Available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/IJMYCO

Identification of nontuberculous mycobacteria by partial gene sequencing and public databases

Ines Joao [a], *Paula Cristovao* [a], *Liliana Antunes* [b], *Baltazar Nunes* [b], *Luisa Jordao* [a,*]

[a] *National Institute of Health, Dr. Ricardo Jorge, Department of Infectious Diseases, Avenida Padre Cruz, 1649-016 Lisboa, Portugal*
[b] *National Institute of Health, Dr. Ricardo Jorge, Department of Epidemiology, Avenida Padre Cruz, 1649-016 Lisboa, Portugal*

ARTICLE INFO

Article history:
Received 4 April 2014
Accepted 7 April 2014
Available online 14 May 2014

Keywords:
Nontuberculous mycobacteria (NTM)
hsp65
16S rRNA
Public databases

ABSTRACT

Background: Nontuberculous mycobacteria (NTM) are a heterogeneous group of microorganisms with distinct clinical relevance. The treatment of NTM infections depends significantly upon the crucial identification of species at this level. The steady increase of mycobacteria species, the use of time-consuming techniques and the lack of standardized identification methods makes the achievement of this goal a demanding challenge. Additionally, inaccurate diagnosis can lead to therapeutic approaches consistent with *Mycobacterium tuberculosis* infection that are useless. In the present study, the performance of public databases in the accurate identification of NTM by sequence analysis of 16S rRNA and *hsp65* genes were evaluated and compared. An algorithm is proposed to achieve an accurate classification of NTM in the geographic region of Portugal (Western Europe).

Methods: Partial sequencing of 16S rRNA and *hsp65* genes of 22 reference strains and 54 clinical isolates was performed. The resulting sequences were analysed by public web databases since their performance is evaluated statistically. The phenotypic characteristics of the isolates were also evaluated.

Results: The use of commercial kits allowed the accurate identification of 57.4% of the clinical isolates. This result was improved either by the use of 16S rRNA (75.9%) and *hsp65* (88.9%) genes analysis alone or combined (96.3%).

Conclusions: Analysis of 16S rRNA gene alone is insufficient for the accurate identification of NTM. A stepwise algorithm combining 16S rRNA and *hsp65* gene analysis by multiple public databases is proposed to identify NTM at the species' level.

Introduction

The *Mycobacterium* genus comprises more than 160 species (http://www.bacterio.net/mycobacterium.html), among them are notable human pathogens, such as *Mycobacterium tuberculosis* and *Mycobacterium leprae*, and ubiquitous mycobacteria known collectively as nontuberculous mycobacteria (NTM) [1].

As the incidence of tuberculosis declines in industrialized countries, e.g., USA, NTM infections gained relevance, although accurate data on incidence is missing since case reports are not mandatory [1–4]. Human infection with NTM became relevant with the emergence of the AIDS pandemic [5,6] being currently recognized as a cause of pulmonary and extra-pulmonary infection in humans.

* Corresponding author. Tel.: +351 217519227.
E-mail addresses: ines.joao@insa.min-saude.pt (I. Joao), paula.cristovao@insa.min-saude.pt (P. Cristovao), liliana.antunes@insa.min-saude.pt (L. Antunes), Baltazar.Nunes@insa.min-saude.pt (B. Nunes), maria.jordao@insa.min-saude.pt (L. Jordao).
http://dx.doi.org/10.1016/j.ijmyco.2014.04.001

NTM infection follows exposure to environmental bacteria rather than human-to-human transmission; as such, there are large geographic differences in terms of species incidence. For example, in immunocompetent individuals, infections caused by *Mycobacterium kansasii* are the most prevalent in England and Wales, whilst in Scotland and northern European countries *Mycobacterium malmoense* takes the lead [3,7,8].

The members of the *Mycobacterium* genus are genetically closer to each other than the microorganisms belonging to other genera making identification a difficult and challenging task [9]. Nevertheless, identification at the species level is important for several reasons. First of all because management, treatment and infection control measures differ dramatically between *M. tuberculosis* and NTM infections [10]. Within NTM it is important to distinguish between sample contaminants and human pathogens. For the last group, it is crucial to discriminate the species in order to initiate an adequate antibacterial therapy [5,11–13].

The conventional biochemical identification methods may be misleading due to phenotypic variations of species, non-standardized test conditions and their being time-consuming [14].

Recent advances in molecular sequencing of conserved genes (16S rRNA, *hsp65*, *rpoB* and the 16S-23S internal transcribed spacer region) brought faster diagnosis. The generalized use of these methodologies also rendered the discovery of new species and the taxonomic reclassification of others increasing the number of NTM species [4]. These methods are dependent on databases. There are currently several web-based public databases for bacterial identification based on 16S rRNA and *hsp65* gene sequences [2,15,16]. Despite the advances in 16S rRNA and *hsp65* gene databases, their performance rarely has been evaluated or compared [2,17].

This study took advantage of sequencing technology to identify NTM based on two highly conserved genes encoding for 16S rRNA (gold standard) and *hsp65*. The aim of this study was to evaluate and to compare the performance of public databases in order to establish an algorithm for the accurate identification, at the species level, of clinically relevant NTM.

