Universiteit Leiden
Opleiding Informatica

Secondary Structure Models and Homology Search for Viral RNA

Name: Alan Zammit

Supervisor: Dr. Alexander P. Goultiaev

2nd Supervisors: Dr. René R.C.L. Olsthoorn and
Prof.dr.ir. Fons J. Verbeek

MASTER'S THESIS

Leiden Institute of Advanced Computer Science (LIACS)
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

Contents

Abstract **3**

1 Introduction and Background **3**

1.1 Overview of the genus *Flavivirus* 4

1.2 General characteristics of flavivirus UTRs 6

1.3 *Flaviviridae* 3' UTR structures and intracellular interactions 9

1.4 Structural RNA and homology search 12

2 Materials and Methods **16**

2.1 High-level overview of algorithmic approach 19

3 Results and Discussion **22**

TR.1 Base-paired model for the 5' DB specific to WNV 22

TR.2 Model for 5' & 3' DB of specific DENV and JEV group members 25

TR.3 Model for 5' & 3' DB structures across MBFV group viruses 27

TR.4 CSSD for SL5' and SL3' structures in MBFV viruses 29

TR.5 Models for viruses pertaining to the TBFV group 33

TR.6 Generalized Y-shaped structure for all *Flavivirus* 3' UTRs 35

Summary of results and discussion 36

4 Further Work and Conclusion **40**

A Prototype Implementation **42**

A.1 Requirements 42

A.2 Scoring algorithm and computational performance 49

A.3 Limitations and enhancements 52

B Sequence/Structure Supplementary Information **53**

B.1 Folded structures for sequences matched by *CSSD A* 53

B.2 Iterations required for *CSSD B* 54

B.3 5'-proximal nucleotide frequency distribution for all sequences used 57

B.4 Extended Test Run 1 to 6 Results 59

B.5 Approximate stacking energy for Test Run 3 to 5 hits 68

References **72**

Abstract

In this study, the use of a topology-based RNA search tool is investigated, with particular application to functional non-coding RNAs in *Flavivirus* genomes. In contrast to current methods based on structural comparative analysis and probabilistic techniques, the approach adopted facilitates the construction and application of a user-defined variable descriptor (a *consensus secondary structure descriptor*) for a given set of homologous sequences, thus allowing for arbitrary secondary structures including pseudoknots to be represented across related genomic regions. The user can subsequently efficiently search against both consensus structure and sequence databases, based on models of functional elements in regions of interest. In this work, of particular interest are structural elements found in the 3' untranslated region of flaviviral genomes.

Preliminary results obtained from experiments run on both Mosquito-Borne (MBFV) and Tick-Borne (TBFV) flaviviral groups are indicative of good levels of sensitivity (with a mean estimate of 82% across both groups), specificity (at 92% mean), and a mean F_1 score of 0.78 when identifying key structural elements of the viral family's 3' untranslated region, ushering the possibility of further study on the method adopted and its application to genomes of interest.

1 Introduction and Background

This first section introduces the *Flavivirus* genus of viruses and some of their defining secondary structural properties, providing a high-level perspective on their significance in biological and epidemiological studies. Characteristics of the 3' untranslated region of flaviviruses are summarily described, highlighting the relevance of non-coding functional elements and tertiary interactions essential to the virulent activity of this genus.

An alternative to mainstream methods for homology search in viral RNA is introduced in the subsequent section, with the objective to allow for identification of signature structures within the genus that are shared between phylogenetically-related groups.

A prototypical implementation of the proposed method is tested in a series of trial runs and which are outlined in the third section along with a brief discussion on results. Within the last section concluding remarks and recommendations for future work are provided.

1.1 Overview of the genus Flavivirus

Flaviviruses are a genus of positive single-stranded RNA viruses belonging to the family *Flaviviridae*, and which include a number of medically relevant pathogens that, in the most part, cycle between mosquito or tick arthropods, and susceptible humans. Amongst the more clinically relevant arthropod-borne viruses (*arboviruses*) are the Dengue virus and West Nile virus. Dengue is perhaps the most important viral disease in humans transmitted by mosquitos (mainly of the species *Aedes aegypti*[1]) and accounts for ~390 million infections per year (Villordo et al., 2016), with no vaccine or antiviral treatment available to date. Levels of pathogenicity vary, even in closely related species – in excess of 27.000 cases have been reported for the highly pathogenic North American strain of the West Nile virus (WNV_{NY99}) since its emergence in 1999, whereas, in contrast, the closely related Australian strain (WNV_{KUN}) does not present itself as highly pathogenic in both humans and animals (Pijlman et al., 2008). Currently, no completely effective therapeutic option exists for any flavivirus infection.

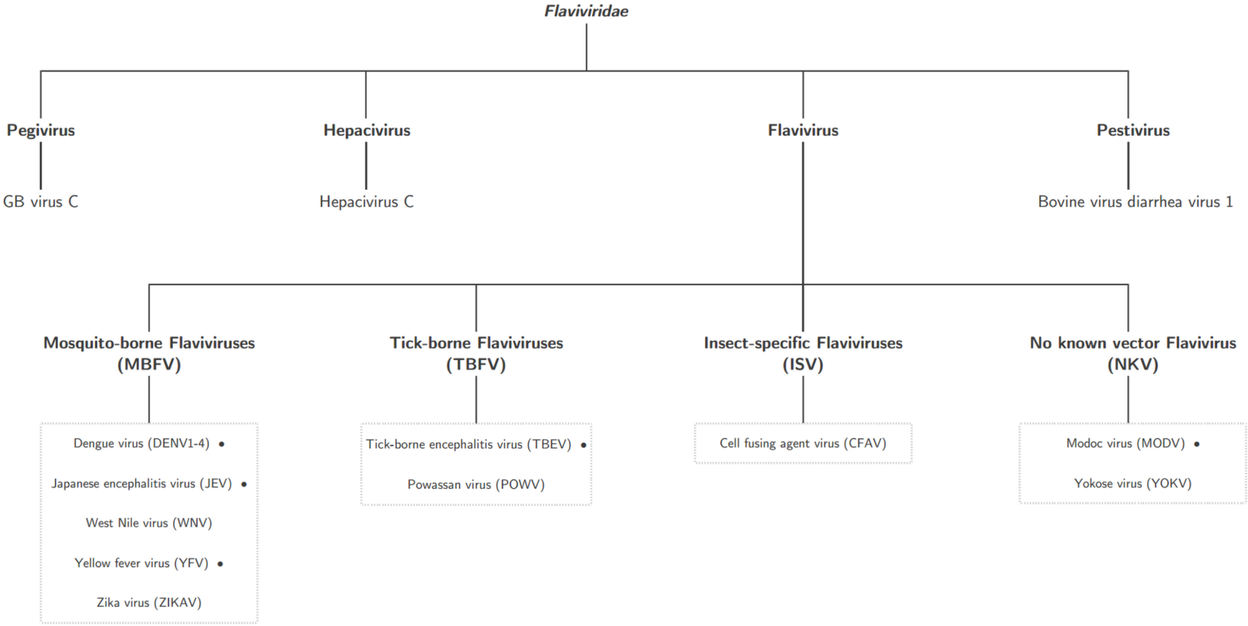

Flaviviridae
Pegivirus
Hepacivirus
Flavivirus
Pestivirus
GB virus C
Hepacivirus C
Bovine virus diarrhea virus 1
Mosquito-borne Flaviviruses (MBFV)
Tick-borne Flaviviruses (TBFV)
Insect-specific Flaviviruses (ISV)
No known vector Flavivirus (NKV)
Dengue virus (DENV1-4) •
Japanese encephalitis virus (JEV) •
West Nile virus (WNV)
Yellow fever virus (YFV) •
Zika virus (ZIKAV)
Tick-borne encephalitis virus (TBEV) •
Powassan virus (POWV)
Cell fusing agent virus (CFAV)
Modoc virus (MODV) •
Yokose virus (YOKV)

Figure 1 Overview of the *Flaviviridae* Family. The four genera stemming from the *Flaviviridae* family are shown including type species listed under each genus. For *Flavivirus*, its four ecological groups are named, in addition to a representative list of member viruses. A selection of well-studied flaviviral subgroups are highlighted, marked by an adjacent •. Most molecular and phylogenetic studies agree on the clear separation between *Flavivirus* and the other genera, with variability reported with respect to the order of evolutionary events – refer to (Lobo et al., 2009) for a phylogenetic reconstruction and analysis based on whole polyprotein sequences, and (Vlachakis et al., 2013) for an assessment based on NS3 protein sequences. A number of studies also suggest association and common origin between the NKV and TBFV/MBFV groups (refer to, for instance, Villordo et al., 2016).

[1] the same species is also responsible for transmission of Chikungunya, Yellow fever, Mayaro and Zika viruses

Of the ~100 species in the *Flaviviridae* family, the *Flavivirus* genus (FV) accounts for more than half of known species (Simmonds et al., 2017), with mammals and birds typically understood to be the primary hosts. The other genera, *Pegivirus*, *Hepacivirus*, and *Pestivirus*, largely share a number of characteristics with *Flavivirus*, including, a wide mammalian host range, a spherically shaped 40-60 nm virion, cytoplasmic replication, and a genome that expresses a single polyprotein which is cleaved into both structural and non-structural proteins. Differentially in Flaviviruses, viral translation occurs directly from genomic RNA that manifests a type 1 cap − a structure which is missing in the other three genera that, however, allow for translation to occur in a cap-independent manner by means of an internal ribosome entry site (IRES) (Dong et al., 2014). Moreover, within the *Flavivirus* genus, though the genomic organization is similar and the mechanism of replication is homogeneous across species, differences in host ranges and viral transmissibility allow for the division of this genus into broad ecological groups. A taxonomic overview[2] of the four *Flaviviridae* genera is provided in **Figure 1**, including the four ecological groups of the *Flavivirus* genus.

The viral genomes for each genus (from left to right, as outlined in **Figure 1**) are approximately of length 8.9-11.3, 8.9-10.5, 9.2-11, and 12.3-13 kilobases (kb) (Simmonds et al., 2017), respectively. The genome of the largest genus, *Flavivirus*, encodes a single open reading frame (ORF) flanked by structured 5' and 3' untranslated regions (UTR) of which a high-level schematic is provided in **Figure 2**. All members of *Flaviviridae* have 5' and 3' UTRs that are embellished with conserved secondary structures, critical for genome replication and translation. In *Pegivirus*, *Hepacivirus*, and *Pestivirus*, specific structures in the 5' UTR act as an IRES site and overall manifest more elaborate structures when compared to *Flavivirus* (Roby et al., 2014). In the latter case, the 5' UTR is ~100 nucleotides (nt) long, whereas the 3' untranslated region ranges between 400 and 700nt, with exceptional cases exceeding 900 nt (Villordo et al., 2016).

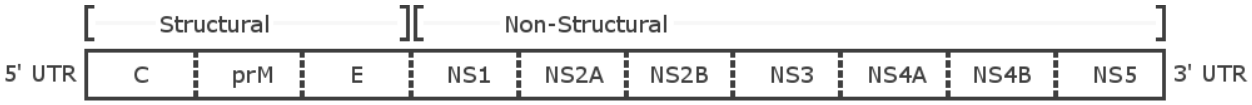


Figure 2 Structure of Flavivirus genome. Schematic diagram showing the structural and non-structural proteins, cleaved by both host and viral proteases, from the single polyprotein encoded by the genome. Flanked by two untranslated regions located at the 5' and 3' ends, the functional properties of the cleaved protein products include, starting from the 5' proximal viral capsid protein, **C**:core protein, **prM**:precursor of M protein, **E**:envelope protein, **NS1**:minus strand synthesis and protection from cell complement, **NS2A**:assembly of replication complex on the ER; **NS2B**:replication complex assembly on the ER and polyprotein protease, **NS3**:polyprotein protease, **NS4A/B**:assembly of replication complex on the ER, **NS5**:replication and cap formation (Fernández-Sanlés et al., 2017)

.

[2] only representative viruses or groups of the ~100 species are shown, for brevity

1.2 General characteristics of flavivirus UTRs

The 5' and 3' untranslated regions of flaviviral genomes are known to function as recognition sites for viral translation and replication, modulation of immune response, and pathogenesis (Bidet et al., 2014); (Gebhard et al., 2011); (Ng et al., 2017); (Pijlman et al., 2008); (Roby et al., 2014). Although all ecological groups contain conserved RNA structures in both 5' and 3' UTRs, *only two RNA structures are conserved in all flaviviral genomes*. At the 5' end, a Y-shaped stem-loop structure (*SLA*) is always present, and in addition, a small hairpin stem-loop (*sHP-3' SL*) is persistently located at the end of the 3' region. These two structures are well-understood to be essential for flaviviral propagation (refer to, for instance, Villordo et al., 2016). Sequence analysis for the DENV group indicates that both sequence and structural features of SLA are highly conserved (Filomatori et al., 2006), with similar findings obtained for the 3' stem-loop (Mohan et al., 1991). Moreover, the acquisition of a circular 'closed loop' topology – essential for the synthesis and replication of intact, full-length genomes – is enabled by means of long distance RNA-RNA interactions[3] between *sequence motifs* at the 5' and 3' proximal ends and which are conserved to varying degrees across all flavivirus genomes (Fernández-Sanlés et al., 2017). Villordo et al. estimated that the four DENV serotypes (DENV1-4) conserve sequence identity at different rates at different genomic locations (*Figure 2*, in Villordo et al., 2016), ranging from ~68% to ~97%, with higher rates observed at the extremeties of both untranslated regions.

