Evaluation of the Integrated Database Network System (IDNS) SmartGene Software for Analysis of 16S rRNA Gene Sequences for Identification of *Nocardia* Species[▽]

Patricia S. Conville,* Patrick R. Murray, and Adrian M. Zelazny

Microbiology Service, Department of Laboratory Medicine, Warren Grant Magnuson Clinical Center, National Institutes of Health, U.S. Department of Health and Human Services, Bethesda, Maryland

Received 1 April 2010/Returned for modification 14 May 2010/Accepted 17 June 2010

16S rRNA gene sequences of 102 *Nocardia* isolates were analyzed using the Integrated Database Network System (IDNS) SmartGene centroid database. A total of 76% of the isolates were correctly identified. Discordant identifications were due to inadequate centroid length (3 species), inaccurate or insufficient entries in the public databases (5 species), and heterogeneous sequences among members of a species (1 species).

Nocardia species are significant human pathogens, especially in immunocompromised patients. Accurate species assignment is important to guide appropriate antibiotic selection, as several species possess high levels of antibiotic resistance (9, 10). 16S rRNA gene sequencing has become the most widely used method for the identification of these organisms.

Noncurated public databases contain a wealth of sequence information but also include a significant amount of information that is compromised by identification errors and low-quality sequence data (8). Searches against such databases and analysis of the results have become increasingly time-consuming, due to the rapidly growing number of submitted sequences. A quality-controlled sequence database that eliminates inaccurate and redundant entries and identifies the best representative sequence for a particular species would facilitate rapid and accurate identification.

SmartGene (SmartGene GmbH, Zug, Switzerland) is a web-based sequence search tool that allows comparison of input sequences with those in two curated databases. Using a proprietary filtering algorithm, the SmartGene 16S rRNA eubacteria database is assembled using sequences deemed acceptable from the public databases. A search using the eubacteria database allows the user to examine the inter- and intraspecies variabilities of a particular species within the SmartGene platform. An additional database, the centroid database, is prepared using a supplementary proprietary algorithm that examines all sequences of a particular species in the eubacteria database and creates a discrete “species group” of sequences for that species. The most representative sequence of that species is designated the “centroid” sequence. Only one sequence per species (the centroid sequence) is included in the centroid database; this sequence may or may not be the type strain of the species, depending on the heterogeneity of that species. Output from a search of the centroid database with an unknown sequence will show the most similar centroid se-

quences sorted by match score or by the number of base mismatches. The results will also indicate the number of sequences in the centroid group.

A study comparing the use of a previous version of the SmartGene software with the use of conventional identification and with that of another proprietary sequence database found SmartGene to be a time- and labor-saving system that provided a larger percentage of accurate genus or species assignments for a diverse group of bacteria than did the proprietary database (7).

We report here a study comparing the usefulness of the SmartGene centroid and eubacteria databases for accurate species-level identification of a variety of clinically relevant *Nocardia* isolates. We also include a careful analysis of discrepant results.

***Nocardia* isolates.** The type strains of 37 *Nocardia* species and 65 clinical isolates of *Nocardia*, representing 22 different species, were examined. The clinical isolates were recovered from patients being treated at the Warren Grant Magnuson Clinical Center of the NIH or were referred to the NIH for identification.

***Nocardia* identification.** All clinical *Nocardia* isolates were identified by 16S rRNA gene sequence analysis (approximately 1,400 bp) using a BLAST search and by comparison to isolates from an in-house database. In addition, the identification of 49 of the clinical isolates was confirmed by *secA1* gene sequencing, as previously described (5).