Material and methods

Bacterial isolates

In the current study 22 reference strains and 54 NTM isolates included in the National Institute of Health Dr. Ricardo Jorge (INSA) Mycobacteria Collection were analysed. The reference strains used were: *M. abscessus* subsp. *bolletii* CIP 108541, *M. alvei* ATCC 51304, *M. avium* subsp. *avium* ATCC 25291, *M. fallax* ATCC 35219, *M. fortuitum* subsp. *fortuitum* ATCC 6841, *M. gordonae* ATCC 14470, *M. intracellulare* ATCC 13950, *M. kansasii* ATCC 12478, *M. malmoense* ATCC 29571, *M. marinum* ATCC 927, *M. monacense* CIP 109237, *M. nonchromogenicum* ATCC 19530, *M. peregrinum* ATCC 14467, *M. porcinum* ATCC 33776, *M. scrofulaceum* ATCC 19981, *M. sherrisii* ATCC BAA-832, *M. simiae* ATCC 25275, *M. smegmatis* ATCC 19420, *M. szulgai* ATCC 35799, *M. terrae* ATCC 15755, *M. triviale* ATCC 23292 and *M. vaccae* ATCC 15483. The mycobacteria deposited in the INSA collection were isolated from clinical samples received during a two-year period (2008–2009). These isolates were identified with commercially available kits, namely AccuProbe (GenProbe, San Diego, USA) and GenoType *Mycobacterium* CM/AS (Hain Lifesciences GmbH, Nehren, Germany). Additionally, their phenotypic characteristics, including growth rate (fast/slow), pigment production, growth at different temperatures (30, 37 and 42 °C) and colony morphology were evaluated.

This study was approved by the INSA ethical committee.

Partial sequencing of 16S rRNA and hsp65 genes

Mycobacterial DNA was extracted with QIAamp DNA Mini kit (Qiagen GmbH, Hilden, Germany) according to the manufacturer's instructions. For 16S rRNA gene (≈976 pb) a forward (AGAGTTTGATCMTGGCTCAG) and a reverse (GTAAGGTTCTK CGCGTTGC) primer, targeting a genome segment within the hyper variable regions A and B, designed with the Beacon Designer 7.0 software were used. The in-house RT-PCR, using SyberGreen (Roche Diagnostics, Manmheim, Germany), was performed in the following conditions: denaturation (10 min at 95 °C), hybridization (5 s at 61 °C, 32 cycles) and elongation (39 s at 72 °C) followed by the melting curve (30 s at 40 °C) using a Light Cycler apparatus. The fluorescence was acquired at 530 nm. In each reaction a negative (water) and positive (*M. avium* ATCC25291) control were used. For *hsp65* gene (440 pb) two primers (Tb11 and Tb12) were used as described previously [18].

The PCR for sequencing was performed with the primers described above. The amplification products were purified using the Jetquick kit (Genomed GmbH, Löhne, Germany) according to the manufacturer's instructions and sequenced. The obtained sequences for 16S rRNA gene were analysed using BioEdit Sequence Alignment Editor 7.0.9. Software (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) and compared with known sequences in the GenBank (http://www.ncbi.nlm.nih.gov), EzTaxon-e (http://eztaxon-e.ezbiocloud.net/) and RIDOM: Ribosomal Differentiation of Medical Microorganisms Database (RIDOM) (http://rdna.ridom.de/) using the BLAST algorithm (http://blast.ncbi.nlm.nih.gov/Blast.cgi). For *hsp65* the obtained sequences were compared with those present in GenBank, Web-Accessible Database of hsp65 Sequences from *Mycobacterium* Reference Strains (Web-Accessible Database) (http://msis.mycobacteria.info) and (hsp65)-BLAST Database (hsp65-BLAST) (http://hsp65blast.phsa.ca/blast/blast.html).

Statistical analysis

The results obtained for the identification of the MNT from each database or combining two databases for each gene were classified in categories. In total 6 categories were established according to the presence/absence of the NTM specie considered correct: (1) Identified only the NTM specie considered correct; (2) Identified one or more species belonging to the same group/complex being the correct specie included; (3) Identified one or more non-related species being the correct specie included; (4) Identified one or more non-related species being the correct specie absent; (5) Identified one or

more species belonging to the same group/complex being the correct specie absent; and (6) Not determined.

Relative frequencies were calculated for category 1, database or combination of databases and for each gene. McNemar test for paired data was used in order to compare the performance of databases or combination of databases for the first category. Cohen's Kappa was also calculated for the agreement between databases for all categories. To compare the results obtained from each gene and combination of two genes, only the databases identified with better performance in the first analysis were used. Two and three databases were used for the 16S rRNA gene and *hsp65* gene, respectively. The results obtained were classified over two categories: Identified and not identified. Relative frequencies were calculated and McNemar test for paired data was used in order to compare the results.

Results

Sample characterization

During the period of the study (2008–2009) 364 mycobacteria were identified by Accuprobe and/or GenoType combined with phenotypic characteristics. A total of 54 mycobacteria isolates identified either as NTM or *Mycobacterium spp* were selected for the present work (Table 1). The percentage of NTM isolated increased from 11.2% in 2008 to 20.7% in 2009. The data available in the collection records shows that 63% of the mycobacteria were isolated in specific media (e.g. Lowenstein-Jensen or Mycobacteria Growth Indicator Tube) before being sent for identification. In these cases the information about the nature of the clinical sample is missing. The majority of the samples for which this data is available were isolated from respiratory samples such as sputum (26%) and bronchia secretions (2%); 7% of the NTM were isolated from hemoculture and 2% from skin biopsies.

