00	*Eggerthella hongkongensis*	93.53	*Eggerthella lenta*	91.74
51	*Granulicatella adiacens*	21	*Granulicatella adiacens*	99.77	*Granulicatella adiacens*	99.77	*Granulicatella elegans*	97.40	*Enterococcus villorum*	94.58
52	*Haemophilus parainfluenzae*	10	*Haemophilus parainfluenzae*	99.53	*Haemophilus parainfluenzae*	99.53	*Haemophilus pittmaniae*	96.51	*Haemophilus haemolyticus*	95.33
53	*Helcococcus kunzii*	34	*Helcococcus kunzii*	99.59	*Helcococcus kunzii*	99.59	*Parvimonas micra*	84.57	*Anaerococcus hydrogenalis*	83.62
54	*Kytococcus sedentarius*		*Kytococcus sedentarius*	98.12	*Kytococcus sedentarius*	98.12	*Dermacoccus nishinomiyaensis*	94.48	*Janibacter sanguinis*	93.24
55	*Lactobacillus casei*	7	*Lactobacillus casei* *Lactobacillus rhamnosus*	100.00 99.51	*Lactobacillus casei*	100.00	*Lactobacillus rhamnosus*	99.51	*Pediococcus pentosaceus*	92.94
56	*Lactobacillus fermentum*	25	*Lactobacillus fermentum*	99.77	*Lactobacillus fermentum*	99.77	*Pediococcus acidilactici*	93.6	*Pediococcus pentosaceus*	93.5
57	*Lactobacillus gasseri*		*Lactobacillus gasseri*	99.92	*Lactobacillus gasseri*	99.92	*Lactobacillus acidophilus*	93.94	*Lactobacillus plantarum*	90.32
58	*Lactobacillus rhamnosus*	7	*Lactobacillus rhamnosus* *Lactobacillus casei*	99.83 99.66	*Lactobacillus rhamnosus*	99.83	*Lactobacillus casei*	99.66	*Pediococcus pentosaceus*	93.31
59	*Lactococcus garvieae*		*Lactococcus garvieae*	99.92	*Lactococcus garvieae*	99.92	*Lactococcus lactis*	92.07	*Streptococcus iniae*	89.32
60	*Lactococcus lactis*		*Lactococcus lactis*	99.52	*Lactococcus lactis*	99.52	*Lactococcus garvieae*	92.75	*Streptococcus agalactiae*	89.34
61	*Laribacter hongkongensis*	27	*Laribacter hongkongensis*	99.93	*Laribacter hongkongensis*	99.93	*Chromobacterium violaceum*	92.39	*Eikenella corrodens*	89.82
62	*Leptotrichia buccalis*		*Leptotrichia buccalis*	99.56	*Leptotrichia buccalis*	99.56	*Streptobacillus moniliformis*	86.68	*Sneathia sanguinegens*	83.96
63	*Microbacterium paraoxydans*		*Microbacterium paraoxydans* *Microbacterium oxydans*	99.49 98.87	*Microbacterium paraoxydans*	99.49	*Microbacterium oxydans*	98.87	*Microbacterium resistens*	98.25

