Towards a genome based taxonomy of Mycoplasmas

Cristiane C. Thompson [a,*], Nayra M. Vieira [b], Ana Carolina P. Vicente [a], Fabiano L. Thompson [b]

[a] Laboratory of Molecular Genetics of Microorganisms, Oswaldo Cruz Institute, FIOCRUZ, Rio de Janeiro, Brazil
[b] Department of Marine Biology, Institute of Biology, Federal University of Rio de Janeiro, UFRJ, Brazil

A B S T R A C T

Mycoplasmas are Gram-positive wall-less bacteria with a wide environmental and host distribution, causing disease in man and in (wild and farmed) animals. The aim of this study was to analyze the usefulness of a genomic taxonomic approach in Mycoplasma systematics. Multilocus Sequence Analysis (MLSA), average amino acid identity (AAI), and Karlin genomic signature allowed a clear identification of species. For instance, Mycoplasma pneumoniae and Mycoplasma genitalium had only 71% MLSA similarity, 67% AAI, and 88 for Karlin genomic signature. Codon usage (Nc) of the phylogenetically distantly related species Mycoplasma conjunctivae and Mycoplasma gallisepticum was identical, in spite of clear differences in MLSA, AAI, and Karlin, suggesting that these two species were subjected to convergent adaptation due to similar environmental conditions. We suggest that a Mycoplasma species may be better defined based on genomic features. In our hands, a Mycoplasma species is defined as a group of strains that share ⩾97% DNA identity in MLSA, ⩾93.9% AAI and ⩽8 in Karlin genomic signature. This new definition may be useful to advance the taxonomy of Mycoplasmas.

© 2011 Elsevier B.V. Open access under the Elsevier OA license.

1. Introduction

Mycoplasmas are one of the smallest and simplest prokaryotes, having only the minimal cellular machinery required for self-replication and survival. They appear to have evolved from Gram-positive bacteria by a process of degenerative evolution towards genome reduction and the loss of a cell wall (Woese et al., 1980; Neimark, 1986). Mycoplasmas are widespread in nature as parasites of humans, mammals, reptiles, fish, arthropods, and plants. They may be symbionts of isopods (Eberl, 2010), songbirds (States et al., 2009), wild and reared fish (Holben et al., 2002), and deep sea Lophelia corals of Gulf of Mexico and Norwegian Fjords (Kellogg et al., 2009). In spite of their reduced genomes, Mycoplasmas have highly versatile transcriptomes (with the production of different transcripts from a single gene and antisense ncRNAs) (Güell et al., 2009) and proteomes (with multifunctional enzymes, post-translational modifications, and signaling molecules) (Kühner et al., 2009). The metabolic network appears to be more linear in Mycoplasmas than in other more complex bacteria. The multifunctionality of enzymes in Mycoplasmas may be a consequence of gene loss and genome reduction. Sequence comparisons among complete Mycoplasma genomes indicated that the 580-Kb genome of Mycoplasma genitalium arose by minimization of the 816-Kb genome of the human respiratory pathogen Mycoplasma pneumoniae (Stein and Baseman, 2006).

Many Mycoplasma species are pathogenic for humans, animals, plants, and insects (Maniloff, 2002). In addition, Mycoplasmas have been a problem as intracellular contaminants in human cell therapy, and in the animal (poultry and swine farming) production as pathogens. Thus, rapid diagnostics and identification of Mycoplasmas is crucial for various activities. Due to the ecological importance of Mycoplasmas, several studies have been performed recently on the diversity of these microbes (Spergser et al., 2010). Online tools for the identification of Mycoplasma-related bacteria have also been developed (Zhao et al., 2009). Identification of Mycoplasmas is based on the current taxonomic standards (Stackebrandt et al., 2002). The species definition and identification for Mycoplasma is based on the general species definition using a combination of 16S rRNA gene sequence analyses, DNA–DNA Hybridization (DDH), serology and phenotypic data. Distinguishing Mycoplasmas by their metabolic characteristics has been generally of limited phylogenetic and taxonomic value. A few exceptions include the ability to ferment glucose, to hydrolyze arginine and urea, and the dependence on cholesterol for growth and anaerobiosis. Cellular localization of reduced nicotinamide adenine dinucleotide (NADH) oxidase activity has also been of help in classification of Mycoplasmas at and above the genus level (Razin, 2006).

Since the 1970s, serology has been established as the most important tool for defining Mycoplasma species (ICSB Subcommittee on the Taxonomy of Mycoplasmatales, 1972; ICSB Subcommittee on the Taxonomy of Mollicutes, 1979, 1995) (Brown et al., 2007). Serological characterization is one of the most widespread used approaches for the identification. These methods are in agreement with DNA–DNA hybridization data and with 16S rRNA

* Corresponding author. Tel./fax: +55 21 38658168.
  E-mail address: thompson@ioc.fiocruz.br (C.C. Thompson).

sequence data. Currently, there are 123 species described in the genus *Mycoplasma* (http://www.bacterio.cict.fr/index.html). Differentiation of closely related species using 16S rRNA gene sequence is very difficult because *Mycoplasma* species may have up to 100% 16S rRNA gene similarity. For instance, the pairs *M. pneumoniae* and *M. genitalium*, *Mycoplasma mycoides* and *Mycoplasma capricolum* have 98% and 99.8% 16S rRNA sequence similarity, respectively. Serology is also very cumbersome, requiring special reagents and the expertise of few international laboratories. The high genomic and phenotypic diversity of Mycoplasmas may also result in cross-reaction or unidentification based on serology.