In 42.6% ($n = 23$) of the cases, the NTM were not classified at the species level, as shown in Table 1. Since this fact can hamper the correct diagnostics and treatment, it was decided to apply other molecular methods based on partial gene sequencing and evaluate their performance.

Molecular methods: partial 16S rRNA and hsp65 gene sequencing and analysis

Two genes, 16S rRNA and *hsp65*, were partially sequenced and analyzed using three public databases for each, being the first gene considered the gold standard. The identification was considered accurate when there were at least two identical results among the databases used and the phenotypic characteristics were concordant. If not, a second gene *hsp65* was assayed using the same criteria.

An accurate identification was obtained for all the reference strains ($n = 22$) independently of the gene or database used.

The results obtained for NTM classification using 16S rRNA partial sequencing alone are shown in Table 2. In all cases the sequence similarity ranged between 99% and 100%. GenBank allowed the classification of 75.9% ($n = 41$), followed by Eztaxon-e with 70.4% ($n = 38$) and RIDOM with 64.8% ($n = 35$) of the isolates.

The results obtained for five isolates using RIDOM were discordant from those obtained with the remaining databases. This was observed for *M. arupense* (E23, E36), *M.*

Table 1 – NTM identification by commercial kits *versus* in-house gene sequencing methods.

Sample ID	Commercial kits	This study
E1, E5, E6, E31, E41, E43–47	MAC	*M. avium*
E53		*M. intracellulare*
E48	*M. intracellulare*	MAC
E24		*M. intracellulare*
E4, E7, E17–19, E28–30, E42, E52	*M. gordonae*	*M. gordonae*
E9, E12, E15, E32, E34, E38	*M. abcessus*	*M. abcessus*
E13, E50–51	*M. chelonae*	*M. chelonae*
E20–21, E40, E54	*M. fortuitum*	*M. fortuitum*
E3		*M. fortuitum complex*
E26	*M. mucogenicum*	*M. fortuitum*
E10	*M. nonchromogenicum*	*M. nonchromogenicum*
E14	*M. xenopi*	*M. xenopi*
E16, E39	*M. lentiflavium*	*M. lentiflavium*
E37	*Mycobacterium spp*	*M. peregrinum*
E8		*M. fortuitum*
E22		*M. nebraskense*
E35		*M. gordonae*
E11		*M. kumamotonense*
E27		*M. monacense*
E23		*M. arupense*
E25		*M. lentiflavium*
E33	Neither MAC nor MTB	*M. frederiksbergense*
E36		*M. arupense*
E2		*M. chelonae*
E49		*M. marinum*

Table 2 – NTM identification by partial sequencing of *16S rRNA* gene.

Sample ID	16S rRNA gene sequencing		
	GenBank	RIDOM	EzTaxon-e
E1, E5, E6, E31, E41, E43-E47	*M. avium*	*M. avium*	*M. avium*
E4, E7, E17-E19, E28-E30, E35, E42, E52	*M. gordonae*	*M. gordonae*	*M. gordonae*
E8, E20, E21, E40, E54	*M. fortuitum*	*M. fortuitum*	*M. fortuitum*
E10	*M. nonchromogenicum*	*M. nonchromogenicum*	*M. nonchromogenicum*
E14	*M. xenopi*	*M. xenopi*	*M. xenopi*
E16, E25, E39	*M. lentiflavium*	*M. lentiflavium*	*M. lentiflavium*
E27	*M. monacense*	*M. monacense*	*M. monacense*
E22	*M. nebraskense*	*M. malmoense*	*M. nebraskense*
E2	*M. chelonae*	*M. chelonae*, *M. abscessus*	*M. franklinii*, *M. salmoniphilum*, *M. chelonae*, *M. abscessus*
E50	*M. chelonae*	*M. chelonae*, *M. abscessus*	*M. franklinii*, *M. chelonae*, *M. abscessus*
E3	*M. fortuitum*, *M. neworleansense*, *M. porcinum*	*M. fortuitum*, *M. porcinum*	*M. neworleansense*, *M. porcinum*
E9, E12, E13, E15, E32, E38, E51	*M. abscessus*, *M. chelonae*	*M. abscessus*, *M. chelonae*	*M. franklinii*, *M. abscessus*, *M. chelonae*
E11	*M. kumamotonense*	*M. terrae*	*M. kumamotonense*
E23, E36	*M. arupense*	*M. terrae*	*M. arupense*
E24	*M. avium complex*, *M.yongonense*	*M. intracellulare*	*M. marseillense*
E26	*M. fortuitum*	*M. fortuitum*	*M. fortuitum*
E33	*M. frederiksbergense*	*M.fluoroanthenivorans*	*M. frederiksbergense*
E34	*M. abscessus*, *M. chelonae*	*M. abscessus*, *M. chelonae*	*M. franklinii*, *M. abscessos*, *M. chelonae*, *M. immunogenum*
E37	*M. peregrinum*, *M. septicum*	*M. peregrinum*, *M. septicum*	*M. septicum*
E48	*M. avium*, *M. marseillense*, *M. yongonense*	*M.intracellulare*	*M. marseillense*
E49	*M. marinum*, *M. ulcerans*	*M. marinum*, *M. ulcerans*	*M. marinum*, *M. ulcerans*
E53	*M. intracellulare*	*M.intracellulare*	*M. chimaera*

frederiksbergense (E33), *M. kumamotonense* (E11) and *M. nebraskense* (E22), which share the particularity of being NTM species identified after the last update of RIDOM database performed in 2005. The other discordant result was generated for E53, which was classified as *M. intracellulare* by GenBank and RIDOM and *M. chimarae* by the EzTaxon-e database.