16S rRNA gene sequence analysis. Organisms were cultivated, and DNA was extracted as described previously (4). Three overlapping regions of the 16S rRNA gene were amplified and sequenced using primers with tails containing M13 binding sites for reactions 1 and 3, as previously described (3, 4). The primers used for reaction 2 were modified as follows (the M13 binding sites are in boldface): Reaction 2 Forward, 5'-GTA-AAA-CGA-CGG-CCA-GCA-ACT-ACG-TGC-CAG-CAG-CCG-C-3'; Reaction 2 Reverse, 5'-CAG-GAA-ACA-GCT-ATG-ACT-GAC-GAC-AGC-CAT-GCA-CCA-CCT-3'. PCR was performed as previously described (3, 4), except for the addition of 0.4 µg/ml bovine serum albumin to the PCR mix. The PCR amplification products from all regions were purified from gel bands, as previously reported (3), or with the Qiaex II

* Corresponding author. Mailing address: 10 Center Drive, MSC 1508, Building 10, Room 2C-385, National Institutes of Health, Bethesda, MD 20892. Phone: (301) 402-1891. Fax: (301) 402-1886. E-mail: pconville@mail.nih.gov.

[▽] Published ahead of print on 23 June 2010.

TABLE 1. *Nocardia* species giving acceptable identification with the centroid database^d

Species	Strain	No. of possible IDs meeting similarity threshold in database ^b :		No. of:	
		Centroid	Eubacteria	Isolates in centroid group	Clinical isolates tested
<i>N. abscessus</i> ^a	DSM 44432	1	2	19	5
<i>N. africana</i> ^a	DSM 44491	1	1	7	0
<i>N. anaemiae</i> ^a	DSM 44821	1	1	1	1
<i>N. araoensis</i> ^a	DSM 44729	1	1	2	0
<i>N. arthritidis</i> ^a	DSM 44731	1	1	8	3
<i>N. asiatica</i> ^a	DSM 44668	1	1	13	2
<i>N. asteroides sensu stricto</i> ^a	ATCC 19247	1	2	37	1
<i>N. brasiliensis</i> ^a	ATCC 19296	1	1	16	6
<i>N. brevicatena</i> ^a	ATCC 15333	1	1	6	0
<i>N. carnea</i> ^a	ATCC 6847	1	1	9	0
<i>N. elegans</i> ^a	JCM 13374	1	1	5	1
<i>N. exalbida</i>	DSM 44883	1	1	2	0
<i>N. farcinica</i>	ATCC 3318	1	2	31	6
<i>N. higoensis</i> ^a	DSM 44732	1	1	2	0
<i>N. inohanensis</i>	DSM 44667	1	1	4	0
<i>N. mexicana</i>	DSM 44952	1	1	4	2
<i>N. niigatensis</i>	DSM 44670	1	1	6	1
<i>N. nova</i>	ATCC 33726	1	2	39	6
<i>N. paucivorans</i>	ATCC BAA-278	1	1	8	0
<i>N. pseudobrasiliensis</i> ^a	ATCC 51512	1	1	14	4
<i>N. puris</i>	DSM 44599	1	1	6	3
<i>N. sienata</i> ^a	DSM 44766	2 ^c	2 ^c	1	0
<i>N. terpenica</i> ^a	DSM 44935	1	2	2	0
<i>N. testacea</i> ^a	DSM 44765	2 ^c	2 ^c	3	2
<i>N. thailandica</i> ^a	DSM 44808	1	1	2	1
<i>N. vermiculata</i> ^a	DSM 44807	1	2	1	0
<i>N. veterana</i>	DSM 44445	1	1	11	4
<i>N. vinacea</i> ^a	JCM 10988	1	1	7	1

^a Centroid is the sequence of the type strain of the species.^b Based on percent similarity of $\geq 99.8\%$. IDs, identifications.^c *N. sienata* and *N. testacea* are known to be 99.9% similar (2 base mismatches).^d Using approximately 1,400-bp sequences. Clinical isolates gave results identical to those of the type strains.

kit (Qiagen, Valencia, CA). Cycle sequencing and sequence analysis were performed as previously reported (3).

