Infection, Genetics and Evolution 12 (2012) 827–831

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

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

Phylogeny of the genus *Mycobacterium*: Many doubts, few certainties

Enrico Tortoli *

Emerging Bacterial Pathogens Unit, San Raffaele Scientific Institute, Milan, Italy

ARTICLE INFO

Article history:
Available online 12 June 2011

Keywords:
Mycobacterium
Phylogeny
Genetic sequencing
Evolution
Multilocus phylogeny

ABSTRACT

The genus *Mycobacterium* is characterized by very limited interspecies genetic variability and this makes the definition of a robust phylogeny problematic. In this study a twofold phylogenetic approach was adopted. Phylogenetic trees were constructed using as targets the almost complete 16S rRNA gene sequences and the concatenated amino acid sequences coded by fragments of *hsp65* and *rpoB* genes. The comparison of the results made it possible to identify clusters of species sharing common phylogenetic pathway but for the majority of mycobacteria the definition of a robust phylogeny remained unreached.

1. Introduction

Phylogenetic studies, initially based on phenotypic similarities (Bergey et al., 1923), made an enormous qualitative leap, from the 1960s onwards, thanks to the dramatic progress of genetic knowledge and, to a minor extent, to the development of mathematical models for the evolutionary analysis of sequences' similarities. Not less important was the contribution of another extraordinary achievement of the modern age, the global network, which makes genetic sequences available in real time on public domain databases.

The 16S rRNA gene has been the first and, for many years, the sole target of genetic sequencing in bacteria. Several unique characteristics make this gene the ideal candidate for mutations analysis. The 16S rRNA gene, devoted to the essential function of protein synthesis, has been present in all organisms since the beginning of evolution. It is furthermore characterized by an evolutionary rate high enough to produce interspecies variability but, at the same time, by a degree of conservation sufficient to minimize the intraspecies variability (Peix et al., 2009). Ribosomal genes are furthermore characterized by an often underestimated unique feature; differently from protein coding genes, whose silent mutations do not undergo the screen of the natural selection, in them every mutation has evolutionary relevance.

GenBank and the partners of the International Nucleotide Sequence Database Collaboration (http://www.insdc.org/) contain nowadays the 16S rDNA genetic sequences of thousands of mycobacterial strains; for many of them, including the type strains of every officially recognized species, the sequence of the almost complete gene is even available.

It is not surprising that almost all phylogenetic studies concerning the genus *Mycobacterium* (Rogall et al., 1990; Stahl and Urbance, 1990; Tortoli, 2003) are based on the 16S rDNA and in particular on the first 500 bp which include its major hypervariable regions. A nonnegligible limit of such phylogenetic reconstructions is however the poor robustness of the trees.

More recently the multilocus approach (Stackebrandt et al., 2002) has been used in two large studies (Devulder et al., 2005; Mignard and Flandrois, 2008) in which four and seven different nucleotide fragments, respectively, were concatenated.

The goal of present study, far from expecting to resolve the discrepancies present in the literature facing *Mycobacterium* phylogenesis, was to produce, targeting the amino acid residues composition instead of nucleotides, further data for future discussion. To increase its reliability two housekeeping genes were selected and concatenated. The idea of including other genes to improve the robustness of the tree, was set aside as the shortage, in the database, of the sequences related to genes other than the ones selected here would imply a dramatic cut of the number of species submitted to analysis. The results of such an approach were then compared, looking for similarities and diversities, with the ones that emerged from the analysis of the 16S rRNA gene.

2. Design and methods

The mycobacterial sequences used for the study were retrieved from GenBank.

The almost complete 16S rRNA gene sequences related to the type strains of different *Mycobacterium* species were collected; they were aligned using the Klustal W program (Thompson 1944) and were trimmed to the longest fragment available for all of them, which corresponded to a stretch of about 1400 bp spanning over the *Escherichia coli*-corresponding positions 53–1460.

* Address: Emerging Bacterial Pathogens Unit, San Raffaele Scientific Institute, San Michele Building, via Olgettina 60, 20132 Milan, Italy. Tel.: +39 02 26435684; fax: +39 02 26435183.
E-mail address: tortoli.enrico@hsr.it

1567-1348/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.meegid.2011.05.025

The hypervariable fragment of *hsp65* (Telenti et al., 1993), a highly conserved housekeeping gene coding for a 65 kDa protein involved in the folding, assembly and transport of proteins, was chosen because of the completeness of the respective database. The sequences of the above mentioned fragment obtained from the type strains of different mycobacterial species were retrieved from GenBank; once aligned, they were trimmed to 399 nucleotides (starting from *Mycobacterium tuberculosis*-homologous position 443) and the corresponding 133 codons were translated to protein sequence.