Next, it was decided to combine more than one database to identify the isolates. Unfortunately, this approach did not improve the NTM identification. The EzTaxon-e/RIDOM allowed the identification of 61.1% ($n = 33$), GenBank/RIDOM 68.5% ($n = 37$) and GenBank/EZTaxon-e 74.1% ($n = 40$) of the isolates. The best combination of databases allowed for the identification of one less isolate. The discordant result was generated for E53 which was classified *M. intracellulare* using the GenBank and RIDOM databases, and *M. chimarae* using the EzTaxon-e database (Table 2).

The McNemar test showed that this result was not significantly different from the results obtained by GenBank database alone ($p = 0.3173$). A value of 0.86 was obtained for Cohen's Kappa coefficient indicating a good agreement between these two databases.

The 16S rRNA gene was not discriminatory for the members of two complexes (MAC and *M. fortuitum* complex), and closely related mycobacteria, such as *M. abscessus*/*M. chelonae* and *M. ulcerans*/*M. marinum*.

In order to overcome this handicap, the partial sequencing of the *hsp65* gene was performed and the results obtained by the three databases (GenBank, Web-Accessible Database and hsp65-BLAST) used are shown in Table 3. A sequence match of 97% was considered as a correct identification [19] for all NTM except *M. gordonae*. In this case, a sequence match of 96% was adopted due to the high polymorphic nature of this specie [20]. The partial sequencing of the *hsp65* gene rendered a better identification rate (88.9%, $n = 48$) with hsp65-BLAST database than 16S rRNA gene sequencing. The other databases used also generated better results with 85.2% and 81.5% identification frequencies for Web-Accessible Database and GenBank, respectively.

Hsp65 proved to be more discriminatory between the closely related mycobacteria species referred to above, but some problems persisted. Although the *hsp65* gene discriminates between *M. chelonae* and *M. abscessus*, GenBank failed in its identification. The same occurred for *M. pseudoshottsi* and *M. marinum* for Web-Accessible Database. For one isolate, the results were inconclusive (E26).

The combination of GenBank either with Web-Accessible Database or hsp65-BLAST rendered the accurate identification of 88.9% ($n = 48$), while 87.0% ($n = 47$) was obtained for Web-Accessible Database/hsp65-BLAST. The combination of databases only improved the results obtained by the database with the worst performance (GenBank). In all cases, according to the McNemar test the results were not significantly different.

A good agreement according to Cohen's test was achieved for the comparison between the best database alone (hsp65-BLAST) and the two combinations (GenBank/hsp65-BLAST

Table 3 – NTM identification by partial sequencing of *hsp65* gene.

Sample ID	*hsp65* gene sequencing		
	GenBank	Web-Accessible Database	(hsp65)-BLAST database
E1, E5, E6, E41, E43-E47	*M. avium*	*M. avium*	*M. avium*
E7, E42, E52	*M. gordonae*	*M. gordonae*	*M. gordonae*
E4, E17-E19, E28- E30	*M. gordonae*	*M. gordonae*[e]	*M. gordonae*
E8, E20, E21, E40, E54	*M. fortuitum*	*M. fortuitum*	*M. fortuitum*
E10	*M. nonchromogenicum*	*M. nonchromogenicum*	*M. nonchromogenicum*
E11	*M. kumamotonense*	*M. kumamotonense*	*M. kumamotonense*
E14	*M. xenopi*	*M. xenopi*	*M. xenopi*
E16, E25, E39	*M. lentiflavium*	*M. lentiflavium*	*M. lentiflavium*
E27	*M. monacense*	*M. monacense*[e]	*M. monacense*[e]
E22	*M. nebraskense*	*M. nebraskense*	*M. nebraskense*
E2,E13, E50, E51	*M. chelonae*	*M. chelonae*	*M. chelonae*
E3	*M. peregrinum*, *M. porcinum*	*M. peregrinum*, *M. porcinum*, *M. boenickei*	*M. peregrinum*, *M. porcinum*, *M. boenickei*
E9, E12, E15, E32, E34, E38	*M. abscessus*, *M. chelonae*	*M. abscessus*	*M. abscessus*
E23	*M. arupense*	*M. arupense*	*M. arupense*
E24	*M. intracellulare*	*M. intracellulare*	*M. yongonense*
E26	*Inconclusive*	*Inconclusive*	*M. ratisbonense*
E31	*M. avium*	*M. avium*	*M. avium*, *M. bouchedurhonense*
E33	*M. neoaurum*, *M. hackensackensee*	*M. neoaurum*	*M. neoaurum*, *M. hackensackensee*
E35	*M. gordonae*	*M. triplex*[e]	*M. gordonae*
E36	*M. arupense*	*M. arupense*	*M. arupense*
E37	*M. peregrinum*	*M. peregrinum*	*M. peregrinum*
E48	*M. avium*, *M. saskatchewanense*	*M. triplex*, *M. seoulens*, *M. saskatchewanense*	*M. avium complex*
E49	*M.marinum*	*M .marinum*, *M.pseudoshottsi*	*M. marinum*
E53	*M. intracellulare*	*M. intracellulare*	*M. intracellulare*

e Less than 97% similarity.

or Web-Accessible Database) with better performances ($k \geqslant 0.68$).

