**vd-sRNA profiler: User focused interface for viroid derived small RNA mapping and profiling**

Charith Raj Adkar-Purushothama1, Pavithran Sridharan Iyer2,#, Teruo Sano3, Jean-Pierre Perreault1,\*

1RNA Group/Groupe ARN, Département de Biochimie, Faculté de Médecine des Sciences de la Santé, Pavillon de Recherche Appliquée au Cancer, Université de Sherbrooke, Sherbrooke, QC, Canada

2Département de Physique, Université de Sherbrooke, Sherbrooke, QC, Canada

3Faculty of Agriculture and Life Science, Hirosaki University, Bunkyo-cho 3, Hirosaki 036-8561, Japan

#Current address: Applied Mathematics, Institute for Quantum Computing, University of Waterloo, Waterloo, ON, Canada

\*Corresponding author: jean-pierre.perreault@usherbrooke.ca (J.-P.P.)

Date of Submission: XX, June 2021

**Abstract**

Viroids are circular, highly structured, single-stranded non-coding RNA pathogens that are known to infect and cause disease in several plant species. Viroid infection known to trigger RNA silencing machinery of the host plant. Detection of viroid derived small RNA (vd-sRNA) in viroid infected host plant opened new avenue in viroid-host pathogenicity. Since then, several viroid research groups studied vd-sRNA retrieved from different viroid-host combinations. Such studies require the segregation of 21- to -24-nucleotide long small RNAs (sRNA) from deep-sequencing databank, followed by separation of vd-sRNA from this sRNA groups that shows sequencing similarity with genomic and anti-genomic strands of the viroid. For visualization, such mapped vd-sRNAs are then profiled on the viroid’s genomic and anti-genomic strand. Although several commercial interfaces are available, they are all programmed for linear RNA molecules. Hence, viroid researchers must develop the computer program to accommodate the vd-sRNAs derived from the circular viroid. This is laborious and often bottleneck for biologists. To overcome this constraint and help the viroid research community, in the present study, a python-based pattern matching interface has been developed to profile and map vd-sRNA on circular genome. Additionally, a feature of matching tolerance has been added to map the vd-sRNA derived from the quasi-species. Efficacy of the software is tested with previously reported deep-sequenced data obtained from two independent studies.

**Introduction**

Viroids are the simplest and smallest known plant pathogens. They consist of a circular single stranded non-coding RNA genome of 246-434 nucleotides (nt) in length [1]. Since viroid do not encode for any known peptides, they rely entirely on their sequence, structure and host factors for their replication and induction of disease symptoms in host plants. Upon infection they induce wide array of symptoms in host plants, such as leaf epinasty, leaf distortion, stunting, and flower distortion [2–4]. Due to their highly base-paired secondary structures and RNA-RNA mode of replication, viroids are both inducers and targets of RNA silencing [5].

RNA silencing (RNA interference [RNAi]), is a multilayer defense system, which protects plants from invading RNA pathogens such as viruses and viroids [6]. Silencing is triggered by the processing of double-stranded or highly structured RNA by host’s RNase III-type ribonucleases (i.e., DICER or DICER-LIKE), resulting in the production of small interference RNAs (siRNAs) of 21–24 nt in length [7]. In 2001, two research groups independently reported the presence of viroid derived small RNAs (vd-sRNA) in plants infected with PSTVd by RNA gel blot assays [8,9], suggesting that viroids are the targets of RNA silencing. Introduction of next-generation sequencing (NGS) technology allowed large scale studies on the accumulation of such vd-sRNA in viroid infected plants [10–13]. The role of vd-sRNA in RNAi mediated down-regulation of the host transcripts and symptom induction has been explored using different host-viroid combinations [14–18]. This has been reviewed elsewhere [19].

The NGS is extensively used by viroid researchers to study the viroid-host interaction emphasizing on vd-sRNA production [20]. Profiling of vd-sRNAs on both the genomic on antigenomic strands of viroid gives several information such as but not limited to: (i) the proportion of vd-sRNA produced from genomic on antigenomic strands of the viroid; (ii) distribution of vd-sRNA on the genomic on antigenomic strands; and (iii) regions of viroid that are susceptible, and resistance for host’s RNA silencing machinery. Such information could be used to understand host-viroid species relationships [11,13,21], viroid quasi-species [22,23], and to develop viroid resistance plants [24]. To get all this information, the first and foremost is mapping of vd-sRNA based on viroid sequence. The biggest hurdle for such studies is the lack of a sRNA mapping software on the circular genome, such as viroids. Although several commercial platforms are available, they are all programmed for linear DNA or RNA molecules. Consequently, either viroid researchers develop their own interface or depend on bioinformaticians to progress in their studies.

Mapping of vd-sRNA to a target viroid’s genomic and/or antigenomic strand sequence is innately tied to the classical example of determining the number of pattern matchings. In computer science, pattern matching is the process of checking a given sequence of characters (in present scenario, nucleotides) exists among the given data. In other words, given two strings and , does string occur as a sub-string of? Concept of pattern matching was first addressed by Knuth, Moris and Pratt in 1970 [8]. Since then, a multitude of proposed solutions to this problem have come into existence that includes but not limited to industrial applications and supply chain performance. Pattern matching is extensively used in the field of bioinformatics to determine the sequence similarity between the subject and query ([2]).

To provide the user friendly and the efficient bioinformatics tool to viroid researchers, in the present study a python language is used to write the pattern matching program. Python is one of the computer languages that is easy to read and implement. Additionally, it is an open-source language and runs on different platforms such as Mac, Windows and Unix. Here developed interface is initially tested on a empirical example and then applied to the previously reported NGS data.