A significant body of work has identified additional RNA elements within UTRs that are essential to the flavivirus life cycle. **Figure 3** outlines typical secondary structure elements found and the relative positioning within the genome (which, in the representation provided, pertain to the DENV genome). The following is a summary of principal findings and respective characteristics:

5' SLB. A short, ~30 nt stem-loop structure is located just upstream of the ORF start codon and is separated from the upstream ~70 nt-long SLA by a poly(U) sequence of typically not less than 10 nt (Lodeiro et al., 2009). SLB is known to manifest variability in both size and shape (Brinton et al., 1988), but invariably overlaps with an AUG translation initiation codon, typically embedded in its stem (Fernández-Sanlés et al., 2017). SLB also contains the 5' Upstream (initiation) AUG Region cyclization sequence (UAR), which, along with a paired cyclization sequence (CS) contained within the ORF, is involved in long-range interactions required for

[3] these interactions enable the correct positioning of NS5 RNA polymerase to initiate complementary strand synthesis. furthermore, in addition to evidence that secondary structure of *SLA* and *sHP-3' SL* is highly conserved, it is known that circularization demands a conformational change in both 5' and 3' UTRs depending on linear versus circular state (Gebhard et al., 2011)

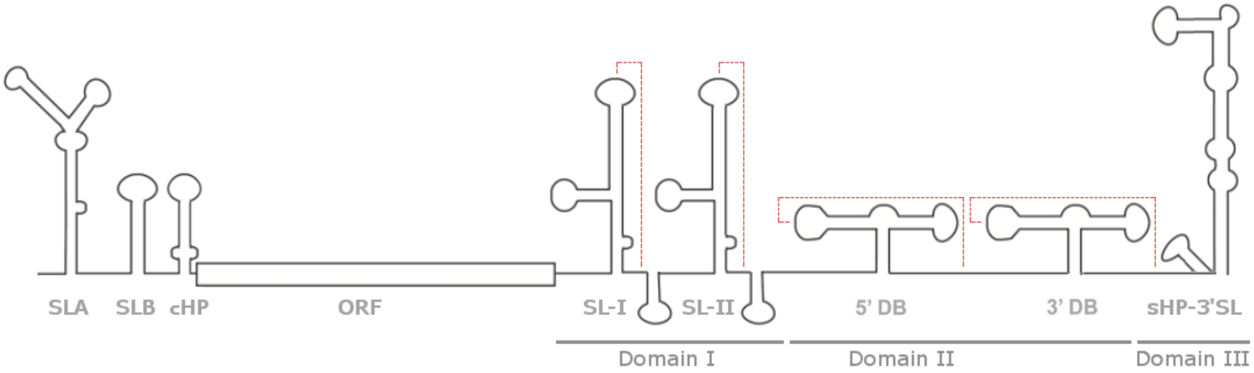


Figure 3 Outline of typical secondary structure elements in flaviviral genome. The schematic diagram shows structural elements found (specific to the DENV group) in the 5' and 3' UTRs flanking the ORF. In this case, duplicated structures can be seen in the 3' UTR of DENV, with both Y-shaped (SL-I and SL-II) and dumbbell (5' DB and 3' DB) showing a pseudoknot configuration, represented by the red dashed lines.

genome replication (Alvarez et al., 2005). Variations exist between different flaviviral groups: in DENV, a downstream AUG region (DAR) motif is involved in genome cyclization and present in the linker between SLB and the downstream *cHP* stem, whereas, two DAR motifs, DAR I and DAR II, act from within the stem and base of SLB in the JEV group - refer to, for example, *Figure 2* for WNV in (Fernández-Sanlés et al., 2017).

5' cHP. A highly-conserved hairpin sits at the 3' end, downstream of the 5' SLB and overlapping the coding region. Although less sequence conservation is manifest, it has been structurally identified in many species across MBFV and TBFV. Interestingly, the ability of this element to direct initiation[4] (by positioning the ribosomal complex) at the correct start codon is found to correlate with thermodynamic stability and is sequence independent (Clyde et al., 2006). Of particular relevance to this work, Clyde and Harris experimentally show that, restoring base pairs (albeit, in a different domain) in mutants that lack replication ability rescues wild-type functionality – highlighting further the notion of sequence independence.

Overview of 3' UTR. The flaviviral 3' UTR is highly structured, with regions conserved across species of varying sequence length, while lacking a poly(A) tail and ending with a conserved CU dinucleotide sequence[5]. Independent of length, all flavivirus 3' UTRs are divided into three domains[6] (I-III). In all species, stem-loop (SL) and dumbbell (DB) structures can be identified

[4] it is discussed in (Fernández-Sanlés et al., 2017) that the presence of stable secondary structural elements such as stem-loops downstream of an AUG start codon increase the likelihood of recognition of the optimal starting triplet by stalling and therefore prolonging contact between the translation machinery and the correct AUG start codon. Moreover, it is shown in (Clyde et al., 2008) that this structure has a *dual-function*, acting in both translation and RNA synthesis processes, highlighting the more general property that multiple functions can be carried out by the same structural unit

[5] this applies to all MBFV and TBFV strains, except for TBEV (Fernández-Sanlés et al., 2017)

[6] referred to in the literature as either domains or regions

along with conserved (repeated) sequences (R)CS, with a presence of duplicated elements that depends on the specific viral group. At the 3' extremity, the sHP-3'SL is essentially conserved across all flaviviruses.

The **first domain** downstream of the coding sequence, consists of a "hypervariable sequence" followed by duplicated stem-loop domains that have comparable structures in all species except in YFV and the TBFV/NKV groups, and in ISFVs (Fernández-Sanlés et al., 2017) where only one stem-loop[7] is present. Pseudoknots (PKs) are present and most typically involved in, or proximal to, apical loops of various structures, ending in downstream sequences.

Domain II is moderately conserved, containing characteristic dumbbell (DB) structures in MBFV and NKV flaviviruses (Fernández-Sanlés et al., 2017) also involving the formation of pseudoknots. Sequence and structure variability can be identified across different groups — for instance, members of the YFV group contain a pseudo-dumbbell (Ψ-DB) possibly derived from what was originally a duplicated DB.

Domain III is highly conserved, characterized by a 3' SL structure at the 3' extremity, identifiable across all species. Upstream of 3' SL is the equally conserved short hairpin (sHP) consisting of a 5 bp stem and conserved 6 nt apical loop, resembling a tetra loop motif known to be involved in RNA-RNA and RNA-protein interactions (Davis et al., 2013). Additionally, partly overlapping with sHP, is a highly conserved 24 nt-long sequence (CS1) demonstrated to be necessary for viral replication and cyclization in DENV (Wengler et al., 1986). The 3' SL stem-loop features a bulge in the upper portion conserved in all species (possibly required for NS5 protein recognition), and also contains several highly-conserved sequence segments. Based on extensive studies, the characteristics of sHP-3'SL are known to be critical for enabling the viral translation initiation and replication (Fernández-Sanlés et al., 2017), also by recruitment of accessory proteins necessary for ribosomal assembly (Polacek et al., 2009) in context of a genome that lacks a poly(A) tail at the 3' end.

[7] it should be noted here that in the literature, nomenclature for 3' UTR elements varies. For example, these duplicated stem-loop structures may be referred to as SL-I/SL-II in DENV, and SL-II/SL-IV in WNV, and more recently, they are referred to by the more functionally-descriptive name 'Xrn1-resistant RNAs' (xrRNAs). In this document, and to the extent possible, the nomenclature used in a given context is of that adopted by the cited work

1.3 Flaviviridae 3' UTR structures and intracellular interactions

Dengue encompasses four distinct serotypes (DENV1-4), each comprising multiple genotypes, and which include distinct lineages or clades (Filomatori et al., 2017). Variability in viral genomes differentially determines viral fitness and epidemic potential[8] (Bennett et al., 2009). The causality for genetic variability and viral displacement in flaviviruses (and in DENV, particularly) might not yet be comprehensively understood, but recent studies show that DENV ability to thrive in multiple hosts (that is, in mosquito and human cells) is linked to sequence variability in its 3' UTR[9]. In other work, it was seminally hypothesized that 3' UTR sequence variation that is also associated with multiple-host adaptation, is a causal factor in the generation of non-coding RNAs (*subgenomic flavivirus RNA*, or sfRNA) – refer to, for instance, (Kieft et al., 2015); (Villordo et al., 2016).

Currently understood to be characteristic of flaviviruses, sfRNAs have a contributing effect on biological mechanisms and virus-host interactions including viral cytopathicity in cells and pathogenicity in mice (Pijlman et al., 2008); (Silva et al., 2010), evasion of anti-viral response (Schuessler et al., 2012); (Moon et al., 2012), and involvement in apoptotic pathways (Liu et al., 2014). Moreover, sfRNAs are known to accumulate in flavivirus-infected cells and interfere in RNA decay pathways (Moon et al., 2012). The highly structured non-coding RNA produced is in the 0,3 - 0,5 kb length range. In (Pijlman et al., 2008), cells infected with mosquito-borne (WNV_{KUN}, WNV_{NY99}, MVEV, AFLV) and tick-borne (SREV) all produced sub-genomic RNA of similar size (~0,5 kb), while YFV- and DENV2-infected cells produced smaller RNA (~0,3 and ~0,4 kb, respectively), in correspondence with the respective 3' UTR sizes[10]. Further analysis of sfRNA production, done by the same authors, shows that sfRNA is amply produced in all the cell types tested (including those of both vertebrate and invertebrate origin). Additionally, it was demonstrated that RNA replication, viral proteins, or the 5' UTR are not essential for sfRNA generation, but rather, cellular proteins were responsible for sfRNA production. The above observations, and, in particular, the ability of viruses to generate different sfRNAs that rapidly change upon host switching, readily pose a number of interesting questions.

[8] the study and control of outbreaks is thus further hampered by complex epidemiological dynamics, such as that seen in the 1990s, where DENV2 from Southeast Asia outplaced the American DENV2, with a significantly higher impact on health in Latin American countries (Filomatori et al., 2017)

[9] sequence analysis of DENV populations obtained from adult mosquitos or mosquito cells showed that mutations mapped onto the 3' UTR were removed after a transfer to human cells (Villordo et al., 2015)

[10] in addition, production of subgenomic RNA was not detected in unrelated virus-infected cells (Semliki Forest virus/*Togaviridae* genus of *Alphavirus* family), further crediting the hypothesis that generation of sfRNA is restricted to and conserved amongst members of the flavivirus genus

A key question that also underscores the motivation for this work is about what drives the different species of sfRNAs to be produced.

In (Filomatori et al., 2017), RNA structural analysis based on human and mosquito viral variants revealed that mutations in a stem loop (xrRNA2), determine[11] the accumulation of shorter species of sfRNAs (referred to as sfRNA3 and sfRNA4) in mosquito-adapted viruses, whereas the longer species sfRNA1 is the main product of viruses that replicate in human cells. The study conducted thus exposes how variability at the 3' UTR of DENV influences and controls the generation of specific patterns of non-coding RNAs (ncRNAs), which in turn is linked to different levels of fitness in each host.

In a series of experiments carried out by Pijlman et al., the authors first rule out the involvement of replication promoters, the 5' UTR, and viral proteins from the production process of sfRNA. Subsequently, also given that sfRNA is known to correspond to 3'-terminal genome fragments, the authors hypothesized that hydrolytic cleavage by cellular ribonuclease (with 5'→3' hydrolyzing action) was the responsible agent for its production. Analysis of sfRNA levels in independent 5'→3' exoribonuclease 1 (*Xrn1*[12]) knockdown experiments validated the hypothesis, showing significant downregulation of sfRNA.

