A multigene approach to phylogenetic analysis using the genus *Mycobacterium* as a model

G. Devulder,¹ M. Pérouse de Montclos² and J. P. Flandrois^{1,2}

Correspondence

G. Devulder

devulder@biomserv.univ-lyon1.fr

¹UMR CNRS 5558, Laboratoire de bactériologie, Faculté de Médecine Lyon-Sud, BP 12, 69921 Oullins Cedex, France

²Laboratoire de bactériologie, CHU Lyon Sud, 69495 Pierre-Bénite Cedex, France

Advances in DNA sequencing and the increasing number of sequences available in databases have greatly enhanced the bacterial identification process. Several species within the genus *Mycobacterium* cause serious human and animal diseases. In order to assess their relative positions in the evolutionary process, four gene fragments, from the 16S rRNA (564 bp), *hsp65* (420 bp), *rpoB* (396 bp) and *sod* (408 bp) genes, were sequenced from 97 strains, including all available type strains of the genus *Mycobacterium*. The results demonstrate that, in this case, the concatenation of different genes allows significant increases in the power of discrimination and the robustness of the phylogenetic tree. The sequential and/or combined use of sequences of several genes makes it possible to refine the phylogenetic approach and provides a molecular basis for accurate species identification.

INTRODUCTION

Within the framework of a broader project of bacterial identification, we were here interested specifically in the identification of members of the genus *Mycobacterium* by molecular methods. The genus *Mycobacterium* contains more than 90 species, as listed in J. P. Euzéby's *List of Bacterial Names with Standing in Nomenclature* (<http://www.bacterio.cict.fr/m/mycobacterium.html>), including organisms that cause serious human and animal diseases. Human infections are caused mainly by slowly growing strains, which form colonies in more than 7 days. There have been increasing numbers of reports of infections caused by mycobacteria other than *Mycobacterium tuberculosis*, especially associated with human immunodeficiency virus infection (Kim *et al.*, 1999). Rapid identification of mycobacteria may greatly improve infection control. Conventional biochemical methods and phenotypic tests for species differentiation are tedious and time-consuming and may require specialized testing that is beyond the capacity of the clinical laboratory.

During the last two decades, the development of molecular

biological tools has led to profound modifications in the classification and methods of identification of these bacteria. Molecular methods using one or several appropriate genes are gaining importance because they yield quick and, in most cases, unequivocal results (Kolbert & Persing, 1999). These methods allow universality of results, which facilitates comparisons of identification between different laboratories. Different phylogenies for the genus have been developed using sequence analysis of the 16S rRNA gene (Kirschner & Bottger, 1998; Rogall *et al.*, 1990; Springer *et al.*, 1996; Cloud *et al.*, 2002). Most mycobacterial identification tools currently available such as the Ribosomal Differentiation of Medical Microorganisms (RIDOM) database (Harmsen *et al.*, 2002) or the Microseq kit (Perkin Elmer/Applied Biosystems) are based on 16S rRNA gene sequence analysis. However, within the genus *Mycobacterium*, the interspecies similarity is relatively high, from 94.3 to 100 %. Some species have a very high degree of similarity or have exactly identical sequences, such as *Mycobacterium kansasii* and *Mycobacterium gastri*, *Mycobacterium senegalense* and *Mycobacterium farcinogenes*, *Mycobacterium marinum* and *Mycobacterium ulcerans*, *Mycobacterium malmoense* and *Mycobacterium szulgai* and members of the *M. tuberculosis* complex. In the last decade, some authors have initiated studies using other sequences such as *recA* (Blackwood *et al.*, 2000), *rpoB* (Kim *et al.*, 1999; Gingeras *et al.*, 1998), ITS (Roth *et al.*, 1998), *hsp65* (Telenti *et al.*, 1993; Brunello *et al.*, 2001; Ringuet *et al.*, 1999), *sod* (Zolg & Philipp-Schulz, 1994) and *gyrB* (Kasai *et al.*, 2000). The use of one gene within the genus *Mycobacterium* does not allow perfect specific discrimination. The combined use of sequences of several genes opens

Published online ahead of print on 20 August 2004 as DOI 10.1099/ij.s.0.63222-0.

The GenBank/EMBL/DDBJ accession numbers for the sequences obtained in this study are listed in a supplementary table in IJSEM Online.

Details of strains and sequence accession numbers and single-gene neighbour-joining trees for *hsp65*, *rpoB* and *sod* and a maximum-likelihood tree for the combined dataset are available as supplementary material in IJSEM Online.

the possibility of increasing discriminatory power. This multigenic approach fulfils the recent recommendations of the ad hoc committee for the re-evaluation of the definition of the bacterial species (Stackebrandt *et al.*, 2002). It recommends that, for the description of novel species, sequences from four or five housekeeping genes besides the 16S rRNA gene are taken into account. We have thus developed a multigene sequence database dedicated to the identification of species within the genus *Mycobacterium*. We have sequenced approximately 400 nt fragments of four genes, 16S rRNA, *hsp65*, *rpoB* and *sod*, for all the cultivable type strains. In this report we analyse the phylogenetic results, compare the phylogenetic trees resulting from these data and show the significance of the use of more than one gene in phylogenetic reconstruction. We show how the concatenation of several genes allows an increase in discriminatory power and provides a more robust phylogenetic tree.