**Materials and methods**

**Setting problem for pattern matching**

The counting version of the standard pattern matching is a problem of enumerating all the occurrences of the string in the string [3,10]. Computational biology provides an archetypal context for counting problems of this sort [2,4,7], where for example, is considered as a mRNA and as a genome. In particular, computing the number of bindings of a mRNA with genome. When a mRNA nucleotide sequence is represented by a string and the genome by a string , the problem of computing the number is bindings of the mRNA sequence immediately admits the counting form of the standard pattern matching problem. Furthermore, we can safely assume that and are drawn from the alphabet . Since this study focuses on solving the binding of vd-sRNA with viroid genome, is considered as a vd-sRNA and as a viroid RNA.

Before laying out the problem precisely, to solve several instances of the above counting problem, each one corresponding to a different viroid RNA nucleotide . Hence, the problem that we would like to solve in this section can be stated as follows.

|  |  |
| --- | --- |
| *Definition1: Viroid bindings problem* |  |

Given the following:

1. Set of N strings where each string has length is sampled from ,
2. Length m string sampled from the alphabet , and
3. Function from strings to to identify a matching. For two equal length strings and , if and only if and match and otherwise.

Compute the size of the sets for each , where:

Equation 1:

We have deliberately left the precise definition of unspecified. The above abstract definition encapsulates a variety of problems in our interest. The particular cases of interest can be characterized using concrete definitions for . In this article, we use four different ways of specifying , these can be divided into two categories, that is (i) matching of in forward and reverse direction with respect to , and, (ii) linear or circular topology for .

(i) matching of in forward and reverse directions with respect to :

The first set of variations occurs from choosing an alignment of the gene relative to the host mRNA. While the natural alignment called the forward matching setting coincides with the definition in “Definition 1”, its counterpart, is referred to as the *reverse matching* setting. The forward and reverse cases differ in their rule for deciding when two equal length strings and, sampled from the same alphabet, are identical. The forward matching setting adopts the natural definition where if and only if , for all , where is the length of . On the contrary, the reverse matching setting places a non-standard requirement: if and only of , for all . In what follows, we will only discuss the case of forward matchings. It is straightforward to adapt our discussion to reverse mappings, without affecting our conclusions.

Setting the topology is crucial to determining the boundaries for a string. The case of linear topology, referred to as linear alignment, refers to a with a finite boundary, that is a length string with following for and no character following . On the contrary, the case of circular topology, referred to as circular alignment is characterized by a gene with periodic boundary conditions, that is, a length string with the character following the character for . It is straightforward to note that the matching problem on a circular aligned gene trivially reduces to the linear alignment case. This can be realized by simply appending the gene string with the length substring of starting at position . So, we will not attribute any conceptual distinction to this setting and continue to analyze out solution for the linear alignment.

**Finding the occurrence of in**

This section presents the solution to the setting of forward matching with linear alignment for the pattern matching problem defined in (Definition 1). First of all, let us introduce the following useful notations.

1. The *length* of a string is simply the number of characters in . We will often use to denote the length of
2. The *Hamming weight* of a binary sequence is the number of ’s in the sequence. We will often use to denote the Hamming weight of

A nave solution to the pattern matching problem entails a systematic search for a match between every sequence in to every substring of . In other words, makes calls to an elementary routine which compared two strings and to identify the number of times occurs in . Consequently, the time complexity of the naive solution is . We want to provide a solution that outperforms the naive one. We will achieve this by mapping the pattern matching problem to a matrix multiplication problem of a similar size. The latter can be solved efficiently by vectorizing individual operations.

The first ingredient of this mapping is the following encoding of each character in into a unique binary sequence of bits.

Equation 2:

The above encoding naturally implies an encoding for strings composed of the characters from , by simply concatenating the binary sequences corresponding to the individual alphabets in the string. Likewise, a nucleotide sequence of length is identified by a binary sequence of bits, denoted by .

It is straightforward to notice that distinct characters in (Equation 2) correspond to orthogonal binary sequences. Therefore, two nucleotide strings and of length each, match exactly if and of only if the number of ones in the dot product: is equal to , i.e., . A similar observation can be made to account for mismatches wherein the the number of ones in should be at least . To summarize, two length $k$ strings and match approximately with at most mismatches if and only if the Hamming weight of is at most .

While we have just described a method to check for a match between two strings of equal length, this can be readily extended to the strings of unequal lengths, i.e., when and for . To realize this, note that can be found in if and only if there are consecutive characters in that match with . In other words, for some . Furthermore, there are exactly substrings of that can be formed by taking consecutive characters, described by the set . Each of these substrings can be encoded into a binary sequence having bits. The substrings that match with will can be identified by checking if is at least .

Lastly, we can combine the tests for matchings between each of for and into a single condition as follows. Let be a matrix whose columns are , respectively. It turns out that the dot product is a vector of size whose th component satisfies if and only if the substring matches with at all but at most locations. In other words, the number of occurrences of in is given by the number of entries in whose value is at least . This inference forms the backbone of our solution for the underlying pattern matching problem.

Let us now summarize our solution to the forward matching problem with linear alignment, in definition 1. Recall that the problem instance is described by a set of sRNA nucleotides and a gene . The encoding in equation 1, leads to two key observations:

1. is mapped to a binary matrix , whose th row denotes the binary encoding of the the nucleotide in the sRNA pool.
2. is mapped to a binary matrix whose th column denotes the binary encoding of the length substring of starting at position .