The advent of rapid genome sequencing and automated genome analyses have opened up a new perspective for the taxonomy of Mycoplasmas. Genome sequences may allow rapid identification and evolutionary inferences. Several recent studies have used whole-genome analysis to determine the taxonomic relationships among bacterial species (Coenye and Vandamme, 2003, 2004; Richter and Rosselló-Móra, 2009; Thompson et al., 2009, 2011). In order to test the feasibility of the genomic taxonomy in Mycoplasmas, several genomic markers were analyzed in a collection of 46 genomes. The availability of whole genome sequences of several closely related species, such as *M. pneumoniae* and *M. genitalium* formed an ideal test case for the establishment of the genomic taxonomy of Mycoplasmas. Disclosing species-specific patterns for the different genome-wide markers would reinforce their usefulness in *Mycoplasma* taxonomy. The aim of this study was to analyze the usefulness of the genomic taxonomic approach in Mycoplasmas, using information of Multilocus Sequence Analysis (MLSA), average amino acid identity (AAI), Karlin genomic signature, codon usage and Genome-to-Genome distances (GGD) for species identification.

2. Material and methods

2.1. Genome sequence data

We analyzed 46 complete genome sequences of Mycoplasmas that were public available for download by June 2nd, 2011 in the National Center for Biotechnology Information (NCBI) under the project accession number indicated in Table 1. The following analyzes were performed according to Thompson et al. (2009) and are briefly described below.

2.2. 16S rRNA gene sequences and Multilocus Sequence Analysis (MLSA)

The 16S rRNA gene sequences and the gene sequences used for MLSA were obtained from GenBank (NCBI). MLSA approach was based on the concatenated sequences of four housekeeping genes (*adk*, *efp*, *gmk* and *gyrB*) that were used in other taxonomic studies. The concatenated sequences were aligned by CLUSTALX and MAFFT alignment methods. Phylogenetic analyses were conducted using MEGA version 5. The phylogenetic inference was based on the neighbor-joining genetic distance method (NJ), and the maximum likelihood method (ML). Distance estimations were obtained by the model of Kimura two-parameters. The reliability of each tree topology was checked by 2000 bootstrap replications (Felsenstein, 1985).

2.3. Average amino acid identity (AAI)

The AAI was calculated as described previously (Konstantinidis and Tiedje, 2005). Conserved genes between a pair of genomes were determined by whole-genome pairwise sequence comparisons using the BLAST algorithm (Altschul et al., 1997). For these comparisons, all protein-coding sequences (CDSs) from one genome were searched against the genomic sequence of the other genome. The genetic relatedness between a pair of genomes was measured by the average amino acid identity of all conserved genes between the two genomes as computed by the BLAST algorithm.

2.4. Determination of dinucleotide relative abundance values

We determined the dinucleotide relative abundance value for each genome. Mononucleotide and dinucleotide frequencies were calculated using COMPSEQ (EMBOSS). Dinucleotide relative abundances ($\rho^{*}XY$) were calculated using the equation $\rho^{*}XY = fXY/fXfY$ where fXY denotes the frequency of dinucleotide XY, and fX and fY denote the frequencies of X and Y, respectively (Karlin et al., 1997; Karlin, 1998). The difference in genome signature between two sequences is expressed by the genomic dissimilarity (δ^{*}), which is the average absolute dinucleotide of relative abundance difference between two sequences. The dissimilarities in relative abundance of dinucleotides between both sequences were calculated using the equation described by Karlin et al. (1997): $\delta^{*}(f,g) = 1/16\Sigma|\rho^{*}XY(f) - \rho^{*}XY(g)|$ (multiplied by 1000 for convenience), where the sum extends over all dinucleotides.

2.5. Codon usage bias

Codon usage bias was calculated for each genome. The effective number of codons used in a sequence (Nc) (Wright, 1990) was calculated using CHIPS (EMBOSS). Nc values range from 20 (in an extremely biased genome where only one codon is used per amino acid) to 61 (all synonymous codons are used with equal probability) (Wright, 1990).

2.6. Genome-to-Genome Distance Calculator (GGDC)

The genome distance was calculated using Genome-To-Genome Distance Calculator (GGDC) (Auch et al., 2010). Distances between a pair of genomes were determined by whole-genome pairwise sequence comparisons using BLAST (Altschul et al., 1990). For these comparisons, algorithms were used to determine high-scoring segment pairs (HSPs) for inferring intergenomic distances for species delimitation. The corresponding distance threshold can be used for species delimitation. Any distance value above the threshold can be regarded as indication that the two genomes analyzed represent two distinct species. Distances are calculated by (i) comparing two genomes using the chosen program to obtain HSPs/MUMs and (ii) inferring distances from the set of HSPs/MUMs using three distinct formulas. Next, the distances are transformed to values analogous to DNA–DNA Hybridizations (DDH). These DDH estimates are based on an empirical reference dataset comprising real DDH values and genome sequences. The regression-based DDH estimate uses parameters from a robust-line fit, whereas the threshold-based DDH estimate applies the distance threshold leading to the lowest error ratio in predicting whether DDH is ⩾70% or <70%.

3. Results

3.1. General features

The complete genomes of the Mycoplasmas comprised a single chromosome (Table 1). The estimated size of the *Mycoplasma* genomes ranged from 580 Kb (*M. genitalium*) to 1,360 Kb (*M. penetrans*). *Mycoplasma* genomes present low GC content. The average GC content of *Mycoplasma* genomes ranged from 21% to 40%. The