DOI: 10.22034/ijpp.2023.2005432.416

بیماری‌های گیاهی / جلد ۵۸ / شماره ۳/ سال ۱۴۰۱: ۲۰۷-۱۹۲

مقاله پژوهشی

ردیابی و واکاوی ناهنجاری در ترادف ژنهای آر ان ای ریبوزومی 16S استرینهای فیتوپلاسمایی

مجید صیام پور[1]**

(تاریخ دریافت: ۱۴۰۲/۰۴/۰۲؛ تاریخ پذیرش: ۱۴۰۲/۰۶/۰۵)

چکیده

تعیین خصوصیات و مشخصات فیتوپلاسماها در درجه اول براساس واکاوی ترادف ژن حفاظت شده آر ان ای ریبوزومی 16S (16S rRNA) انجام می گیرد. حتی تغییرات اندک در ترادف این ژن می تواند نشان دهنده رخدادهای تکاملی بسیار طولانی مدت باشد که همراه با آن خصوصیات اکولوژیکی در جمعیت باکتری ها نیز تغییر یافته است. وجود هر گونه خطا و ناهنجاری در ترادف ژن های 16S rRNA باعث بروز خطاهای بارز در تحلیل های بعدی از جمله مطالعات تبازرایی و تاکسونومی خواهد شد. در این مطالعه با ابزارهای بیوانفورماتیک ناهنجاری های معمول از جمله خطا در تعیین ترادف و یا تشکیل کایمر در ترادف ژن 16S rRNA مربوط به فیتوپلاسماهای نماینده از بیش از ۱۷۰ زیرگروه در ۴۰ گروه آر ان ای ریبوزومی 16S بررسی شد. نتایج امکان وجود چنین ناهنجاری هایی را در هشت استرین فیتوپلاسمایی تایید کرد. اغلب این استرین ها در درخت تبارزایی روی شاخه هایی که بطور غیر معمول طویل بودند قرار گرفتند. از بین استرین هایی که وجود ناهنجاری در ژن 16S rRNA آن ها ردیابی شد می توان به استرین های مرجع مربوط به 'Candidatus Phytoplasma wodyetiae'، 'Candidatus Phytoplasma allocasuarinae' و 'Candidatus Phytoplasma lycopersici' و نیز استرین های نماینده گروههای ریبوزومی 16SrXXVI و 16SrXXVII اشاره کرد. نتایج این مطالعه همچنین پیشنهاد کرد که ناهنجاریهای مربوط به این ژن در فیتوپلاسماها احتمالاً محدود به هشت استرین مشخص شده در این مطالعه نیست.

کلمات کلیدی: کایمر، خطای تعیین ترادف، استرین مرجع، گروههای آر ان ای ریبوزومی

---

* بخشی از طرح تحقیقاتی نویسنده در دانشگاه شهرکرد

** مسئول مکاتبات، پست الکترونیکی: siampour@sku.ac.ir

۱ دانشیار، گروه گیاهپزشکی، دانشکده کشاورزی، دانشگاه شهرکرد.

DOI: 10.22034/ijpp.2023.2005432.416

Research Article

Detection and Characterization of Anomalies in the 16S rRNA Gene Sequence of Phytoplasma Strains

Majid Siampour[1]**

(Received: 23.06.2023; Accepted: 27.08.2023)

Abstract

Characterization of phytoplasmas is primarily based on sequence analysis of their highly conserved 16S rRNA gene sequences. Even minor changes in the sequence of 16S rRNA gene can elucidate long-term evolutionary events along with modification of ecological characteristics within bacterial communities. The presence of any error and anomalies in 16S rRNA gene sequences can confound downstream analyses such as taxonomy and phylogenetic analyses. The most common sequence anomalies in the bacterial 16S rRNA gene sequences are chimera and sequencing errors. This investigation employed bioinformatics tools to examine the presence of such anomalies in the 16S rRNA gene sequence of representative phytoplasma strains from more than 170 subgroups within 40 16Sr groups. The findings suggested that the 16S rRNA gene sequences of eight phytoplasma strains contained anomalies, characteristics of chimeras, or some sorts of sequencing errors. Most of these strains were resolved on atypically elongated branches in the phylogenetic tree. The most notable strains with likely anomalous 16S rRNA gene sequences were reference strains of '*Ca.* P. wodyetiae, '*Ca.* P. allocasuarinae' and '*Ca.* P. lycopersici' as well as representative strains of the 16Sr groups XXVI and XXVII. The findings of this study suggest that anomalous 16S rRNA gene sequences are probably not restricted to the eight strains detected by the bioinformatics tools employed in this study.

Keywords: Chimera, Sequencing error, Reference strain, 16S r groups

[*]This article is part of the author's research project at Shahrekord University

[**] Corresponding author's E-mail: siampour@sku.ac.ir

1. Associate Prof. of Plant Phathology, Department of Plant Protection, Shahrekord University, Shahrekord, Iran

Introduction

Phytoplasmas constitute a monophyletic group of plant pathogens confined to the phloem; their transmission is facilitated by specific homopteran insects in a persistent propagative manner. Given the unattainability of axenic phytoplasma cultures, their recognition relies mostly upon the analysis of conserved gene sequences (Hogenhout *et al.* 2008; Bertaccini *et al.* 2022). Central to the characterization of phytoplasmas is the scrutiny of the 16S rRNA gene sequence, among other conserved genes. Even subtle variations in this sequence can reveal prolonged evolutionary divergence and ecological uniqueness. (Ochman *et al.* 1999). Despite considerable diversity, all phytoplasmas are unified under the single genus designation '*Candidatus* Phytoplasma'. Current guidelines stipulate the recognition of a novel '*Ca*. Phytoplasma' species when its 16S rRNA gene sequence exhibits less than 98.65% similarity with the 16S rRNA gene sequences of existing '*Ca*. Phytoplasma' species and their related strains. To date, the definition of 49 distinct '*Ca*. Phytoplasma' species has been achieved, primarily through the 16S rRNA gene sequence identity scores (Bertaccini *et al.* 2022). Additionally, phytoplasmas have been classified into 39 16Sr groups and more than 170 subgroups by their distinct 16S rRNA gene sequence RFLP profiles (Bertaccini and Lee 2018; Wei and Zhao 2022). Each '*Ca*. Phytoplasma' species, as well as each 16Sr group/subgroup, is demarcated by a single 16S rRNA gene sequence derived from a reference or representative phytoplasma strain (Bertaccini *et al.* 2022; Wei and Zhao 2022).

Numerous sequence abnormalities have been identified within bacterial 16S rRNA genes stored in repository databases. For instance, a study conducted by Ashelford *et al*. (2005) revealed that approximately 5% of the 16S rRNA gene sequences present in public databases (as of 2005) exhibited various anomalies. Failing to adequately account for these significant anomalies can mislead our understanding of bacterial taxonomy, diversity, and underlying evolutionary processes. Two notable types of sequence anomalies are chimeras and sequencing errors. Chimeras are sequences composed of distinct sequence fragments that don't naturally exist. Typically, chimeric sequences arise during PCR reactions contaminated with different but closely related sequence templates (Paabo et al., 1990; Sze and Schloss, 2019). The incidence of chimera formation in specific PCR amplifications of gene libraries could be up to 30% or even higher; thereby posing serious challenges to subsequent analyses (Wang and Wang, 1996; Ashelford et al., 2005). In addition to chimeras-, the repository databases of 16S rRNA gene sequences also frequently contain sequencing errors, arising from deficient sequencing methodologies or insufficient sequencing repetitions (Sze and Schloss, 2019; Scholes *et al.* 2011). Detecting these anomalies, particularly chimeras, poses a formidable task, prompting the development of various computational methods to facilitate their identification. In this study, the 16S rRNA gene sequences of the reference or representative phytoplasma strains were analyzed to ascertain the presence of chimeras and sequencing errors. In this regard, two software programs, Mallard and Pintail, originally designed for anomaly detection in 16S rRNA gene sequences, were employed sequentially (Ashelford *et al.* 2005, 2006). The limitation of the method used for the detection of anomalies is also discussed.