METHODS

Bacterial strains. Ninety-seven strains of mycobacteria from two collections, the DSMZ (Deutsche Sammlung von Mikroorganismen und Zellkulturen) and the CIP (Collection de l'Institut Pasteur), were used in this study; details are given in a supplementary table in IJSEM Online. The type strain of *Mycobacterium genavense* (ATCC 51234^T) is not available from the ATCC. Strain DSM 44424, generally considered as the reference strain, was used as a substitute in this study. An additional strain, *Nocardia abscessus* DSM 44432^T, was used as the phylogenetic root. Strains were grown in different culture media as recommended by the CIP or DSMZ.

Genes used. The 16S rRNA gene was the first gene used for bacterial identification (Rogall *et al.*, 1990). It is generally accepted that the 16S rRNA gene is the best target for studying phylogenetic relationships at the species level, because it is present in all bacteria, is functionally constant and is composed of highly conserved as well as more variable domains. Moreover, many sequences are available in public databases. Sequencing a 16S rRNA gene fragment by using the PCR technique and a selection of appropriate primers provides a phylogenetic framework that serves as the backbone for modern microbial taxonomy.

The gene *hsp65*, which is present in all mycobacteria, belongs to the family of heat-shock protein (Hsp) genes. These proteins are highly immunogenic, with an exceptional degree of evolutionary conservation. They have a function in intracellular protein folding, assembly and transport. Their expression is upregulated under cellular stress. The gene *rpoB* encodes the β -subunit of RNA polymerase, an oligomeric enzyme responsible for RNA synthesis. Mutations within a limited region of *rpoB* are known to be related to rifampicin resistance in *M. tuberculosis* (Telenti *et al.*, 1993). The gene *sod* encodes metalloenzymes that constitute one of the major defence mechanisms against oxidative stress.

Preparation of DNA and PCR. Chromosomal DNA was released from bacterial suspensions by a simple boiling method according to Afghani & Stutman (1996). Amplification of the 16S rRNA gene was done with primers g2R and rM582R (De Beenhouwer *et al.*, 1995). This fragment corresponds to positions 50–582 in *Mycobacterium bovis* (Suzuki *et al.*, 2001). All strains could be amplified with these primers. The amplification was done with a 70 μ l reaction mixture containing 35 μ l Roche PCR master mix (1 $\times$ PCR buffer, 1.25 U *Taq* polymerase, 0.2 mM dNTPs, 10 mM Tris/HCl, 50 mM KCl, 1.5 mM MgCl₂), 2.8 μ l Roche 25 mM MgCl₂ solution (1 mM),

0.14 μ l of each primer (0.2 μ M), 3.5 μ l 5% DMSO and 26.42 μ l water and 2 μ l DNA. The thermal profile involved initial denaturation for 10 min at 90 °C, 40 cycles of denaturation for 25 s at 94 °C, annealing for 30 s at 60 °C and extension for 45 s at 72 °C followed by 1 cycle at 72 °C for 10 min.

Amplification of the *hsp65* gene was done with primers derived from primers Tb11 and Tb12 (Ringuet *et al.*, 1999) targeting positions 396–836 of the published sequence from *M. tuberculosis* (Shinnick, 1987). The same experimental protocol was used for amplification except that annealing was performed at 55 °C.

Amplification of *rpoB* was done with two sets of primers, MF (Kim *et al.*, 1999) and TBB, which target positions 1713–2108 of the published *rpoB* sequence from *M. tuberculosis* H37RV, and primer GrpoB1, derived from primer MF, and reverse primer GrpoB2, derived from the sequence alignment.

Amplification of *sod* was done with forward primer Z205 and two different reverse primers, Z212 (Zolg & Philippi-Schulz, 1994) and GSOD2. Sixteen of the initial 97 strains could not be sequenced, although various primers and various experimental conditions were used. Primers sets suitable for the purpose of this research are listed in Table 1.

Sequence analysis. Sequence data were edited using the TED software from the Staden package. Sequences were aligned using CLUSTAL W (Thompson *et al.*, 1994). Phylogenetic analysis was done and trees were calculated using PHYLO_WIN (Galtier *et al.*, 1996) and NJplot (Perrière & Gouy, 1996). Trees were calculated using the neighbour-joining method under the global gap removal option of PHYLO_WIN and Kimura's two-parameter model as the substitution model (Kimura, 1980). Phylogenetic trees were also calculated by the maximum-likelihood method with PHYML software (Guindon & Gascuel, 2003), which allows a discrete-gamma model to be used to accommodate rate variation among sites. PHYML is an extremely fast and accurate method especially developed to estimate large phylogenies by maximum-likelihood. Bootstrap values of the supertree were computed by resampling 500 times. RRTree was used to compare the substitution rate between DNA sequences grouped in a phylogenetic lineage (Robinson-Rechavi & Huchon, 2000).