Among the remaining nucleotide-databases related to mycobacterial genes, which are widely incomplete, the choice fell upon *rpoB*, probably the target most investigated in recent years. Of such a housekeeping gene, which codes for the β-subunit of RNA polymerase, two different regions have been proposed as target of genetic sequencing in mycobacteria (Adékambi et al., 2003; Kim et al., 1999), the largest and more variable, starting at *Mycobacterium smegmatis*-corresponding position 2593, was chosen here. Such sequences were available in GenBank for type strains belonging to 88 mycobacteria only; once aligned and trimmed to the largest fragment available, their length turned out to be variable, in different species, from 687 to 657 nucleotides corresponding to 229–219 codons; the respective amino acid sequences were then determined.

For each of the 88 species whose *rpoB* and *hsp65* sequences were both available in GenBank the translated amino acid strings were concatenated in a single stretch. The *hsp65*, and the 16S rDNA, sequences of the species other than the 88 above were no further investigated.

The phylogenetic analysis was conducted separately on the two sets of 88 aligned sequences: the nucleotide one, corresponding to 16S rDNA, and the protein one corresponding to concatenated *rpoB* and *hsp65* fragments. To each of them was added the corresponding sequence of *Nocardia farcinica* which was used as the outgroup. Trees were constructed using the neighbour-joining method under the total gap remotion and Kimura's two-parameter substitution model (Kimura, 1980) and visualized using the MEGA 5 software package (Tamura et al., 2011). The bootstrap analysis was conducted on 1000 resamplings.

3. Results

The phylogenetic tree inferred from the sequences of the almost complete 16S rDNA was in agreement with information obtained previously (Fig. 1) on the basis of the first third of such gene (Rogall et al., 1990; Stahl and Urbance, 1990). Rapid growers were clearly separated from slow growers and within the latter, two major clusters were evidently detectable. A group of slowly growing mycobacteria including *Mycobacterium simiae* presented the helix 18 of rRNA 12 nucleotides shorter than the majority of others while the group including the members of *Mycobacterium terrae* complex presented the helix 18 two nucleotides longer. A further well defined cluster was the one including the species of the *Mycobacterium avium* complex (MAC). Among rapid growers, two major clusters included the species related to *Mycobacterium fortuitum* and *Mycobacterium chelonae* respectively, with other mycobacteria scattered in a number of minor branches. A cluster including thermotolerant rapidly growing mycobacteria, a group tentatively proposed on the basis of an extra cytosine et *E. coli*-homologous position 184 (Kirschner et al., 1993; Springer et al., 1996) was hardly recognizable. The tree was characterized by poor robustness with less than a quarter of the nodes supported by bootstrap > 80%.

The position of three species only, among the 88 investigated, revealed somehow unexpected: the slowly growing *Mycobacterium tusciae* clustered with rapid growers, *Mycobacterium triviale* was

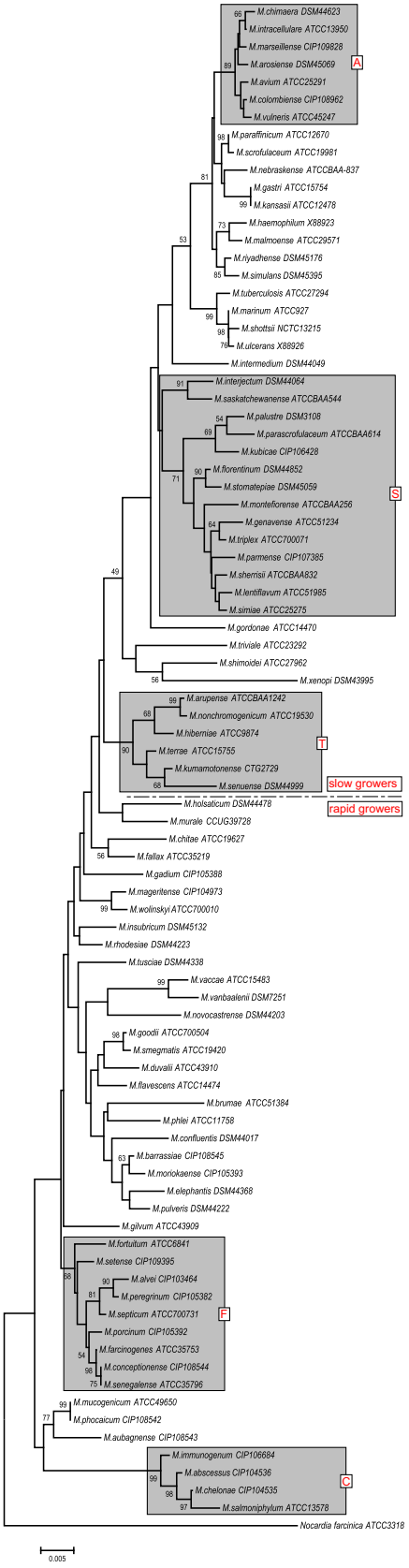

Fig. 1. Phylogenetic tree based on 16S rRNA gene almost complete sequences and constructed using the neighbour-joining method and the MEGA 5 package. The significance of branches (when >50) is indicated by bootstrap values calculated on 1000 replicates. Bar, 5 substitutions per 1000 nucleotides. Major clusters: (A) MAC; (C) *M. chelonae* group; (F) *M. fortuitum* group; (S) *M. simiae* group; and (T) *M. terrae* complex.