The use of the two genes together improved the identification of NTM. For 16S rRNA, excluding the results of the obsolete RIDOM database, an accurate classification was obtained for 74.1% ($n = 40$) of the isolates. The *hsp65* reached a score of 88.9% ($n = 48$) that was improved by the combination of the two genes (96.3%, $n = 52$). When using one gene alone for NMT identification, *hsp65* showed a better performance than 16S rRNA. The difference is significant according to McNemar test ($p = 0.008$).

Nevertheless, the combination of the two genes improved their performance significantly when compared with 16S rRNA ($p = 0.001$) but not to *hsp65* ($p = 0.157$).

Discussion

The introduction of molecular biology methods in mycobacteria identification, back in the 1990s, led to either the discovery or taxonomic reclassification of relevant human pathogens [9]. Commercially available kits – although effective in the identification of the most frequent clinical mycobacteria species – failed to identify rare and/or new species and are expensive [21]. In this context it is mandatory to find alternatives in order to achieve an accurate identification affordable for the majority of the laboratories.

To accomplish an accurate identification of NTM at the species level more than one gene must be used as a genetic marker [22]. Often the genotypic analysis must be complemented with the phenotypic analysis. This study took advantage of sequencing technology to identify mycobacteria based on two highly conserved genes (16S rRNA and *hsp65*) and of several public databases to perform the data analysis.

The 16S rRNA is universally used for bacteria identification; nevertheless, it has several limitations. Although being rare, in some cases (*M. terrae* and *M. celatum* complex), two copies of the 16S rRNA gene, with different sequences, can be found in the same mycobacteria [23,24]. The analysis of the *hsp65* gene, known as the single copy gene, can solve this issue. Another advantage over the 16S rRNA gene is the ability to differentiate between intimately related species, such as *M. abscessus*/*M. chelonae*, *M. szulgai*/*M. malmoense* and *M. kansasii*/*M. gastri* [25]. This differentiation is difficult to achieve by 16S rRNA analysis due to the small variation in the sequence [26]. Analysis of 16S rRNA sequences does not allow for distinguishing between the following MNT species: *M. abscessus*/*M. chelonae*; *M. avium*/*M. intracellulare*; *M. peregrinum*/*M. septicum*; *M. kansasii*/*M. gastri* and *M. marinum*/*M. ulcerans* [27–30].