Xrn1-mediated decay is known to be an important pathway for mRNA decay in cells (Georg et al., n.d.) and thus mRNA stability (Moon et al., 2012). In other work, Moon et al. argue that "between 20 and 50% of gene expression may be regulated post-transcriptionally at the level of mRNA decay" (Moon et al., 2015), clearly highlighting the importance of this pathway. Filomatori et al. note that mutations at xrRNA2 impeding PK formation are sufficient to impair the xrRNA1/2 function of stalling Xrn1, thus accruing shorter sfRNAs. Additionally, the authors note that the relative thermodynamic stabilities of xrRNA1/2 – known to be contingent on PK formation (Funk et al., 2010) – offers an explanation for characteristics of different flaviviruses: in WNV, a higher stability of xrRNA1 results in efficient halting of Xrn1 - see also, (Pijlman et al.,

[11] multiple experiments where carried out by Filomatori et al., involving both destructive and reconstitutive mutations, and different xrRNAs, indeed showing a more subtle, complex pattern of functional dependence between different structures and their impact on sfRNA production

[12] eukaryotes embody two 5'→3' exoribonucleases, Xrn1 and Xrn2. the latter is bound within the nucleus and involved in activities such as RNA maturation and transcription termination, whereas, Xrn1 is predominantly cytoplasmic and is required for degradation of 5' monophosphorylated RNA. the structure of Xrn1 is known for various organisms, including *Drosophila melanogaster* and *Homo sapiens*. It is reported in (Jones et al., 2012) that Xrn1 is highly conserved between these two organisms in the nuclease domain, whereby processive degradation occurs by means of a helical structure (known as the *tower domain*) that acts like a "ratchet-like mechanism" . It is also stated that *such structural characteristics largely explain the specificity* of Xrn1 for 5' monophosphorylated RNA and its processivity

2008), whereas in Zika virus (ZIKV), comparable stabilities of the same two structures results in the accumulation of similar amounts of subgenomic flavivirus RNA.

Interestingly, experimental results by MacFadden et al. show that xrRNAs can halt exoribonucleases other than Xrn1[13], and that the halting site of the enzyme depends on the interface between the RNA and the enzyme (possibly relating to the distance between the active site and the surface of each enzyme) – the overall implication being that xrRNAs may operate as general 'mechanical blocks' and therefore enable flaviviruses to thrive in a variety of hosts and vectors (MacFadden et al., 2018). Moreover, it was found that xrRNA mutated to modify the folded structure ablated resistance in all three enzymes used in the experimental setup, indicating that resistance hinges on RNA topological arrangement and not on enzyme specific characteristics. In addition, it was shown that mutations involving tertiary structure interactions (base-triples, pseudoknots) negatively impacted the ability to resist Xrn1. It is also shown experimentally that members of ISFV and NKV, whilst having divergent sequences to those highly conserved in MBFVs, exhibit structural characteristics including long-range interactions similar to MBFV.

Also recently, other members of the *Flaviviridae* family [Hepatitis C Virus (HCV) and Bovine Viral Diarrhea Virus (BVDV)] were found to stall Xrn1 at the *5' UTR*, and interfere with its enzymatic activity, though to a lesser degree when compared to FV (Moon et al., 2015). Whereas the general structure required for stalling Xrn1 at the *3' UTR* of flaviviruses is well known (Chapman et al., 2014), functional and structural studies relating the two regions with respect to Xrn1 are not yet available, however, the functional importance of PKs in both regions is a relatively well-established fact (Moon et al., 2015).

[13] more fundamentally, the bacterial and yeast enzymes employed experimentally to challenge Xrn1 structures are not naturally exposed to the virus

1.4 Structural RNA and homology search

The accumulation of large swathes of genomic data has provided the scientific community with valuable raw data within which to mine, allowing for a better understanding of how biological entities use the information encoded within genomes. A fundamental step towards this goal is to identify and determine the roles of functional elements within sequences and detecting *homology* - the patterns of similarity indicative of shared evolutionary ancestry. In this section, computational methods for detecting structural RNA homology are succinctly reviewed with an eye to the respective merits and disadvantages.

RNAs functionally depend on, and thus conserve, a specific three dimensional structure that is energetically favourable — where, many times conservation is observed to prevail across evolutionary timescales. An RNA's structure is contingent on intra-molecular interactions between residues in its polynucleotide chain as well as by inter-molecular interactions with neighboring RNAs or proteins. Most interactions are based on hydrogen bonds formed through Watson-Crick base-pairing of two residues. The most common (and thermodynamically favorable) pairs are the so-called canonical Watson-Crick (WC) G-C and A-U base-pairs[14]. Also common is the G-U *wobble* pair, albeit typically less stable than WC pairing. Base-pairs normally stack up in groups (*stems*) that form thermodynamically favourable helices, and the set of which define an RNA's *secondary structure*. Also critical, though less frequently occurring, are *pseudoknot* (PK) structures. Whereas classical base-paired structures follow well-nested topologies, base-pairing in pseudoknotted structures allows for overlaps in sequence position. **Figure 4** shows an archetypal secondary structure representation of RNA including a simple example of a pseudoknot.

The level of structural conservation varies between different RNAs and between different functional (structural) elements of a given RNA class or family. Generally, the level of conservation signifies biofunctional importance. With reference to the previous example of the larger sub-unit of ribosomal RNA, the structure is highly conserved, with maximal persistence in the regions that are most functionally critical – at the surface, where the structure interacts with the small sub-unit; and for substrate binding, factor binding, and catalytic activity.

Notably in regions of biological importance, sub-sequence conservation may also be maximal. It is thus perhaps unsurprising that highly conserved structures that cut across all kingdoms of life,

[14] in this document, G-C and A-U may refer to Guanine-Cytosine and Adenine-Uracil base-pairing, respectively, or equivalently to their converse pairs

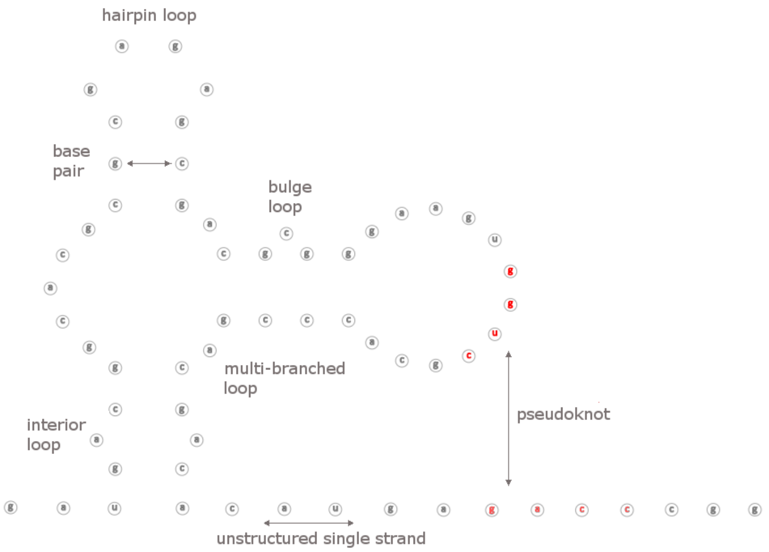


Figure 4 Archetypal structure representation of RNA

such as rRNAs, have been used extensively to deduce phylogenetic relationships across diverse taxa (refer to, for example, Kumar, 1996). Other techniques are available for determining the atomic and molecular structure of target molecules, including *X-ray crystallography* and *Nuclear Magnetic Resonance* (NMR). The previously mentioned techniques are both costly and time-consuming, and in some cases - for example when using NMR - the technique is not suitable for large RNA structures. Alternatively, thermodynamic-based prediction of secondary structure may be employed based on *free energy minimization* computational approaches. Although current knowledge of RNA thermodynamics is not complete, it is estimated that for sequences of up to ~700 nt in length, about 70% of secondary structure can be predicted using thermodynamics alone (Mathews, 2004). However, folding space for RNA is significant in size, where an arbitrary sequence of length n is determined to have 1.8^n possible secondary structures (Zuker et al., 1984), prompting the need for adding constraints and enabling a more tractable approach.

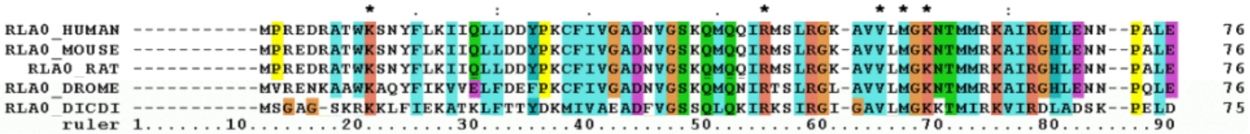
Biochemical experimental data may readily provide an information base for constraining folding space. The most commonly used methods introduce chemical modifications to bases or ribose sugars to differentiate between nucleotides in base pairs and unstructured regions (refer to, for instance, Mathews et al., 2010), with chemical reagents used such as Dimethyl sulfate (DMS), 1,1-Dihydroxy-3-ethoxy-2-butanone (Kethoxal), or the multiple reagents used in Selective 2-hydroxyl acylation analyzed by primer extension (SHAPE), where different probing techniques vary by relative method cost and accuracy. An approximative and generally cost effective, alternative method for inferring RNA structure is based on sequence analysis wherein examples of evolutionarily-related RNAs from different biological entities are compared – on assumption of *conserved sequence/structure* – and for which a brief overview is provided in the following.

Structural comparative analysis. Structural conservation in RNA of related species naturally allows for comparative methods to be recruited. Popular methods adopting this approach include the following:

- **Production of multiple alignment followed by folding**. In this approach, a multiple sequence alignment (MSA, refer to the example in **Figure 5**) is first constructed solely based on the given sequences, and then the lowest free energy structure common to all is predicted. This method scales up well with the number of sequences as the more computationally demanding folding step is done once for the entire alignment, however prediction quality hinges strictly on the quality of MSA
- **Folding of all sequences, followed by alignment**. More than one structure is predicted and the lowest free energy structure common to all sequences is selected. Though more resilient to error than the first approach, this method is computationally demanding and intractable for large datasets
- **Simultaneous folding and alignment**. This method mimics a typical expert-driven process of iteratively folding and aligning a set of sequences. In more recent implementations, additional constraints, such as the exclusion of base-pairs that would inherently lead to a high folding free energy construct, are used to guide the alignment/folding process
- **Covariance Models**. A covariance model (CM) is a model based on stochastic context-free grammars that extends linear modelling techniques (such as Hidden Markov Models) to incorporate base-pairing information (Nawrocki et al., 2009). An alignment may be directly produced by the method, or an existing one employed if available, to produce a secondary structure model based on co-varying nucleotide information at specific positions in the alignment

[15] refer to http://www.clustal.org/ for an overview of the tool. note that in this specific case, the original alignment which was featured in https://en.wikipedia.org/wiki/Multiple_sequence_alignment was produced using ClustalX, the companion tool of ClustalW that allows a user to manipulate/produce alignment via a graphical user interface

[16] https://en.wikipedia.org/wiki/UniProt



```
                               *   .      :       .          .   *         * * *      :
RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
  RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE   76
RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE   76
RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD   75
     ruler 1.......10........20........30........40........50........60........70........80........90
```

Figure 5 Example multiple sequence alignment. Partial representation of a protein MSA produced with the progressive multiple alignment tool ClustalW[15]. Each sequence, displayed individually as a single row, is an instance of the 60S acidic ribosomal protein P0 encoded by the *RPLP0* gene, and retrieved from the freely available UniProt knowledge base (UniProtKB/Swiss-Prot[16]). The sequence data corresponds to homologues for *Homo sapiens*; *Rattus norvegicus*; *Mus musculus*; *Drosophila melanogaster*; *Dictyostelium discoideum.* Only the first 90 residues positions of the alignment are displayed. Note that the colours represent the amino acid conservation based on the characteristics and frequency distribution of residues in each column, as per the colour scheme adopted by ClustalX. Moreover, various symbols are adopted by ClustalW to signal conservation status. For example, in the case of *K* and *R*, the headlining * (asterisk) symbol above each of the two columns indicates a MSA position that sits a single, fully conserved residue

A significant amount of alternative options exist for modeling and allowing subsequent search based on sequence/structure information. Some methods such as the ones listed above are generic in nature, whereas others are designed to address a more restricted class of problems – for example, in (Giegerich et al., 2004), the notion of representative 'abstract RNA shapes' is used to restrict search within a given energy range and by specific clusters of structures. A challenge that is common to most of the above approaches is the limited ability in handling complex topological structures involving tertiary interactions such as pseudoknots. In a number of approaches, PKs are not included in the model, whereas in others PKs are 'broken down' into other model components that are supported. Whenever allowed to be included, pseudoknots generally present a significant computational challenge to the method adopted, often hindering the efficient use of the model by the end-user.

Here, we address the problem of identifying functional structures in the 3' UTRs of flavivirus genomes using *consensus secondary structure descriptors* (termed as such, and abbreviated as *CSSD*). Because of inherent complexity in the structures involved, the available homology modeling approaches are not feasible.

Principal research question: Can a search method strictly based on CSSD allow for the efficient retrieval of complex RNA topologies, inclusive of pseudoknots and other tertiary interactions, in a given dataset of sequences?

2 Materials and Methods

The notation adopted for consensus secondary structure descriptor (CSSD) modeling follows closely that of *Infernal* (Nawrocki et al., 2009), a popular RNA homology search and multiple alignment tool that is based on stochastic context-free grammars (*covariance models*). *Infernal*, in turn, annotates RNA secondary structures in "WUSS format" (the Washington University Secondary Structure notation), a representation based on the common bracket notation for RNA secondary structures, where matching bracket or parentheses symbols denote base pairing partners. Other symbols are used in this approach, generally following the WUSS notation[17] and summarized as follows:

base pairs. <> and **()** symbols denote base pairing. Analogous to their use in *Infernal*, <> (angled brackets) represent base pairs of a terminal stem structure; likewise, **()** (parentheses) are used for helices closing a multi-branched structure, that is, the point of connection between different double-stranded segments. Up to 3 terminal stems are allowed within a single multi-branch structure. In the current setup, Watson-Crick and G-U wobble pairs can be represented by these pairing symbols.

hairpin, bulge, and interior loops. **_** (underscore) denotes a hairpin loop residue, whereas – (dash) is used for positions in a bulge or interior loop. In interior loops, dash symbols account for 5' and 3' strand nucleotides independently.

pseudoknots. PKs are represented by matching **{}** or [] pairs. In the current version of the tool implemented (refer to **Annex A**), up to 2 PK interactions can be modeled, noting that further tertiary interactions can be easily incorporated by extension of the technical setup used.