RESULTS AND DISCUSSION

Phylogenetic analysis

We studied the phylogeny of the set of mycobacteria using four genes: 16S rRNA (564 bp), *hsp65* (420 bp), *rpoB* (396 bp) and *sod* (408 bp). Ninety-seven strains of the genus *Mycobacterium* were subjected to sequencing for each gene, but the *sod* gene of 16 strains could not be sequenced. Our final tree therefore included only 81 strains. For each gene, a tree was calculated in order to check the global compatibility between the different trees. Clusters supported by bootstrap values greater than 80% in the four phylogenetic trees were identified. Strains included in these clusters were gathered in identical clusters supported by bootstrap values greater than 80% in others trees. Missing clusters in other trees correspond to strains not significantly associated with other strains. The congruence of these trees allows the concatenation of different sequences to construct a global phylogeny. The third codon position was not saturated on the tree resulting from synonymous changes according to the value of K_s (ratio of synonymous

Table 1. Primers used in PCR

F and R respectively represent forward and reverse primers.

Gene	Primer	Direction	Sequence (5'–3')
16S rRNA	g2R	F	GAGAATTCGTGCTTAACACATGCAAGTCG
	rM582R	R	ATGGATCCGTGAGATTTACGAACAACGC
<i>hsp65</i>	Tb11	F	ACCAACGATGGTGTGTCCAT
	Tb112	R	CTTGTCGAACCGCATACCCT
<i>rpoB</i>	MF	F	CGACCACTTCGGCAACCG
	TBB rpoB2	R	TACGGCGTCTCGATGAASCC
	GrpoB1	F	ATCGACCACTTCGGCAACCGCC
	GrpoB2	R	GGTACGGCGTCTCGATGAASCCG
<i>sod</i>	Z205	F	ACGTTCACCACAGCAAGCACCA
	Z212	R	TCGGCCCAGTTCACGACGTT
	GSOD2	R	TCGGCCAGTTCACGACGTTCCA

substitutions to the number of synonymous sites). All sites were taken into account for genes encoding proteins (*hsp65*, *rpoB*, *sod*). The global phylogeny includes 1788 bp.

Four phylogenetic trees were computed starting from these genes and another tree resulted from the concatenation of these sequences. The numbers of strains included in these trees were respectively 97, 96, 95, 82 and 81 for the 16S rRNA, *hsp65*, *rpoB* and *sod* trees and the tree resulting from the concatenation. Each species is represented by its type strain except *M. bovis*, for which we also included the BCG strain, and *M. genavense*, as discussed above.

Single-gene trees

In this study, the phylogenetic analysis based on the 16S rRNA gene includes all the established cultivable species described in the genus *Mycobacterium*. We have sequenced around 564 bp from the 16S rRNA gene that includes the two variable areas A and B (Tortoli, 2003). These areas correspond to *Escherichia coli* positions around 130–210 and 430–500, respectively. We observed relatively high similarity (95.5 %) and a low G + C content (58.8 mol%), whereas the whole genome G + C content of this genus is between 59 and 66 mol%. A phylogenetic tree was calculated that includes type strains of 96 species or subspecies (Fig. 1). The tree resulting from this analysis is consistent with phylogenetic trees described in previous studies (Tortoli, 2003). This gene allows the discrimination of most species within the genus *Mycobacterium*. We have examined classical phenotypic and biochemical properties in order to explore the congruence between properties and phylogeny. None of pathogenicity, pigmentation, nitrate reductase, urease, arylsulfatase or tolerance to 5 % NaCl seemed to agree perfectly with the phylogeny. However, as expected, two distinct groups were highlighted by this phylogenetic tree: rapidly growing and slowly growing mycobacteria (Stahl & Urbance, 1990). Only two species are not well positioned in the phylogenetic tree: *Mycobacterium doricum* and *Mycobacterium holsaticum*. *M. doricum* is close to the rapid-growers group, whereas it

is known as a slow-grower, forming colonies in about 2 weeks. The nucleotide sequence of the 16S rRNA gene of *M. doricum* is characterized by a short helix 18 and a single cytosine insertion in helix 10. The combination of such features is considered as the genetic signature of thermo-tolerant rapid growers (Tortoli, 2003). *M. holsaticum*, described by Richter *et al.* (2002) as a rapidly growing strain, is associated with the slowly growing cluster. This species forms visible colonies within 1 week, but its 16S rRNA gene is characterized by an extraordinary similarity to the *M. tuberculosis* sequence within hypervariable region A. This observation explains its poor phylogenetic position. Except for these two species, the distinction between slow and rapid growers is clear.

