Crop Protection 107 (2018) 93–103

Contents lists available at ScienceDirect

Crop Protection

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

Revealing new information from existing genomic data for pepper mild mottle virus pathotype determination

Bojana Banović Đeri[a,*], Vesna Pajić[b], Dragana Dudić[b]

[a] Institute of Molecular Genetics and Genetic Engineering, University of Belgrade, Belgrade, Serbia
[b] Faculty of Agriculture, University of Belgrade, Belgrade, Serbia

ARTICLE INFO

Keywords:
Pepper mild mottle virus
PMMoV
Monitoring
Capsicum
Data mining
Pathotype

ABSTRACT

Primary goals of 21st century science involve eco-friendly solutions for detection, control and suppression of plant viruses. Even though we are accumulating knowledge and data on plant viruses' nucleotide sequences, we are still using a minimum of information available from the collected data. Applying bioinformatics tools and data mining approach to viral sequences is extremely useful in revealing the hidden knowledge, giving guidelines for further biological/bioinformatics studies and developing novel environmental-friendly virus specific defense strategies in crop protection. In this paper we tested to what extent modern bioinformatics methods are able to reveal new information that would bring us closer to our primary goals. On the date of the search (March 2015) we extracted all available PMMoV entries from publically available databases, represented by heterogeneous data set containing 231 nucleotide sequences covering different parts of the PMMoV genome, that were of different geographical origin, related to different time periods, associated with different pathotypes, and were not previously compared to each other. Results revealed that nucleotide content at genomic positions 552, 565, 639, 666, 708, 5921, 5975 and 6002 can be used to discern three distinct PMMoV genotype variants and their association to one of two virus pathotypes, $P_{1,2}$ or $P_{1,2,3}$. These sites have never been reported as informative before, probably because by being silent mutations they escaped usual research scrutiny of looking for pathotype determinants among nonsense, missense mutations and indels. Our model was further tested in predicting pathotype of ten newly deposited PMMoV sequences and the successful outcome of the test supported the model as an usefull asset for discrimination among pathotypes $P_{1,2}$ and $P_{1,2,3}$ according to distinct nucleotide content in replicase and coat protein encoding genes. Based on the presented results, we also suggested new tests for fast and cost-effective screening of PMMoV pathotypes and eventually for inducing plant resistance against pepper mild mottle virus.

1. Introduction

Plant viruses' impact on crop production has always been one of the major focuses of scientific interest. In the fight against plant viruses a number of different strategies like predicting plant variety-virus pathotype interactions, developing different monitoring methods (i.e. PCR, RFLP, PFGE, and MST), etc. are being applied. All of these strategies involve collecting numerous plants' and viral data including their nucleotide sequences so that the amount of nucleotide data has been rapidly increasing in correlation with the expansion of high throughput sequencing technology. The accumulation of sequence data calls obviously for development of new bioinformatics tools and data analysis approaches to process them. However, the magnitude of the biological data production currently vastly exceeds the extent of data processing, hence the knowledge on various biological phenomena remains hidden in databases.

In this paper we chose pepper mild mottle virus (PMMoV) as *in silico* plant virus model to investigate what kind of information one can extract from the publically available sequences by using bioinformatics tools and data mining approach. PMMoV is of interest to study because it might inflict a great agricultural damage loss by significantly reducing yields in pepper (*Capsicum* spp., both sweet and hot pepper cultivars) and by infecting tomato throughout Australia, China, Europe, Japan, North America, North Africa and Taiwan (Wang et al., 2006; Jarret et al., 2008; Çağlar et al., 2013; Milošević et al., 2015). The main obstacle for farmers to suppress PMMoV is that the virus forms very

Abbreviations: coat protein, (CP); expectation-maximization class algorithm, (EM algorithm); kilodalton, (kDa); microscale thermophoresis, (MST); open reading frame, (ORF); pepper mild mottle virus, (PMMoV); polymerase chain reaction, (PCR); pulse field gel electrophoresis, (PFGE); purine bases, (Pur); pyrimidine bases, (Pyr); ribonucleic acid, (RNA); restriction fragment length polymorphism, (RFLP); single nucleotide variants, (SNVs); untranslated region, (UTR)