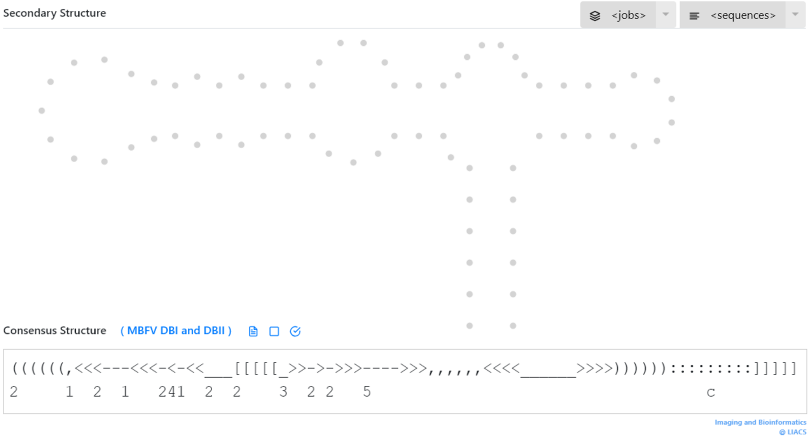
single stranded positions. a **:** (colon) symbol represents unstructured nucleotide positions.

unpaired nucleotides in multi-branch loops. denoted by a **,** (comma) symbol.

positional variability. the ability to annotate for multiplicity of a symbol at any given position is included, whereby, positions in the CSSD containing one of the above symbols may be annotated with a *positional variable*, specifying the effective number of instances allowed for that symbol when matching a consensus against one or more sequences[18].

[17] refer to http://eddylab.org/infernal/Userguide.pdf for an overview of WUSS notation as used in *Infernal* and for comparison with respect to the notation adopted here

[18] by way of example, the positional variable '**2**' placed directly under a **:** (colon), signals that at that position in the CSSD where the colon is found, 0, 1, or 2 unstructured nucleotides will match the consensus, provided that all other symbols and



```
((((((,<<<---<<<-<-<<___[[[[[_>>->->>>---->>>,,,,,,<<<<______>>>>))))))::::::::::]]]]]
2     1  2  1   241  2  2    3  2 2   5                                    c
```

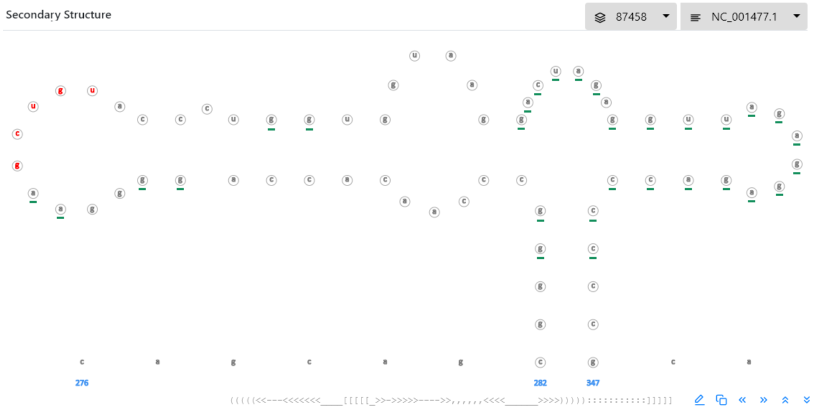
Figure 6 Example CSSD for MBFV 5' and 3' DB. A CSSD for both dumbbells characteristic of region II in MBFV 3' UTRs is shown in the bottom part. The second line of the CSSD allows for positional variables associated with symbols in the first line. The top part outlines the structure, excluding visualization of a) positional variables; that is, the visualization corresponds to an explicit realization of the CSSD' first line, and b) the 3' extremity including the downstream PK sequence

Figures 6 and **7** illustrate typical use of CSSD notation for building and applying a secondary structure model. In the example provided, taken from *TR.2* of **Section 3**, a single CSSD covers duplicated (5' and 3') dumbbell (DB) structures present in most Mosquito-borne Flaviviruses. The DB model used in **Figure 6** includes a base stem (using parentheses notation) that encloses two terminal structures stemming from a multi-branch (-junction), each denoted by matching angled brackets. The right-hand stem-loop sub-structure, represented by the CSSD sub-string '<<<<______>>>>' has a stem of length 4, a loop of length 6, and is not annotated for positional variables. The left-hand sub-structure includes 3 internal loops interspersed within a helical structure of 9 base pairs. This sub-structure's loop is of length 9 and incorporates 5 nucleotide positions involved in a PK interaction (indicated by the opening square brackets). The sub-string representing this left-hand structure,

```
<<<---<<<-<-<<___[[[[[_>>->->>>---->>>
  2  1   241  2  2    3  2 2   5
```

recruits 11 positional variables, where each contiguous string of the same symbol (< or >, −, _, and [ ) is designated either one or no positional variable. The manner in which a consensus structure is 'expanded' to represent multiple, unique, secondary structures is agnostic to where (i.e. under which specific instance of the repeated symbol) a positional variable is placed; the only strict constraints being a) up to one positional variable per sub-structure is allowed – for

corresponding positional variables in the CSSD also match



```
(((((<<---<<<<<<<____[[[[[_>>->>>>>---->>,,,,,,<<<<______>>>>))))):::::::::::]]]]]
```

Figure 7 Matching consensus for a DENV1 strain. This figure visualizes the 3' DB structure matching a DENV1 strain (NC_001477.1), at position 282 relative to the stop codon, in response to a search placed against the CSSD provided in **Figure 6**. It should be noted that amongst the ~25M possible secondary structures expressed by this CSSD, one or more (unique) structures may result in a match for this sequence at any given position. At nt position 282, 4 matches were returned for this sequence, one of which is rendered above. The nucleotides highlighted in red are involved in a PK interaction, wherein this figure only the sequence in the apical loop is shown. Markings in green are additionally provided here, corresponding to highly conserved sequences or base pairs in DENV group members. The matching structure is also represented linearly at the bottom of the figure

example – when annotating a *contiguous* stack of base pairs, and b) in case of paired symbols, such as [] and **()**, only 1 positional variable is allowed on either 5' or 3' strands. PK positions may be located anywhere within the structure. A more comprehensive, technical overview of the CSSD method and its implementation is provided in **Annex A**, however the following constraints may be worth noting here:

- the length of a consensus secondary structure descriptor is hard limited to 1000 symbols, and which by far exceeds the requirement for local structure descriptors, even if the effective limit depends on the tool platform's memory and computational capacity. Within these limits, any number of consecutive secondary sub-structures may be defined, separated by unstructured positions (using the colon symbol **:**), and

- at any given CSSD position, designated positional variables may lie within the range '**1**'-'**p**', where alphabetic characters '**a**' through '**p**' are effectively translated to consecutive variable values 10 to 25.

In comparing sequences and consensus secondary structures, a mechanism is put in place that allows for a scoring function to operate on sequence/CSSD pairs, taking into account the position, and implicitly, the respective sequence and linear secondary structure descriptor 'prefix'[19] at which

[19] the analogy of a prefix here is used to refer to either a) the RNA sub-sequence preceding a mismatched position, or b) the

a mismatch occurs, if any. In principle, the scoring mechanism is arbitrary and can accommodate biologically meaningful constraints such as minimum free energy calculations. However, for the purpose and the objectives set out in this work, a binary score (**1** - for an exact match, or **0** - failure to match) is preferred. A simple example follows, in order to illustrate the archetypal process of CSSD-based search. Searches can be placed bidirectionally, that is, with either a CSSD or a sequence used as the query term, and therefore targeting either a dataset of sequences or consensus secondary structures, respectively.

```
<<<<____>>>>
1
```

(**step 1**) small hairpin structure with 1 positional variable at the first stem base pair, submitted as a query CSSD

```
uaagaaaacccccaacagggugccaaacggcaacacgccu
```

(**step 2**) a sequence fragment submitted as a target sequence

```
uaagaaaacccccaacagggugccaaacggcaacacgccu
            cccaacaggg                      <<<____>>>     at position 13
                      ugccaaacggca          <<<<____>>>>   at position 23
                       gccaaacggc           <<<____>>>     at position 24
```

(**step 3**) search result: the target sequence shown juxtaposed with 3 exact matches found, along with the matching structure

2.1 High-level overview of algorithmic approach

Annex A provides a more detailed account of the technical requirements adopted when implementing the prototypical system for CSSD-based search and the computational performance achieved. A significant part in providing an effective search tool deals with the efficient design and implementation of a multi-component, distributed system, wherein, providing a detailed and complete account of the system is outside the scope of this document. In this subsection, however, a high-level overview of the core algorithmic approach is provided, in order to highlight the general search approach adopted. **Listing 1** below represents such an approach, where, based

sub-structures, scanned in the 5' → 3' direction, that exactly match a given sequence up to a mismatch position. For the latter case, this includes both single- and double-stranded regions that have precisely and entirely matched the sequence under consideration on the architecture described in **Annex A**, a specific (cluster) computation node is assumed to have been assigned a subset of sequences S_n, along with a designated CSSD subset C_n.

```
$\varepsilon(C_n)$ ← expand_cssd ($C_n$)                                   ▷ expand $C_n$ into a set of unique,
                                                                           ▷ ordered, fixed-position descriptors.
for each $s \in S_n$ do
    $pos_s$ ← fivep_nt ($s$)                                               ▷ set position to 5' nucleotide of $s$,
    while ($pos_s \leq$ threep_nt ($s$)) do                                ▷ and sweep in 5' → 3' direction.
        clear ($pos_{bt}$)
        push ($pos_{bt}$, $-1$)                                            ▷ set last backtracking position.
        $skip\_position$ ← false
        for each $c \in \varepsilon(C_n)$ do                               ▷ match descriptors in order.
            $score_c$ ← 0                                                  ▷ initialize each descriptor's score.
            if not sat_constraints ($c$, $s$, $pos_s$) then
                continue
            else
                $pos_c$ ← 0                                                ▷ initialize descriptor's position.
                while ($pos_c < |c|$ and not $skip\_position$) do
                    if is_unpaired ($c$, $pos_c$) then                     ▷ no pairing required.
                        $score_c$ ← update_c_score ($score_c$, $c$, $pos_c$, $s$, $pos_s$)
                        $pos_c$ ← $pos_c + 1$
                        push ($pos_{bt}$, $pos_c$)                         ▷ update backtracking.
                    else                                                   ▷ check paired structure.
                        if is_compatible ($c$, $pos_c$, $s$, $pos_s$) then
                            $score_c$ ← update_c_score ($score_c$, $c$, $pos_c$)
                            $pos_c$ ← update_c_position ($c$, $pos_c$)
                            push ($pos_{bt}$, $pos_c$)                     ▷ update backtracking.
                        else
                            $pos_c$ ← pop ($pos_{bt}$)
                            if ($pos_c \geq 0$) then
                                $c$ ← skip_structure ($c$, $\varepsilon(C_n)$, $pos_c$, $s$, $pos_s$)
                            else
                                $skip\_position$ ← true                    ▷ no further backtracking.
                            end if
                        end if
                    end if
                end while
                if $skip\_position$ then
                    break
                end if
            end if
        end for
        $pos_s$ ← $pos_s + 1$
    end while
end for
```

Listing 1 core search algorithm

In the following, additional comments are provided in relation to **Listing 1** above:

- in *line 1*, the assigned subset of CSSD, C_n, is expanded in an outwards-inwards fashion such that failure to match at an 'outer level' enables fast trimming of the search space through backtracking. This is also achieved by ensuring that helices are expanded starting from lowest (positionally variably-defined) number of base pairs to the highest

- each sequence position pos_s is scanned only once for any size of $|C_n|$ (*line 5*)

- at each sequence position, a stack of backtracking positions pos_{bt} is maintained in order to fallback to the 'last known good structure' matched, in cases where a nested structure fails to match (*lines 19,24*)

- prior to matching a descriptor c against sequence s at a given position pos_s, any general constraints defined for the expanded descriptor are checked (*line 11*). An example of such a constraint is a base-triple interaction, which, although unused in the reported test runs, has been implemented experimentally in the current prototype. Backtracking on such 'general constraints' also allows for fast failures to the next available descriptor

- for both unpaired descriptor symbols (for example, a **:** single stranded position), as well as paired symbols (for example, when specifying a terminal stem), the current score $score_c$ is updated using both descriptor c and sequence s information (*lines 17,22*). This allows for arbitrary scoring functions (such as free energy calculations) to be included in the scoring scheme

- if a paired-structure-match fails (*lines 26-31*), the algorithm backtracks to the last known good structure/position and identifies the next available substructure (*line 28*); otherwise the current sequence position is skipped (*line 30*). By way of example, in the following snippet of CSSD, upon failure to match at the multi-branch structure of the first descriptor, **skip_structure** at *line 28* allows for search to proceed (skip) to the fourth descriptor:

```
⋮
:::((<<<___>>>,<<<___>>>))
:::(((<<<___>>>,<<<___>>>)))
:::((((<<<___>>>,<<<___>>>))))
::::((<<<___>>>,<<<___>>>))
⋮
```

3 Results and Discussion

A series of test runs (**TR.1** to **TR.6**) were conducted with the goal to assess preliminary performance of the concept introduced in the previous section. Specifically, structural models for functional elements in the flaviviral 3' UTR were progressively established, in order to provide a first indication of model sensitivity and specificity. All viral 3' UTR sequences used were retrieved from NCBI's genome repository[20], and the prototypical tool described in **Annex A** was employed to search for the modeled structures in targeted 3' regions. Additional information about individual test run results and the verification of the results is made available in **Annex B.4**.