Continued on following page

TABLE 1—*Continued*

Strain no.	Identification by polyphasic approach	Reference	Identification results reported by 16SpathDB		Best match in 16SpathDB	Nucleotide identity (%)	Second-best match in 16SpathDB	Nucleotide identity (%)	Third-best match in 16SpathDB	Nucleotide identity (%)
			Bacterial species	Nucleotide identity (%)						
64	*Micrococcus luteus*	7	*Micrococcus luteus* *Micrococcus lylae* *Micrococcus antarcticus*	98.11 97.78 97.32	*Micrococcus luteus*	98.11	*Micrococcus lylae*	97.78	*Micrococcus antarcticus*	97.32
65	*Moraxella osloensis*		*Moraxella osloensis*	99.53	*Moraxella osloensis*	99.53	*Moraxella lincolnii*	92.49	*Moraxella lacunata*	92.41
66	*Mycobacterium chelonae*	7	*Mycobacterium chelonae* *Mycobacterium abscessus* *Mycobacterium immunogenum*	99.79 99.79 98.97	*Mycobacterium chelonae*	99.79	*Mycobacterium abscessus*	99.79	*Mycobacterium immunogenum*	98.97
67	*Mycobacterium marinum*		*Mycobacterium marinum* *Mycobacterium ulcerans*	100.00 100.00	*Mycobacterium marinum*	100.00	*Mycobacterium ulcerans*	100.00	*Mycobacterium asiaticum*	98.71
68	*Mycobacterium nonchromogenicum*	7	*Mycobacterium nonchromogenicum*	98.13	*Mycobacterium nonchromogenicum*	98.13	*Mycobacterium terrae*	96.69	*Mycobacterium cookii*	95.56
69	*Mycoplasma hominis*	7	*Mycoplasma hominis*	100.00	*Mycoplasma hominis*	100.00	*Brevibacillus parabrevis*	76.75	*Clostridium ramosum*	76.07
70	*Nocardia cyriacigeorgica*		*Nocardia cyriacigeorgica*	100.00	*Nocardia cyriacigeorgica*	100.00	*Nocardia abscessus*	98.73	*Nocardia paucivorans*	98.33
71	*Olsenella uli*	7	*Olsenella uli*	100.00	*Olsenella uli*	100.00	*Atopobium vaginae*	92.15	*Atopobium parvulum*	91.97
72	*Pantoea dispersa*		*Pantoea dispersa*	99.92	*Pantoea dispersa*	99.92	*Enterobacter cancerogenus*	97.49	*Kluyvera cryocrescens*	97.49
73	*Propionibacterium acnes*	7	*Propionibacterium acnes*	99.80	*Propionibacterium acnes*	99.80	*Propionibacterium propionicum*	93.46	*Propionibacterium avidum*	93.27
74	*Propionibacterium avidum*		*Propionibacterium avidum* *Propionibacterium propionicum*	99.45 99.30	*Propionibacterium avidum*	99.45	*Propionibacterium propionicum*	99.30	*Propionibacterium acnes*	96.10
75	*Providencia stuartii*		*Providencia stuartii*	99.77	*Providencia stuartii*	99.77	*Providencia rettgeri*	98.75	*Providencia rustigianii*	98.13
76	*Pseudomonas oryzihabitans*		*Pseudomonas oryzihabitans*	99.79	*Pseudomonas oryzihabitans*	99.79	*Pseudomonas pseudoalcaligenes*	96.84	*Pseudomonas aeruginosa*	96.21
77	*Salmonella enterica*	23	*Salmonella enterica*	99.93	*Salmonella enterica*	99.93	*Citrobacter farmeri*	98.47	*Enterobacter cloacae*	98.46
78	*Shewanella algae*		*Shewanella algae*	99.92	*Shewanella algae*	99.92	*Shewanella putrefaciens*	95.02	*Aeromonas hydrophila*	90.90
79	*Solobacterium moorei*	8	*Solobacterium moorei*	98.33	*Solobacterium moorei*	98.33	*Bulleidia extructa*	93.20	*Holdemania filiformis*	88.86
80	*Staphylococcus aureus*	7	*Staphylococcus aureus*	99.63	*Staphylococcus aureus*	99.63	*Staphylococcus epidermidis*	97.66	*Staphylococcus caprae*	97.64
81	*Staphylococcus epidermidis*	7	*Staphylococcus epidermidis* *Staphylococcus caprae* *Staphylococcus capitis*	100.00 99.21 99.02	*Staphylococcus epidermidis*	100.00	*Staphylococcus caprae*	99.21	*Staphylococcus capitis*	99.02
82	*Staphylococcus lugdunensis*	18	*Staphylococcus lugdunensis* *Staphylococcus haemolyticus*	99.93 99.05	*Staphylococcus lugdunensis*	99.93	*Staphylococcus haemolyticus*	99.05	*Staphylococcus hominis*	98.84
83	*Streptococcus anginosus*	33	*Streptococcus anginosus*	99.77	*Streptococcus anginosus*	99.77	*Streptococcus intermedius*	97.37	*Streptococcus constellatus*	96.62
84	*Streptococcus dysgalactiae*	31	*Streptococcus dysgalactiae*	99.62	*Streptococcus dysgalactiae*	99.62	*Streptococcus iniae*	97.47	*Streptococcus agalactiae*	97.33
85	*Streptococcus iniae*	7	*Streptococcus iniae*	99.62	*Streptococcus iniae*	99.62	*Streptococcus porcinus*	95.21	*Streptococcus dysgalactiae*	95.11
86	*Streptococcus porcinus*		*Streptococcus porcinus*	99.76	*Streptococcus porcinus*	99.76	*Streptococcus agalactiae*	96.79	*Streptococcus iniae*	96.79
87	*Streptococcus pyogenes*	12	*Streptococcus pyogenes*	100.00	*Streptococcus pyogenes*	100.00	*Streptococcus canis*	98.24	*Streptococcus agalactiae*	96.93
88	*Tsukamurella pulmonis*	20	*Tsukamurella pulmonis* *Tsukamurella tyrosinosolvens* *Tsukamurella strandjordii* *Tsukamurella inchonensis* *Tsukamurella paurometabola*	99.13 99.05 98.90 98.74 98.66	*Tsukamurella pulmonis*	99.13	*Tsukamurella tyrosinosolvens*	99.05	*Tsukamurella strandjordii*	98.90
89	*Tsukamurella tyrosinosolvens*	20	*Tsukamurella tyrosinosolvens* *Tsukamurella pulmonis* *Tsukamurella strandjordii* *Tsukamurella inchonensis* *Tsukamurella paurometabola*	100.00 99.69 99.69 99.39 99.24	*Tsukamurella tyrosinosolvens*	100.00	*Tsukamurella pulmonis*	99.69	*Tsukamurella strandjordii*	99.69
90	*Veillonella atypica*		*Veillonella atypica* *Veillonella dispar*	98.95 98.42	*Veillonella atypica*	98.95	*Veillonella dispar*	98.42	*Veillonella parvula*	97.71
91	*Vibrio furnissii*		*Vibrio furnissii* *Vibrio fluvialis*	99.62 98.85	*Vibrio furnissii*	99.62	*Vibrio fluvialis*	98.85	*Vibrio vulnificus*	97.02

important bacteria using 16S rRNA gene sequencing. In this database, 16SpathDB, we sought to create an automated user-friendly platform that indicated the most likely identity of the 16S rRNA gene sequence of a medically important bacterium, as well as other medically important bacteria with similar 16S rRNA gene sequences that may be alternative identities, which the user should be aware of. In this article, we describe this comprehensive database of 16S rRNA gene sequences of medically important bacteria and its use for automated identification of these bacteria.

MATERIALS AND METHODS

Database design. 16SpathDB is a Web-based 16S rRNA gene sequence database for identification of medically important bacteria. MySQL was employed as the database back end to store the 16S rRNA gene sequence information, and PHP was used to generate HTML web pages for the user interface. The most representative 16S rRNA gene sequence of each medically important bacterial species listed in the most current edition of the *Manual of Clinical Microbiology* (14) was retrieved from GenBank by manual inspection according to the following criteria. First, strains with good phenotypic and/or genotypic characterization (e.g., type strains and strains with complete genomes sequenced) were preferred. Second, strains isolated from humans were preferred. Third, sequences with fewer undetermined bases were preferred. Fourth, longer sequences, especially those with better coverage of the 5′ end, were preferred. For bacterial species with >2% intragenomic difference in their 16S rRNA gene sequences and those with intervening sequences in their 16S rRNA genes, more than one 16S rRNA gene sequence for each species was included.