Methods

Dataset

The 16S rRNA gene sequences of 173 phytoplasma strains, representing 172 subgroups within 39 established 16Sr groups (Bertaccini and Lee 2018; Wei and Zhao 2022), were retrieved (in 5'to 3' orientation) from the NCBI database (www.ncbi.nlm.nih.gov) (Table S1).

Sequence alignment and phylogenetic tree reconstruction

Pairwise and multiple sequence alignments were performed using CLUSTAL W program implemented in MEGA 11 software package (Tamura *et al.* 2021). The alignment was used to reconstruct a Neighbor-Joining phylogenetic tree

in MEGA 11. The topology of the phylogenetic tree was assessed through a bootstrap analysis of 1000 replicates.

Variable and conserved regions of sequences

The phytoplasma 16S rRNA gene sequences were analyzed to identify the variable and conserved regions using an entropy-based approach in the DAMBE 6 software (Xia 2017). A sliding window size of 50 bp was used to plot the entropy values of sequence variation along the sequence. Sites containing gaps were excluded from the window length.

Detection and characterization of anomalies

Mallard and Pintail programs (Ashelford *et al.* 2005; 2006) were employed to analyze the phytoplasma 16S rRNA gene dataset for potential chimeras and sequencing errors. By conducting pairwise comparisons of each sequence against all others in the dataset a DE value (Deviation from the Expectation) was generated for each comparison. Plotting the DE values against their related sequence difference allowed creating a global plot of DE values (%) versus sequence differences (%). By applying a cut-off value of 95% or higher for DE values, the outlier sequences were identified. Subsequently, BLASTn (https://www.ncbi.nlm.nih.gov) and Pintail program were used to screen each outlier (potential anomalous sequence) for error (false positives). In this regard, the following approach was adopted. Initially, BLASTn was utilized to identify the nearest neighbor sequence (subject sequence) to the potential outlier (query sequence). Comparisons were then made to ensure that the subject sequence was reliable (Ashelford *et al.* 2006). Lastly, the subject sequence was compared with the query sequence using the Pintail program. The Pintail-generated plot was evaluated to identify any anomalies along the query sequence. The position of potential breakpoints in the anomalous sequence was estimated using Pintail and BLASTn.

Prediction of 16S rRNA secondary structure preservation

A template-based approach was utilized to predict base-pairings in 16S rRNA gene sequences, relying on the secondary structure of the *Escherichia coli* 16S rRNA gene as the model (Sweeney *et al.* 2021). This methodology was employed to assess the preservation of the secondary structure resulting from mutations in the 16S rRNA gene sequence of '*Ca*. P. caricae'; one of the phytoplasmas exhibiting unusual phylogenetic branching (Kirdat *et al.* 2023).

Results and Discussion

Dataset characteristics and phylogenetic tree

Among the 40 phytoplasma 16Sr groups (172 subgroups), 24 groups are comprised of only one subgroup (groups 16SrVIII, 16SrXVI-XXI, 16SrXXIII-XXXI and 16SrXXXIII- XL) (Table S1). The '*Ca*. Phytoplasma' species designated in the majority of these groups were considered 'orphan species', signifying the lack of any related strains (Kirdat *et al.* 2023). Moreover, some delineated 16Sr groups including 16SrXXIII to 16SrXXVIII, XXXIV, and XXXV had no named '*Ca*. Phytoplasma' species (Table S1). The 16S rRNA gene sequences of 173 phytoplasma strains were used to infer a comprehensive phylogenetic tree of phytoplasma strains. As shown in Fig. 1, the resulting tree revealed three major clades (I-III). The majority of the phytoplasma 16Sr groups that had only one subgroup were resolved within clade I. As shown, the branching pattern displayed certain '*Ca*. Phytoplasma species (e.g., '*Ca*. P. wodyetiae', '*Ca*. P. lycopercici', '*Ca.* P. graminis' and '*Ca*. P. caricae'), along with representative strains of various 16Sr groups (e.g., 16SrXXV, 16SrXXVI, 16SrXXVIII and 16SrXII-I) appeared to be atypical. This is characterized by elongated terminal branches diverging from the ancestral node. The uncertain phylogenetic position and atypical branching pattern of these '*Ca*. Phytoplasma' species were also speculated elsewhere (Kirdat *et al*. 2023). Several factors

including missing sequences for certain taxa (Darriba *et al.* 2016), evolutionary rate variation, and error in sequence data (including chimeric sequences) could contribute to this atypical branching in the phylogenetic trees (Fukatsu *et al.* 2007; Mai and Mirarab 2017).

Sequence variability along the phytoplasma 16S rRNA gene sequences

Fig. 2 depicts the variability pattern of phytoplasma 16S rRNA gene sequences, as determined by estimating the entropy variation with respect to base position. As the variability at each nucleotide position increases, so does the estimated entropy, and vice versa. The estimated entropy plot encompassed eight peaks corresponding to the hypervariable regions of the phytoplasma16S rRNA gene sequences. As revealed, the variable and conserved regions were well distributed throughout the sequence. The phytoplasma 16S rRNA gene sequence variability pattern was comparable with that in other bacteria (Ashelford *et al.* 2005).

Detection of anomalies

Using the dataset of 173 16S rRNA gene sequences, with that of OAY (M30790) utilized as the reference, a total of 14706 DE values were generated and plotted against their corresponding genetic difference. By a cut-off value of 95%, the majority of DE values were suitably clustered together; however, certain sequences were found to be outliers (Fig. 3A). These outliers were 16S rRNA genes of phytoplasma strains representing 16Sr groups/subgroups I-Y, I-AD, III-N, VI-E, XXVI-A, XXVII-A, XXXIII-A and XXXVI-A (Table1; Figs 4B-D). According to the literature, no other phytoplasma strains have been assigned to these eight 16Sr groups/subgroups. Moreover, the majority of these phytoplasma strains exhibited an unusually elongated branch on the phylogenetic tree, which may be attributed to the anomalies in their 16S rRNA gene sequences (Fig. 1). The two notable anomalous sequences were the reference strains of '*Ca*. P. lycopersisi' (strain THP, subgroup 16SrI-Y, accession no. EF199549) and '*Ca*. P. wodyetiae' (strain FPYD Bangi-2, subgroup 16SrXXXVI-A, accession no. KC844879), which were anomalous even when the conservative cut-off of 100% was applied (Table 1; Figs 3B-C).

Chimeric sequences

Using Pintail and BLASTn programs, the eight potentially anomalous sequences were further compared to their reliable nearest neighbor sequences (subject sequences). Results showed that four of the eight sequences representing 16Sr groups XXXVI (KC844879), XXXIII (AY135523), XXVI (AJ539179), and XXVII (AJ539180) were chimeric, i.e., composed of fragments from distinct phytoplasma strains (Table 1; Fig. 4). The 16S rRNA gene sequence of '*Ca*. P. wodyetiae' (KC844879; 16SrXXXVI) was a two-fragment chimera made up of major and minor parents belonging to the subgroups 16SrXIV-A ('*Ca*. P. cynodontis') and 16SrI-B ('*Ca*. P. asteris') (Figs. 4A and 4B; Table 1). '*Ca*. P. wodyetiae' was detected and identified in the foxtail palm plant, co-infected with strains related to these parental phytoplasmas (Naderali *et al.* 2017). This is consistence with identification of this phytoplasma 16S rRNA gene as a chimera.