These observations imply that the dot product yields an ( binary matrix, where is equal to if and only if the th sRNA nucleotide binds to the gene, at position . Finally, taking the sum of all the rows of , yields a vector of size , where It is easy to realize that is nothing but the number of sRNA nucleotides in which bind to the gene at location . In other words, , from (Equation 2). This completely specifies our solution to the pattern matching problem. Casting the pattern matching problem as matrix multiplication is instrumental [6] to its low runtimes for large number of sRNA nucleotides. Table 1 shows the runtimes for computing the solutions to the pattern matching problem in definition 1, for a few different sizes of the sRNA pool.

**Profiling of vd-sRNA on viroid genomic and antigenomic strand**

The key inputs to the software are the genome sequence, specified in a text file without line-breaks and the pool of vd-sRNA, specified in a text file wherein each new line identifies a nucleotide sequence composed of “A”, “T”, “G” and “C” characters. If there is any nucleotide sequences with “N”, it is ignored. To accommodate various formats, software accept an integer which specifies the number of lines to skip before reading every nucleotide sequence in the pool file. Additional settings for the matching problem include (i) tolerance: the maximum number of mismatches allowed to accommodate the vd-sRNA derived from the quasi-species, (ii) topology of the gene, i.e., an integer that takes the value 0 for linear matching and 1 for circular matching, and (iii) the number of cores to be used by the software. For example, “gene.txt pool.txt 2 1 1 1” is a complete input specification, indicating that the problem of computing bindings, in the circular topology while allowing for at most one mismatch, between the sequence in pool.txt with those in the pool.txt. All such instances of the matching problem can be gathered in a text file, placed in vbind/data/input. The output format for the solution of the matching problem is an array integer for every distinct length of vd-sRNA sequences in the pool. Each array is of length equal to the gene sequence, where its i-th component specifies the number of matchings amongst sequences of a given length in the pool, with the sequence in the gene starting at positon 1.

**Viroid small RNA data**

Previously published two sRNA data obtained from tomato plants infected with variants of PSTVd is used in the present study for the analysis. Specifically, the total sRNA sequence obtained from the tomato plants infected with PSTVd-I (GenBank Acc. No.: AY937179) and PSTVd-RG1 (GenBank Acc. No.: U23058) having the GEO Acc. No.: GSM1695657 and GSM1717894, respectively [16,25]

**Software accessibility and instructions**

The script is available at <https://github.com/paviudes/vbind>. An example input and running instructions as shown in data/input/example.txt

**Results**

**Overview of sRNA mapping and profiling**

In order to verify the working of current developed tool, a 10-nt long DNA string used as “genome”, an equivalent to the viroid genomic strand (Fig. 1A). From this DNA string, a pool of sRNAs were prepared randomly such that pool have at least one non-matching sRNA, genomic sRNA and antigenomic sRNA of 3- and 4-nt in length. Additionally, couple of sRNAs taken between position 8 and 3 of the DNA string to imitate the sRNA derived from circular genomic strand (Table 1). This pool is equivalent of NGS data of sRNA. Running the patterning matching to find the matching of sRNAs in the pool with genome strand, summarizes the results in the terminal as shown in Fig. 1A. Summary of obtained data includes how many sRNAs are present in the pool of a given length, and how many of that are forward matching (sRNA matched with genomic strand), how many are of reverse matching (sRNA matched with antigenomic strand) and time taken to complete the analysis. A window below shows the matching of small RNA length and position at which it showed matching and how many sRNAs matched. For example, in Fig. 1B, under the forward matching [[3 0012001000]….] indicates, in 3-nt long sRNA pool, 0 mismatches at position 1 and 2; 1 sRNA matches at position 3, 2 sRNA matches at position 4 and so on. That said, the tool developed could find the matching small RNA on both the genomic and antigenomic strand of the RNA.

To profile the sRNA on the genomic and antigenomic strand of the genome, above matched sRNAs are run on the genome strand. As presented in Fig. 1C, the X-axis indicates the length of viroid RNA whereas Y-axis indicates the number of matching sRNAs. By testing the above example, it is clear that, the here developed tool is able to segregate and output the total sRNA count, the sRNA matching genomic strand and antigenomic strand, thus addressing the fundamental questions of mapping the sRNA from the NGS data on given genome sequence. Further, profiling of mapped sRNAs on the empirical DNA strand, explains that, here developed software could be applied to analyse vd-sRNA obtained from viroid infected plants.

**Mapping of sRNA obtained PSTVd-I infected plants**

To map and profile the sRNA on genomic and antigenomic strand of viroid RNA, sRNA pool of 15- to 37-nt obtained from PSTVd-I infected tomato plant was retrieved from GEO. The data set was processed with the 359-nt long PSTVd-I genome sequence with the tolerance of zero, in order to segregate genomic and antigenomic matching vd-sRNA.

As, vd-sRNAs of 21- to 24-nt are the one researchers interested in, the matching obtained for these sizes are presented in Table 2. Out of 4,316,543 sRNA of 21- to 24-nt, a total of 488,176 vd-sRNAs showed 100% sequence similarity with PSTVd-I sequence accounting for 11.3% of total recovered sRNAs. Specifically, 380,731 (8.8%) and 107,445 (2.5%) vd-sRNAs of 21- to 24-nt derived for the genomic and anti-genomic strand of PSTVd-I, respectively. In order to have a better picture on the expression level of individual sized vd-sRNA, obtained vd-sRNAs were normalized to per million reads. Detailed analysis showed that, 22-nt long sRNA had maximum number of vd-sRNA (34.3 %) whereas, least expressed vd-sRNA is 24-nt (1.2 %). Interestingly, analysis of the vd-sRNA based on individual sizes revealed that the genomic strand produced more 22-nucleotide sRNAs (29.3%), while the anti-genomic strand produced more 21-nucleotide sRNAs (7.9%). Overall, more genomic strand derived sRNAs were covered when compared to anti-genomic strand of the PSTVd-I. In other words, (+) vd-sRNA expressed 3.5 times more than (-) vd-sRNA. This can be attributed to the higher recovery of the genomic strand of viroid in infected plants as described elsewhere [26]. These data are in agreement with the previous report where same sRNA pool was used for the analysis [16], confirming the reproducibility of the here developed software.