Identification of clinical bacterial isolates using 16SpathDB. To evaluate the usefulness of 16SpathDB, the 16S rRNA gene sequences of 91 nonduplicated medically important bacterial isolates we collected in our clinical microbiology laboratory in the past 10 years were input to the database for analysis (Table 1) (7–12, 18, 20–25, 27–31, 33–35). The exact identities of these isolates were determined by a polyphasic approach using a combination of phenotypic tests and 16S rRNA gene sequencing as described in our publications.

Availability and update of 16SpathDB. 16SpathDB is available at no charge at http://147.8.74.24/16SpathDB and will be updated periodically for every new edition of the *Manual of Clinical Microbiology*.

RESULTS

16SpathDB and functionality of the database. As of November 2010, 16SpathDB contained 1,014 16S rRNA gene sequences from 1,010 unique bacterial species.

Identification of medically important bacteria. The main goal for setting up 16SpathDB was to provide a convenient and efficient platform for identification of medically important bacterial isolates using their 16S rRNA gene sequences. The interfaces of the database are simple and user friendly. From the home page, one can enter the "query page" by clicking the "identify bacteria by 16S rRNA gene sequence" hyperlink (Fig. 1a). From this page, users can input one or more query 16S rRNA gene sequences of 50 bp to 1,800 bp by pasting the sequences in FASTA format in the textbox or uploading a file that contains the sequences in FASTA format by clicking the "browse" button. By clicking the "begin identification" button, the query sequence(s) will be aligned with each of the 16S rRNA gene sequences in 16SpathDB using the pairwise global alignment algorithm with free end gap penalty. The percent nucleotide identity calculated from the alignment between the query sequence and each of the sequences in 16SpathDB is then used for the evaluation of the identity of the query sequence, using the algorithm depicted in Fig. 2.

The results of the comparison will be shown on the results page. If there is one species in 16SpathDB with >98.0% nucleotide identity to the query sequence, this bacterial species, as well as the percent nucleotide identity between the query sequence and the sequence of the most likely bacterial species, will be reported (category 1) (Fig. 1b and 2). If there is more than one species in 16SpathDB with >98.0% nucleotide identity to the query sequence, the species that showed the highest nucleotide identity to the query sequence ("best match in 16SpathDB"), as well as those with 16S rRNA gene sequences having <1% difference from the "best match in 16SpathDB," will be reported to alert the user that further tests, such as biochemical tests or sequencing additional genes, may be necessary to distinguish between the most probable identities (category 2) (Fig. 1c and 2). If there are no species in 16SpathDB with >98.0% nucleotide identity to the query sequence but there are one or more species in 16SpathDB with >96.0% nucleotide identity to the query sequence, only the genus will be reported (category 3) (Fig. 1d and 2). The user will also be reminded that further tests are necessary for definite species identification. If there is no species in 16SpathDB with >96.0% nucleotide identity to the query sequence, the results page will show "no species in 16SpathDB was found to be sharing high nucleotide identity to your query sequence" (category 4) (Fig. 1e and 2). This indicates that the query sequence may represent a bacterial species not included in the *Manual of Clinical Microbiology* or a novel bacterial species. Users are advised to perform a BLAST search against the GenBank nr database to differentiate between the two possibilities. By clicking the "run BLAST" hyperlink, users can enter the "query page" of the NCBI BLAST website. The cutoffs of 98.0% for reporting the identity at the species level and 96.0% for reporting the identity at the genus level were found to be the optimal cutoffs by using 400 16S rRNA gene sequences of bacteria isolated from patients found in the GenBank database as the query sequences (data not shown).

From the results page, users can click the bacterial species name and go to the page in GenBank that contains detailed information on the representative 16S rRNA gene sequence selected for that bacterial species in 16SpathDB. From the results page, users can also click "show top 10 matches" to show the 10 most closely matched sequences from 16SpathDB compared to the query sequence (Fig. 1f).

Browsing 16S rRNA gene sequences of medically important bacteria. In addition to identification of 16S rRNA gene sequences, users can also inspect the detailed contents of the database, as well as the information on the individual sequences. From the home page, one can enter the "sequence information" page by clicking the "browse bacteria 16S rRNA gene information" hyperlink (Fig. 3). From this page, the user can go to the page in GenBank that contains the detailed information on the 16S rRNA gene sequence selected for the corresponding bacterial species by entering the name of the bacterial species in the text box and click "view sequence information" and then the GenBank accession number. In addition, users can visualize the 10 bacterial species with 16S rRNA gene sequences having the highest nucleotide identities with the input bacterial species and