The tree based on the variable fragment of the 16S rRNA gene is not very robust. There are only 34 nodes that are supported by bootstrap values greater than 50 % (35.05 % of nodes), of which 16 are greater than 80 % (Table 2), and very few deep nodes are supported by high bootstrap values. The monophyletic slow-growth cluster is not supported. Moreover, some species could not be discriminated at the species level. Sequences analysis shows that some species have identical sequences within the considered fragment: *M. kansasii*–*M. gastri* (Cloud *et al.*, 2002), *M. ulcerans*–*M. marinum* (Stinear *et al.*, 2000), *Mycobacterium tokaiense*–*Mycobacterium murale*, *Mycobacterium chelonae*–*Mycobacterium abscessus*, *Mycobacterium septicum*–*Mycobacterium peregrinum*, *Mycobacterium vanbaalenii*–*Mycobacterium vaccae* and the *M. tuberculosis* complex. It was also impossible to characterize subspecies of *Mycobacterium fortuitum* and *Mycobacterium avium*. This led us to think that, although widely used in taxonomy and identification, this gene is not really suitable in our case: the resulting tree is not very robust and does not allow the discrimination of all strains at the species level. We decided to integrate three additional housekeeping genes. These genes are relatively variable and are present in all mycobacteria species.

The same analysis was thus carried out for three other genes,

Table 2. Gene features

Mean pairwise distances (*p*-distance) and mean pairwise distances calculated using Kimura's two-parameter model of substitution (K2P distance) are given. Numbers of nodes giving bootstrap values (B) greater than the threshold indicated, together with the percentage of all nodes that this represents, are also given.

Gene	Fragment size (bp)	Species included (<i>n</i>)	Mean <i>p</i> -distance	Mean K2P distance	G + C content (mol%)	B > 50 % (%)	B > 80 % (%)
<i>hsp65</i>	420	96	0.090	0.096	65.9	39 (40.63)	26 (27.08)
<i>rpoB</i>	396	95	0.082	0.087	67.1	24 (25.26)	14 (14.74)
<i>sod</i>	408	82	0.163	0.186	64.1	30 (36.59)	22 (26.83)
16S rRNA	564	97	0.045	0.047	58.8	34 (35.05)	16 (16.49)
All	1788	81	0.092	0.099	63.3	49 (60.49)	31 (38.27)

hsp65, *rpoB* and *sod* (trees are available as Supplementary Figs A–C in IJSEM Online), respectively including 96, 95 and 82 strains. The third codon position being not saturated on the tree resulting from synonymous changes according to the value of K_s , all sites were taken into account for genes encoding proteins (*hsp65*, *rpoB*, *sod*).

The *hsp65* gene is more variable than the 16S rRNA gene (Ringuet *et al.*, 1999). From a 420 bp fragment, we observed an average rate of similarity close to 91.0 % in the 96 strains included, with a G + C content close to 66 mol%. All species studied were discriminated from each other except the members of the *M. tuberculosis* complex and the species *M. farcinogenes* and *M. senegalense*. Slowly and rapidly growing groups were clearly separated except *M. doricum*, which was not grouped with slowly growing species. Moreover, 39 nodes were supported by bootstrap values greater than 50 % (40.63 % of nodes), against 34 for the 16S rRNA gene (Table 2). This tree shows a good power of discrimination, with a relative robustness.

The *rpoB* sequences (396 bp) were used to calculate a phylogenetic tree including 95 strains. These sequences showed 91.8 % similarity and a high G + C content (67.1 mol%). Slowly and rapidly growing groups were clearly separated except *M. doricum*, which was not grouped with slowly growing species. Each species was differentiated as a distinct entity in the phylogenetic tree. Twenty-four nodes were supported by bootstrap values greater than 50 % (25.26 % of nodes). *rpoB* sequence analysis was strongly recommended as an additional method to 16S rRNA gene-based identification of mycobacteria (Kim *et al.*, 1999). This tree shows a good power of discrimination, but does not seem to be very robust.

The *sod* gene analysis included 82 strains with a sequence

length of 408 bp. Of the 97 strains studied, 15 could not be sequenced. These sequences show 83.7 % similarity. This constitutes the most variable gene of this study. The great variability observed may explain the difficulties in sequencing this gene. The G + C content was 64.1 mol%. Each species was differentiated as a distinct entity and the two groups of slow and rapid growers were well separated. As described in other phylogenetic studies, *M. doricum* was not positioned in the slowly growing group. Thirty nodes were supported by bootstrap values greater than 50 %, although there were fewer species in this phylogeny than in the others (37.04 % of nodes).

For all the tested genes, the separation between the slowly growing and rapidly growing mycobacteria is clear, although not supported by a high bootstrap value. Moreover, *M. doricum* is always badly positioned in the phylogenetic tree. Biochemical and phenotypic properties could not be superimposed on the different phylogenies. For these different gene analyses, we have checked the congruence of each one with the others. For each tree, we have listed all the species included in nodes supported by bootstrap values greater than 80 %. Species included within clusters supported by strong bootstrap values in one of these trees are associated in the same strong clusters in other trees or are not associated in another cluster supported by high bootstrap values. The congruence of these various trees allowed us to perform the concatenation of these sequences in order to increase the precision of mycobacterial phylogenetic relationships using the variable regions of a greater number of sites (1788 bp).

