Systematic and Applied Microbiology 36 (2013) 539–548

Contents lists available at ScienceDirect

Systematic and Applied Microbiology

journal homepage: www.elsevier.de/syapm

Genome wide sequence analysis grants unbiased definition of species boundaries in "*Candidatus* Phytoplasma"

Giuseppe Firrao[a,b,*], Marta Martini[a], Paolo Ermacora[a], Nazia Loi[a], Emanuela Torelli[a], Xavier Foissac[c,d], Patricia Carle[c,d], Bruce C. Kirkpatrick[e], Lia Liefting[e,1], Bernd Schneider[f], Cristina Marzachì[g], Sabrina Palmano[g]

[a] *Dipartimento di Scienze Agrarie ed Ambientali, Università di Udine, Udine, Italy*
[b] *Istituto Nazionale di Biostrutture e Biosistemi, Interuniversity Consortium, Italy*
[c] *INRA, UMR-1332 Biologie du Fruit et Pathologie, Villenave d'Ornon, France*
[d] *Université Bordeaux Ségalen, UMR-1332 Biologie du Fruit et Pathologie, Villenave d'Ornon, France*
[e] *Department of Plant Pathology, University of California, Davis, CA, USA*
[f] *Julius Kühn-Institut, Dossenheim, Germany*
[g] *Istituto di Virologia Vegetale, CNR, Strada delle Cacce 73, Torino, Italy*

ARTICLE INFO

Article history:
Received 2 March 2013
Received in revised form 8 July 2013
Accepted 18 July 2013

Keywords:
Candidatus Phytoplasma
Genomics
Taxonomy
MLSA
ANI

ABSTRACT

The phytoplasmas are currently named using the *Candidatus* category, as the inability to grow them *in vitro* prevented (i) the performance of tests, such as DNA-DNA hybridization, that are regarded as necessary to establish species boundaries, and (ii) the deposition of type strains in culture collections. The recent accession to complete or nearly complete genome sequence information disclosed the opportunity to apply to the uncultivable phytoplasmas the same taxonomic approaches used for other bacteria. In this work, the genomes of 14 strains, belonging to the 16SrI, 16SrIII, 16SrV and 16SrX groups, including the species "*Ca.* P. asteris", "*Ca.* P. mali", "*Ca.* P. pyri", "*Ca.* P. pruni", and "*Ca.* P. australiense" were analyzed along with *Acholeplasma laidlawi*, to determine their taxonomic relatedness. Average nucleotide index (ANIm), tetranucleotide signature frequency correlation index (Tetra), and multilocus sequence analysis of 107 shared genes using both phylogenetic inference of concatenated (DNA and amino acid) sequences and consensus networks, were carried out. The results were in large agreement with the previously established 16S rDNA based classification schemes. Moreover, the taxonomic relationships within the 16SrI, 16SrIII and 16SrX groups, that represent clusters of strains whose relatedness could not be determined by 16SrDNA analysis, could be comparatively evaluated with non-subjective criteria. "*Ca.* P. mali" and "*Ca.* P. pyri" were found to meet the genome characteristics for the retention into two different, yet strictly related species; representatives of subgroups 16SrI-A and 16SrI-B were also found to meet the standards used in other bacteria to distinguish separate species; the genomes of the strains belonging to 16SrIII were found more closely related, suggesting that their subdivision into *Candidatus* species should be approached with caution.

Introduction

Phytoplasmas are a large and diverse group of bacterial plant pathogens that have not yet been cultivated *in vitro*. Since their discovery [11] the research on these bacteria has been dominated by the efforts toward their classification. With the introduction of PCR, a taxonomic approach based on the analysis of selectively amplified 16S rRNA genes permitted the definition of a phylogenetically sound classification. Currently, phytoplasmas can be subdivided according of their 16S rDNA sequence into 20 clades [55], or, in a more comprehensive scheme based on observed or virtual RFLP pattern of 16S rDNA, into more than 30 groups, several of which can be further subdivided into subgroups, thus comprising more than 100 so called subgroups lineages [69]. This 16S rDNA based classification scheme is readily applicable in identification and classification of novel strains, and has found very wide application in epidemiological studies. However, it cannot be translated into a binomial nomenclatural system as commonly used in bacterial systematic. Despite its enormous power in resolving phylogenies at

* Corresponding author at: Dipartimento di Scienze Agrarie ed Ambientali, Università di Udine, via Scienze 208, 33100 Udine, Italy. Fax: +390432558501.
E-mail address: firrao@uniud.it (G. Firrao).
[1] Present address: Plant Health and Environment Laboratory, Ministry for Primary Industries, Auckland, New Zealand.

0723-2020/$ – see front matter © 2013 Elsevier GmbH. All rights reserved.
http://dx.doi.org/10.1016/j.syapm.2013.07.003

the higher hierarchic levels, the 16S rDNA sequence alone is not sufficient to define and circumscribe species [14,58,66]. As phytoplasmas have not yet beeen cultivated *in vitro* they cannot be deposited in culture collections and cannot be investigated for some properties that are regarded as necessary for their classification at the species level [3,57]. Standing the current inability to comply with the minimal standard for their formal naming [3], the category of *Candidatus*, implemented out of the Bacteriological Code to record the properties of putative taxa of procaryotes [42], has been adopted for their nomenclature [60]. To date, several "*Candidatus* Phytoplasma" species have been described, in most cases congruently with the 16S rDNA based schemes. In several cases, however, phytoplasmas sharing high similarity in their 16S rDNA sequence have been described as different *candidatae* species due to their distinct biological properties. Although the subdivision into *candidatae* species has been agreed by a large panel of a dedicated working team [60], it implies a large degree of subjectivity due to the paucity and uncertain taxonomic significance of the properties that could be scored for the non cultivable phytoplasmas in contrast to other cultivable mollicutes.