TR.1 Base-paired model for the 5' DB specific to WNV

The dumbbell (DB) structure is characteristic of MBFV viruses, and is manifest in domain II of the 3' UTR as a conserved and typically duplicated structure, with the involvement of pseudoknot interactions. For the first trial, an *initial* consensus structure model of the 5' DB specific to members of the **West Nile virus** [**Japanese encephalitis virus group**] was constructed on the basis of models for conserved RNA elements published in (Villordo et al., 2016).

To achieve this goal, out of a total selection of 74 sequences, the 5' DB of 3 WNV strains were used as 'template structures' on which to 'train' the CSSD, whereas the remaining 71 strains were used as test sequences. Of the 71 test sequences, 41 strains were chosen for their evolutionary origin in the MBFV, TBFV, NKV, and ISV groups – listed below. The other 30 sequences were composed of equally-partitioned, random shuffles of the 3 template WNV sequences. A consensus model of the 5' DB specific to the 3 WNV sequences was initially constructed *excluding pseudoknots*, in part guided by structural information obtained from Villordo et al. (refer to *Figure 5II* in Villordo et al.). It should be noted that under this scenario, adopting the predicted structures from Villordo et al. obviated the need to make direct use of a MSA to elucidate secondary structure.

In the following is a definition of the resulting CSSD, followed by a list of the real sequences used in this as well as the following trials. Search results for this run are provided in **Table 1** of which a summary plot is shown in **Figure 8**.

```
(((((,<<<-<<<<-<<<___________>>>->>>>--->>>,,,,,,<<<<______>>>>)))))
              1        2        1
```

CSSD A. BP model for the 5' DB based on 3 selected WNV strains

[20] https://www.ncbi.nlm.nih.gov/genome/

The following 44 flavivirus sequences (along with the respective accession numbers) were used across all test runs 1-6:

[**JEV** group]
WNV Kunjin (L24512.1), WNV 385-99 (EF571854.1), WNV PT6.16 (AJ965626.2), WNV (M12294.2), Alfuy (AY898809.1), Murray Valley encephalitis (NC_000943.1), St. Louis encephalitis (NC_007580.2), Japanese encephalitis (GQ304752.1), Usutu (NC_006551.1)

[**DENV** group]
DENV1-4 (NC_001477.1, NC_001474.2, NC_001475.2, NC_002640.1), Kedougou (NC_012533.1)

[**YFV** group]
Sepik (NC_008719.1), Wesselsbron (NC_012735.1), Yellow Fever (NC_002031.1)

[**Ntaya** group]
Rocio (MF461639.1), Ilheus (AY632539.4), Tembusu (JF895923.2), Bagaza (AY632545.2)

[**Kokobera** group]
Kokobera (AY632541.4)

[**Aroa** group]
Bussuquara (NC_009026.2), Iguape (AY632538.4)

[**Spondweni** group]
Zika (NC_012532.1)

[**TBEV** group]
TBEV Neudoerfl (U27495.1), TBEV Oshima (AB753012.1), TBEV Vasilchenko (L40361.3), TBEV IR99 (AB049398.1), TBEV Hypr_IC (KP716974.1), Powassan (NC_003687.1), Louping ill (NC_001809.1), Langat (NC_003690.1), Alkhurma hemorrhagic fever (NC_004355.1), Karshi (NC_006947.1), TBEV Sofjin (JX498940.1), Omsk hemorrhagic fever (AY193805.1)

[**Entebbe** group]
Entebbe bat (NC_008718.1), Yokose Oita (AB114858.1)

[**Modoc** group]
Apoi (AF452050.1), Modoc (NC_003635.1)

[**Rio Bravo** group]
Montana myotis leukoencephalitis (AJ299445.1), Rio Bravo (JQ582840.1)

[**ISV** group]
Cell fusing agent virus (NC_001564.2)

For the first test run, 3 West Nile virus sequences (Kunjin, 385-99, PT6.16) were used as template sequences, with the remaining sequences used as test examples.

Strain	Accession	3’ UTR Length	Matching positions	PK (Apical loop)	Matching Consensus
Kunjin	L24512.1	627	358	ugguguu	(((((,<<<-<<<<<<<____________>>>>>>>--->>>,,,,,,<<<<______>>>>)))))
385-99	EF571854.1	634	365	ugguguu	(((((,<<<-<<<<-<<<__________>>>->>>>--->>>,,,,,,<<<<______>>>>)))))
PT6.16	AJ965626.2	609	365	ugguguu	(((((,<<<-<<<<<<<____________>>>>>>>--->>>,,,,,,<<<<______>>>>)))))

Table 1 Results for base-paired model of specific WNV strains (as matched against ***CSSD A***). Matching nucleotide positions for exact matches found between the model and a given sequence are shown, indexed from nucleotide position 1. In this table, only sequences that yielded a match are included, for brevity, given that none of the other 71 test sequences matched

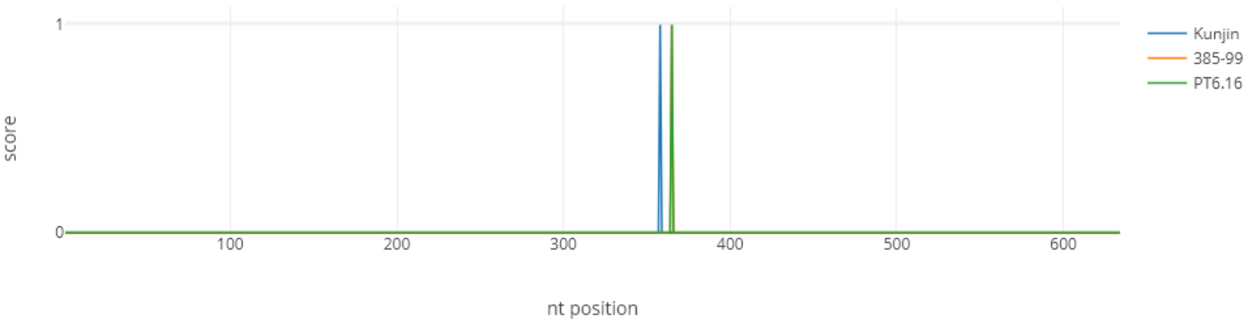


Figure 8 TR.1 - Results Overview. For the 3 WNV strains used to build the initial model, 5’ nucleotide positions for all exact mathces (that is, a search score equal to 1) against *Consensus Structure A* are given. X-axis values yielding a peak score thus correspond to the matching positions listed in column 4 of **Table 1**

In **Table 1**, nt positions for matches corresponding to 5’ DB are shown in column 4. The position provided is for the 5’ nucleotide of the respective subsequences matching the consensus. (By way of example, for the WNV$_{KUN}$ strain with accession L24512.1 as shown in **Table 1**, the matched consensus "(((((,<<<-<<<<<<<____________>>>>>>>--->>>,,,,,,<<<<______>>>>)))))" of length 66 nt, corresponds to the subsequence starting (ending) at position 358 (424), relative to the sequence’s stop codon.) No test sequence yielded a match and are not shown for brevity. It should be noted that the 5’ DB for test sequence WNV (M12294.2) differs from the local structural conformation adopted by the 3 templates. Relative to the templates, M12294.2 extends the internal loop below the apical loop by 1 nt on either strand, and in addition, has a shorter apical loop (by 2nt). Adjusting the CSSD accordingly would yield 4 exact matches including the 3 template sequences and M12294.2. Moreover, although PK interactions were not incorporated into the model, the sequences involved in the PK (at the apical loop) can be observed in the matching sequence – refer to column 5 in **Table 1**. The specific secondary structure descriptor matching a given sequence was extracted from the CSSD, and provided in the last column for reference. Verification of matching positions (**Figure 8**) was done manually using the 3 sequences, as well as confirmation of the highly conserved sequences shown in *Figure 5II* in Villordo et al. Folded structures for the 3 WNV sequences, as matched by this model, are shown in **Annex B**.

TR.2 Model for 5’ & 3’ DB of specific DENV and JEV group members

In a second trial, *CSSD A* was progressively modified to *simultaneously* accommodate 5’ and 3’ DBs of specific DENV and JEV group members, *inclusive of one PK interaction*. Thus, highly conserved sequences and PK base pairs derived from Villordo et al. were utilized to guide a more general *CSSD B*. In addition, *mfold*-predicted[21] structures for DENV1-4, Japanese encephalitis, and Usutu, where used to explicate the minimum modifications, i.e. the structural variability necessary for *CSSD B* to incorporate DB structures from the two groups. The iterative process adopted when using *mfold* is briefly outlined in **Annex B**, with the finalized consensus shown below.

```
((((((,<<<---<<<-<-<<___[[[[[_>>->->>>---->>>,,,,,,<<<<______>>>>)))))):::::::::]]]]]
1     1  2  1   241  1  1    3  2 2   1                                    c
```

CSSD B. Model for DENV, JEV groups 5’ and 3’ DBs

Strain	Accession	3’ UTR Length	Matching positions (5’ DB)	PK sequences	PK distance	Matching positions (3’ DB)	PK sequences	PK distance
Kunjin	L24512.1	627	357,358	ugguguu; aacacca	79	434,435	gcugu; acagc	19
385-99	EF571854.1	634	365	ugguguu; aacacca	80	441,442	gcugu; acagc	19
PT6.16	AJ965626.2	609	365	ugguguu; aacacca	80	441,442	gcugu; acagc	11
DENV1	NC_001477.1	465	197,198	gcugu; gcagc	10	281,282	gcugu; acagc	10
DENV2	NC_001474.2	454	184,185	gcugu; gcagc	9	271,272	gcugu; acagc	9
DENV3	NC_001475.2	443	176,177	gcugu; gcagc	10	260,261	gcugu; acagc	8
DENV4	NC_002640.1	387	114,115	gcugu; gcagc	14	205	gcugu; acagc	9
Japanese encephalitis	GQ304752.1	585	313,314	ugca; ugcg	11	394,395	gcugu; acagc	9
Usutu	NC_006551.1	668	389,390	gaug; cguu	20	468,469	gcugu; acagc	17

Table 2 Results for 5’, 3’ DB model for DENV and JEV groups (*CSSD B*). Matching nucleotide positions for matches found between *CSSD B* and target sequences are listed in columns 4, 7. *PK distance* refers to the number of nucleotides from, but not including, the 3’ nucleotide at the base of a DB and the 5’ nucleotide of the downstream sequence involved in a PK

Besides introducing an essential pseudoknot interaction, *CSSD B* underlines a significant departure in the interpretation of what it represents, as compared to *CSSD A* from *TR.1*. The latter denotes a collection of 12 structures that are very closely related: any two structures ‘recognized’ by *CSSD A* may differ by up to 1 nucleotide at either strand of an internal loop, and in the limit, by 2 nucleotides at the apical loop. *CSSD B*, on the other hand, accesses a shape

[21] http://unafold.rna.albany.edu/?q=mfold

space of ∼2,5M structures[22], with a correspondingly dynamic thermodynamic range. Hence, the motivation for selecting the specific sequences and CSSD for this trial straddles multiple lines of inquiry: a) how do the exact match results for WNV, in *TR.1*, change under significant structural *generalization* of the initial consensus; b) can a manually-derived CSSD, based on available predicted structures, yield precise and accurate search results across related groups of MBFV viruses; and also c) can a single DB CSSD reliably detect more than one DB structure.

Table 2 incorporates results[23] for DENV1-4, Japanese encephalitis, and Usutu – i.e. the sequences used to augment *CSSD B* (listed in the bottom part of the table) – along with the 3 WNV sequences used for constructing *CSSD A*. As summarized in **Figure 9** below, the previous results obtained in *TR.1* for WNV strains are unchanged in **Table 2**, except for a double hit (i.e. two neighbouring matches) at 5' DB WNV$_{KUN}$, which is within expectation given the variability now afforded at the base stem of the consensus structure[24]. Validation of the results was done in a similar fashion to the first trial, but including verification of PK interactions as per Villordo et al.