The 16S rRNA gene sequence of the '*Ca*. P. allocasuarinae' reference strain AlloY (AY135523; subgroup16SrXXXIII-A) was a potential bipartite chimera. The AlloY phytoplasma strain was detected and identified in *Allocasuarina muelleriana* in Australia (Gibb *et al.* 2003; Marcone *et al.* 2004). The best matches for major and minor parents were '*Ca.* P. rhamni' (AJ583009; 16SrXX-A) and '*Ca.* P. australasia' (Y10096, 16SrII-D) (Figs. 4C and 4D). The breakpoint was approximated in the nucleotide positions ~770 (corresponding to position ~ 1145 of the OAY phytoplasma; Table 1). There were 11 and 35 SNPs between the 16S rRNA gene sequences of '*Ca*. P. allocasuarinae' and '*Ca.* P. rhamni' at the 5' and 3' regions of the breakpoint, respectively. On the contrary, there was only one SNP between the 16S rRNA gene sequences of the '*Ca.* P. allocasuarinae' and '*Ca.* P. australasia' at the 3' region of the breakpoint.

Table 1: Phytoplasma strains with potential anomalies in their 16S rRNA gene sequences detected by unusually high DE values (difference from Expectation) in Mallard software, with three cut-offs found within a library of 173 phytoplasma strains.

GenBank query accession	Name/16Sr subgroup classification	Associated disease/Strain	Highest DE difference (Cut off 95-100)$	No. of outliers (Cut off 95)#	Location of Anomaly relative to OAY strain (base position)	Description
EF199549	'*Ca.* P. lycopersici/ 16SrI-Y	Tomato "brote grande"/ THP	2.21***	83	179-451	Likely anomalous near 5' end; likely sequencing error; Major parent: '*Ca*. P. asteris' subgroup IB
KC844879	'*Ca.* P. wodyetiae'/16SrXXXVI-A	foxtail palm yellow decline/ Bangi-2	1.61***	37	952 (or 846)†	Likely two fragment chimera with 5' and 3'ends originated from '*Ca*. P. cynodontis' subgroup 16SrXIV-A and *Ca*. P. asteris' subgroup IB, respectively
AJ539179	No new species described/XXVI-A	Mauritius sugar cane yellows/ D3T1	0.91**	37	~420- ~705	Likely three fragment chimera with middle fragment originated from *Ca*. P. asteris' subgroup IB, and side fragments derived from '*Ca*. P. palmae' subgroup 16SrIV-A
AY135523	'*Ca.* P. allocasuarinae'/ XXXIII-A	allocasuarina yellows/AlloY	0.63**	8	~1145	Likely two fragment chimera with 5' and 3' ends more similar to '*Ca*. P. rhamni'subgroup 16SrXX-A and '*Ca*. P. australasia' subgroup 16SrII-D, respectively
AJ539180	No new species described /XXVII_A	Mauritius sugar cane yellows/D3T2	0.36*	7	487 (or 542)*	Likely two fragment chimera with 5' and 3'ends similar to '*Ca*. P. asteris'16SrI-B and '*Ca*. P. palmae'16SrIV-A, respectively
DQ286577	16SrI-AD	Basil (*Ocimum basilicum*) little leaf	0.33**	7	51, 649-675	Potential anomaly at 5'-end and potentially in ~30 nucleotides in the middle, likely sequencing error
AY270156	16SrVI-E	*Centaurea solstitialis* Virescence/ CSVI	0.09*	2	~365	Potential anomaly near 5'end, likely sequencing error
GU004365	16SrIII-N	Potato purple top/ AKpot6	0.13*	5	892-907	A 16 bp deletion in conserved region. Likely sequencing error

$: the query sequence was anomalous with defined cut-offs 100% (****); 99% (**) and 95% (*)

number of sequences in the dataset which generated unusually high DE values in pairwise comparison with the query sequence

†nucleotide sequences between these positions were conserved in parental strains

The two other chimeric sequences were 16S rRNA gene accessions AJ539179 and AY539180 representing the 16Sr groups XXVI and XXVII, respectively (Figs 4E-4H). Both of these sequences were derived from phytoplasmas detected in sugarcane (Wei *et al.* 2007). Comparison of the accession AJ539179 with the 16S rRNA gene of '*Ca*. P. palmae' (16SrIV-A, AF498307), as the subject sequence, revealed anomalies in its middle region between base positions ~270-560 (Table 1; Figs 4G and 4H). While the side fragments matched the 16S rRNA gene sequence of '*Ca*. P. palmae' the middle fragment matched that of '*Ca*. P. asteris'. Such sequence anomaly was suggestive of three-fragment chimeras in which the middle and side fragments match different subject sequences. The sequence accession AY539180 was a two-fragment chimera composed of partial 16S rRNA gene sequences from '*Ca*. P. palmae' and '*Ca*. P. asteris' (Table 1; Figs. 4E and 4F). Designation of the two phytoplasma groups 16Sr XXVI and 16SrXXVII has been based solely on the virtual RFLP analysis of these two potentially anomalous sequence records (Wei *et al.* 2007; Bertaccini and Lee 2018).

Sequencing errors

The remaining four anomalous sequences (accession nos. EF199549, DQ286577, AY270156, and GU004365) with biased DE values revealed anomalies that could be caused by sequencing errors (Table 1; Fig. 5). Most sequencing errors occur at the beginning and ends of the sequence reads, with no significant similarity with other reliable sequences. (Ashelford *et al.* 2005). Among these four sequences, the most anomalous was 16S rRNA gene of '*Ca*. P. lycopersici' (EF199549; subgroup16SrI-Y) (Arocha *et al.* 2007). Analyses showed that this sequence record was anomalous at base positions 174-144 with only 88% similarity with the subject sequence from '*Ca*. P. asteris'. The rest of the sequence was highly similar between the query and the subject sequences. (Fig 5A). In other words, a significant sequence variability was concentrated in a small part of the 16S rRNA gene sequence, which is in contrast with the sequence variability pattern observed along the phytoplasma 16S rRNA gene sequences (Fig. 2).

The 16S rRNA gene sequence accession DQ286577, representing the subgroup 16SrI-AD, revealed signals of anomalies in the first 50 nucleotides at the 5'-end and also in base positions 630-666. The rest of the sequence was nearly identical to 16S rRNA gene of '*Ca*. P. asteris' (Fig. 5B).

The phytoplasma strain of the subgroup 16SrVI-E (AY270156) was also predicted to contain errors at the 5'-end (370 bases) of its 16S rRNA gene sequence. Comparison with the subject sequence from subgroup 16SrVI-D (AF228053) revealed significant sequence disparity at the 5'-end of their 16S rRNA gene (Fig. 5C). Finally, analyses revealed that the potential anomaly in the phytoplasma strain of subgroup III-N (GU004365) was due to a unique 16-nucleotide deletion in a highly conserved region of its 16S rRNA gene sequence (corresponding to bases 892-907 of OAY strain; Fig. 2) (Table1, Fig 5D).