**Profiling mapped sRNA on PSTVd-I**

In order to examine the regions of the PSTVd, that produced more vd-sRNA, the above mapped vd-sRNA was plotted on the genomic and anti-genomic strand of the PSTVd-I. Hence, each 21- to 24-nt long vd-sRNA and cumulative 21- to 24-nt long vd-sRNAs are profiled on the PSTVd-I (Fig. 2).

The data presented here clearly shows the regions that produces more vd-sRNA on both the genomic and anti-genomic strand and this is in agreement with the previous report [16], thus confirming the reproducibility of the here developed profiling tool.

**Accommodating vd-sRNA of PSTVd-I quasi-species**

Viroids are known to form quasi-species in host plant [22,23,27–29]. That said, a single sequence type could give rise to several hundreds or thousands of sequence variants in host plant. In order to accommodate the vd-sRNA derived from the PSTVd-I sequence variants, in the present analysis, vd-sRNA mapping against the genomic and anti-genomic strand of the PSTVd-I was decreased by allowing 1 mismatch. This decreased stringency increased overall matching by 0.4% compared with the matching at 0 mismatch (also called 0 tolerance), specifically from 11.3 % to 11.7 % as presented in Table 3. In other words, 113,094 and 117,011 vd-sRNAs of 21- to 24-nt long are present per million sRNA with zero tolerance and 1 mismatch, respectively. Allowing one mismatch increased the percentile of matching as low as 0.1 % (for 24-nt long vd-sRNA) to much as 1.1 % (for 22-nt long vd-sRNA) compared to zero mismatches. However, overall distribution of (+) vd-sRNA to (-) vd-sRNA remained almost the same as with that of zero mismatches. Further, as with zero tolerance mapping, 1 mismatch too had highest genomic derived vd-sRNAs for 22-nt long sRNAs whereas anti-genomic derived vd-sRNAs of 21-nt long. To evaluate the regions of the PSTVd, that produced more vd-sRNA, the above mapped vd-sRNA with 1 mismatch is plotted on the genomic and anti-genomic strand of the PSTVd-I as described before (Fig. 3).

**Evaluating the mapping tool on vd-sRNA obtained from PSTVd-RG1 infected plants**

To increase the confidentiality of the programing tool developed here, one more sRNA data set that is publicly available is analyzed. Specifically, the sRNA sequence obtained from tomato plants infected with PSTVd-RG1 variant is used for the analysis. Both PSTVd-I and PSTVd-RG1 are 359-nt long in their genome size. PSTVd-I and PSTVd-RG1 have 3 mismatches in their genome and they induce intermediate and lethal disease symptoms in tomato cultivar Rutgers (*Lycopersicum esculantum* cv Rutgers) [18]. As described earlier, a total of 730,499 sRNAs of 21- to 24-nt are mapped on the PSTVd-RG1 in both forward and reverse direction to segregate genomic (+) and anti-genomic (-) vd-sRNA at 0 and 1 mismatch tolerance. Obtained data is normalized per million reads (Table 4). Results showed that total of 102,555 vd-sRNA are recovered per million reads which contained 88,493 (8.8 %) and 14062 (1.4%) of genomic and anti-genomic strand derived vd-sRNA that showed 100% sequence similarity with PSTVd-RG1. This accounted for 10.3% of total sequenced 21- to 24-nt long sRNA. Out of all 21- to 24-nt long vd-sRNAs, 21-nt long vd-sRNA species is the highly expressed (25.8%) whereas 24-nt long vd-sRNA is the least expressed (1.2%) for both genomic and anti-genomics strand of PSTVd-RG1.

In order to allow the vd-sRNA derived from sequence varians of PSTVd-RG1, one mismatch allowed in the mapping. Decreasing the stringency increased the number for vd-sRNA by 3.3 % (from 102,555 per million reads with zero mismatch to 135,654 per million reads with one mismatch) (Table 5). This attributed to 2.9% and 0.4% increase in genomic and anti-genomic strand vd-sRNA compared to zero tolerance mapping. Analyzing the vd-sRNA derived from genomic and anti-genomic strands revealed that genomic strand derived sRNAs expressed at least 6.4 times higher than anti-genomic strand derived sRNAs.

To visualize the regions of the PSTVd-RG1 that produced vd-sRNA in both genomic and anti-genomic strands, the mapped vd-sRNA with zero and 1 mismatch is plotted on the genomic and anti-genomic strand of the PSTVd-RG1 (Figure 4). The profiles obtained for 21-, 22-, 23-, 24-nt long vd-sRNA and the cumulative 21- to 24-nt long vd-sRNAs are presented in Fig. 4A and Fig. 4B, for zero and 1 mismatch, respectively.