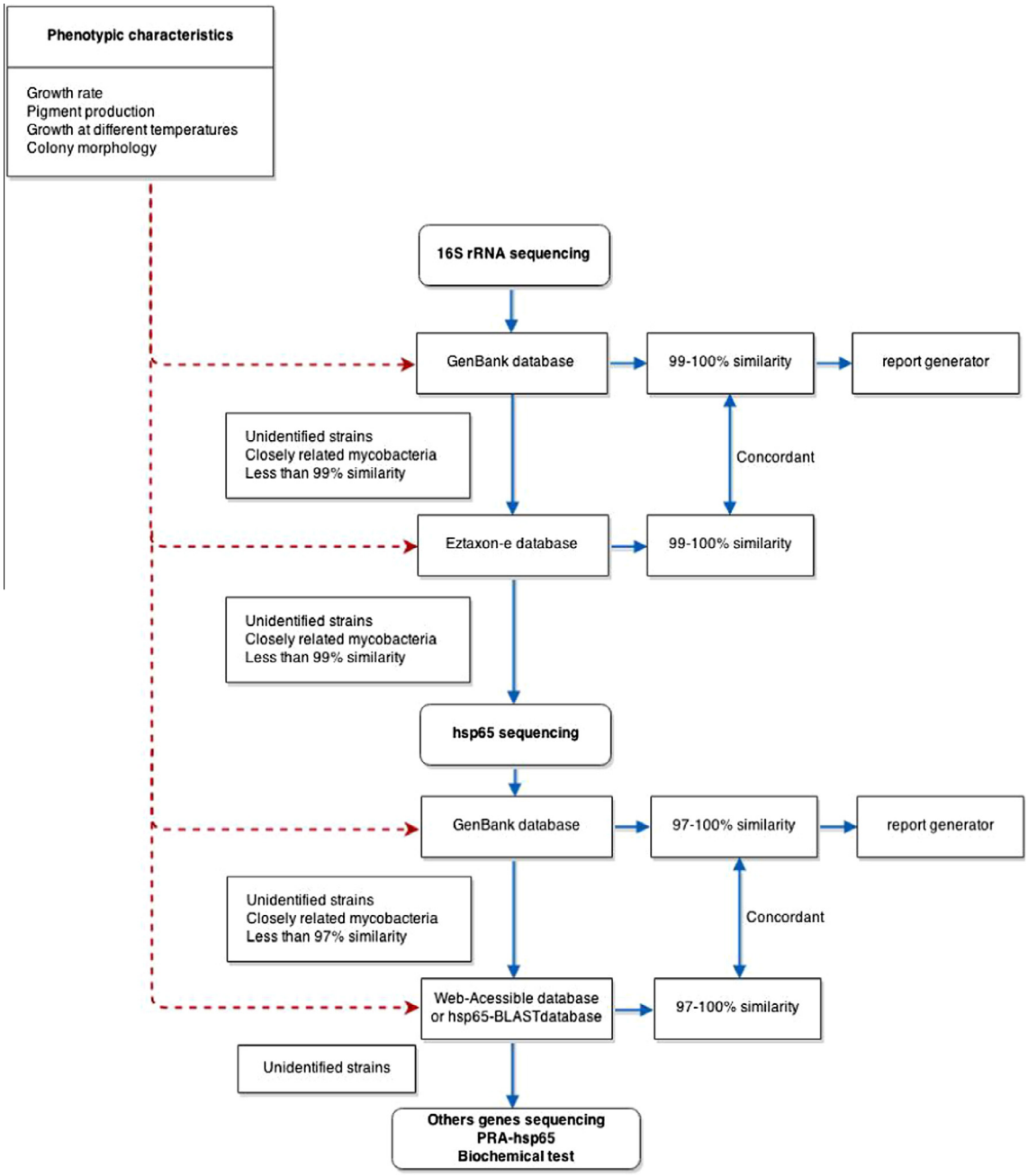

Fig. 1 – Flow chart for the identification of *Mycobacterium* species. The 16SRNA sequences were analyzed by GenBank and Ezotaxon databases. If a 99–100% sequence similarity is achieved and the phenotypic characteristics are concordant, the NTM is identified. Otherwise, *hsp65* gene sequencing is performed using a similarity cut-off above 97%. If the NTM is not identified at the species level, other approaches, including gene sequencing and biochemical tests must be performed. Closely related mycobacteria were shown as follows: *M. abscessus/M. chelonae*; *M. ulcerans/M. marinum*; *M. pseudoshottsi/M. marinum*; *M. fortuitum/M. neworleansense/M. porcinum* and *M. peregrinum /M. septicum*.

Despite all these advantages, *hsp65* also has pitfalls. In the present work *M. gordonae*, whose heterogeneity has been genetically established [26], raised several issues [18,19,31,32]. The nucleotide sequences of *hsp65* obtained for seven isolates (E4, E17–E19 and E28–E30) had 99–100% sequence homology between each other, but only 96% homology with *M. gordonae* ATCC 14470. A similar problem has been reported previously by McNabb and colleagues [19]. This author suggested a 97% similarity cut-off for *hsp65* sequences for an accurate identification. Nevertheless, to achieve these standards, the group was forced to add extra sequences to the database. It was decided to use another approach accepting a 96% cut-off and complementing the *hsp65* sequencing with PCR-restriction enzyme analysis, PRA-*hsp65* ([18] and Appendix: Table A1 PRA-*hsp65*). A ninety-six similarity percentage was achieved, in the Web-Accessible Database, for the seven isolates previously mentioned which share a common PRA pattern classified as type 3. The *M. gordonae* (E35),

Table A1 – Mycobacteria identification by PRA-*hsp65*.

Sample ID	Bands with		PRA patterns
	BstEII (pb)	HaeIII (pb)	
E7, E42, E52	235-120-85	160-11-60	*M. gordonae* type 1
E4, E17–E19, E28- E30	235-120-100	130-115	*M. gordonae* type 3
E35	320-115	130-110-70-60	*M. gordonae* type 8

PRA-hps65 is based on the amplification of a 441-bp segment from *hsp65* gene followed by digestion with restriction enzyme (BstEII and HaeIII). The pattern of restriction was compared to a published algorithm [31].

which was erroneously identified as *M. triplex* by the same database, exhibited a PRA pattern of type 8. The remaining isolates, whose classification was unambiguous in all databases (>97%) belong to PRA pattern type 1 classification. Collectively, these data provide evidence to support the low sequence homology between *M. gordonae* isolates.

The databases used in the analysis also generate problems with different databases yielding different identifications for the same sequence. Another issue, particularly relevant for new species, was the lack of database updates. This was especially evident for GenBank and RIDOM databases [33].

According to Tortoli's description, genetic databases can be categorized in two major groups: controlled and uncontrolled [33]. In this study both categories were used. GenBank represents the first group, while the remaining (RIDOM, EzTaxon-e and Web-Accessible Database) are members of the second group. Sequences submitted to GenBank lack validation both in terms of sequence quality and updatedness of mycobacteria taxonomic classification. Among the deficiencies found, the existence of "species" with unvalidated labels, misclassified species and incomplete classifications to the genus level [22,32,33] were highlighted in this study. The lack of control over the sequences submitted to this database allowed for identical sequences to yield different percentages of similarity. A major problem arising from this is that low quality sequences could generate perfect identifications which will be accepted if the analyst is not familiarized with the problem.

Additionally, this database holds a huge number of flawed sequences, which can lead to identification errors, because even if the sequences are correct, the highest score may be attributed to a sequence with 99% similarity if the size of the submitted fragment is bigger [22].