Conservation of 16S rRNA gene mutations in secondary structure

No anomalies were found in the 16S rRNA gene sequence of several phytoplasma strains, despite their unusual terminal branching in the phylogenetic tree. '*Ca*. P. caricae' (AY725234; 16SrXVII-A.) and '*Ca*. P. graminis' (AY725228; 16Sr XVI-A) (Kirdat *et al.* 2023), as well as the representative strain of the group 16SrXXVIII (AY744945) are among others (Fig. 1). These three phytoplasma strains were reported from Cuba, however, they were neither highly similar in their 16S rRNA gene sequence nor shared a close common ancestor. Among these strains, only a few closely related strains to '*Ca*. P. graminis' were reported (from Cuba by the same authors). Multiple alignment and BLAST analyses revealed a unique 6-25 nucleotide insertion in the 3' region of their 16S rRNA gene sequences positioned in a conserved region (corresponding to base positions 1347-1367 of OAY). The 16S rRNA gene sequence of '*Ca*. P. caricae' and that of the subject strain '*Ca*. P. solani' (16SrXII-A, KF751387) were compared at the secondary structure level. Results showed that the inserted nucleotides in '*Ca*. P. caricae' 16S rRNA gene were placed in a small loop region of a hairpin; thereby elongating the loop by 25 nucleotides (Fig. 6). This small loop was found to be conserved, also present in the 16S

rRNA gene sequence of *E.coli* (Gray *et al.* 1984). Moreover, a hairpin structure at base positions 1039-1055 of '*Ca*. P. caricae' was less preserved than in '*Ca*. P. solani' (i.e., included more non-canonical unstable base pairings) (Fig. 6). Further investigations are required to ascertain the impact of these structural modifications on the function of 16S rRNA gene in '*Ca*. P. caricae'. Similarly, the 16S rRNA gene representing the subgroup 16SrXXVIII (AY744945) contained mismatches and sequence insertions resulted in significant modifications of the secondary structure (data not shown). Re-amplification and direct sequencing could rule out the presence of any sequencing errors in the 16S rRNA gene sequences of these reference strains.

Further Limitations of Sequence Anomaly Detection

Eight of 173 16S rRNA gene sequences evaluated in this study displayed indications of anomalies that were discernible by the Mallard and Pintail software. However, it appeared that the methods of this study were not capable of identifying all potential chimeras or sequencing errors. As an illustration, an *in silico* chimeric sequence, created by joining the 5' and 3' halves of the 16S rRNA gene sequences from '*Ca*. P. australasia' (subgroup 16SrII-D, Y10096) and '*Ca.* P. aurantifolia' (subgroup 16SrII-B, U15442), couldn't be detected by the methodologies utilized in this study (data not shown). Moreover, available software may be incompetent for detecting sporadic nucleotide changes caused by sequencing errors. For instance, the 16S rRNA gene of '*Ca.* P. aurantifolia' reference strain WBDL (U15442) contained four unique base changes that were never confirmed in numerous other analyzed strains (Zreik *et al*. 1995; Siampour *et al.* 2019). The initial assignment of '*Ca.* P. aurantifolia' (WBDL phytoplasma) to the subgroup 16SrII-B was due to one of these potential sequencing errors occurring at the *Hpa*II recognition site (Lee *et al.* 1998). The sensitivity of the employed bioinformatics tools was low to validly identify such sparse sequencing errors.

In some 16S rRNA gene sequences with suspected anomalies, a comparative analysis of their secondary structure could be helpful to ensure that the concerned base modifications preserve stability of the secondary structure. Indeed, it is expected that the mutations in the 16S rRNA gene sequences will be conserved, i.e., located on non-conserved loops or on stems with mutual mutations, to preserve the stability of the structure (Pei *et al.* 2010). As shown, however, such criteria were not fully satisfied in the case of '*Ca*. P. caricae'. Altogether, it is reasonable to conclude that the anomalous phytoplasma 16S rRNA genes are not limited to the eight records reported in this study.

Among other strains with anomalous 16S rRNA genes, were the reference strains of three species including '*Ca*. P. lycopersici', '*Ca.* P. wodyetiae', and '*Ca*. P. allocasuarinae'. This shows the considerable adverse effects of 16S rRNA gene sequence anomalies on the description of new phytoplasma species. Moreover, the 16S rRNA gene sequence is highly conserved, with an average substitution rate of 1-2% per 50 million years (Ochman *et al.* 1999). Accordingly, the presence of anomalous 16S rRNA gene sequences among reference phytoplasma strains can extremely mislead the evolutionary inference. Hence, it is imperative to exercise the utmost methodological precautions, as outlined by Kim *et al*. (2014), to ensure accurate sequencing of this highly conserved gene. The overall findings of this study suggest a reevaluation of the 16Sr groups/subgroups or '*Ca*. Phytoplasma' species designated by potentially anomalous 16S rRNA gene sequences.

Note added upon submission of manuscript

Before this manuscript was submitted for publication, a paper was published on the identification of chimeric 16S rRNA sequences in almost all phytoplasma strains deposited in GenBank (Tiwarekar *et al*. 2023). Identification of the 16S rRNA gene from the reference strains of '*Ca.* P. wodytiae' and '*Ca.* P. allocasuarinae' agrees with results of the present study. Consistent with the results of this study, the atypical divergence of '*Ca.* P. graminis', '*Ca.* P. caricae' and '*Ca.* P. lycopersici' and the identification of atypical sequences in their 16S rRNA sequences are also shown.

Acknowledgment

This work has been financially supported by the research deputy of Shahrekord University .The grant number was 141/1402/25.