**Discussion**

Though viroids are single stranded and circular, due to (i) sequence complementarity, they form highly base-paired secondary structures, and (ii) replication through asymmetric rolling circle mechanism, viroid act as both the inducers of RNA and targets of host’s RNA silencing machinery (reviewed in [30]). That said, upon infection, all the viroids trigger RNA silencing and that in-turn cleaves viroid RNA into sRNAs of 21- to 24-nt. The accumulation of such [vd-sRNAs](javascript:;) has been extensively studied in different viroid-host combinations [10,11,13]. Since, viroids are non-coding pathogens, recent works were directed towards understanding the role of such vd-sRNA in its pathogenicity and symptom induction [14–16,21,31–34]. This works including predicted the vd-sRNA:target mRNA duplex formation and then, looking for the number of vd-sRNA in the viroid infected plant that potentially binds predicted target mRNA. For latter studies, it is important to map the vd-sRNA from NGS data obtained from viroid infected plants. Profiling obtained vd-sRNA on both the genomic and antigenomic strand of the viroid helps to understand that regions of viroid that produces more vd-sRNA and the region that produces less vd-sRNA. That indirectly suggests which region of the viroid is more susceptible and resistance to host RNA silencing. Such information helps in developing RNAi mediated viroid resistance transgenic plant development [35]. That said, mapping of vd-sRNA and profiling on viroid is very important in understanding host-viroid interaction.

Viroids being circular in nature, a specific computer software is required to map the sRNAs derived from the junction of first and last nucleotide of the viroid. This feature is not available in commercially available software. Our Python-based software tool is tailormade to address this issue along with the ~~To address this issue along with~~ detection of both genomic and antigenomic strand derived sRNAs. Besides its capabilities for detecting mappings, our tool is attractive for its efficiency: table 6 shows that large instances of the viroid binding problem involving as many as 5 million nucleotide sequences can be solved in under 40 minutes on a desktop computer with a single core. This eliminates the need for high-end hardware requirements which are often the case for commercial software. Lastly, since our software is developed on Python, it is straightforward to customize it for various scientific purposes. ~~reduced computation runtime and easy-to-use, a python-interface is specifically selected in this work.~~

The solution to the instance of a viroid binding problem, computed by our in-house software is represented by two matrices with integer entries. The first matrix describes the solution for forward matchings – number of sRNA nucleotides in the pool that match with every substring of the gene. The second, describes the solution for reverse matchings. In each of the two cases, the matchings are specified by a matrix -- $M^{+}$ for the forward matchings and $M^{-}$ for the reverse matchings – of size $(L \times n)$ where $L$ is the number of distinct nucleotide lengths in the pool, and $m$ is the length of the gene. Each entry $M^{+}\_{\ell,j}$ (or, $M^{-}\_{\ell,j}$) is the number of sRNA sequences in the pool, of length indexed by $\ell$ that have a forward (or reverse) matching with a substring of the gene starting at position $j$.

|  |  |  |  |
| --- | --- | --- | --- |
| **Size of the sRNA pool** | **Size of the gene** | **Tolerance** | **Runtime in seconds** |
| 5875050 | 359 | 0 | 1450 |
| 5875050 | 359 | 1 | 1800 |
| 730499 | 359 | 0 | 128 |
| 730499 | 359 | 1 | 131 |

Table 6 The above table shows the runtime of our inhouse tool for computing solutions to the viroid binding problem (in definition 1). Note that large instances, i.e., for a pool with around 5.8 million nucleotides are solved in under 30 minutes. This is significantly faster than a naïve solution (Earlier paper where we used the slow version of this code, and [\*]. These runtimes are reported on a single core of a 3.2 GHz Intel i7 processor. Improving these runtimes comes at the cost of a higher memory usage.

[\*]: Article:  
Title: Pattern Matching: Overview, Benchmark and Comparison with F-Transform General Matching Algorithm  
Year: 2017  
Issue date: July 2017  
Publisher: Springer-Verlag  
Volume: 21  
Number 13  
DOI: *https://doi.org/10.1007/s00500-017-2618-3*

In addition to solving the viroid bindings problem, the tool developed in this work is also capable of a visual representation of the output, by profiling the vd-sRNA on the genomic and antigenomic strand of the viroid strand (Fig. 2; Fig.3, and Fig. 4). Here, the X-axis shows the positions on the gene sequence. While the forward matching solutions are represented on the positive Y-axis, the reverse matchings are on the negative Y-axis. For each integer $i$ on the x-axis, the corresponding value on the positive Y-axis is simply $M^{+}\_{\ell, i}$ for a fixed length $\ell$, while the value on the negative Y-axis is $M^{-}\_{\ell, i}$. The area under the curve is filled for a visual appeal. It is important to note that the message conveyed in figures XX is identical to the results of [Adkar 2015, 2017], albeit the later adopting a slightly different visual representation for the plot. In previous studies [Adkar 2015, 2017], the height of the curve at a given position $i$ on the X-axis is the number of vd-sRNA nucleotides that match with a substring containing the character of the gene at position $i$. This can be derived from the $M^{+}, M^{-}$ matrices of our tool followed by a simple postprocessing routine.

To provide the user friendly and the efficient bioinformatics tool to viroid researchers, vd-sRNA profiler is developed in the present study by pattern matching using python language. This software seamlessly allows researchers to map and plot vd-sRNAs on their parent viroid molecule. Choosing the different matching tolerance allows the user to consider and visualize the vd-sRNA derived from viroid quasi-species. This software will help viroid researchers to use in their studies to evaluate the production of vd-sRNA and its profiling on their target viroid species.