On the other hand, the controlled databases include excellent quality sequences which are correctly characterized but with a limited scope. Their application is hampered by the lack of sequence updates (RIDOM), and by the exclusion of sequences that do not belong to reference strains (Web-Accessible Database).

Conclusion

In conclusion, the workflow presented in Fig. 1 is suggested to achieve an accurate identification of NTM. The first gene to be analysed is 16S rRNA since it is the gold standard and has shown a superior performance in the identification of the most frequent NTM (MAC and *M. gordonae*) both in this sample and in Western European settings [3]. Since the rapid-growing *Mycobacterium* species (*M. abscessus*, *M. chelonae* and *M. fortuitum*) are the third most frequent in these settings and the first two species are hardly identified by this gene, the use of *hsp65* is suggested. In parallel it is recommended that phenotypic tests (growth rate, pigment production, etc.) be used to complement the analysis.

When this two step approach did not render an accurate identification, the sequencing of additional genetic markers (e.g. *rpoB*) can improve the results. Another option is the use of the PRA-*hsp65* gene. This approach is particularly useful for polymorphic mycobacteria such as *M. gordonae* and has the additional advantage of being an affordable technique.

Conflict of interest

None declared.

Author contributions

Conceived and designed the experiments and data analysis: Ines Joao, Liliana Antunes, Baltazar Nunes and Luisa Jordao. Performed experiments: Ines Joao and Paula Cristovao. Performed the data analysis and wrote the paper: Ines Joao, Liliana Antunes, Baltazar Nunes and Luisa Jordao.

Acknowledgements

We are grateful to the INSA's Technology and Innovation Unit for technical support. This work received financial support from INSA, and LJ is a recipient of a Portuguese Science Foundation Grant No. C2008-C2008_P2.

Appendix A

Table A1.

REFERENCES

[1] E. Tortoli, Clinical manifestations of nontuberculous mycobacteria infections, Clin. Microbiol. Infect. 15 (2009) 906–910.

[2] J. Dai, Y. Chen, M. Lauzardo, Web-Accessible Database of *hsp65* Sequences from *Mycobacterium* Reference Strains, J. Clin. Microbiol. 49 (2011) 2296–2303.

[3] W. Hoefsloot, J. van Ingen, C. Andrejak, K. Angeby, R. Bauriaud, P. Bemer, et al, The geographic diversity of nontuberculous mycobacteria isolated from pulmonary