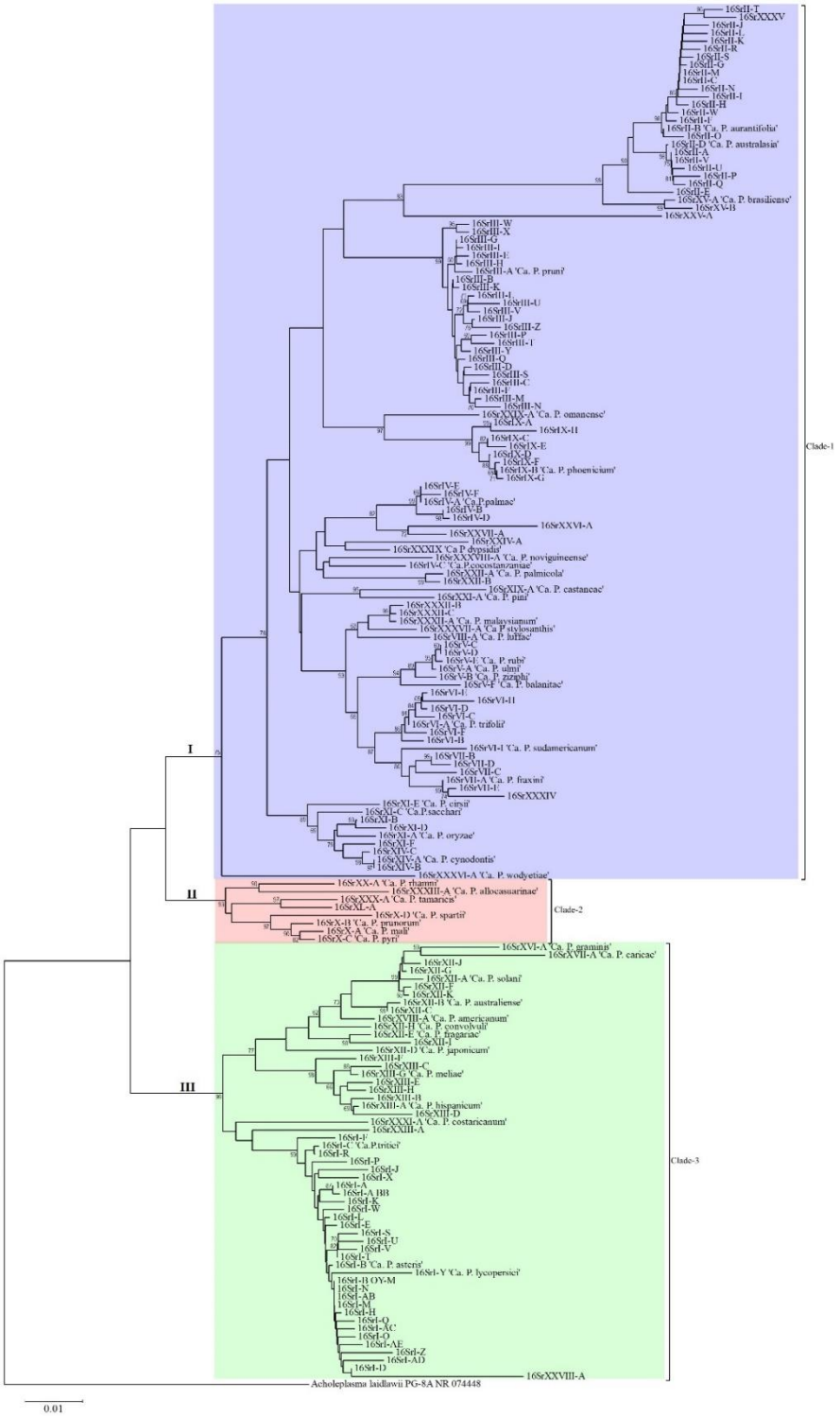

Fig. 1: Phylogenetic tree comprising 173 phytoplasma strains belonging to 40 16Sr groups which include the reference strains of 49 '*Ca*. Phytoplasma' species. The phylogenetic tree was inferred from the 16S rRNA gene sequence analysis using the Neighbor-Joining method in MEGA 11. The reliability of the tee topology was evaluated by bootstrap test with 1000 replicates. Bootstrap values of >60 are shown as percentage next to the branches. The 16S rRNA gene sequence of *Acholeplasma laidlawii* (strain PG-8A) was usedas outgroup to root the tree. The three main clades I-III of phytoplasma strains are shown on branches.

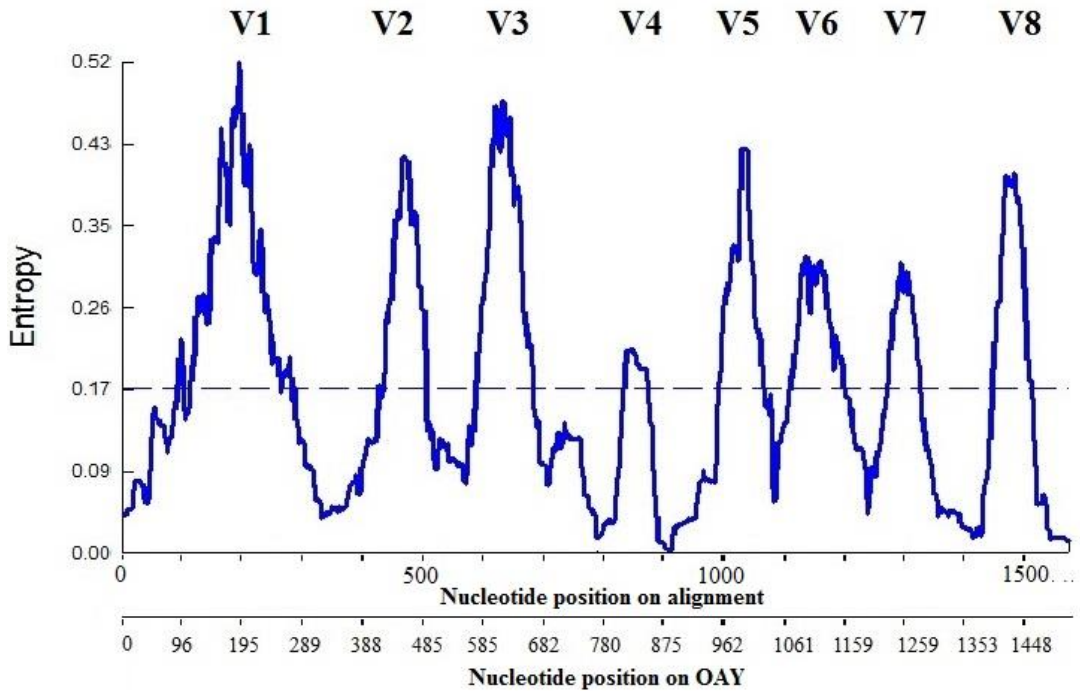

Fig.2: Entropy plot showing conserved and hypervariable regions of phytoplasma 16S rRNA gene sequences. Entropy values were calculated based on sequence variability on nucleotide positions in the alignment with a sliding window size of 50 nucleotides. Sites with alignment gaps were not counted in the window length. Hypervariable regions were defined as V1- V8. The global mean entropy value is shown by the dotted line. The nucleotide positions in the alignment and also with respect to the 16S rRNA gene of OAY phytoplasma (M30790) are shown.