1. F. Le Gall, "Faster Algorithms for Rectangular Matrix Multiplication," 2012 IEEE 53rd Annual Symposium on Foundations of Computer Science, 2012, pp. 514-523, doi: 10.1109/FOCS.2012.80.
2. Title: Some string matching problems from Bioinformatics which still need better solutions,  
   Author: Gaston H. Gonnet,  
   Journal: Journal of Discrete Algorithms,  
   Volume: Volume 2, Issue 1,  
   DOI: https://doi.org/10.1016/S1570-8667(03)00062-5.  
   Pages: 3-15  
   URL: <https://www.sciencedirect.com/science/article/pii/S1570866703000625>
3. Book  
   Title: Combinatorial Pattern Matching  
   Subtitle: 16th Annual Symposium, CPM 2005, Jeju Island, Korea, June 19-22, 2005. Proceedings  
   Editors: Alberto Apostolico, Maxime Crochemore, and Kunsoo Park  
   Series title: Lecture Notes in Computer Science  
   DOI: <https://doi.org/10.1007/b137128>  
   Publisher: Springer-Verlag
4. Article:  
   Title: An efficient algorithm for the blocked pattern matching problem  
   Authors: Fei Deng, Lusheng Wang and Xiaowen Liu  
   Journal: *Bioinformatics*, Volume 31, Issue 4, 15 February 2015, Pages 532–538,  
   DOI: https://doi.org/10.1093/bioinformatics/btu678
5. Book:  
   Title: Pattern Matching Algorithms  
   Editors: Alberto Apostolico, Zvi Galil  
   Publisher: Oxford University Press, 1997
6. Same as 1
7. Title: Computing similarity between RNA strings.  
   Authors: Bafna V., Muthukrishnan S., Ravi R.  
   In series: Combinatorial Pattern Matching. Lecture Notes in Computer Science.  
   Volume: 937  
   Pages: 1-16  
   Year: 1995  
   Publisher: Springer  
   DOI: <https://doi.org/10.1007/3-540-60044-2_30>
8. Article:  
   Title: Fast Pattern Matching in Strings  
   Authors: Knuth, Donald E. and Morris, Jr., James H. and Pratt, Vaughan R.  
   Journal: SIAM Journal on Computing  
   Volume: 6  
   Number: 2  
   Pages: 323-350  
   Year: 1977  
   DOI: <https://doi.org/10.1137/0206024>
9. In proceedings:  
   Title: Fast Searching in Packed Strings  
   Author: Philip Bille  
   Booktitle: Combinatorial Pattern Matching  
   Year: 2009  
   Editors: Gregory Kucherov and Esko Ukkonen  
   Pages: 116—126  
   Publisher: Springer  
   DOI: <https://doi.org/10.1007/978-3-642-02441-2_11>
10. PhD thesis:  
    Title: Topics in combinatorial pattern matching  
    Author: Hjalte Wedel Vildhøj  
    Year: 2015  
    Publisher: Technical University of Denmark  
    Series: DTU Compute PHD-2014  
    Number: 348

References: [[I will add them using some bibliography software, later.]]

1. Fast algorithms for matrix multiplication (unpublished version of Computing similarity between RNA strings: https://link.springer.com/chapter/10.1007/3-540-60044-2\_30 in 7): https://arxiv.org/abs/1204.1111
2. Need for efficient string-matching solutions in bioinformatics: https://www.sciencedirect.com/science/article/pii/S1570866703000625
3. Combinatorial pattern matching: https://link.springer.com/book/10.1007/b137128
4. Block pattern matching: https://academic.oup.com/bioinformatics/article/31/4/532/2748202
5. Introduction to Pattern matching book: https://www.amazon.ca/Pattern-Matching-Algorithms-Alberto-Apostolico/dp/0195113675
6. Fast algorithms for matrix multiplication: https://ieeexplore.ieee.org/document/6375330
7. Computing similarity between RNA strings: https://link.springer.com/chapter/10.1007/3-540-60044-2\_30
8. Knuth Pratt pattern matching algorithm: https://epubs.siam.org/doi/10.1137/0206024
9. Fast pattern matching in packed strings: https://link.springer.com/chapter/10.1007/978-3-642-02441-2\_11
10. Counting version of the pattern matching problem: <https://orbit.dtu.dk/en/publications/topics-in-combinatorial-pattern-matching>.

1. Kitabayashi, S.; Tsushima, D.; Adkar-Purushothama, C.R.; Sano, T. Identification and molecular mechanisms of key nucleotides causing attenuation in pathogenicity of Dahlia isolate of potato spindle tuber viroid. *Int. J. Mol. Sci.* **2020**, *21*, doi:10.3390/ijms21197352.

2. Ding, B. The biology of viroid-host interactions. *Annu. Rev. Phytopathol.* **2009**, *47*, 105–131, doi:10.1146/annurev-phyto-080508-081927.

3. Tsagris, E.M.; Tsagris, E.M.; Martínez de Alba, A.E.; Martínez de Alba, A.E.; Gozmanova, M.; Gozmanova, M.; Kalantidis, K.; Kalantidis, K. Viroids. *Cell. Microbiol.* **2008**, *10*, 2168–79, doi:10.1111/j.1462-5822.2008.01231.x.

4. Flores, R.; Hernández, C.; Alba, A.E.M. de; Daròs, J.-A.; Serio, F. Di Viroids and Viroid-Host Interactions. *Annu. Rev. Phytopathol.* **2005**, *43*, 117–139, doi:10.1146/annurev.phyto.43.040204.140243.

5. Pallas, V.; Martinez, G.; Gomez, G. The interaction between plant viroid-induced symptoms and RNA silencing. *Methods Mol. Biol.* **2012**, *894*, 323–343.

6. Ding, B. Viroids: Self-replicating, mobile, and fast-evolving noncoding regulatory RNAs. *Wiley Interdiscip. Rev. RNA* 2010, *1*, 362–375.