Concatenated dataset

The concatenation of genes has been shown to be extremely sensitive to the addition or withdrawal of genes for a given

Fig. 1. Phylogenetic tree of the genus *Mycobacterium* computed from 16S rRNA gene sequences by the neighbour-joining method and Kimura's two-parameter model as the substitution model. The tree includes 97 strains and was rooted using *N. abscessus* DSM 44432^T. Strain details and accession numbers are given in a supplementary table available in IJSEM Online; unless the strain name is given, sequences were obtained from type strains. Other single-gene trees are available as Supplementary Figs A–C in IJSEM Online.

number of species and Teichmann & Mitchison (1999) concluded that there was an absence of phylogenetic signal in bacterial genes. Another way has been proposed to

compute the genome tree: the supertree method (Daubin *et al.*, 2002). If two genes are informative at different levels of the phylogeny, their concatenation will attenuate this

Fig. 2. Phylogenetic tree of the genus *Mycobacterium* computed from the concatenation of 16S rRNA gene, *hsp65*, *sod* and *rpoB* sequences by the neighbour-joining method and Kimura's two-parameter model as the substitution model. The tree includes 81 strains and was rooted using *N. abscessus* DSM 44432^T.

information rather than bring it out. In this study, we have checked the congruence for each tree before performing concatenation. The tree resulting from this concatenation is presented in Fig. 2. This tree includes the 81 strains for which we were able to sequence the four genes (16S rRNA, *hsp65*, *rpoB*, *sod*). The global G + C content resulting from the concatenation is 63.4 mol%, which is close to the average G + C content of members of the genus *Mycobacterium*.

The discriminatory power of this last tree is much more significant than that observed with the single 16S rRNA gene. Each species was differentiated as a distinct entity, except members of the *M. tuberculosis* complex. The introduction of the *gyrB* gene could be used to differentiate these species (Niemann *et al.*, 2000). The missing discriminatory power of some genes related to some species is compensated for by others. For example, *hsp65* discriminates *M. chelonae* from *M. abscessus* perfectly (Ringuet *et al.*, 1999; Devallois *et al.*, 1997), whereas they have the same 16S rRNA gene sequence. In the same way, *M. senegalense* and *M. farcinogenes* (Hamid *et al.*, 2002) have exactly the same *hsp65* sequence, whereas they have different 16S rRNA gene sequences. The use of several genes allowed the discrimination of each strain at the species level. The mean distance from these concatenated sequences was 9.9 %, against 4.5 % with the 16S rRNA gene alone, and the distance matrix included a smaller number of low values. Most close species had distances greater than 2 %, and the distances observed were relatively homogeneous. Only six species showed very low pairwise distances similar to the distances observed for subspecies: *M. farcinogenes*–*M. senegalense*, *M. marinum*–*M. ulcerans* and *M. murale*–*M. tokaiense*, and we did not observe intermediate distances as seen in the 16S rRNA gene tree. The sequence concatenation allowed us to obtain a very good species discrimination with relatively homogeneous distances.

The increase in sequence size led to a considerable increase in the tree robustness. The progressive concatenation

showed an increase in deep-node bootstrap values. The percentage of bootstrap values greater than 50 % is 60.49 % compared with 35.05 % for the 16S rRNA gene sequences (Table 2). Bootstrap values of nodes that were well supported in at least two datasets and also observed for each other gene are reported in Table 3. A significant increase in bootstrap values was observed in the tree resulting from the concatenation. For example, the cluster *Mycobacterium smegmatis*–*Mycobacterium goodii*, supported by bootstrap values of 88, 98, <50 and 93 %, respectively, for the single 16S rRNA, *hsp65*, *rpoB* and *sod* genes, is strongly supported in the final tree (100 %). Data concatenation therefore provides a good means of increasing the robustness of the final tree. This strong increase in bootstrap values demonstrates that phylogenetic trees calculated from several different genes such as housekeeping genes may considerably improve the phylogenetic relevance. The slow-growers node is supported by a bootstrap value of 72 % in the final tree whereas, in the 16S rRNA gene tree, a bootstrap value of 10 % was observed. Slowly and rapidly growing groups are clearly separated except for *M. doricum*, which is not included within the slow-growers group. We noticed that the slow-growers cluster is strictly monophyletic and is derived from a rapidly growing ancestor, whereas the rapid-growers cluster is polyphyletic; *Mycobacterium fallax* and *Mycobacterium brumae* are not associated in the rapid-growers cluster. An identical topology calculated with PHYML was observed for the tree calculated using maximum-likelihood with a discrete-gamma model (available as Supplementary Fig. D in IJSEM Online). The cluster *M. fallax*–*M. brumae* also seems independent of the two others. This division between slowly and rapidly growing mycobacteria is related to the rRNA operon copy number, respectively equal to one or two (Ji *et al.*, 1994; Domenech *et al.*, 1994). However, in some cases, the number of rRNA operons does not correlate with growth rate. For example, *M. chelonae*, which is known to have a single rRNA operon, is classified as a fast grower. The organization of rRNA operons within the genome may

Table 3. Bootstrap values for clusters that were observed in all four trees and supported in at least two datasets by bootstrap values greater than 80 %

Values are bootstrap percentages from the trees indicated for each cluster.