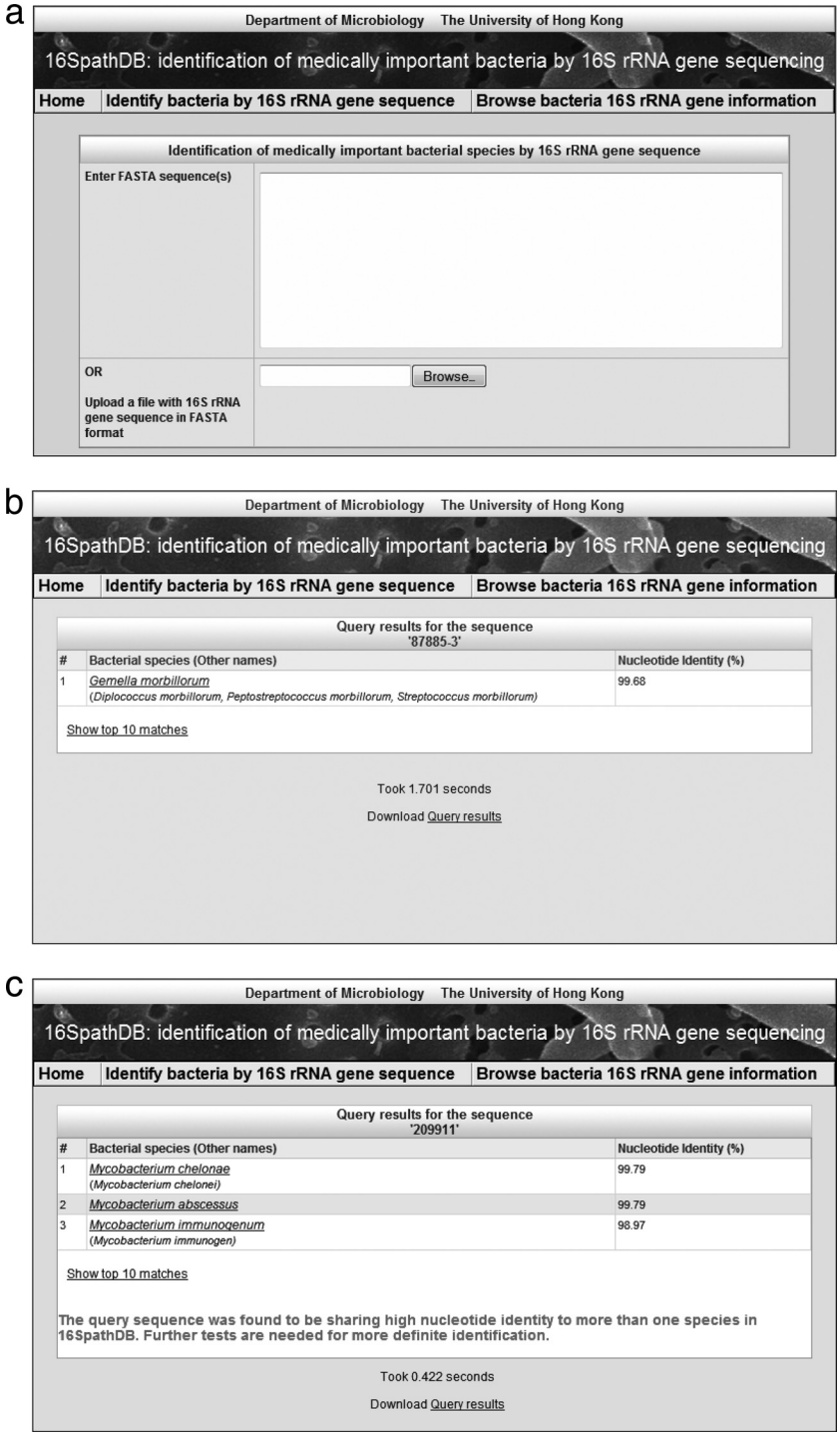
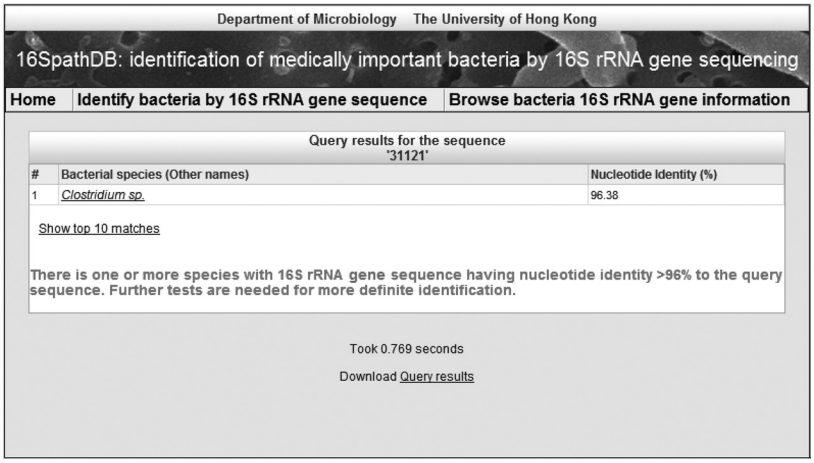
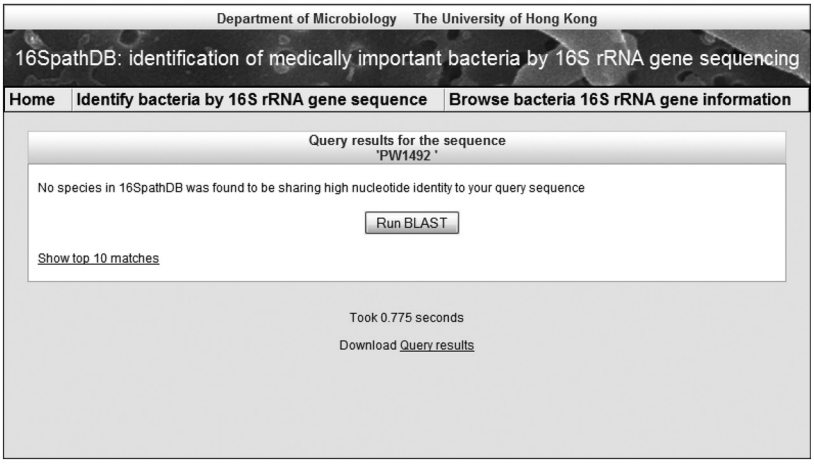


FIG. 1. (a) "Query page" of 16SpathDB. (b) Example of query results showing one species (*Gemella morbillorum*) in 16SpathDB with >98.0% nucleotide identity to the query sequence. (c) Example of query results showing more than one species (*Mycobacterium chelonae*, *Mycobacterium abscessus*, and *Mycobacterium immunogenum*) in 16SpathDB with >98.0% nucleotide identity to the query sequence. (d) Example of query results showing no species in 16SpathDB with >98.0% nucleotide identity to the query sequence but one or more species in 16SpathDB with >96.0% nucleotide identity to the query sequence. (e) Example of query results showing no species in 16SpathDB with >96.0% nucleotide identity to the query sequence. (f) Example of query results showing the identity of the query sequence (*Clostridium paraputrificum*) and the 10 most closely matched sequences from 16SpathDB compared to the query sequence.