SmartGene evaluation. 16S rRNA gene sequences of *Nocardia* type strains and clinical isolates (approximately 1,400 bp) were individually analyzed using the Integrated Database Network System (IDNS) centroid and 16S rRNA gene eubacteria databases, with results sorted by identities/mismatches. A percent similarity of $\geq 99.8\%$ (value rounded off) to a centroid sequence (IDNS centroid database) or type strain sequence (eubacteria database) was considered an acceptable identification. This cutoff value was chosen based on our experience with *Nocardia*; 16S rRNA gene sequences of isolates belonging to a particular species usually are $\geq 99.8\%$ similar (2). To assess the use of the centroid database with partial 16S rRNA gene sequences, 500-bp sequences (from the 5' end of the 16S rRNA gene) of all *Nocardia* type strains were also analyzed.

Results and discussion. For 28 of 37 (76%) *Nocardia* type strains whose 16S rRNA gene sequences were tested using 1,400-bp regions against the centroid database, the closest acceptable centroid was the correct species for that sequence (Table 1). For 6 of 28 species, the centroid database gave more precise results than the eubacteria database, as invalid species or sequences with incorrect identifications present in the eubacteria database were eliminated in the centroid database (Table 1). The results obtained with the clinical isolates of each species were identical to the results obtained with the type

strains of those species. Standard BLAST search results were essentially identical to those obtained with the eubacteria database (data not shown).

Several discordant results were obtained using the centroid database and are summarized in Table 2. In the first error type (error type 1), three type strains were misidentified due to centroids with sequences shorter than those of the input sequences. As with other sequence matching software, one of the criteria for scoring matches is match length; a centroid with a sequence shorter than the input sequence receives a lower score than a longer (although incorrect) sequence. The selection of short centroids is related to the small number of available sequences in the public databases for some species; this error might be corrected by the selection of only full-length (or nearly full-length) sequences as centroid sequences, if such sequences exist.

The second error type resulted in discordant results for 5 species. These errors were due to inaccurate or insufficient entries in the public database. The first example (error type 2A) resulted in erroneous identification of the *Nocardia transvalensis* type strain and is a direct result of inaccurate information in the public database. The centroid for *N. transvalensis* is based on numerous *Nocardia wallacei* sequences erroneously identified as *N. transvalensis* in the public database; the *N. transvalensis* centroid sequence is identical to the sequence of

TABLE 2. Discordant results from searches of type strain sequences with the centroid database^c

Species	Strain	No. of:		Results				
		Clinical isolates tested	Isolates in centroid group	Centroid database				Best match(es) in eubacteria database
				Best match(es)	% similarity of match	Source(s) of discrepant results	Error type ^b	
<i>N. aobensis</i>	DSM 44805	0	6	<i>N. africana</i>	99.2	Short centroid	1	<i>N. aobensis</i>
<i>N. cyriacigeorgica</i>	DSM 44484	5	52	<i>N. abscessus</i>	98.8	Short centroid	1	<i>N. cyriacigeorgica</i>
<i>N. shimofusensis</i> ^a	DSM 44733	0	4	<i>N. higoensis</i>	99.2	Short centroid	1	<i>N. shimofusensis</i>
<i>N. transvalensis</i>	ATCC 6865	0	28	<i>N. blacklockiae</i>	98.8	Incorrect centroid	2A	<i>N. transvalensis</i>
				<i>N. wallacei</i>	98.6			
				<i>N. transvalensis</i>	98.6			
<i>N. wallacei</i> ^a	ATCC 49873	4	2	<i>N. wallacei</i>	100	Incorrect centroid for <i>N. transvalensis</i>	2B	<i>N. wallacei</i>
<i>N. beijingensis</i>	DSM 44636	4	24	<i>N. transvalensis</i>	99.9			
				<i>N. beijingensis</i>	100	Probable incorrect identification of <i>S. gardneri</i> in public database	2B	<i>N. beijingensis</i>
				<i>S. gardneri</i>	99.9			<i>S. gardneri</i>
<i>N. ninae</i> ^a	DSM 44978	0	1	<i>N. ninae</i>	99.7	Few sequences in public database; errors in those sequences	2C	<i>N. ninae</i>
<i>N. pneumoniae</i> ^a	DSM 44730	1	2	<i>N. pneumoniae</i>	99.4	Few sequences in public database; errors in those sequences	2C	<i>N. pneumoniae</i>
<i>N. otitidiscaviarum</i>	ATCC 14629	2	24	<i>N. otitidiscaviarum</i>	99.7	Heterogeneous sequences in species group	3	<i>N. otitidiscaviarum</i>

^a Centroid is the sequence of the type strain of the species.