Species	16S rRNA	<i>hsp65</i>	<i>rpoB</i>	<i>sod</i>	All
<i>M. tuberculosis</i> complex	100	100	100	100	100
<i>M. tokaiense</i> – <i>M. murale</i>	100	100	100	100	100
<i>M. ulcerans</i> – <i>M. marinum</i>	100	100	100	100	100
<i>M. farcinogenes</i> – <i>M. senegalense</i>	85	94	88	100	100
<i>M. xenopi</i> – <i>M. heckeshornense</i>	88	100	84	100	100
<i>M. kansasii</i> – <i>M. gastri</i>	100	92	55	99	100
<i>M. smegmatis</i> – <i>M. goodii</i>	88	98	<50	93	100
<i>M. immunogenum</i> – <i>M. chelonae</i> – <i>M. abscessus</i>	98	<50	95	91	100
<i>M. rhodesiae</i> – <i>M. aichiense</i>	66	94	64	100	100
<i>M. nonchromogenicum</i> – <i>M. hiberniae</i>	65	96	53	81	99

explain this type of discordance (Menendez *et al.*, 2002). Indeed, this division was observed with all analysed genes and therefore seems to be linked to a whole evolutionary process.

Concatenation of the genes significantly increased the resolution and robustness of the tree. The phylogeny became more precise with fewer unresolved nodes when we increased the number of genes included in the analysis. These different genes from diverse locations on the whole genome seem to have evolved in the same way. We noticed the increasing value of the node that separates the slowly growing from the rapidly growing mycobacteria. This led us to think that speciation within this genus was progressive and relatively homogeneous across the whole genome. As observed with the 16S rRNA gene, the biochemical and phenotypic properties did not allow us to identify clusters.

From the two main clusters, the relative rate test was performed to compare the substitution rate between the rapidly and slowly growing mycobacteria using RRTree (Robinson-Rechavi & Huchon, 2000). The mean substitution rates relative to the outgroup were 0.114 and 0.118, respectively, for the rapidly growing and slowly growing mycobacteria. The probability associated with the test was 0.30. This *P* value is highly non-significant; the two groups have thus evolved at the same rate. Species with a short generation time, which therefore copy their DNA more frequently per time unit, did not accumulate more copying errors than species with a longer generation time. This observation implies that the evolutionary rate does not seem to be correlated to the number of generations but to the time. If we consider this hypothesis as true, the mutation rate estimated per generation would be higher for slowly growing mycobacteria, in which fewer DNA duplications occurred. However, we do not know how these species grow in the natural environment. The growth rate depends on the temperature, the nutrient composition and many other factors. We have also ignored what has happened during the last several million years.

With the aim of evaluating the discriminatory capacity of the phylogenetic trees, we specified the analysis of the distance matrix during the progressive sequence concatenation. At first, we selected sequences of the 81 strains retained in the last tree. Only pairwise distances from different species were analysed. The distances between subspecies of *M. avium* and *M. fortuitum* and the *M. tuberculosis* complex species were not evaluated. Only 74 strains were considered as different species. The progressive concatenation of new sequences provides a significant decrease in the number of *p*-distance values lower than 0.03. The number of *p*-distances lower than 0.03 varies from 490 to 12 for one to four genes of a total of 2701 pairwise distances. We observed an important decrease in the number of values lower than 0.03 when we added the first protein-coding gene. By comparison, the decreases resulting from other additions of protein-coding genes were less important. In the final tree, only 12 distance values lower than 0.03 were observed,

including only three pairwise distances lower than 0.01 concerning very close species: *M. farcinogenes*–*M. senegalense* (0.003), *M. marinum*–*M. ulcerans* (0.005) and *M. murale*–*M. tokaiense* (0.005).

From a taxonomic point of view, the DNA–DNA relatedness parameter and, whenever determinable, ΔT_m remain the acknowledged standard for species delineation. The phylogenetic definition of species would generally include strains with approximately 70 % or more DNA–DNA relatedness and with a ΔT_m of 5 °C or less. Species having 70 % or greater DNA–DNA relatedness generally have at least 97 % sequence identity in the 16S rRNA gene primary structure (Stackebrandt & Goebel, 1994). In fact, this value seems very low compared to our results. The mean distance observed from the 16S rRNA gene tree is 4.5 % and analysis of the distance matrix shows that most close species have distances much lower than 3 % (for the 81 strains retained in the final analysis, 490 pairwise distances are lower than 3 %). Moreover, these distances were calculated from the hyper-variable fragment of the 16S rRNA gene. We may suppose that the same analysis based on the whole 16S rRNA gene would give a larger number of small distances. As described above, the progressive concatenation of new genes allowed us to reduce significantly the number of small distances and thus increase the discriminatory power. Within the framework of identification, the discriminatory capacity of a gene may be defined as its ability to discriminate closely related species. The existence of no null distances is a first criterion. In order to avoid ambiguous identification, it would also be interesting to consider a minimal distance threshold. Indeed, in our study, only type strains were used (with one exception), preventing the integration of intraspecific variability. The differentiation between these species and, more largely, the intraspecific variability may constitute the limit of this approach. Some species like *M. kansasii* include several types, of which several are genetically close to *M. gastri*.