In the 15 years lasted since the meeting when the basis of the classification scheme for phytoplasma were discussed, thanks to the large amount of studies on phytoplasmas their chromosomal and non chromosomal characteristics, their interaction with the plant and insect host, their epidemiology and spread, their biology had become much more clearly understood [12,21,30,50,59]. The *Candidatus* species had become more a solid reference than simply a method to record the properties of an organism known only for its ribosomal sequence. According to the paper concerning its implementation, the category of *Candidatus* apply to organisms that "can be recognized by their molecular structures but cannot be assigned to a known genus because of the lack of enough distinguishing characteristics. Formal recognition will come when new observations allow;" [42]. New observations actually accumulated for phytoplasmas. In view of the formal recognition of species of phytoplasmas, although not yet allowed due to the persisting inability to deposit a live type strain in a culture collection, the scientific community needs to be concerned that the criteria used for the definition of "*Candidatus* Phytoplasma" species meet the standards used for other bacteria. In recent times, the increased knowledge at the genome level of microorganisms opened unprecedented opportunities. As genome wide comparative sequence analysis is becoming the elective method to unveil the natural classification of prokaryotes, its application to phytoplasmas would prevent the use of arbitrarily chosen boundaries when delimitating *Candidatus* species and would help defining the *Candidatus* as a provisional status for formal species. If the very recent claim of axenic cultivation of phytoplasmas [8] will be confirmed and the deposition of strains in culture collections will become possible, then a congruence in classification with widely accepted taxonomic criteria would greatly facilitate the shift from the *Candidatus* to formal nomenclature.

In this study we analyzed the genomes of 14 different phytoplasmas and showed that the information extracted from their genome sequence can be used to critically evaluate the robustness of *Candidatus* species definition and contribute to the delineation of species boundaries with methods congruent with those used for formal species definition.

Material and methods

Sequence sources

Most of the genome sequence used for this work were retrieved from public databases (either EBI or Molligen). In addition, the unpublished genome draft sequences of the "*Ca.* P. asteris" strains L163 (S. Palmano and coworkers, unpublished) and CY (C. Marzachì and coworkers, unpublished), "*Ca.* P. pruni" strain WX (B.C. Kirkpatrick and coworkers, unpublished), "*Ca.* P. pyri" strain PD (B. Schneider and coworkers, unpublished), the flavescence dorée phytoplasma strains FD92 (X. Foissac and coworkers, unpublished) and FD-Piedmont (C. Marzachì and coworkers, unpublished) were included. The genome sequences analyzed in this work belong to phytoplasma that were assigned to different clades, groups and subgroups as detailed in Table 1 along with reference of other genome sequences used in this work.

Sequence alignments

In order to provide a solid alignment of DNA sequence a multistep procedure was set up with the development of a set of *ad hoc* Perl scripts. First the protein sequence of all genomes were compared in order to identify families of orthologs by a modification of the reciprocal smallest distance algorithm [48,65]: after establishment by reciprocal best BLASThits, families were evaluated using MAFFT [26] and CLUSTALW2 [34] generated alignments; the families that did not pass a quality check (*i.e.* with a mean < .7 or a standard deviation < .05 in the identity values calculated between all pairs of predicted proteins) were revised manually, splitting unconsistent families when necessary. Families that did not comprise one protein per each genome or that contained more than one protein for at least one genome (paralogs) were not retained. Then the alignments were split, sorted and re-merged in order to identify and exclude alignment regions that contained a number of gaps higher than a cutoff (10 gaps/50 aa. positions) and that could therefore be of uncertain alignment. The alignments of the remaining protein families were converted into DNA alignments of their corresponding ORFs. The alignments were analyzed individually and as a concatenated sequence for both protein and nucleic acid sequences independently. Alignment inspection and preliminary analyses were carried out with SEAVIEW [18].

Phylogenetic analysis

Several different phylogenetic analyses were carried out: for the 16SrRNA sequences of 15 *Acholeplasmatales* (14 phytoplasmas and one acholeplasma) 1406 nucleotide sites (200 informative sites) were included in the analysis. For the MLSA of 15 *Acholeplasmatales* based on a concatenation of 107 gene/protein sequences a total of 85,014 nucleotide sites (39,098 informative) or 28,338 a. a. sites (14,333 informative) were included in the analysis. For the MLSA inclusive of 27 bacteria (15 *Acholeplasmatales*, 10 *Mycoplasmatales* and 3 *Firmicutes*, 9642 a. a. sites (5862 informative) were included. The genes are listed in the supplementary table and grouped in functional categories in the supplementary figure.

Maximum likelihood analysis was carried out with PHYML [19], using GTR or LC as a substitution model for DNA and protein sequence analysis, respectively. Tree topologies were estimated using the better topology obtained using Nearest Neighbor Interchange (NNI) or Subtree Pruning and Regrafting (SPR). A most parsimonious tree was used as input tree. The support of the data for each internal branch of the phylogeny was estimated using non-parametric bootstrap with 1000 replicated for 16S rDNA analysis and 100 replicates for MLSA.

Ribosomal DNA and concatenated gene sequence data were also analyzed using split networks with the aid of the software SPLITSTREE4 [24]. Split networks are used to represent incompatible and ambiguous signals in a data set. The distance-based network used here, Neighbor net, is depicted as a tree with additional edges, so that the distance between two taxa is equal to the length of the