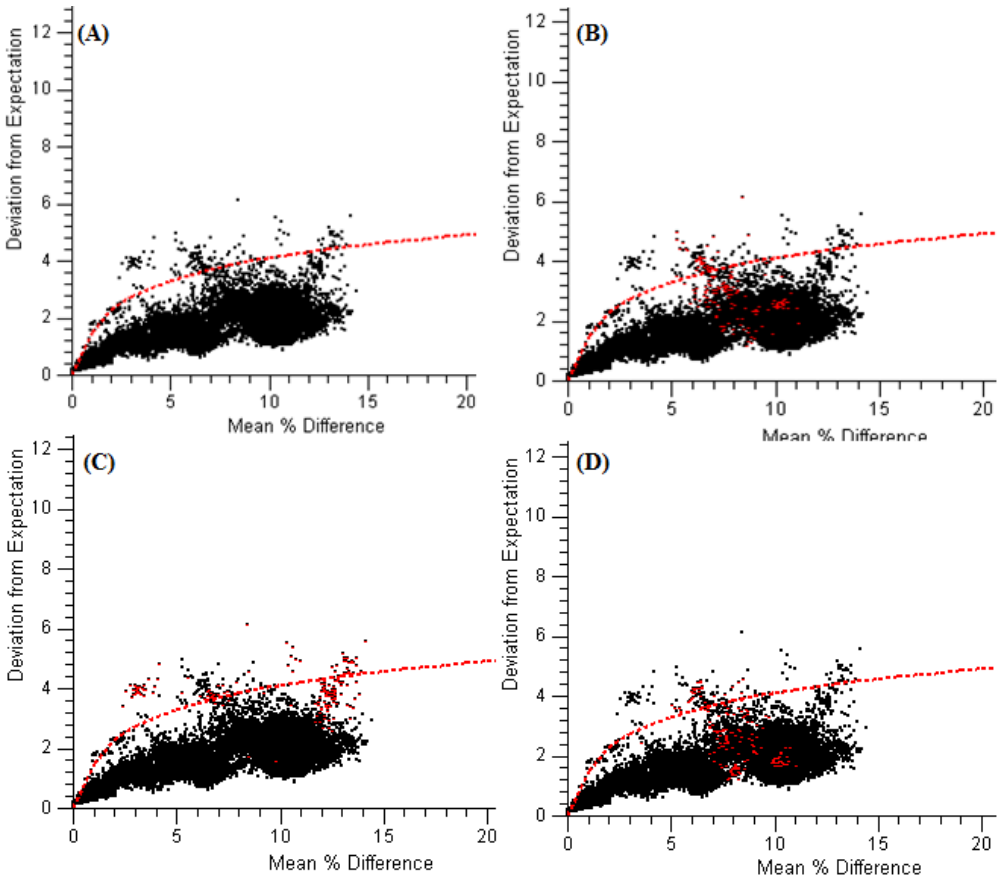

Fig.3. Deviation from the expectation (DE) values generated using Mallard program by pairwise comparisons of 16S rRNA gene sequences from 173 phytoplasma strains (OAY phytoplasma was used as reference strain). A) DE values from all 14706 pairwise comparisons were plotted against their corresponding sequence distance. The 95% cut-off value is shown as a dotted curve. Sequences with potential anomalies are those with unusually high DE values superimposed on the cut-off curve. The DE values shown in red dots in figures B, C and D are comparisons involving the 16S rRNA gene sequences of '*Ca.* P. wodyetiae' (16SrXXXVI-A, KC844879), '*Ca.* P. lycopersici' (16SrI-Y, EF199549) and representative phytoplasma strain of the subgroup 16SrXXVI-A (AJ539179), respectively. See Tables 1 and S1 for more details.

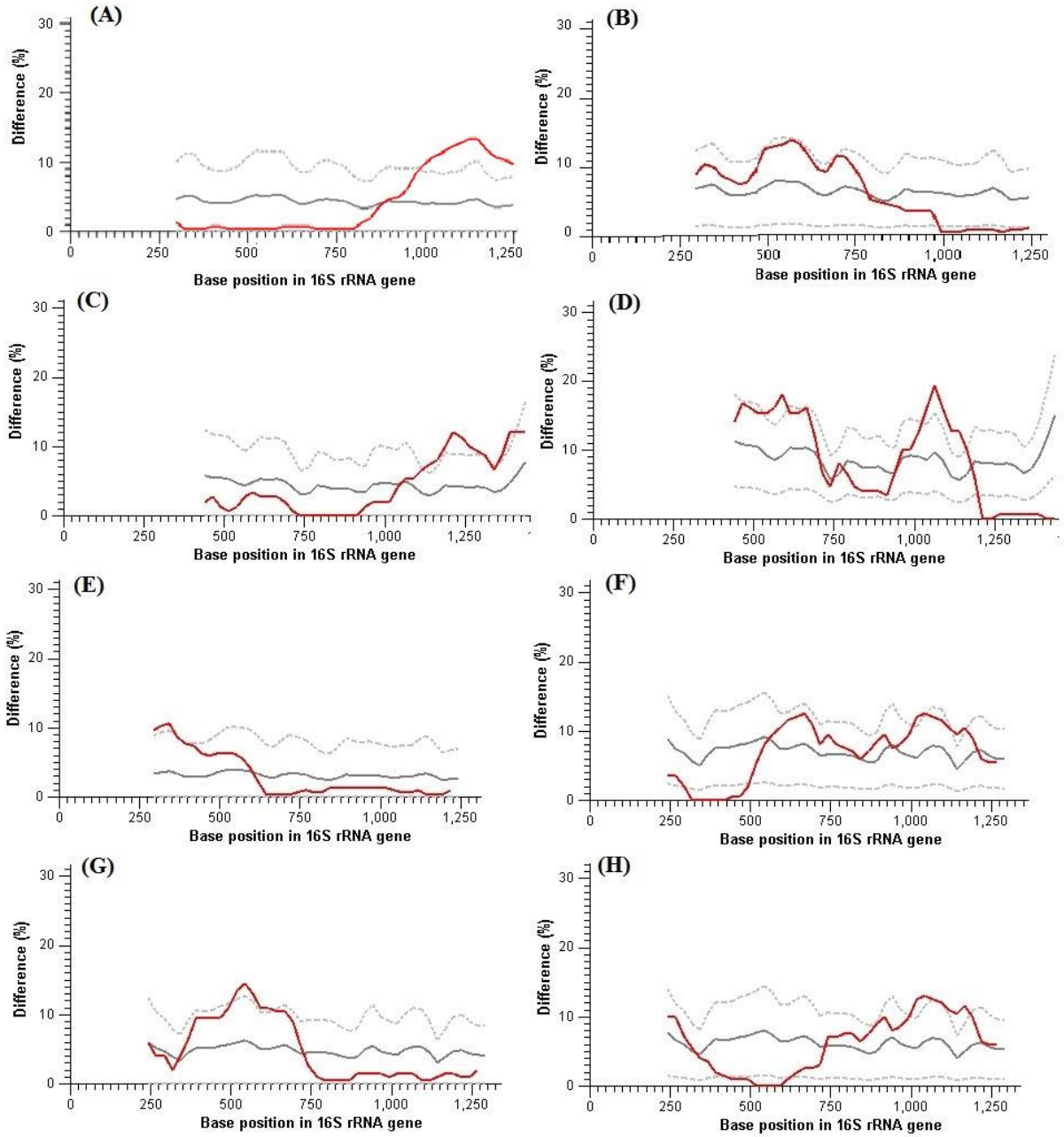

Fig.4. Plots generated using Pintail program comparing potential chimeric versus error-free 16S rRNA gene sequences of phytoplasma strains. A and B represent the comparisons of chimeric 16S rRNA gene sequence of '*Ca.* P. wodyetiae' (KC844879, 16SrXXXVI) with major (16SrXIV-A, AJ550984) and minor (16SrI-B, M30790) parents, respectively. C and D represent pairwise comparisons of chimeric 16S rRNA gene sequence of '*Ca.* P. allocasuarinae' (AY135523, 16DrXXXIII) with major (16SrXXA, AJ583009) and minor (16SrII-D, Y10096) parents, respectively. E and F represent the comparisons of chimeric 16S rRNA gene from representative strain of 16SrXXVII-A (AY539180) with major and minor parents of the subgroups 16SrIV-A and 16SrI-B, respectively. G and H represent the comparisons of chimeric 16S rRNA gene from representative strain of XXVI-A (AJ539179) with major and minor parents of the subgroups 16SrIV-A and 16SrI-B, respectively. The solid red line represents changes in genetic distance (% difference) against base position in the alignment between the two sequences in comparison. The solid gray line and ± 5% confidence intervals (dotted gray lines) represent the expected differences between the two sequences if they are error-free.