Conclusions

The arrival of molecular tools, particularly the analysis of 16S rRNA gene sequences, deeply modified the taxonomic approach. Today, even this tool shows limits because of the low polymorphism of the 16S rRNA. It appears essential to propose new alternatives. We have thus developed a multi-gene sequence database incorporating four genes (16S rRNA, *hsp65*, *rpoB*, *sod*) within the genus *Mycobacterium*. The sequential and/or combined use of sequences of several genes easily leads to the refinement of the phylogenetic approach, to understand better the complexity of evolution within bacterial groups in order to identify and characterize novel species. This multigenic approach is consistent with the recent recommendations of the ad hoc committee for the re-evaluation of the definition of the bacterial species (Stackebrandt *et al.*, 2002). The concatenation of congruent genes provides reliable phylogenetic trees, more discriminatory and robust. The final phylogenetic tree agrees with pre-existing, generally accepted

phylogenetic relationships; in particular, the partition between slowly and rapidly growing mycobacteria is strongly supported. The genes studied, with different locations on the whole genome, evolved in the same way. These results encourage us to follow this strategy in order to refine the phylogeny. In this study, only type strains were integrated (with one exception), which does not allow intraspecies variability to be taken into account. Integration of a large number of strains within a species would be required for a better evaluation of the robustness of this approach.

ACKNOWLEDGEMENTS

We thank particularly Catherine Pichat for her assistance in this work.

REFERENCES

- Afghani, B. & Stutman, H. R. (1996). Polymerase chain reaction for diagnosis of *M. tuberculosis*: comparison of simple boiling and a conventional method for DNA extraction. *Biochem Mol Med* 57, 14–18.
- Blackwood, K. S., He, C., Gunton, J., Turenne, C. Y., Wolfe, J. & Kabani, A. M. (2000). Evaluation of recA sequences for identification of *Mycobacterium* species. *J Clin Microbiol* 38, 2846–2852.
- Brunello, F., Ligozzi, M., Cristelli, E., Bonora, S., Tortoli, E. & Fontana, R. (2001). Identification of 54 mycobacterial species by PCR-restriction fragment length polymorphism analysis of the hsp65 gene. *J Clin Microbiol* 39, 2799–2806.
- Cloud, J. L., Neal, H., Rosenberry, R., Turenne, C. Y., Jama, M., Hillyard, D. R. & Carroll, K. C. (2002). Identification of *Mycobacterium* spp. by using a commercial 16S ribosomal DNA sequencing kit and additional sequencing libraries. *J Clin Microbiol* 40, 400–406.
- Daubin, V., Gouy, M. & Perrière, G. (2002). A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res* 12, 1080–1090.
- De Beenhouwer, H., Liang, Z., De Rijk, P., Van Eekeren, C. & Portaels, F. (1995). Detection and identification of mycobacteria by DNA amplification and oligonucleotide-specific capture plate hybridization. *J Clin Microbiol* 33, 2994–2998.
- Devallois, A., Goh, K. S. & Rastogi, N. (1997). Rapid identification of mycobacteria to species level by PCR-restriction fragment length polymorphism analysis of the hsp65 gene and proposition of an algorithm to differentiate 34 mycobacterial species. *J Clin Microbiol* 35, 2969–2973.
- Domenech, P., Menendez, M. C. & Garcia, M. J. (1994). Restriction fragment length polymorphisms of 16S rRNA genes in the differentiation of fast-growing mycobacterial species. *FEMS Microbiol Lett* 116, 19–24.
- Galtier, N., Gouy, M. & Gautier, C. (1996). SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* 12, 543–548.
- Gingeras, T. R., Ghandour, G., Wang, E. & 7 other authors (1998). Simultaneous genotyping and species identification using hybridization pattern recognition analysis of generic *Mycobacterium* DNA arrays. *Genome Res* 8, 435–448.
- Guindon, S. & Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52, 696–704.
- Hamid, M. E., Roth, A., Landt, O., Kroppenstedt, R. M., Goodfellow, M. & Mauch, H. (2002). Differentiation between *Mycobacterium farcinogenes* and *Mycobacterium senegalense* strains based on 16S-23S ribosomal DNA internal transcribed spacer sequences. *J Clin Microbiol* 40, 707–711.
- Harmsen, D., Rothganger, J., Frosch, M. & Albert, J. (2002). RIDOM: ribosomal differentiation of medical micro-organisms database. *Nucleic Acids Res* 30, 416–417.
- Ji, Y., Colston, M. J. & Cox, R. A. (1994). The ribosomal RNA (*rrn*) operons of fast-growing mycobacteria: primary and secondary structures and their relation to *rrn* operons of pathogenic slow-growers. *Microbiology* 140, 2829–2840.
- Kasai, H., Ezaki, T. & Harayama, S. (2000). Differentiation of phylogenetically related slowly growing mycobacteria by their gyrB sequences. *J Clin Microbiol* 38, 301–308.
- Kim, B. J., Lee, S. H., Lyu, M. A., Kim, S. J., Bai, G. H., Chae, G. T., Kim, E. C., Cha, C. Y. & Kook, Y. H. (1999). Identification of mycobacterial species by comparative sequence analysis of the RNA polymerase gene (*rpoB*). *J Clin Microbiol* 37, 1714–1720.
- Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* 16, 111–120.
- Kirschner, P. & Bottger, E. C. (1998). Species identification of mycobacteria using rDNA sequencing. *Methods Mol Biol* 101, 349–361.
- Kolbert, C. P. & Persing, D. H. (1999). Ribosomal DNA sequencing as a tool for identification of bacterial pathogens. *Curr Opin Microbiol* 2, 299–305.
- Menendez, M. C., Garcia, M. J., Navarro, M. C., Gonzalez-y-Merchand, J. A., Rivera-Gutierrez, S., Garcia-Sanchez, L. & Cox, R. A. (2002). Characterization of an rRNA operon (*rrnB*) of *Mycobacterium fortuitum* and other mycobacterial species: implications for the classification of mycobacteria. *J Bacteriol* 184, 1078–1088.
- Niemann, S., Harmsen, D., Rusch-Gerdes, S. & Richter, E. (2000). Differentiation of clinical *Mycobacterium tuberculosis* complex isolates by gyrB DNA sequence polymorphism analysis. *J Clin Microbiol* 38, 3231–3234.
- Perrière, G. & Gouy, M. (1996). WWW-Query: an on-line retrieval system for biological sequence banks. *Biochimie* 78, 364–369.
- Richter, E., Niemann, S., Gloeckner, F. O., Pfyffer, G. E. & Rüsch-Gerdes, S. (2002). *Mycobacterium holsaticum* sp. nov. *Int J Syst Evol Microbiol* 52, 1991–1996.
- Ringuet, H., Akoua-Koffi, C., Honore, S., Varnerot, A., Vincent, V., Berche, P., Gaillard, J. L. & Pierre-Audigier, C. (1999). *hsp65* sequencing for identification of rapidly growing mycobacteria. *J Clin Microbiol* 37, 852–857.
- Robinson-Rechavi, M. & Huchon, D. (2000). RRTree: relative-rate tests between groups of sequences on a phylogenetic tree. *Bioinformatics* 16, 296–297.
- Rogall, T., Flohr, T. & Böttger, E. C. (1990). Differentiation of *Mycobacterium* species by direct sequencing of amplified DNA. *J Gen Microbiol* 136, 1915–1920.
- Roth, A., Fischer, M., Hamid, M. E., Michalke, S., Ludwig, W. & Mauch, H. (1998). Differentiation of phylogenetically related slowly growing mycobacteria based on 16S-23S rRNA gene internal transcribed spacer sequences. *J Clin Microbiol* 36, 139–147.
- Shinnick, T. M. (1987). The 65-kilodalton of *Mycobacterium tuberculosis*. *J Bacteriol* 169, 1080–1088.
- Springer, B., Stockman, L., Teschner, K., Roberts, G. D. & Bottger, E. C. (1996). Two-laboratory collaborative study on identification of mycobacteria: molecular versus phenotypic methods. *J Clin Microbiol* 34, 296–303.
- Stackebrandt, E. & Goebel, B. M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis

in the present species definition in bacteriology. *Int J Syst Bacteriol* **44**, 846–849.

Stackebrandt, E., Frederiksen, W., Garrity, G. M. & 10 other authors (2002). Report of the ad hoc committee for the re-evaluation of the species definition in bacteriology. *Int J Syst Evol Microbiol* **52**, 1043–1047.

Stahl, D. A. & Urbance, J. W. (1990). The division between fast- and slow-growing species corresponds to natural relationships among the mycobacteria. *J Bacteriol* **172**, 116–124.

Stinear, T. P., Jenkin, G. A., Johnson, P. D. & Davies, J. K. (2000). Comparative genetic analysis of *Mycobacterium ulcerans* and *Mycobacterium marinum* reveals evidence of recent divergence. *J Bacteriol* **182**, 6322–6330.

Suzuki, M. T., Beja, O., Taylor, L. T. & Delong, E. F. (2001). Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal marine bacterioplankton. *Environ Microbiol* **3**, 323–331.

Teichmann, S. A. & Mitchison, G. (1999). Is there a phylogenetic signal in prokaryote proteins? *J Mol Evol* **49**, 98–107.

Telenti, A., Marchesi, F., Balz, M., Bally, F., Bottger, E. C. & Bodmer, T. (1993). Rapid identification of mycobacteria to the species level by polymerase chain reaction and restriction enzyme analysis. *J Clin Microbiol* **31**, 175–178.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**, 4673–4680.

Tortoli, E. (2003). Impact of genotypic studies on mycobacterial taxonomy: the new mycobacteria of the 1990s. *Clin Microbiol Rev* **16**, 319–354.

Zolg, J. W. & Philippi-Schulz, S. (1994). The superoxide dismutase gene, a target for detection and identification of mycobacteria by PCR. *J Clin Microbiol* **32**, 2801–2812.