d



e



f

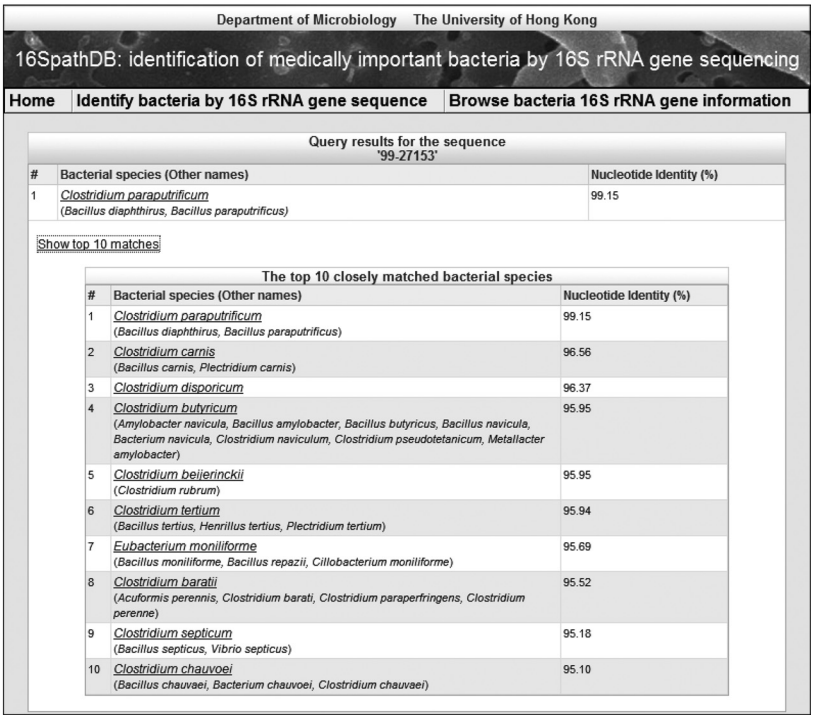


FIG. 1—*Continued.*

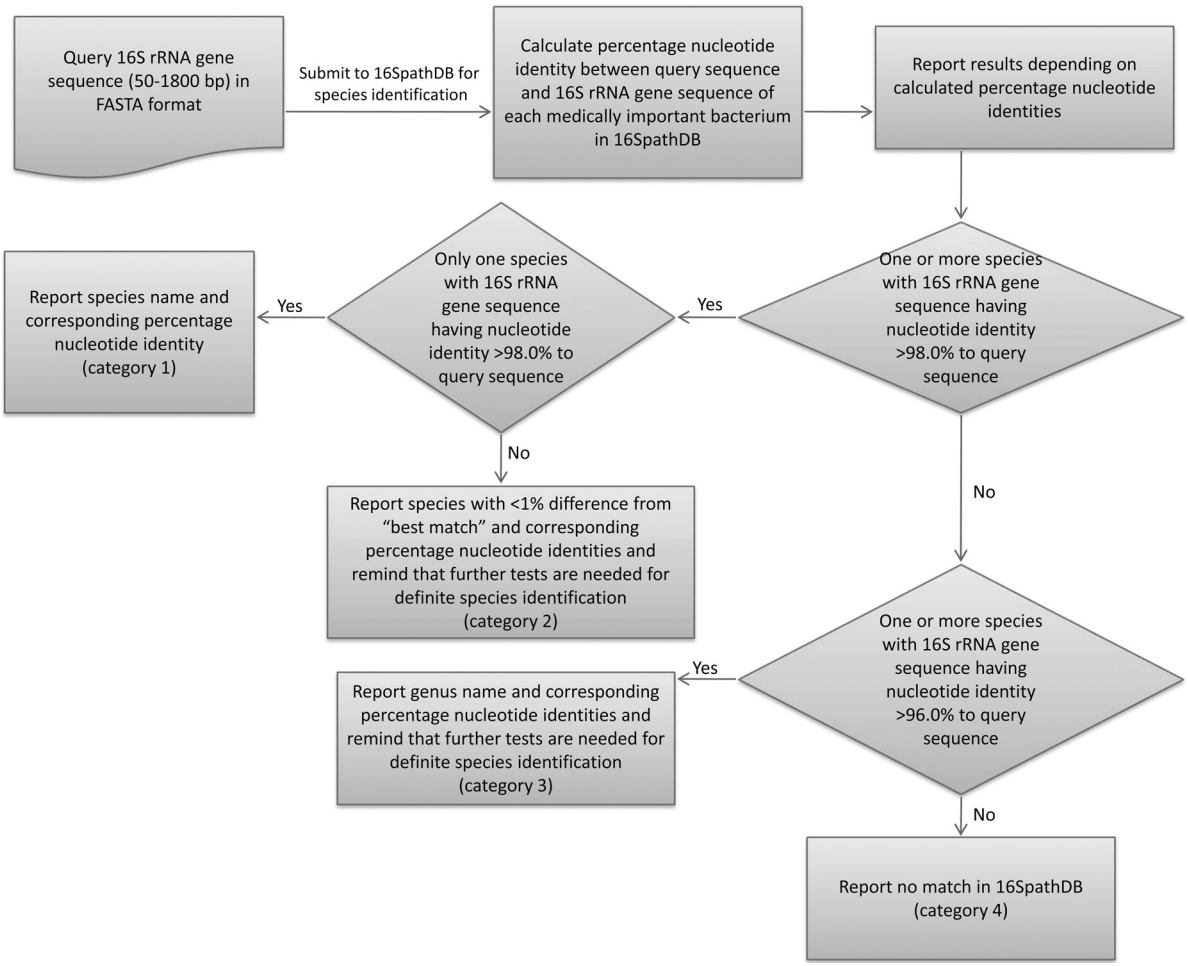


FIG. 2. Algorithm for reporting results in 16SpathDB.

inspect the detailed information on their 16S rRNA gene sequences in GenBank accordingly (Fig. 3).