7. Dadami, E.; Boutla, A.; Vrettos, N.; Tzortzakaki, S.; Karakasilioti, I.; Kalantidis, K. DICER-LIKE 4 But Not DICER-LIKE 2 may have a positive effect on potato spindle tuber viroid accumulation in nicotiana benthamiana. *Mol. Plant* 2013.

8. Itaya, A.; Folimonov, A.; Matsuda, Y.; Nelson, R.S.; Ding, B. Potato spindle tuber viroid as inducer of RNA silencing in infected tomato. *Mol. Plant. Microbe. Interact.* **2001**, *14*, 1332–1334, doi:10.1094/MPMI.2001.14.11.1332.

9. Papaefthimiou, I.; Hamilton, A.; Denti, M.; Baulcombe, D.; Tsagris, M.; Tabler, M. Replicating potato spindle tuber viroid RNA is accompanied by short RNA fragments that are characteristic of post-transcriptional gene silencing. *Nucleic Acids Res.* **2001**, *29*, 2395–2400, doi:10.1093/nar/29.11.2395.

10. Bolduc, F.; Hoareau, C.; St-Pierre, P.; Perreault, J.-P. In-depth sequencing of the siRNAs associated with peach latent mosaic viroid infection. *BMC Mol. Biol.* **2010**, *11*, 16, doi:10.1186/1471-2199-11-16.

11. Ivanova, D.; Milev, I.; Vachev, T.; Baev, V.; Yahubyan, G.; Minkov, G.; Gozmanova, M. Small RNA analysis of potato spindle tuber viroid infected phelipanche ramosa. *Plant Physiol. Biochem.* **2014**, *74*, 276–282, doi:10.1016/j.plaphy.2013.11.019.

12. Navarro, B.; Pantaleo, V.; Gisel, A.; Moxon, S.; Dalmay, T.; Bisztray, G.; Di Serio, F.; Burgyán, J. Deep sequencing of viroid-derived small RNAs from grapevine provides new insights on the role of RNA silencing in plant-viroid interaction. *PLoS One* **2009**, *4*, e7686, doi:10.1371/journal.pone.0007686.

13. Tsushima, D., Adkar-Purushothama, C.R., Taneda, A., and Sano, T. Changes in relative expression levels of viroid-specific small RNAs and microRNAs in tomato plants infected with severe and mild isolates of Potato spindle tuber viroid. *J. Gen. Plant Pathol* **2015**, *81*, 49–62.

14. Navarro, B.; Gisel, A.; Rodio, M.E.; Delgado, S.; Flores, R.; Di Serio, F. Small RNAs containing the pathogenic determinant of a chloroplast- replicating viroid guide the degradation of a host mRNA as predicted by RNA silencing. *Plant J.* **2012**, *70*, 991–1003, doi:10.1111/j.1365-313X.2012.04940.x.

15. Eamens, A.L.; Smith, N.A.; Dennis, E.S.; Wassenegger, M.; Wang, M.B. In Nicotiana species, an artificial microRNA corresponding to the virulence modulating region of Potato spindle tuber viroid directs RNA silencing of a soluble inorganic pyrophosphatase gene and the development of abnormal phenotypes. *Virology* **2014**, *450*–*451*, 266–277, doi:10.1016/j.virol.2013.12.019.

16. Adkar-Purushothama, C.R.; Brosseau, C.; Giguère, T.; Sano, T.; Moffett, P.; Perreault, J.-P. Small RNA Derived from the Virulence Modulating Region of the Potato spindle tuber viroid Silences callose synthase Genes of Tomato Plants. *Plant Cell* **2015**, *27*, 2178–94, doi:10.1105/tpc.15.00523.

17. Adkar-Purushothama, C.R.; Sano, T.; Perreault, J.-P. Viroid derived small RNA induces early flowering in tomato plants by RNA silencing. *Mol. Plant Pathol.* **2018**, doi:10.1111/mpp.12721.

18. Adkar-Purushothama, C.R.; Perreault, J.-P. Alterations of the viroid regions that interact with the host defense genes attenuate viroid infection in host plant. *RNA Biol.* **2018**, 1–12, doi:10.1080/15476286.2018.1462653.

19. Adkar-Purushothama, C.R.; Perreault, J. Current overview on viroid-host interactions. *Wiley Interdiscip. Rev. RNA* **2020**, *11*, e1570, doi:10.1002/wrna.1570.

20. Adkar-Purushothama, C.R.; Perreault, J.P. Impact of nucleic acid sequencing on viroid biology. *Int. J. Mol. Sci.* **2020**, *21*, doi:10.3390/ijms21155532.

21. Diermann, N.; Matoušek, J.; Junge, M.; Riesner, D.; Steger, G. Characterization of plant miRNAs and small RNAs derived from potato spindle tuber viroid (PSTVd) in infected tomato. *Biol. Chem.* **2010**, *391*, 1379–1390, doi:10.1515/BC.2010.148.

22. Brass, J.R.J.; Owens, R.A.; Matoušek, J.; Steger, G. Viroid quasispecies revealed by deep sequencing. *RNA Biol.* **2017**, *14*, doi:10.1080/15476286.2016.1272745.

23. Adkar-Purushothama, C.R.; Bolduc, F.; Bru, P.; Perreault, J.-P. Insights Into Potato Spindle Tuber Viroid Quasi-Species From Infection to Disease. *Front. Microbiol.* **2020**, *11*, 1235, doi:10.3389/fmicb.2020.01235.