* Corresponding author.
E-mail address: bojanabanovic@imgge.bg.ac.rs (B. Banović Đeri).

https://doi.org/10.1016/j.cropro.2018.01.017
Received 21 October 2017; Received in revised form 23 January 2018; Accepted 26 January 2018
Available online 07 February 2018
0261-2194/

stable and easily spread viral particles, which remain infective even in the irrigation/river/sea-water, soil, compost, plant debris, etc. (Rosario et al., 2009; Hamza et al., 2011). In the past, PMMoV outbreaks were prevented by methyl bromide fumigation of soil, but Montreal Protocol implementation banned such chemicals in 2005, so new eco-friendly strategies based on the knowledge of the virus as well as the plant resistance mechanisms are needed to reduce the frequency and severity of the viral outbreaks.

Our environmental-friendly strategy is established on testing the usefulness of designing PMMoV genome-specific molecular characterization and monitoring tests, based on the comparative analysis of PMMoV sequences' data available in the public databases by applying modern bioinformatics tools and data mining approach. At the same time we want to determine to what extent these sequences can be used for obtaining new knowledge about the virus and how the extracted knowledge could be applied to suggest future directions for crop viruses' research.

1.1. PMMoV

PMMoV is a single-stranded RNA virus of the genus *Tobamovirus* and its genome is represented by a positive-sense single-stranded RNA app. 6357 nucleotides long, which contains four open reading frames (ORFs) from the 5′ end toward the 3′ end (Alonso et al., 1991). The first ORF comprises nucleotides from 70 to 3423 encoding for the small replicase subunit (126 K) of 1117 amino acids, which is essential for the viral replication, while the second ORF extends from 70th to 4908th nucleotide, with a read-trough of an amber stop codon, encoding for the large replicase subunit (183 K) of 1612 amino acids, which is also essential for the viral replication. The third ORF extends from 4909th to 5682nd nucleotide encoding for the cell-to-cell movement protein (30 K) of 257 amino acids, which is important for viral particles spreading. The fourth ORF prolongs from 5685th to 6158th nucleotide encoding for the coat protein (CP) of 156 amino acids, which form a viral capsid, considered as the main elicitor of plant resistant-genes mediated hypersensitive response. While the 126 K and 183 K proteins are translated directly from the viral RNA, the 30 K and CP proteins are translated from subgenomic RNAs. Each PMMoV viral particle contains approx. 2130 copies of the 17–18 kDa coat protein subunits that form a helix around the viral RNA.

Plant resistance against viruses was shown to be effective only for some period of time because the viruses manage to overcome plant defenses through viral genome mutations and recombination (García-Arenal and McDonald, 2003; Genda et al., 2007). Even though CP was primarily shown to elicit plant defense response through four genes of peppers' *L* - locus (designated as *L1-L4*; Boukema, 1980, 1982, 1984), Yoon et al. (2006) showed that the replicase gene and 3′ non-coding RNA region are also major pathogenicity determinants in PMMoV. Of all tobamoviruses, PMMoV is considered to be the most pathogenic for *L* genes, which is in concordance with the reports on numerous PMMoV strain specific mutations causing changes in the virus phatotype (Tsuda et al., 1998; Hagiwara et al., 2002; Hamada et al., 2002, 2007; Velasco et al., 2002; Genda et al., 2007; Antignus et al., 2008; Nakazono-Nagaoka et al., 2011; Choi et al., 2013). Specifically, some strains of PMMoV were able to overcome *L1-L4* gene mediated resistance in pepper (Gilardi et al., 1998; Genda et al., 2007; Antignus et al., 2008) and according to their increased pathogenicity they were classified into five subgroups of pathotypes P_0, P_1, $P_{1,2}$ (e.g. Spanish strain PMMoV-S), $P_{1,2,3}$ (e.g. Italian strain PMMoV-I from Sicily) and $P_{1,2,3,4}$ (e.g. Japan strain PMMoV-L4BV) (Boukema, 1984; Sawada et al., 2004; Fraile et al., 2011), with pathotype P_0 displaying the weakest and $P_{1,2,3,4}$ the strongest pathogenicity.