Identification of clinical bacterial isolates using 16SpathDB. Among the 91 nonduplicated medically important bacterial isolates we collected in our clinical microbiology laboratory in the past 10 years (Table 1), 71 (78%) were reported by 16SpathDB as a single bacterial species having >98.0% nucleotide identity with the query sequence (category 1), 19 (20.9%) were reported by 16SpathDB as more than one bacterial species having >98.0% nucleotide identity with the query sequence (category 2), none was reported by 16SpathDB to the genus level (category 3), and 1 (1.1%) was reported by 16SpathDB as "no species in 16SpathDB was found to be sharing high nucleotide identity to your query sequence" (category 4). For the 71 bacterial isolates reported by 16SpathDB as a single bacterial species, all results were identical to the true identities of the isolates as determined by the polyphasic approach. For the 19 bacterial isolates reported by 16SpathDB as more than one bacterial species, all results contained the true identities of the isolates as determined by the polyphasic approach. In fact, all 19 of these isolates had their true identities as the "best match in 16SpathDB."

DISCUSSION

Rapid and accurate interpretation of 16S rRNA gene sequence results is the cornerstone for identification of medically important bacteria by 16S rRNA gene sequencing. During the process of using 16S rRNA gene sequencing for identification of a large number of medically important bacteria in the past 5 years, we have developed an automated, user-friendly, and comprehensive database, 16SpathDB, of the most representative 16S rRNA gene sequences of all medically important bacteria listed in the most current edition of the *Manual of Clinical Microbiology* (14). In contrast to RDP-II and SmartGene IDNS software (Table 2), 16SpathDB includes only 16S rRNA gene sequences of medically important bacteria. Since more than 99.9% of the bacterial strains recovered from patients were included in the *Manual of Clinical Microbiology* (unpublished data), the inclusion of other bacterial species that have never been isolated from patients would create ambiguity during data interpretation, as the target users of 16SpathDB are technicians and clinical microbiologists who work on 16S rRNA gene sequencing for clinical isolates. As for the MicroSeq databases (Table 2), one of their main drawbacks for iden-

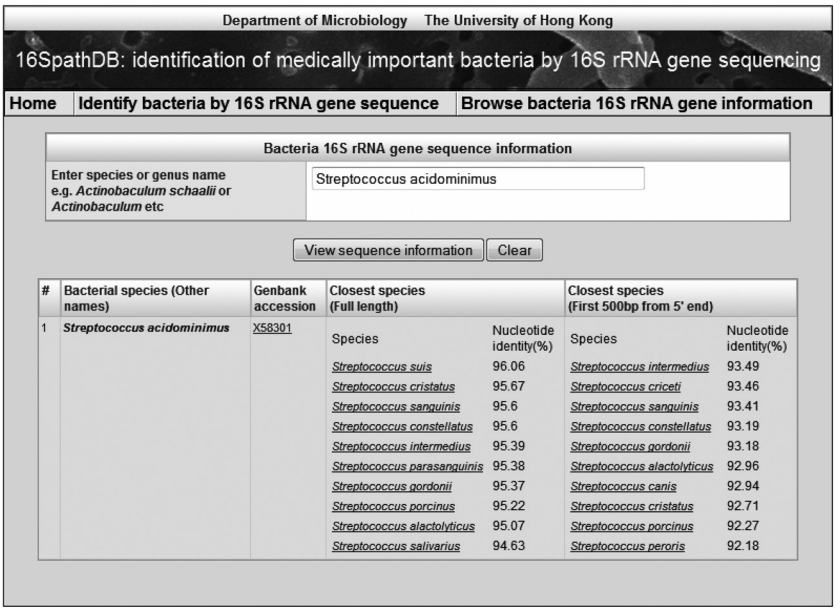


FIG. 3. Example showing the 10 bacterial species with 16S rRNA gene sequences having the highest nucleotide identities to the input bacterial species (*Streptococcus acidominimus*).

tification of medically important bacteria is that the databases do not include a significant number of medically important bacteria that 16S rRNA gene sequencing is able to identify. For example, 98 to 108 (53.3 to 67.1%), 38 to 39 (22.7 to 37.3%), and 23 to 39 (19.8 to 41.9%) medically important anaerobic, aerobic Gram-positive, and aerobic Gram-negative bacteria that should be confidently identified by 16S rRNA gene sequencing are not included in the full- and 500-MicroSeq databases, respectively (19, 32). If the query sequence is from one of these bacteria, the MicroSeq databases would automatically give wrong results. Furthermore, the results from the MicroSeq databases show only a single "identity" of the query sequence. Other bacterial species with similar 16S rRNA gene sequences, which may also be the identity of the isolate, are ignored. For example, it is well known that 16S rRNA gene sequencing is not useful for distinguishing some bacterial species, such as *Streptococcus pneumoniae*, *Streptococcus pseudopneumoniae*, *Streptococcus mitis*, and *Streptococcus oralis*, as their 16S rRNA gene sequences show more than 99% identity. In 16SpathDB, in addition to the species that shows the highest nucleotide identity to the query sequence, those species with 16S rRNA gene sequences having less than 1% difference from the species that showed the highest nucleotide identity to the query sequence will also be reported (Fig. 1c). This helps to alert the user that further tests have to be carried out in order to distinguish between these probable identities.