samples: an NTM-NET collaborative study, Eur. Respir. J. 42 (2013) 1604–1613.
[4] C. Piersimoni, C. Scarparo, Pulmonary infections associated with non-tuberculous mycobacteria in immunocompetent patients, Lancet Infect. Dis. 8 (2008) 323–334.
[5] D.E. Griffith, T. Aksamit, B.A. Brown-Elliott, A. Catanzaro, C. Daley, F. Gordin, et al, An official ATS/IDSA statement: diagnosis, treatment, and prevention of nontuberculous mycobacterial diseases, Am. J. Respir. Crit. Care Med. 175 (2007) 367–416.
[6] Management of opportunist mycobacterial infections: Joint Tuberculosis Committee Guidelines 1999. Subcommittee of the Joint Tuberculosis Committee of the British Thoracic Society, Thorax 55 (2000) 210–218.
[7] J.M. Grange, M.D. Yates, Infections caused by opportunist mycobacteria: a review, J. R. Soc. Med. 79 (1986) 226–229.
[8] B. Henriques, S.E. Hoffner, B. Petrini, I. Juhlin, P. Wåhlén, G. Källenius, Infection with *Mycobacterium malmoense* in Sweden: report of 221 cases, Clin. Infect. Dis. 18 (1994) 596–600.
[9] E. Tortoli, Impact of genotypic studies on mycobacterial taxonomy: the new mycobacteria of the 1990s, Clin. Microbiol. Rev. 16 (2003) 319–354.
[10] S. Hofmann-Thiel, L. Turaev, T. Alnour, L. Drath, M. Mullerova, H. Hoffmann, Multi-centre evaluation of the speed-oligo Mycobacteria assay for differentiation of *Mycobacterium* spp. in clinical isolates, BMC Infect. Dis. 11 (2011) 353.
[11] K.A. Nash, B.A. Brown-Elliott, R.J. Wallace, A novel gene, erm(41), confers inducible macrolide resistance to clinical isolates of *Mycobacterium abscessus* but is absent from *Mycobacterium chelonae*, Antimicrob. Agents Chemother. 53 (2009) 1367–1376.
[12] W.J. Koh, O.J. Kwon, K.S. Lee, Nontuberculous mycobacterial pulmonary diseases in immunocompetent patients, Korean J. Radiol. 3 (2002) 145–157.
[13] J. van Ingen, R. de Zwaan, M. Enaimi, P.N.R. Dekhuijzen, M.J. Boeree, D. van Soolingen, Re-analysis of 178 previously unidentifiable *Mycobacterium* isolates in the Netherlands in 1999–2007, Clin. Microbiol. Infect. 16 (2010) 1470–1474.
[14] Y. Kazumi, S. Mitarai, The evaluation of an identification algorithm for *Mycobacterium* species using the 16S rRNA coding gene and *rpoB*, Int. J. Mycobacteriol. 1 (2012) 21–28.
[15] O.S. Kim, Y.J. Cho, K. Lee, S.H. Yoon, M. Kim, H. Na, et al, Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species, Int. J. Syst. Evol. Microbiol. 62 (2012) 716–721.
[16] D. Harmsen, J. Rothgänger, M. Frosch, J. Albert, RIDOM: Ribosomal Differentiation of Medical Micro-organisms Database, Nucleic Acids Res. 30 (2002) 416–417.
[17] K.S. Park, C.S. Ki, C.I. Kang, Y.J. Kim, D.R. Chung, K.R. Peck, et al, Evaluation of the GenBank, EzTaxon, and BIBI services for molecular identification of clinical blood culture isolates that were unidentifiable or misidentified by conventional methods, J. Clin. Microbiol. 50 (2012) 1792–1795.
[18] A. Telenti, F. Marchesi, M. Balz, F. Bally, E.C. Böttger, T. Bodmer, Rapid identification of mycobacteria to the species level by polymerase chain reaction and restriction enzyme analysis, J. Clin. Microbiol. 31 (1993) 175–178.
[19] A. McNabb, D. Eisler, K. Adie, M. Amos, M. Rodrigues, G. Stephens, et al, Assessment of partial sequencing of the 65-kilodalton heat shock protein gene (*hsp65*) for routine identification of *Mycobacterium* species isolated from clinical sources, J. Clin. Microbiol. 42 (2004) 3000–3011.
[20] S. Itoh, Y. Kazumi, C. Abe, M. Takahashi, Heterogeneity of RNA polymerase gene (*rpoB*) sequences of *Mycobacterium gordonae* clinical isolates identified with a DNA probe kit and by conventional methods, J. Clin. Microbiol. 41 (2003) 1656–1663.
[21] E. Tortoli, M. Pecorari, G. Fabio, M. Messino, A. Fabio, Commercial DNA probes for mycobacteria incorrectly identify a number of less frequently encountered species, J. Clin. Microbiol. 48 (2010) 307–310.
[22] C.Y. Turenne, L. Tschetter, J. Wolfe, A. Kabani, Necessity of quality-controlled 16S rRNA gene sequence databases: identifying nontuberculous *Mycobacterium* species, J. Clin. Microbiol. 39 (2001) 3637–3648.
[23] U. Reischl, K. Feldmann, L. Naumann, B.J. Gaugler, B. Ninet, B. Hirschel, et al, 16S rRNA sequence diversity in *Mycobacterium celatum* strains caused by presence of two different copies of 16S rRNA gene, J. Clin. Microbiol. 36 (1998) 1761–1764.
[24] B. Ninet, M. Monod, S. Emler, J. Pawlowski, C. Metral, P. Rohner, et al, Two different 16S rRNA genes in a mycobacterial strain, J. Clin. Microbiol. 34 (1996) 2531–2536.
[25] H. Kim, S.H. Kim, T.S. Shim, M. Kim, G.H. Bai, Y.G. Park, et al, Differentiation of *Mycobacterium* species by analysis of the heat-shock protein 65 gene (*hsp65*), Int. J. Syst. Evol. Microbiol. 55 (2005) 1649–1656.
[26] P. Kirschner, B. Springer, U. Vogel, A. Meier, A. Wrede, M. Kiekenbeck, et al, Genotypic identification of mycobacteria by nucleic acid sequence determination: report of a 2-year experience in a clinical laboratory, J. Clin. Microbiol. 31 (1993) 2882–2889.
[27] T. Rogall, J. Wolters, T. Flohr, E.C. Böttger, Towards a phylogeny and definition of species at the molecular level within the genus *Mycobacterium*, Int. J. Syst. Bacteriol. 40 (1990) 323–330.
[28] C. Turenne, P. Chedore, J. Wolfe, F. Jamieson, K. May, A. Kabani, Phenotypic and molecular characterization of clinical isolates of *Mycobacterium elephantis* from human specimens, J. Clin. Microbiol. 40 (2002) 1230–1236.
[29] K.J. Williams, C.L. Ling, C. Jenkins, S.H. Gillespie, T.D. McHugh, A paradigm for the molecular identification of *Mycobacterium* species in a routine diagnostic laboratory, J. Med. Microbiol. 56 (2007) 598–602.
[30] D.A. Stahl, J.W. Urbance, The division between fast- and slow-growing species corresponds to natural relationships among the mycobacteria, J. Bacteriol. 172 (1990) 116–124.
[31] E. Chimara, L. Ferrazoli, S.Y.M. Ueky, M.C. Martins, A.M. Durham, R.D. Arbeit, et al, Reliable identification of mycobacterial species by PCR-restriction enzyme analysis (PRA)-*hsp65* in a reference laboratory and elaboration of a sequence-based extended algorithm of PRA-*hsp65* patterns, BMC Microbiol. 8 (2008) 48.
[32] B. Häfner, H. Haag, H.K. Geiss, O. Nolte, Different molecular methods for the identification of rarely isolated non-tuberculous mycobacteria and description of new *hsp65* restriction fragment length polymorphism patterns, Mol. Cell. Probes. 18 (2004) 59–65.
[33] E. Tortoli, Standard operating procedure for optimal identification of mycobacteria using 16S rRNA gene sequences, Stand. Genomic Sci. 3 (2010) 145–152.