Nowadays, with more PMMoV sequence data available, bioinformatics analyses have a great potential to reveal the hidden knowledge contained in viral nucleotide sequences (i.e. to determine single nucleotide variants (SNVs) and their connection with the occurrence of braking-resistance pathotypes), which can be extremely useful in developing easy to use/cost-effective monitoring methods and learning to induce resistance in plants, and so improve the breeding efforts and pepper-crop yields.

2. Results

2.1. Nucleotide content

Nucleotide content of PMMoV profile sequence favors T and A nucleotides, with the following percentages: A (29.6%), T (28%), G (23.9%) and C (18.4%). The overall ratio of (A,T)/(G,C) was around 7:5 (1.36) and the ratio of Pur vs. Pyr nucleotides was 1.15%. In the region from 1st to 200th nucleotide (containing the 5′ untranslated region (5′ UTR) and the first 130 nucleotides of 126K/183 K replicase genes) adenine dominates with 40%, while in 3400th-4400th nucleotide region (corresponding to the second ORF encoding 183 K protein) it dominates with 35–37%. In the rest of the genome the content of adenine is lower, except in the region around 5400th nucleotide (corresponding to the third ORF encoding for the movement protein) where adenine content also exhibits high value (38%). The distribution of all four nucleotides over the part of the genome corresponding to the fourth ORF (coat protein) is quite stable (under 30%).

While the entire PMMoV genome has GC content of 51.9%, this content varies in groups within the range 37.36%–49.47% (being the lowest in the Group 2 and the highest in the Group 4.3).

When compared to the consensus sequence all groups of sequences do not show more than 27.7% similarity per group, which highlights considerable sequence diversity.

2.2. Genome polymorphism and SNVs

2.2.1. Whole genome sequences analysis

We identified 524 sites containing nucleotide variants in respect to the profile sequence. SNVs with the highest number were transitions A > G (177), C > T (216), G > A (206) and T > C (193). The most of the variations occurred in a single position, but there were several variations of length two or three nucleotides: one site of SNVs of length 2 (Japan-2005-P0|AB113117.1, Japan-2005-P0|AB113116.1, China-2006-P12|AY859497), two sites of SNVs of length 2 (SouthKorea-2005-P12|AB126003.1, Japan-2007-P12|AB254821.1), three sites of SNVs of length 2 (Japan-2003-P1234|AB276030.1) and 6 sites of SNVs of length 2 and two sites of SNVs of length 3 (Spain-2002-P123|AJ308228).

2.2.2. Groups' 1.1–4.3 analysis

Results of the analysis of variations in nucleotide content per groups of sequences revealed the following percentages of SNVs per group: 14.5% (Group 1.1), 10.15% (Group 1.2), 17.89% (Group 2), 8.86% (Group 3), 8.86% (Group 4.1), 25.42% (Group 4.2) and 46.41% (Group 4.3). As expected, it can be observed that sequences covering the fourth ORF (encoding for the coat protein) have the highest percentage of positions in which nucleotide content differs. Distribution of all SNVs per groups is given in Table 2.

Of all possible SNVs, variations comprising all four types of single nucleotide transitions A > G, C > T, G > A, and T > C dominate across all groups. Also, we observed an increase in number of variations related to two types of tranversions T > A and A > T in Group 4. Further, the presence of some variants were quite permanent throughout all groups (for example, SNV T > C ranging from 15.7% to 24.1%), while others displayed different distribution in different groups (i.e. A > G is 40% represented in Group 2 but only 2.2% in Group 4.3).

2.3. Determination of genotype variants

After the comprehensive analysis of nucleotide content at all positions in the whole genome and partial genome sequences, several