^b Error type 1, short sequence selected as centroid; error type 2, inaccurate or insufficient entries in the public database (error types 2A and 2B, inaccurate information in the public database; error type 2C, possible sequence errors in type strains); error type 3, heterogeneous sequences of members of a species.

^c Clinical isolates have results identical to those of the type strains.

the *N. wallacei* type strain and is only 98.6% similar to a reliable sequence of the type strain of *N. transvalensis*.

The second example of discordant results due to inaccurate or insufficient entries in the public database resulted in the return of centroids of 2 species within the acceptable percent similarity range for searches with *N. wallacei* and *Nocardia beijingensis* (error type 2B). The search with the sequence of the type strain of *N. wallacei* returned identifications of both *N. wallacei* and *N. transvalensis*; this error is the result of the inaccurate centroid for *N. transvalensis*, as noted above. The centroid database search for the type strain of *N. beijingensis* returned centroids of *N. beijingensis* and *Streptomyces gardneri*. The *S. gardneri* sequence is most likely an incorrect identification in the public database (Table 2).

In the third example of discordant results due to erroneous or insufficient entries in the public database (error type 2C), the top matches in the searches using the sequences of *Nocardia ninae* and *Nocardia pneumoniae* were of the correct species but were <99.8% similar to the sequences of the centroid, resulting in an inadequate identification. There are few entries for these species in the public database; the centroids selected for these species were derived from sequences containing some possible sequence errors, as determined by comparison of those sequences with reliable recent GenBank sequence submissions for the same type strains. This illustrates the need for examining the number of sequences included in the species group when analyzing ambiguous results with the centroid database. The developers of this software can eliminate these types of discrepancies by continually monitoring public databases for new entries and modifying centroid sequences as necessary.

The third error type was related to the heterogeneity of sequences assigned to the species *Nocardia otitidiscaviarum* (error type 3). There are numerous sequences in the public databases for this species, which has been shown to have a high level of intraspecies heterogeneity (6). A sequence different from that of the type strain was chosen as the centroid for that species (Table 2), resulting in an identification for the type strain with a percent similarity that was below the customary threshold.

For all but one of the nine discordant results, searches using the eubacteria database (Table 2) and GenBank (data not shown) returned acceptable identifications.

Because 500-bp 16S rRNA gene sequences are frequently used for bacterial identification, especially with commercial sequencing products, 500-bp sequences of all type strains were also tested using the centroid database. Using these shortened sequences, ambiguous results were obtained for 4 species groups comprised of 10 species. This is due to insufficient base heterogeneity in the first 500 bases of the 16S rRNA genes of some *Nocardia* species, resulting in inadequate discrimination of some species, as has been previously reported (1). Ambiguous identifications were obtained for the following *Nocardia* species groups: *N. abscessus*/*N. asiatica*/*N. arthritis*, *N. africana*/*N. elegans*/*N. veterana*, *N. exalbida*/*N. gamkensis*, and *N. higoensis*/*N. shimofusensis*.

The SmartGene centroid database has the potential to offer a significant improvement over public databases for assigning *Nocardia* species identifications based on 1,400-bp sequences of the 16S rRNA gene. Many of the centroids selected for *Nocardia* species are sequences identical or nearly identical to those of the type strain of the species. The advantage of the

centroid approach is that only one entry per species is presented; the search output is relatively short and easily evaluated for sequence similarity. In addition, sufficient information is provided to allow a user with limited experience in *Nocardia* taxonomy to arrive at a correct identification as long as certain minimal standards of sequence similarity are achieved. For the *Nocardia* isolates tested using the centroid database, the lowest percentage of similarity indicating conspecificity may be as low as 99.4% (Table 2), although care should be used in the evaluation of species that show <99.8% similarity to a given species. More experienced users may find the eubacteria database useful for examining the intraspecies variability of a particular species.