When applied to the remaining 35 sequences, *CSSD B* manifests resistance to false positives. No unrelated sequence yielded a match, except for Cell fusing agent which returned a single DB. Although other members of the ISV group are known to be MBFV-related and also render a single DB, supporting evidence in literature could not be identified for validation of this specific match. Moreover, the 3' DB of Ilheus, Zika, WNV (M12294.2), Tembusu, Rocio, and St. Louis encephalitis, as well as the single DB for Modoc were returned as a match, and validated[25].

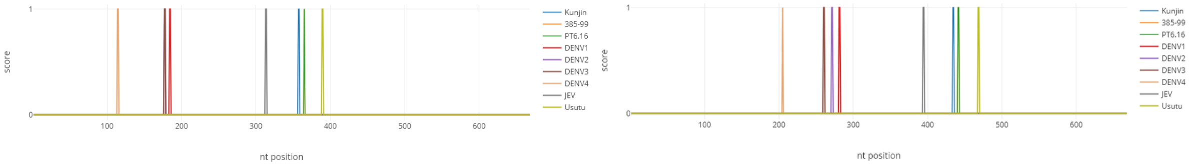


Figure 9 **TR.2 - Results Overview**. DENV/JEV group summary of results for 5' (left) and 3' (right) DBs

[22] whereas this figure does not take into account important considerations such as minimum free energies, and therefore, viability of structures, the cardinality of the respective shape spaces (Schuster, 1995) is used here for lightweight comparison

[23] **additional notes on results in Table 2**: **a**) for multiple matching positions, only the first position's PK sequences and distances are shown in columns 5,6,8, and 9; **b**) the full length PK sequences are shown for the WNV strains, whereas *CSSD B* can only retrieve PK structures with shorter sequences. For brevity, the longer model used to match the 3 WNV strains is not shown; **c**) also with respect to the 3 WNV strains, the longer stretch of unstructured nucleotides required to match PK distances of 79 and 80 are not shown in the model. It should be noted however (also with reference to future work proposed in **Section 4**) that this not impact the sensitivity and specificity rates reported; and **d**) specifically for Japanese encephalitis and Usutu, the PK sequences identified could not be verified in literature, however, all other conserved sequences expected were correctly identified within the structure, providing high confidence in the two being 'true positive' matches

[24] the '**1**' variable modifier at the first position of *CSSD B* is likely to yield two "nested" matches at adjacent positions in most sequences. Although algorithmically simple to correct for, as a matter of general approach adopted in this work, it was elected to return any and all matches solely using base-pairing, and no other rule, constraint, or correction

[25] the nt positions identified in these matches are identical to those reported for *TR.3*

TR.3 Model for 5’ & 3’ DB structures across MBFV group viruses

Ten sequences were progressively added to the nine previously used for *TR.1* and *TR.2*, and a new CSSD built to cover both DBs of members of Yellow Fever (Yellow Fever), Ntaya (Tembusu, Bagaza, Ilheus, Rocio), JEV (St. Louis encephalitis, Alfuy, Murray Valley encephalitis), Kokobera (Kokobera), and Spondweni (Zika) groups. The selection of viruses was governed by the need to maximize CSSD coverage of MBFV viruses[26], along with having access to viral reference structures for validation purposes, whilst keeping the shape space unfolded by the new CSSD relatively similar to that of *TR.2*. The new consensus structure is shown below, followed by **Figure 10** and **Table 3** that summarize search results for the new template sequences used:

```
((((((,<<<---<<<-<-<<____[[[__>>->->>>---->>>,,,,,,<<<<______>>>>))))))::::::::]]]
2     1  2  2   2411 2       3  2 2   5                                    c
```

CSSD C. Model for MBFV group 5’ and 3’ DBs

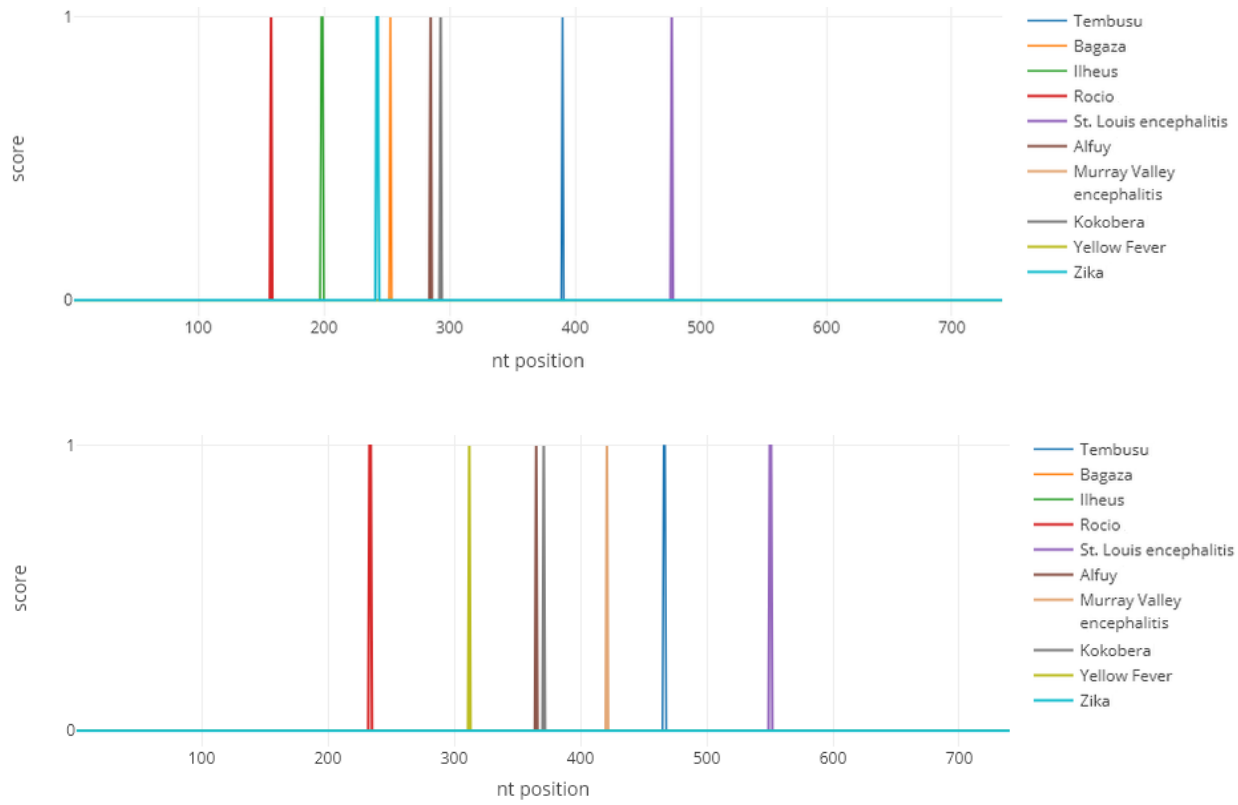


Figure 10 **TR.3 - Results Overview**. MBFV group summary of search results for 5’ (top) and 3’ (bottom) DBs

[26] the selection includes Tembusu, Ilheus, Rocio, St. Louis encephalitis and Zika for which only a single DB was retrievable in *TR.2*

Strain	Accession	3' UTR Length	Matching positions (5' DB)	PK distance	Matching positions (3' DB)	PK distance	Note
Tembusu	JF895923.2	660	390,391	19	466,467	19	
Bagaza	AY632545.2	731	N/A	N/A	N/A	N/A	
Ilheus	AY632539.4	391	N/A	N/A	197,198,199	18	missed 5' DB
Rocio	MF461639.1	427	158	N/A	232,233,234	20	no PK for 5' DB found[27]
St. Louis encephalitis	NC_007580.2	740	478	21	549,550,551	15	
Alfuy	AY898809.1	565	284	20	363,364,365	23	
Murray Valley encephalitis	NC_000943.1	617	N/A	N/A	420,421	16	
Kokobera	AY632541.4	561	292	20	371	20	
Yellow Fever	NC_002031.1	511	N/A	N/A	312	12	FP at 209 (8nt)
Zika	NC_012532.1	431	N/A	N/A	240,241,242	9	

Table 3 Results for model incorporating 5' and 3' DB structures across the MBFV group

The 9 template sequences used in the previous test run returned similar results to those shown in **Table 2** for *TR.2*, however, in this run, Usutu also yielded 3 false positives at positions 5, 14, and 56. For the 10 new templates (**Table 3**), Ilheus and Rocio did not yield a 5' DB structure as was to be expected, whereas Yellow Fever captured a false positive at position 209, suggesting a need to refine the CSSD to better address conformational requirements. The relatively high number of false positives for Usutu may suggest that nucleotide content may influence the behaviour of the method applied here – where, relative to the other sequences, the 3' UTR of Usutu contains the highest (lowest) percentage content of U (C) nucleotides in its 5'-proximal AU-rich region, and thus amenable to fit a variety of secondary structure models. *CSSD C* was applied to the remaining 25 sequences that were not involved in building the model, with results summarized as follows:

- **valid, exact matches** (7 sequences, 9 true positives)
  DBs for MBFVs WNV (289,290,366,367), Iguape (293, 367), Bussuquara (226,227), and Sepik (266) were matched and verified. Moreover, members of the NKV group Yokose (212-214), Modoc (128), Rio Bravo (210) returned a match for their single DB structure

[27] the correct structure at position 158 corresponding to the 5' DB of Rocio can be identified if a search using this model is performed only after removing the PK constraint

- **invalid matches** (1 sequence, 1 false positive)
  a single DB structure was identified as incorrect for the ISV group's Cell fusing agent (110)
- **missed matches** (3 sequences, 3 false negatives)
  Kedougou (DENV group/MBFV) and Wesselsbron (Yellow Fever group/MBFV) did not yield a match as expected for a single DB structure, whereas Bussuquara (Aroa group/MBFV) only revealed one of two structures
- **no match** (21 sequences, 33 true negatives)
  3 NKV group (Entebbe bat, Apoi, Montana myotis leukoencephalitis), all 12 TBFV group members, 1 DENV group (Kedougou), and 1 YFV group (Wesselsbron) sequences did not return any matches. Moreover, 1 ISV group (Cell fusing agent), 1 RBV group (Rio Bravo), 1 Entebbe group (Yokose), and 1 YFV group (Sepik) viruses did not yield a second match

For completeness, it should be noted that the 3' DB of Ilheus, Zika, WNV (M12294.2), Tembusu, Rocio, St. Louis encephalitis, and the single DB returned for Modoc, all of which were matched by *CSSD B* in *TR.2*, were confirmed as yielding the same matching positions in this run.

TR.4 CSSD for SL5' and SL3' structures in MBFV viruses

For this trial, two CSSD for SL-I/II and SL-II/IV 3' UTR stem-loop structures of region I in MBFVs were constructed, following structure[28]/sequence data provided in *Figures 2,4,5* in Villordo et al. The models follow below, in addition to template search results in **Table 4** and **Figure 11**:

```
<<<<<-----<<<-<<<<-<_[[[[__>>>>>>>>>>>>>::]]]]
        5         123  1  4              5
```

CSSD D1. SL5' model for DENV1-4, Kokobera

```
(((((<<<_____>>>,<<<<<__[[[[____>>>>>,,))))):::]]]]
       3    3        2 2   4   2              j
```

CSSD D2. SL3' model for remaining MBFV group members

It may be reasonable to suggest that, based on previous trials, the two CSSD developed might be unified into a single consensus with further modeling iterations, however, for the purpose of assessing general search performance the current setup may suffice – while noting that, a unified (hence, more general) consensus model might increase false positive rates to a limited degree. The same, multiple model approach, was also adopted for the next test scenario.