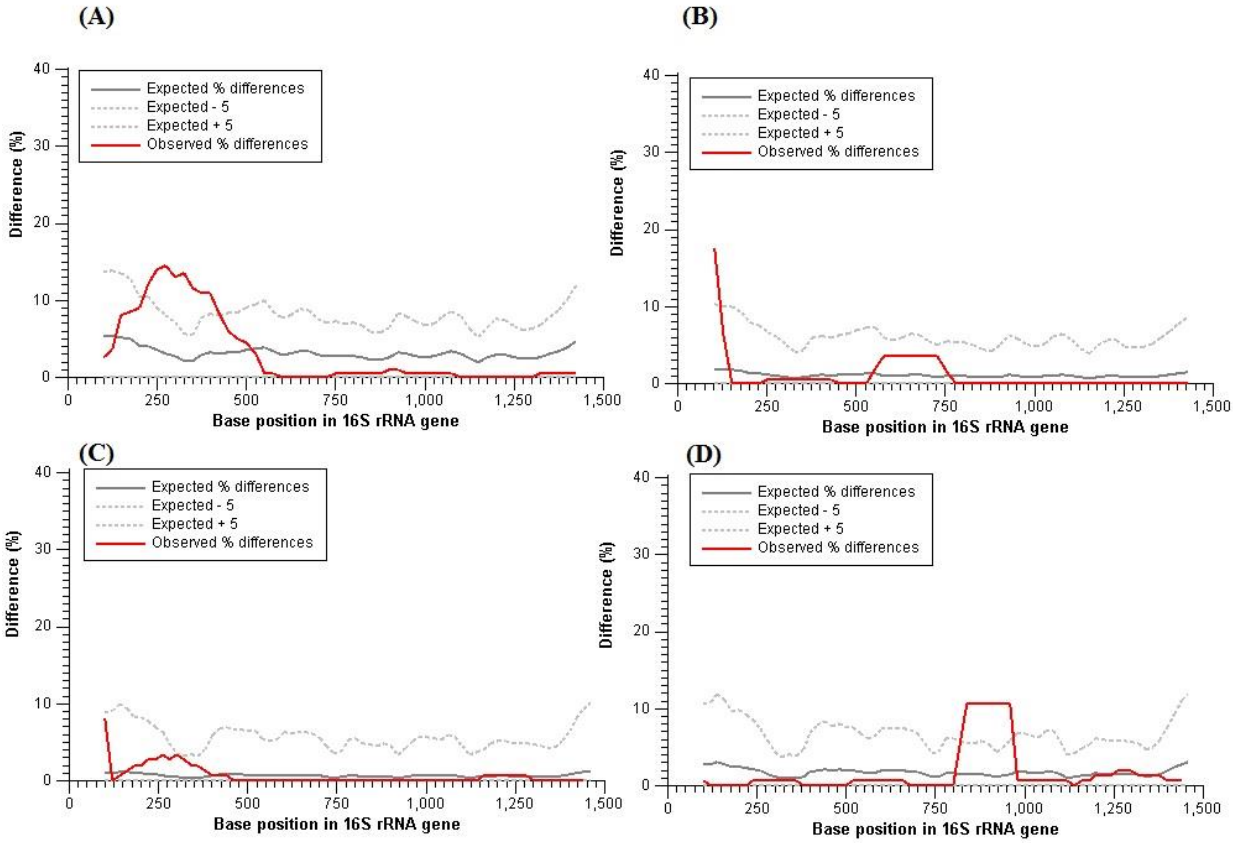

Fig.5. Plots generated using Pintail program showing phytoplasma 16S rRNA gene sequences with potential sequencing errors. A-D: Comparisons of anomalous accession records EF199549 (‘*Ca.* P. lycopersici’; 16SrI-Y), DQ286577 (subgroup 16SrI-AD), AY270156 (subgroup 16SrVI-E) and GU004365 (subgroup 16SrIII-N) versus error-free neighboring sequence records M30790 (16SrI-B), M30790 (16SrI-B), AF228053 (16SrVI-D) and AF510724 (III-F), respectively. See the legend of Fig. 3 for more information.

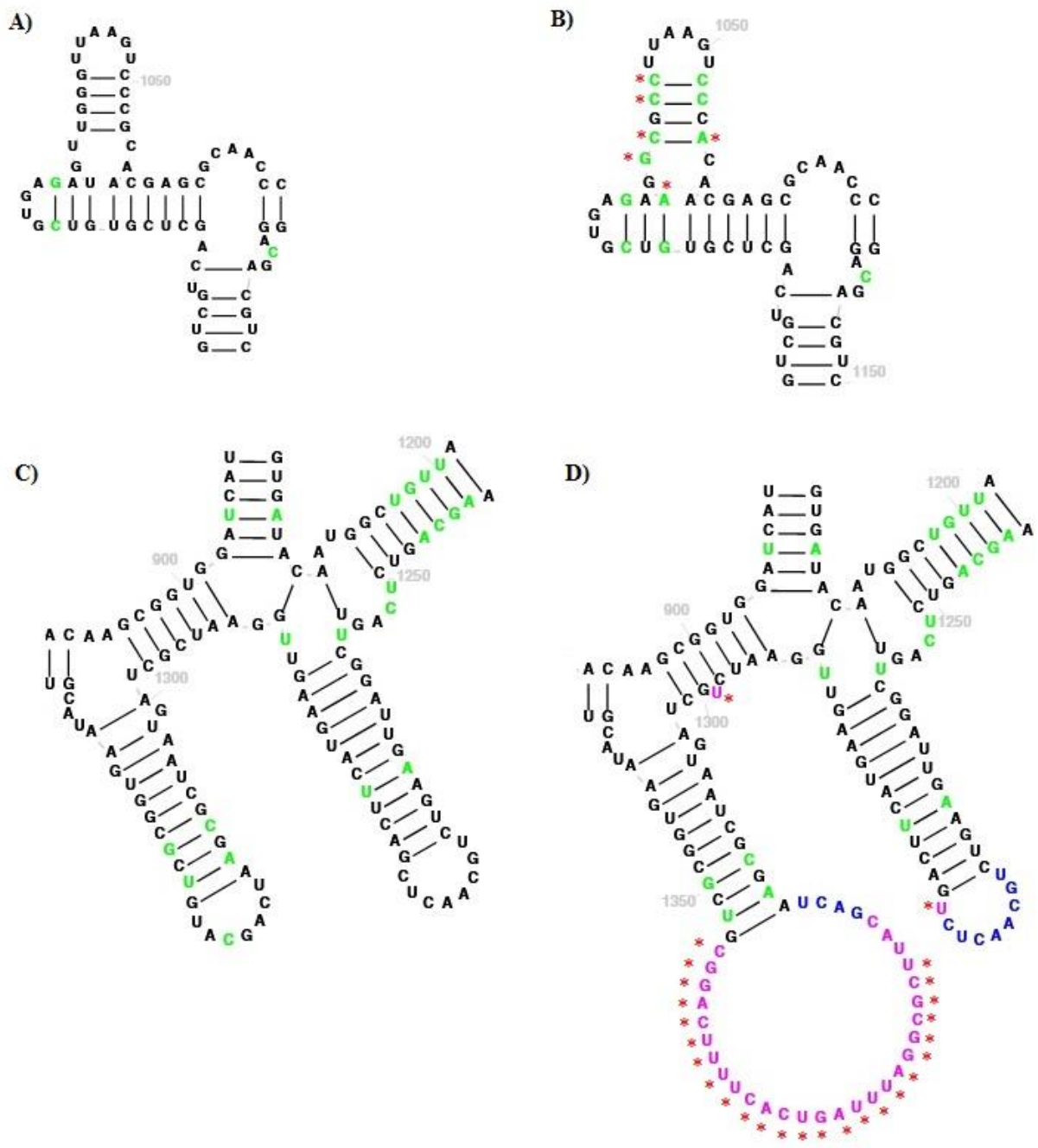

Fig.6. Comparisons of conservation and covariation of nucleotide changes in predicted 16S rRNA gene secondary structure of '*Ca.* P. solani' (16SrXII-A, AF248959) and '*Ca.* P. caricae' (16SrXVII-A, AY725234). Part of the predicted 16S rRNA secondary structure of '*Ca.* P. solani' (A and C) were compared with corresponding regions from that of '*Ca.* P. caricae'. The nucleotide changes in '*Ca.* P. caricae' compared to '*Ca.* P. solani' are asterisked. Nucleotides in black and green are, respectively, identical and different between the query sequence and the model (from *E.coli*). The pink nucleotides represent inserted nucleotides compared to the model (from *E.coli*).