Continued monitoring of public databases by SmartGene developers and the input of nearly full-length 16S rRNA gene sequences by the user can easily remedy the difficulties encountered with the centroid database for the identification of *Nocardia* species. This system represents a useful method for sequence comparison by laboratorians at all levels of experience with sequence interpretation.

We thank Frank G. Witebsky, Department of Laboratory Medicine, Warren Grant Magnuson Clinical Center, NIH, for critically reviewing the manuscript.

The views expressed herein are our views and should not be construed as those of the U.S. Department of Health and Human Services.

REFERENCES

1. Cloud, J. L., P. S. Conville, A. Croft, D. Harmsen, F. G. Witebsky, and K. C. Carroll. 2004. Evaluation of partial 16S ribosomal DNA sequencing for identification of *Nocardia* species by using the MicroSeq 500 system with an expanded database. *J. Clin. Microbiol.* **42**:578–584.
2. Conville, P. S., J. M. Brown, A. G. Steigerwalt, J. W. Lee, V. L. Anderson, J. T. Fishbain, S. M. Holland, and F. G. Witebsky. 2004. *Nocardia kruczakaie* sp. nov., a pathogen in immunocompromised patients and a member of the “*N. nova* complex.” *J. Clin. Microbiol.* **42**:5139–5145.
3. Conville, P. S., J. M. Brown, A. G. Steigerwalt, J. W. Lee, D. E. Byrer, V. L. Anderson, S. E. Dorman, S. M. Holland, B. Cahill, K. C. Carroll, and F. G. Witebsky. 2003. *Nocardia veterana* as a pathogen in North American patients. *J. Clin. Microbiol.* **41**:2560–2568.
4. Conville, P. S., S. H. Fischer, C. P. Cartwright, and F. G. Witebsky. 2000. Identification of *Nocardia* species by restriction endonuclease analysis of an amplified portion of the 16S rRNA gene. *J. Clin. Microbiol.* **38**:158–164.
5. Conville, P. S., A. M. Zelazny, and F. G. Witebsky. 2006. Analysis of *secA1* gene sequences for identification of *Nocardia* species. *J. Clin. Microbiol.* **44**:2760–2766.
6. Patel, J. B., R. J. Wallace, Jr., B. A. Brown-Elliott, T. Taylor, C. Imperatrice, D. G. B. Leonard, R. W. Wilson, L. Mann, K. C. Jost, and I. Nachamkin. 2004. Sequence-based identification of aerobic actinomycetes. *J. Clin. Microbiol.* **42**:2530–2540.
7. 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.
8. Turenne, C. Y., L. Tschetter, J. Wolfe, and A. Kabani. 2001. Necessity of quality-controlled 16S rRNA gene sequence databases: identifying nontuberculous *Mycobacterium* species. *J. Clin. Microbiol.* **39**:3637–3648.
9. Wallace, R. J., Jr., M. Tsukamura, B. A. Brown, J. Brown, V. A. Steingrube, Y. Zhang, and D. R. Nash. 1990. Cefotaxime-resistant *Nocardia asteroides* strains are isolates of the controversial species *Nocardia farcinica*. *J. Clin. Microbiol.* **28**:2726–2732.
10. Wilson, R. W., V. A. Steingrube, B. A. Brown, Z. Blacklock, K. C. Jost, Jr., A. McNabb, W. D. Colby, J. R. Biehle, J. L. Gibson, and R. J. Wallace, Jr. 1997. Recognition of a *Nocardia transvalensis* complex by resistance to aminoglycosides, including amikacin, and PCR-restriction fragment length polymorphism analysis. *J. Clin. Microbiol.* **35**:2235–2242.