[28] reference is also made to (Göertz et al., 2016)

Strain	Accession	3' UTR Length	Matching positions (SL5')	PK distance	Matching positions (SL3')	PK distance	Note
Kunjin	L24512.1	627	106	2	266	6	FP at 8 (4nt), 26 (2nt)
385-99	EF571854.1	634	113	2	273	6	
PT6.16	AJ965626.2	609	113	2	273	6	
DENV1	NC_001477.1	465	57	6	130	2	
DENV2	NC_001474.2	454	37	6	110	2	
DENV3	NC_001475.2	443	33	6	109	2	
DENV4	NC_002640.1	387	37	3	N/A	N/A	
Japanese encephalitis	GQ304752.1	585	70	3	230	3	
Usutu	NC_006551.1	668	150	2	309	3	FP at 10 (6nt), 39 (6nt), 104 (7nt)
Tembusu	JF895923.2	660	141	3	299	11	
Bagaza	AY632545.2	731	286	2	448	10	FP at 206 (16nt)
Ilheus	AY632539.4	391	34	3	N/A	N/A	
Rocio	MF461639.1	427	70	3	N/A	N/A	
St. Louis encephalitis	NC_007580.2	740	233	2	N/A	N/A	SL-IV not found
Alfuy	AY898809.1	565	44	2	204	3	
Murray Valley encephalitis	NC_000943.1	617	100	2	262	3	
Kokobera	AY632541.4	561	41	4	203	5	
Yellow Fever	NC_002031.1	511	N/A	N/A	N/A	N/A	FP at 449 (5nt)
Zika	NC_012532.1	431	23	6	107	18	

Table 4 Results for SL5' and SL3' search across the MBFV group

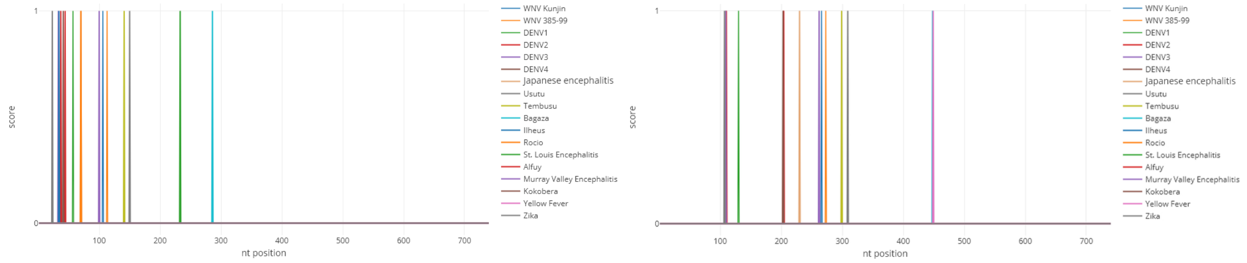


Figure 11 **TR.4 - Results Overview**. MBFV group summary of search results for SL5' (left) and SL3' (right)

The same 19 sequences used as templates in the previous run were used to build and verify correct usage of a consensus model under this scenario. A search based on *CSSD D1/2* was run against the remaining (25) sequences unused when building the CSSD, with results summarized as follows:

- **valid, exact matches** (6 sequences, 8 true positives)
  SL structures for MBFV viruses WNV (M12294.2, 38,198), Iguape (37,205), Bussuquara (36), Sepik (152), Kedougou (135), Wesselsbron (167) were exactly matched and verified
- **invalid matches** (4 sequences, 5 false positives)
  SL structures were incorrectly identified for TBEV Neudorf (667), TBEV Hypr_IC (629), NKV group's Yokose (126,235), and Modoc (142)
- **missed matches** (1 sequence, 1 false negative)
  Yokose (Entebbe group, NKV/MBFV-related) did not yield a match as expected for a single DB structure
- **no matches**[29] (22 sequences, 32 true negatives)
  4 NKV (Entebbe bat, Apoi, Montana myotis leukoencephalitis, and Rio Bravo), the ISV group's Cell fusing agent, and 10 of 12 TBFV viruses did not return any matches. Moreover, 1 DENV group (Kedougou), 2 YFV group (Sepik, Wesselsbron), 1 Aroa group (Bussuquara), 1 Modoc group (Modoc), and 2 TBEV group (Neudoerfl, Hypr_IC) viruses did not yield a second match

The outcome of test runs 1-4 turns out a triplet of consensus structures (*CSSD C*, *CSSD D1*, and *CSSD D2*) that recognize structures compatible with 5' and 3' DB and SL elements, respectively, in MBFV viruses. **Figure 12** below summarizes the exact mathces found for these structures in MBFV sequences. For clarity, only exact matches are shown.

[29] true negatives are considered as such *relative to the model adopted* and based on the cited literature

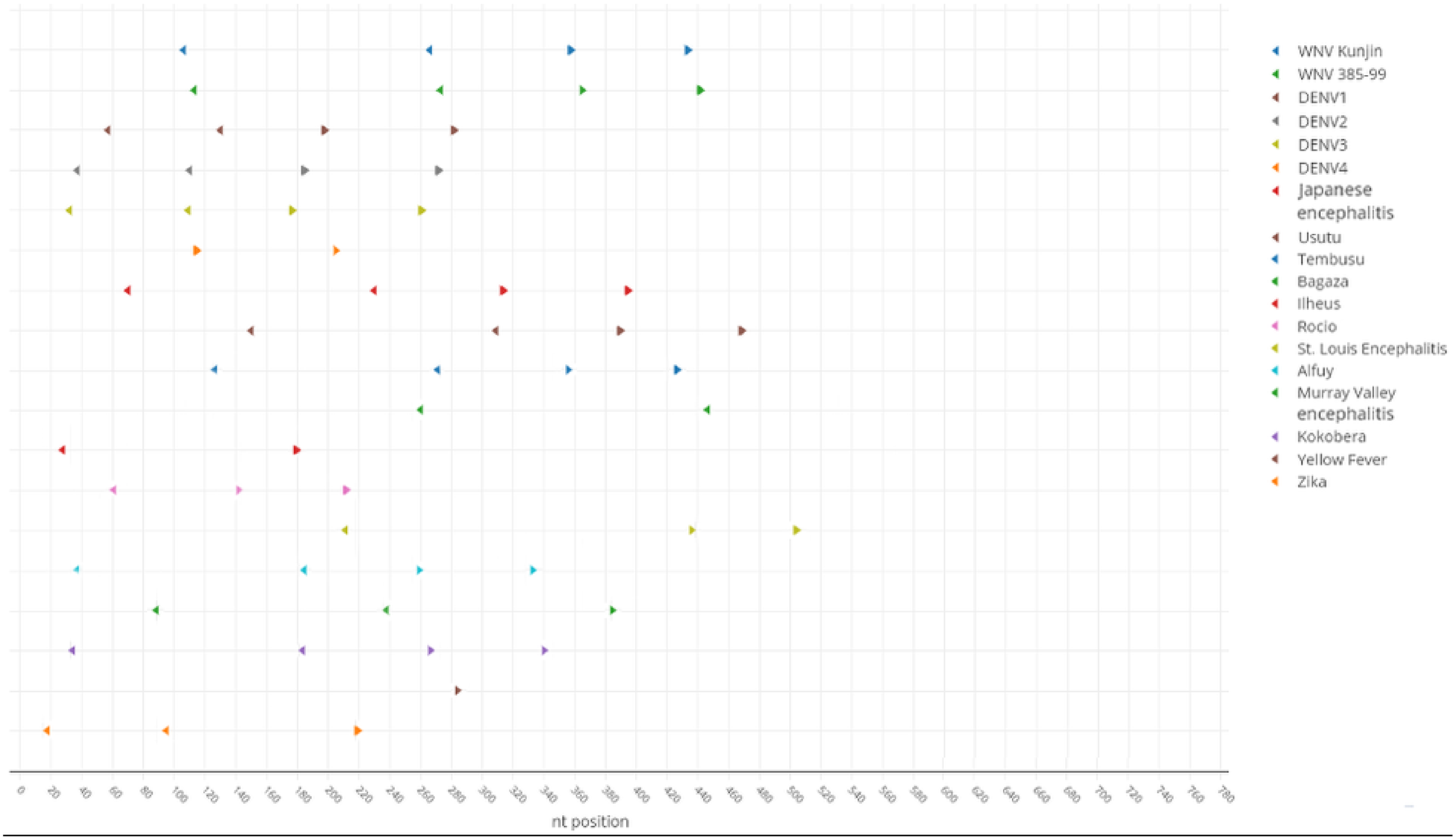


Figure 12 Graphical overview of exact matches for MBFV group. ◄ and ► symbols denote SL and DB structures, respectively

TR.5 Models for viruses pertaining to the TBFV group

Two typically duplicated structures characterizing TBFV virus 3' UTRs – a Y-shaped and a 2nd stem loop – require alternative modeling. In addition, a third stem loop structure (AU-SL) is included, thus modeling all conserved structures in the 3' UTR except for the 3' proximal sHP-3'SL. For this trial, four separate models were constructed from the aligned sequences and predicted structures in (Gritsun et al., 2006); (Gritsun et al., 1997), respectively, and using the work by Villardo et al. as additional reference. Provided below are the models, followed by search results in **Table 5**, summarized in **Figure 13** below:

```
(-(((,,,,,<<<_____>>>,<<<<_____[[[[___>>>>,)))-):::::::::::::::::::::::]]]]
 2     2       7            1                 1           4
```

CSSD E1. TBFV Y-SL

```
<<--<-<<-<____[[[[____>->>->-->>:::::::::]]]]
```

CSSD E2. TBFV 5' GC-SL

```
<<<-<<-<<<<<____[[[[___>>>>>->>----->>>::]]]]]:<<<<<_______>>>>>
   1  2       1             1    2      2    2        1
```

CSSD E3. TBFV 3' GC-SL including hairpin loop

```
((((((<<<<<<____[[[[____>>>>>>,<<<____>>>,)))))):::::::::::::]]]]
            3       2                                  6
```

CSSD E4. TBFV AU-SL

The 11 sequences listed in **Table 5** were used to construct the 4 consensus structures based on the multiple alignment (of the same sequences) in Gritsun and Gould and also using the predicted structures in Gritsun et al. A single TBFV sequence (Karshi) was used as a test sequence which returned a correct hit for both Y-SL structures at positions 21, 156 (and with no other matches yielded for the remaining 3' UTR models). The remaining 32 sequences from the other flavivaral groups did not return any matches, except for a single false positive for Alfuy at position 41.

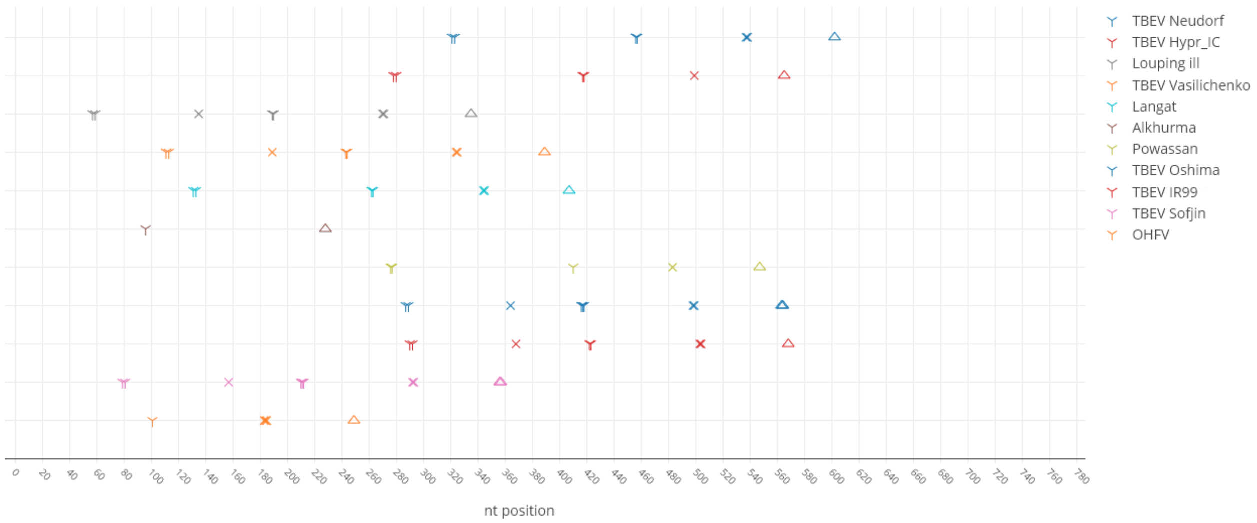


Figure 13 Overview of matches for TBFV. Y, X, and △ denote Y-SL, GC-SL, and AU-SL structures, respectively

Strain	Accession	3' UTR Length	Matching positions (5' Y-SL)	Matching positions (5' GC-SL)	Matching positions (3' Y-SL)	Matching positions (3' GC-SL)	Matching positions (AU-SL)
TBEV Neudorf	U27495.1	767	321,323	N/A	456,457	537,538	603
TBEV Hypr_IC	KP716974.1	729	278,280	N/A	417,418	499	565
Louping ill	NC_001809.1	500	57,59	135	189,190	270,271	335
TBEV Vasilchenko	L40361.3	533	111,113	189	243,244	324,325	389
Langat	NC_003690.1	571	131,133	N/A	262,263	344,345	407
Alkhurma	NC_004355.1	323	96	N/A	N/A	N/A	228
Powassan	NC_003687.1	712	276,277	N/A	410	483	547
TBEV Oshima	AB753012.1	727	287,289	364	416,417,418	498,499	563,564
TBEV IR99	AB049398.1	733	291,293	369	423,424	504,505	569
TBEV Sofjin	JX498940.1	521	80,82	157	210,211,212	292,293	357,358
Omsk hemorrhagic fever	AY193805.1	413	102	N/A	N/A	183,184,185	249