24. Adkar-Purushothama, C.R.; Kasai, A.; Sugawara, K.; Yamamoto, H.; Yamazaki, Y.; He, Y.-H.; Takada, N.; Goto, H.; Shindo, S.; Harada, T.; et al. RNAi mediated inhibition of viroid infection in transgenic plants expressing viroid-specific small RNAs derived from various functional domains. *Sci. Rep.* **2015**, *5*, 17949, doi:10.1038/srep17949.

25. Adkar-Purushothama, C.R.; Iyer, P.S.; Perreault, J.-P. Potato spindle tuber viroid infection triggers degradation of chloride channel protein CLC-b-like and Ribosomal protein S3a-like mRNAs in tomato plants. *Sci. Rep.* **2017**, *7*, 8341, doi:10.1038/s41598-017-08823-z.

26. Hutchins, C.J.; Keese, P.; Visvader, J.E.; Rathjen, P.D.; McInnes, J.L.; Symons, R.H. Comparison of multimeric plus and minus forms of viroids and virusoids. *Plant Mol. Biol.* **1985**, *4*, 293–304, doi:10.1007/BF02418248.

27. Gago, S.; Elena, S.F.; Flores, R.; Sanjuán, R. Extremely high mutation rate of a hammerhead viroid. *Science* **2009**, doi:10.1126/science.1169202.

28. López-Carrasco, A.; Ballesteros, C.; Sentandreu, V.; Delgado, S.; Gago-Zachert, S.; Flores, R.; Sanjuán, R. Different rates of spontaneous mutation of chloroplastic and nuclear viroids as determined by high-fidelity ultra-deep sequencing. *PLOS Pathog.* **2017**, *13*, e1006547, doi:10.1371/journal.ppat.1006547.

29. Tessitori, M.; Rizza, S.; Reina, A.; Causarano, G.; Di Serio, F. The genetic diversity of Citrus dwarfing viroid populations is mainly dependent on the infected host species. *J. Gen. Virol.* **2013**, *94*, 687–693, doi:10.1099/vir.0.048025-0.

30. Pallás, V., Martínez, G., and Gómez, G. The interaction between plant viroid induced symptoms and RNA silencing. *Methods Mol. Biol.* **2012**, *894*, 323–343.

31. Bao, S.; Owens, R.A.; Sun, Q.; Song, H.; Liu, Y.; Eamens, A.L.; Feng, H.; Tian, H.; Wang, M.-B.; Zhang, R. Silencing of transcription factor encoding gene StTCP23 by small RNAs derived from the virulence modulating region of potato spindle tuber viroid is associated with symptom development in potato. *PLoS Pathog.* **2019**, *15*, e1008110, doi:10.1371/journal.ppat.1008110.

32. Adkar‐Purushothama, C.R.; Perreault, J. Current overview on viroid–host interactions. *WIREs RNA* **2020**, *11*, doi:10.1002/wrna.1570.

33. Thibaut, O.; Claude, B. Innate immunity activation and RNAi interplay in citrus exocortis viroid-tomato pathosystem. *Viruses* **2018**, *10*, 587, doi:10.3390/v10110587.

34. Aviña-Padilla, K.; Rivera-Bustamante, R.; Kovalskaya, N.Y.; Hammond, R.W. Pospiviroid infection of tomato regulates the expression of genes involved in flower and fruit development. *Viruses* **2018**, *10*, 516, doi:10.3390/v10100516.

35. Adkar-Purushothama, C.R.; Kasai, A.; Sugawara, K.; Yamamoto, H.; Yamazaki, Y.; He, Y.-H.; Takada, N.; Goto, H.; Shindo, S.; Harada, T.; et al. RNAi mediated inhibition of viroid infection in transgenic plants expressing viroid-specific small RNAs derived from various functional domains. *Sci. Rep.* **2015**, *5*, 17949, doi:10.1038/srep17949.

**Figure legends**

**Figure 1**. Schematic representation of software workflow

**Figure 2**. Profiling of the vd-sRNAs recovered from PSTVd-I infected tomato plants at zero mismatches. Sequence profiles of the (+) and the (−) PSTVd-sRNA populations recovered from the leaf tissues of infected tomato plants. Panels (A), (B), (C), (D) and (E) represent the profiles of the (+) and the (−) PSTVd-I derived sRNAs of sizes 21-, 22-, 23-, 24 and cumulative of 21- to 24-nt, respectively. Data is normalized to reads per million.

**Figure 3**. Profiling of the vd-sRNAs recovered from PSTVd-I infected tomato plants at 1 mismatch. Sequence profiles of the (+) and the (−) PSTVd-sRNA populations recovered from the leaf tissues of infected tomato plants. Panels (A), (B), (C), (D) and (E) represent the profiles of the (+) and the (−) PSTVd-I derived sRNAs of sizes 21-, 22-, 23-, 24 and cumulative of 21- to 24-nt, respectively. Data is normalized to reads per million.

**Figure 4**. Profiling of the vd-sRNAs recovered from PSTVd-I infected tomato plants at zero and 1 mismatches. Sequence profiles of the (+) and the (−) PSTVd-sRNA populations recovered from the leaf tissues of infected tomato plants. Panels (A), (B), (C), (D) and (E) represent the profiles of the (+) and the (−) PSTVd-I derived sRNAs of 0 mismatches of sizes 21-, 22-, 23-, 24 and cumulative of 21- to 24-nt while, Panels (F), (G), (H), (I) and (J) represent the profiles of the (+) and the (−) PSTVd-I derived sRNAs of of 1 mismatches of sizes 21-, 22-, 23-, 24 and cumulative of 21- to 24-nt, respectively. Data is normalized to reads per million.