Reference

Arocha, Y., Antesana, O., Montellano, E., Franco, P., Plata, G. and Jones, P. 2007. 'Candidatus Phytoplasma lycopersici', a phytoplasma associated with 'hoja de perejil'disease in Bolivia. International Journal of Systematic and Evolutionary Microbiology 57: 1704-1710.

Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J. and Weightman, A.J. 2005. At least 1 in 20 16S rRNA sequence records currently held in public repositories is estimated to contain substantial anomalies. Applied and Environmental Microbiology 71:7724-7736.

Ashelford, K.E., Chuzhanova, N.A., Fry, J.C., Jones, A.J. and Weightman, A.J. 2006. New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras. Applied and Environmental Microbiology 72:5734-5741.

Bertaccini, A. and Lee, I.M. 2018. Phytoplasmas: an update. pp. 1-29. In G. Rao A. Bertaccini N. Fiore and W. Liefting (Eds) Phytoplasmas: plant pathogenic bacteria-I: Characterisation and Epidemiology of Phytoplasma-associated Diseases. Springer. Singapore.

Bertaccini, A., Arocha-Rosete, Y., Contaldo, N., Duduk, B., Fiore, N., Montano, H.G., Kube, M., Kuo, C.H., Martini, M., Oshima, K. and Quaglino, F. 2022. Revision of the 'Candidatus Phytoplasma'species description guidelines. International Journal of Systematic and Evolutionary Microbiology 72:005353.

Darriba, D., Weiß, M. and Stamatakis, A. 2016. Prediction of missing sequences and branch lengths in phylogenomic data. Bioinformatics 32:1331-1337.

Fukatsu, T., Koga, R., Smith, W.A., Tanaka, K., Nikoh, N., Sasaki-Fukatsu, K., Yoshizawa, K., Dale, C. and Clayton, D.H. 2007. Bacterial endosymbiont of the slender pigeon louse, Columbicola columbae, allied to endosymbionts of grain weevils and tsetse flies. Applied and Environmental Microbiology 73:6660-6668.

Gibb, K.S., Tran-Nguyen, L.T.T. and Randles, J.W. 2003. A new phytoplasma detected in the South Australian native perennial shrub, Allocasuarina muelleriana. Annals of Applied Biology 142:357-364.

Gray, M.W., Sankoff, D. and Cedergren, R.J. 1984. On the evolutionary descent of organisms and organelles: a global phylogeny based on a highly conserved structural core in small subunit ribosomal RNA. Nucleic Acids Research 12:5837-5852.

Hogenhout, S.A., Oshima, K., AMMAR, E.D., Kakizawa, S., Kingdom, H.N. and Namba, S. 2008. Phytoplasmas: bacteria that manipulate plants and insects. Molecular Plant Pathology 9:403-423.

Kim, M., Oh, H.S., Park, S.C. and Chun, J. 2014. Towards a taxonomic coherence between average nucleotide identity and 16S rRNA gene sequence similarity for species demarcation of prokaryotes. International Journal of Systematic and Evolutionary Microbiology 64:346-351.

Kirdat, K., Tiwarekar, B., Sathe, S. and Yadav, A. 2023. From sequences to species: Charting the phytoplasma classification and taxonomy in the era of taxogenomics. Frontiers in Microbiology 14: 1123783.

Lee, I.M., Gundersen-Rindal, D.E., Davis, R.E. and Bartoszky, I.M. 1998. Revised classification scheme of phytoplasmas based on RFLP analyses of 16S rRNA and ribosomal protein gene sequences. International Journal of Systematic Bacteriology 48: 1153-1169.

Mai, U. and Mirarab, S. 2018. TreeShrink: fast and accurate detection of outlier long branches in collections of phylogenetic trees. *BMC Genomics*, *19*(5):23-40.

Marcone, C., Gibb, K.S., Streten, C. and Schneider, B. 2004. 'Candidatus Phytoplasma spartii','Candidatus Phytoplasma rhamni' and 'Candidatus Phytoplasma allocasuarinae', respectively associated with spartium witches'-broom, buckthorn witches'-broom and allocasuarina yellows diseases. International Journal of Systematic and Evolutionary Microbiology, *54*(4):1025-1029.

Naderali, N., Nejat, N., Vadamalai, G., Davis, R.E., Wei, W., Harrison, N.A., Kong, L., Kadir, J., Tan, Y.H. and Zhao, Y. 2017. 'Candidatus Phytoplasma wodyetiae', a new taxon associated with yellow decline disease of foxtail palm (*Wodyetia bifurcata*) in Malaysia. International Journal of Systematic and

Evolutionary Microbiology 67:3765-3772.

Ochman, H., Elwyn, S. and Moran, N.A. 1999. Calibrating bacterial evolution. Proceedings of the National Academy of Sciences 96:12638-12643.

Pääbo, S., Irwin, D.M. and Wilson, A.C. 1990. DNA damage promotes jumping between templates during enzymatic amplification. Journal of Biological Chemistry 265:4718-4721.

Pei, A.Y., Oberdorf, W.E., Nossa, C.W., Agarwal, A., Chokshi, P., Gerz, E.A., Jin, Z., Lee, P., Yang, L., Poles, M. and Brown, S.M. 2010. Diversity of 16S rRNA genes within individual prokaryotic genomes. Applied and Environmental Microbiology 76:3886-3897.

Schloss, P.D., Gevers, D. and Westcott, S.L. 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PloS one 6:e27310.

Siampour, M., Izadpanah, K., Martini, M. and Salehi, M. 2019. Multilocus sequence analysis of phytoplasma strains of 16SrII group in Iran and their comparison with related strains. Annals of Applied Biology 175:83-97.

Sweeney, B.A., Hoksza, D., Nawrocki, E.P., Ribas, C.E., Madeira, F., Cannone, J.J., Gutell, R., Maddala, A., Meade, C.D., Williams, L.D. and Petrov, A.S. 2021. R2DT is a framework for predicting and visualising RNA secondary structure using templates. Nature Communications 12:3494.

Sze, M.A. and Schloss, P.D. 2019. The impact of DNA polymerase and number of rounds of amplification in PCR on 16S rRNA gene sequence data. mSphere 4: e00163-19.

Tamura, K., Stecher, G. and Kumar, S. 2021. MEGA11: molecular evolutionary genetics analysis version 11. Molecular Biology and Evolution 38:3022-3027.

Tiwarekar, B., Kirdat, K., Sathe, S., Foissac, X. and Yadav, A. 2023. Chimera alert! The threat of chimeric sequences causing inaccurate taxonomic classification of phytoplasma strains. bioRxiv 2023.04.

Wang, G.C. and Wang, Y. 1996. The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. Microbiology 142:1107-1114.

Wei, W. and Zhao, Y. 2022. Phytoplasma taxonomy: nomenclature, classification, and identification. Biology 11:1119.

Wei, W., Davis, R.E., Lee, I.M. and Zhao, Y. 2007. Computer-simulated RFLP analysis of 16S rRNA genes: identification of ten new phytoplasma groups. International Journal of Systematic and Evolutionary Microbiology 57:1855-1867.

Xia, X. 2017. DAMBE6: new tools for microbial genomics, phylogenetics, and molecular evolution. Journal of Heredity 108:431-437.

Zreik, L., Carle, P., BOVé, J.M. and Garnier, M. 1995. Characterization of the Mycoplasmalike Organism Associated with Witches'-Broom Disease of Lime and Proposition of a Candidatus Taxon for the Organism, "Candidates Phytoplasma aurantifolia". International Journal of Systematic Bacteriology 45:449-453.