16SpathDB offers efficient and accurate analysis of 16S rRNA gene sequences from medically important bacteria. In the present study, all 91 bacteria recovered in our clinical microbiology laboratory in the past 10 years were successfully identified using 16SpathDB (Table 1). These 91 isolates included 55 aerobic and 36 anaerobic bacteria, and 62 Gram-positive bacteria, 28 Gram-negative bacteria, and one *Mycoplasma hominis* isolate. Among these 91 bacteria, 20.9% showed multiple possible identities, which reflected an inherent limitation of using 16S rRNA gene sequencing for bacterial identification. For this 20.9% of the strains, phenotypic tests or sequencing of additional gene loci should be performed to differentiate among the reported bacterial species. For example, when *Tsukamurella pulmonis* or *Tsukamurella tyrosinosolvens* (strain no. 88 and 89) (Table 1) were the exact identities of the query isolates, 16S rRNA gene sequencing was not able to distinguish them from the other medically important *Tsukamurella* spp. The API 20C AUX, and API 50 CH systems (bioMérieux, Lyon, France) have to be used to distinguish among these *Tsukamurella* spp. (20).

One limitation of 16SpathDB is that our database includes only the bacterial species that are known to be associated with infections, as described in the *Manual of Clinical Microbiology*. This was deliberately done because if those bacterial species that have never been reported to cause infections were also included, as in the RDP-II and SmartGene IDNS software, the proportion of results that showed multiple possible identities would be markedly increased. This would defeat the original purpose of designing the database, which is used for identification of medically important bacteria in clinical microbiology laboratories. For example, the 16S rRNA gene sequence of strain no. 50 (Table 1) was reported as "no species in 16SpathDB was found to be sharing high nucleotide identity to your query sequence." This is because *Gordonibacter pamelaeae* has never been reported to be associated with human disease and was not included in the *Manual of Clinical Microbiology*. In such a situation, users should perform a BLAST search against the GenBank nr database for the identification of the 16S rRNA gene sequence (30). Although the database will be updated regularly whenever there is a new edition of the *Manual of Clinical Microbiology*, users should bear in

TABLE 2. Comparison of 16SpathDB and other commonly used software for bacterial identification using 16S rRNA gene sequencing

Software	Yr of first description	Company/organization	Website	Partial/full 16S rRNA gene sequence included	Database size	Source of sequences	Cost	Quality control	Updates
Ribosomal Database Project (RDP)	1992	Michigan State University, East Lansing, MI	http://rdp.cme.msu.edu/	Partial and full	1,418,497 (release 10.22)	GenBank	No	Partial	Periodically
MicroSeq Microbial Identification System[a]	1998	Applied Biosystems, Foster City, California	http://www.microseq.com	Partial and full[a]	1,834 and 1,261 in the two databases, respectively[a]	Sequence of 16S rRNA gene of one strain from each species	Yes	All type strains from culture collections	Periodically
Ribosomal Differentiation of Medical Microorganisms (RIDOM)	1999	Ridom GmbH, Würzburg, Germany	http://rdna2.ridom.de/	Partial	23[b]	Sequences of 16S rRNA genes of medically relevant bacteria, mainly belonging to the *Neisseriaceae*, *Moraxellaceae*, and genus *Mycobacterium*	No	All strains from culture collections	Periodically
SmartGene IDNS software	2006	SmartGene GmbH, Switzerland	http://www.smartgene.com/modbacteria.html	Partial and full	243,000	GenBank	Yes	Partial	Daily
16SpathDB	2010	Department of Microbiology, University of Hong Kong, Hong Kong	http://147.8.74.24/16SpathDB	Partial and full	1,010	GenBank	No	All sequences manually selected from GenBank	Periodically

[a] Contains two databases, MicroSeq ID 16S rDNA 500 Library v2.2, which contains sequences from the 5′-end 527 bp of 16S rRNA genes, and MicroSeq ID 16S rDNA Full Gene Library v2.0, which contains full 16S rRNA gene sequences.
[b] Number from from http://rdna2.ridom.de/ridom2/servlet/link?page = list (8 Nov 2010).

mind that those bacterial species that have never been reported to be associated with infections may still have the potential to be so associated.

ACKNOWLEDGMENTS

This work is partly supported by the HKSAR Research Fund for the Control of Infectious Diseases of the Health, Welfare and Food Bureau, Research Grants Council Grant, and University Development Fund, The University of Hong Kong.

We thank Kit-Wah Leung and Ami M. Y. Fung for their critical comments on the database.

REFERENCES

1. **Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and D. L. Wheeler.** 2008. GenBank. Nucleic Acids Res. **36:**D25–D30.
2. **Cole, J. R., et al.** 2005. The Ribosomal Database Project (RDP-II): sequences and tools for high-throughput rRNA analysis. Nucleic Acids Res. **33:**D294–D296.
3. **Cole, J. R., et al.** 2009. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic Acids Res. **37:**D141–D145.
4. **Harmsen, D., J. Rothganger, M. Frosch, and J. Albert.** 2002. RIDOM: Ribosomal Differentiation of Medical Micro-organisms Database. Nucleic Acids Res. **30:**416–417.
5. **Harmsen, D., J. Rothganger, C. Singer, J. Albert, and M. Frosch.** 1999. Intuitive hypertext-based molecular identification of micro-organisms. Lancet **353:**291.
6. **Harmsen, D., et al.** 2001. Diagnostics of neisseriaceae and moraxellaceae by ribosomal DNA sequencing: ribosomal differentiation of medical microorganisms. J. Clin. Microbiol. **39:**936–942.
7. **Lau, S. K., et al.** 2006. Usefulness of the MicroSeq 500 16S rDNA bacterial identification system for identification of anaerobic Gram positive bacilli isolated from blood cultures. J. Clin. Pathol. **59:**219–222.
8. **Lau, S. K., et al.** 2006. Bacteremia caused by *Solobacterium moorei* in a patient with acute proctitis and carcinoma of the cervix. J. Clin. Microbiol. **44:**3031–3034.
9. **Lau, S. K., et al.** 2004. Anaerobic, non-sporulating, Gram-positive bacilli bacteraemia characterized by 16S rRNA gene sequencing. J. Med. Microbiol. **53:**1247–1253.
10. **Lau, S. K., et al.** 2004. Characterization of *Haemophilus segnis*, an important cause of bacteremia, by 16S rRNA gene sequencing. J. Clin. Microbiol. **42:**877–880.
11. **Lau, S. K., et al.** 2004. *Eggerthella hongkongensis* sp. nov. and *Eggerthella sinensis* sp. nov., two novel *Eggerthella* species, account for half of the cases of *Eggerthella* bacteremia. Diagn. Microbiol. Infect. Dis. **49:**255–263.
12. **Lau, S. K., P. C. Woo, T. C. Yim, A. P. To, and K. Y. Yuen.** 2003. Molecular characterization of a strain of group A streptococcus isolated from a patient with a psoas abscess. J. Clin. Microbiol. **41:**4888–4891.
13. **Maidak, B. L., et al.** 1999. A new version of the RDP (Ribosomal Database Project). Nucleic Acids Res. **27:**171–173.
14. **Murray, P. R., E. J. Baron, J. H. Jorgensen, M. R. Louise, and M. A. Pfaller (ed.).** 2007. Manual of clinical microbiology, 9th ed. American Society for Microbiology, Washington, DC.
15. **Patel, J. B., et al.** 2000. Sequence-based identification of *Mycobacterium* species using the MicroSeq 500 16S rDNA bacterial identification system. J. Clin. Microbiol. **38:**246–251.
16. **Simmon, K. E., A. C. Croft, and C. A. Petti.** 2006. Application of SmartGene IDNS software to partial 16S rRNA gene sequences for a diverse group of bacteria in a clinical laboratory. J. Clin. Microbiol. **44:**4400–4406.
17. **Tang, Y. W., et al.** 2000. Identification of coryneform bacterial isolates by ribosomal DNA sequence analysis. J. Clin. Microbiol. **38:**1676–1678.
18. **Tse, H., et al.** 2010. Complete genome sequence of *Staphylococcus lugdunensis* strain HKU09-01. J. Bacteriol. **192:**1471–1472.
19. **Woo, P. C., et al.** 2007. *In silico* analysis of 16S ribosomal RNA gene sequencing-based methods for identification of medically important anaerobic bacteria. J. Clin. Pathol. **60:**576–579.
20. **Woo, P. C., et al.** 2009. First report of *Tsukamurella* keratitis: association between *T. tyrosinosolvens* and *T. pulmonis* and ophthalmologic infections. J. Clin. Microbiol. **47:**1953–1956.
21. **Woo, P. C., et al.** 2003. *Granulicatella adiacens* and *Abiotrophia defectiva* bacteraemia characterized by 16S rRNA gene sequencing. J. Med. Microbiol. **52:**137–140.
22. **Woo, P. C., A. M. Fung, S. K. Lau, E. Hon, and K. Y. Yuen.** 2002. Diagnosis of pelvic actinomycosis by 16S ribosomal RNA gene sequencing and its clinical significance. Diagn. Microbiol. Infect. Dis. **43:**113–118.
23. **Woo, P. C., A. M. Fung, S. S. Wong, H. W. Tsoi, and K. Y. Yuen.** 2001. Isolation and characterization of a *Salmonella enterica* serotype Typhi variant and its clinical and public health implications. J. Clin. Microbiol. **39:**1190–1194.
24. **Woo, P. C., et al.** 2005. *Clostridium* bacteraemia characterised by 16S ribosomal RNA gene sequencing. J. Clin. Pathol. **58:**301–307.

25. **Woo, P. C., et al.** 2007. Surgical site abscess caused by *Lactobacillus fermentum* identified by 16S ribosomal RNA gene sequencing. Diagn. Microbiol. Infect. Dis. **58:**251–254.
26. **Woo, P. C., S. K. Lau, J. L. Teng, H. Tse, and K. Y. Yuen.** 2008. Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. Clin. Microbiol. Infect. **14:**908–934.
27. **Woo, P. C., et al.** 2009. The complete genome and proteome of *Laribacter hongkongensis* reveal potential mechanisms for adaptations to different temperatures and habitats. PLoS Genet. **5:**e1000416.
28. **Woo, P. C., et al.** 2004. Bacteremia due to *Clostridium hathewayi* in a patient with acute appendicitis. J. Clin. Microbiol. **42:**5947–5949.
29. **Woo, P. C., D. M. Tam, S. K. Lau, A. M. Fung, and K. Y. Yuen.** 2004. *Enterococcus cecorum* empyema thoracis successfully treated with cefotaxime. J. Clin. Microbiol. **42:**919–922.
30. **Woo, P. C., et al.** 2010. First report of *Gordonibacter pamelaeae* bacteremia. J. Clin. Microbiol. **48:**319–322.
31. **Woo, P. C., et al.** 2003. Analysis of a viridans group strain reveals a case of bacteremia due to Lancefield group G alpha-hemolytic *Streptococcus dysgalactiae* subsp. *equisimilis* in a patient with pyomyositis and reactive arthritis. J. Clin. Microbiol. **41:**613–618.
32. **Woo, P. C., et al.** 2009. Guidelines for interpretation of 16S rRNA gene sequence-based results for identification of medically important aerobic Gram-positive bacteria. J. Med. Microbiol. **58:**1030–1036.
33. **Woo, P. C., et al.** 2004. "*Streptococcus milleri*" endocarditis caused by *Streptococcus anginosus.* Diagn. Microbiol. Infect. Dis. **48:**81–88.
34. **Woo, P. C., et al.** 2005. Life-threatening invasive *Helcococcus kunzii* infections in intravenous-drug users and *ermA*-mediated erythromycin resistance. J. Clin. Microbiol. **43:**6205–6208.
35. **Woo, P. C., G. K. Woo, S. K. Lau, S. S. Wong, and K. Yuen.** 2002. Single gene target bacterial identification. *groEL* gene sequencing for discriminating clinical isolates of *Burkholderia pseudomallei* and *Burkholderia thailandensis.* Diagn. Microbiol. Infect. Dis. **44:**143–149.