Table 5 Results for 3' UTR structures in TBFV group

TR.6 Generalized Y-shaped structure for all Flavivirus 3' UTRs

A generalized secondary structure model for functional elements in flaviviral 3' UTR was implemented, in consultation with Dr. René Olsthoorn at the Leiden Institute of Chemistry, and which is reproduced below with kind permission. This model represents the second, key departure from the previously used models and trial runs, in the following ways:

a) whereas the consensus structures used in test runs 1-5 are based on reliable sources (including multiple sequence alignments and structure predictions), the resulting CSSD were built using a (manual) process which is somewhat open to interpretation and error – and this is especially the case for the more elaborate structures, such as in *TR.4* (SL5' and SL3' for MBFV), where, the modality adopted was a trade-off between model generality (i.e. variation) and coverage. In contrast, the CSSD implemented for this trial is directly based on a high-confidence model derived from expert-curated structural models, thus inherently reducing modeling uncertainty; and

b) the CSSD for this test run represents the largest search space of all RNA secondary structures explored, which is in excess of an order of magnitude in size increment, from ~27M structures in *TR.3*, to ~1B structures in this run

The implemented model was applied to the set of all (real) sequences used in previous runs and cutting across all flaviviral groups, with results clearly showing exact matches at multiple locations, for each test sequence; the results are summarized in **Figure 14** below:

```
(((,,[[[<<____>>,<<<<_{{{_>>>>,))):]]]::}}}
  3 2 3  7  6   1  5 4 7 5    1   m    2
```

CSSD F. Generalized Y-shaped structure for all *Flavivirus* 3' UTR elements

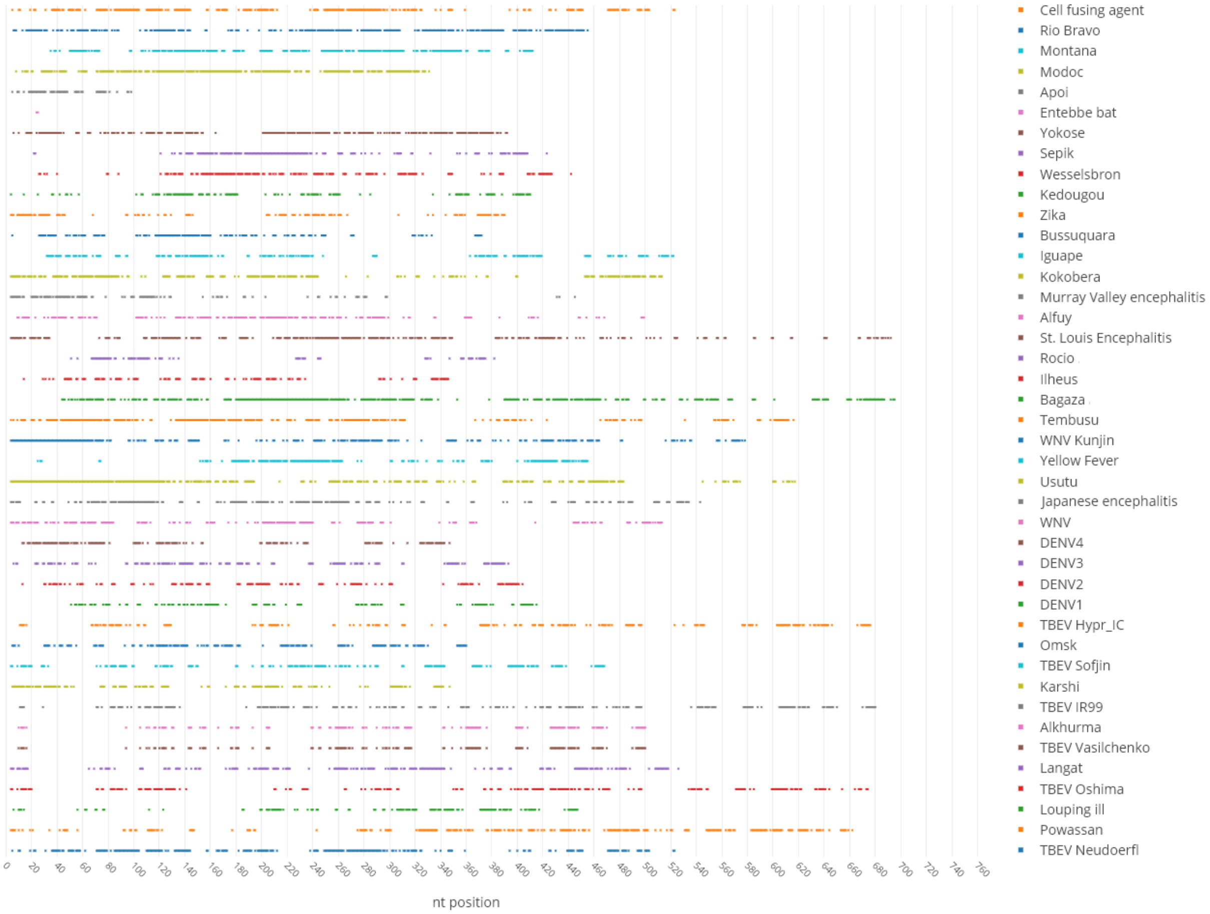


Figure 14 Overview of generalized Y-shaped structure search results

Summary of results and discussion

Results for test runs 1 through 3 support the notion that a topological approach to search in flaviral RNA 3' UTRs gives rise to a good modeling and search response cycle. More specifically, as from the second trial, by inclusion of DENV and JEV group members into the initial consensus based on WNV, a model-generalization opportunity arises, which yields partial (single DB) matches for closely-related strains in JEV (WNV, M12294.2 and St. Louis encephalitis), as well as in related groups Ntaya (Ilheus, Tembusu, Rocio), Spondweni (Zika), and in the MBFV-related NKV (Modoc). It might also be relevant noting the following:

- given that the 5' DB of JEV group members exhibit longer PK contact distances relative to DENV group members, whereby the downstream PK sequence of the former overlaps with the 3' proximal DB structure, significant consensus structure variability is required to include both DB (PK) conformations. This variability, however, does not seem to have a serious impact on CSSD reliability;

- with regards to model specificity and MBFV groups, inclusion of pseudoknot interactions carries weight as a relevant constraint, boosting discrimination between unrelated groups and reducing false positive rates. For example, in a separate trial not documented above, when using a CSSD identical to *CSSD C* in *TR.3* but excluding the PK, Usutu (JEV group) returns spurious matches in the 3' UTR hypervariable region (at positions 13-15, and 56). It might therefore be worth exploring the action of additional PKs, or base triple interactions (refer to, for instance, MacFadden et al., 2018), on the performance of the CSSD method; and

- in relation to the false positives (FP) and false negatives (FN) noted in **Table 3** for *TR.3*, it is tempting to assert that these are simply artifacts of the manual process adopted in building *CSSD C*. While this might well be the case, further verification and testing is required to determine the root cause of these inexact ('training') results and determine the full potential of the CSSD method. If revisited and further improved, the CSSD could potentially yield even better sensitivity and specificity metrics than the ones shown below

Taking the results of *TR.3* as a benchmark for detecting duplicated DB structures across MBFV viruses, a measure of sensitivity/specificity can be procured using customary formulation, where positive/negative hits are tracked on the basis of *single* RNA elements returned in search, shown below in subscript:

$$\left.\begin{array}{lllll}\text{(True Positive Rate)} & \textbf{Sensitivity}_{\textit{CSSD C}} & = & TP_{[\,9\,]} \; / \; (TP_{[\,9\,]} + FN_{[\,3\,]}) & \simeq \; \mathbf{75\%} \\ \text{(True Negative Rate)} & \textbf{Specificity}_{\textit{CSSD C}} & = & TN_{[33]} \; / \; (TN_{[33]} + FP_{[1]}) & \simeq \; \mathbf{97\%} \\ \text{(+ve Predictive Value)} & \textbf{Precision}_{\textit{CSSD C}} & = & TP_{[\,9\,]} \; / \; (TP_{[\,9\,]} + FP_{[\,1\,]}) & \simeq \; \mathbf{90\%} \end{array}\right\} \quad (1)$$

Two CSSD were used in *TR.4*, even if in principle, a single CSSD could be constructed to accommodate both stem-loop structures. For this work, it might be deemed sufficient to operate using multiple CSSD, given that the principal aim is to assess initial performance of CSSD-based processes. Similar sensitivity and specificity metrics were estimated for this trial run, taking into account both consensus descriptors:

$$\left.\begin{array}{lllll}\text{(True Positive Rate)} & \textbf{Sensitivity}_{\textit{CSSD } D_{1,2}} & = & TP_{[\,8\,]} \; / \; (TP_{[\,8\,]} + FN_{[\,1\,]}) & \simeq \; \mathbf{89\%} \\ \text{(True Negative Rate)} & \textbf{Specificity}_{\textit{CSSD } D_{1,2}} & = & TN_{[32]} \; / \; (TN_{[32]} + FP_{[5]}) & \simeq \; \mathbf{87\%} \\ \text{(+ve Predictive Value)} & \textbf{Precision}_{\textit{CSSD } D_{1,2}} & = & TP_{[\,8\,]} \; / \; (TP_{[\,8\,]} + FP_{[\,5\,]}) & \simeq \; \mathbf{62\%} \end{array}\right\} \quad (2)$$

Summarizing the above results in terms of F_1 score, defined as $2 \times \frac{\text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}}$, one obtains a score of **0.82** and **0.73** for *TR.3* and *TR.4*, respectively.

While noting that the performance indicators summarized in **Relations 1** and **2** are encouraging, an improvement that would apply to all test runs is the inclusion of larger test datasets. However, even with the limited dataset used in the above trials, the utility of topological constraints, and in particular, the inclusion of pseudoknots is already very detectable. A dramatic example of this can be obtained from *TR.5*. When including the hairpin structure and pseudoknot constraints as given in *CSSD E3*, the results obtained are both highly specific and sensitive: all but one (Alkhurma) sequences are retrieved and no false positives returned. Removal of both the pseudoknot and hairpin structure, retrieves the original TBFV sequences but also yields a large number of false positives: DENV1-4, Japanese encephalitis, Usutu, Yellow Fever, WNV Kunjin, Bagaza, Kokobera, Iguape, Wesselsbron, Sepik, Yokose, Apoi, Modoc, and Rio Bravo.

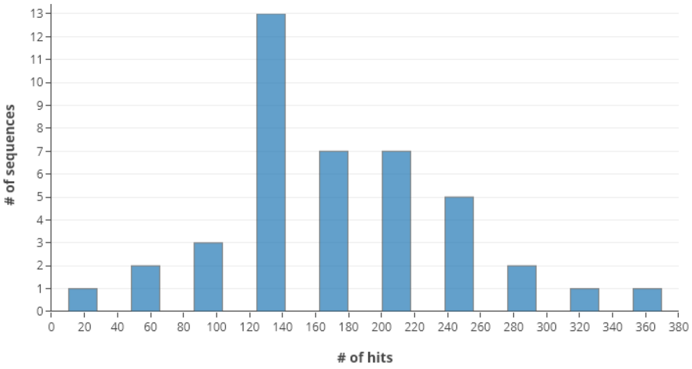


Figure 15 Distribution of hit for sequences in *TR.6*

CSSD F for *TR.6* is based on a high-confidence model, made readily available for effecting search. As no sequences were used as templates, this run involved the largest set of test sequences available. Importantly, given the large number of hits resulting from this search (**Figure 15**) it is sufficiently clear the without further constraint (such as, additional tertiary interactions, inclusion of highly conserved sequences, or thermodynamic constraints) the number of false positives would be exceedingly high for large shape spaces (~1B structures for *CSSD F*). Indeed, an initial test setup that extends the current work to include base stacking energies[30] as part of the scoring function, reveals that it is possible to separate true and false positives using appropriate *scoring thresholds*. High-level results using the updated scoring function for test runs 3-5 are shown in **Annex B.5**.

In conclusion, after careful review of the results obtained in the 6 principal trail runs conducted, it is hypothesized that further study of this method is warranted, by executing searches on a significantly larger scale than that undertaken in this preliminary work. In such a case, it might be feasible to consider search using a single consensus structure for single/duplicated elements such

[30] nearest neighbour parameters for stacking in a helix were obtained from `http://rna.chem.rochester.edu/`

as dumbbells and stem-loops. As an initial estimate, based on the work carried out and literature reviewed, it is estimated that in the order of 10 consensus structures would be an adequate set to cover the prevalent structures in flaviviral 3' UTR regions.

4 Further Work and Conclusion

The concluding remarks provided at the end of the previous section are restated here, with an eye to laying out options for potential future work based on the current study. The following list summarily provides some recommendations that could be part of such an undertaking:

1. **test datasets, putative structures, and phylogenetic comparative analysis**. In the order of a hundred viruses have been classified as flaviviruses, of which significantly less than a half have been used in this study as templates or test subjects. Indeed, although perhaps enough sequences and structures have been adopted by way of preliminary assessment, a next stage could involve an increase in sequence coverage in both breadth of viral groups, as well as the number of strains used. The higher-level of confidence acquired could promote search for putative structures in uncharacterized (3' UTRs of) flaviviral genomes. An interesting object of further study might observe the interplay of (clustered) CSSD-search results and current phylogenetic studies of flaviviruses

2. **detailed assessment of large-space consensus structures**. Though useful in assessing the overall behaviour of wide-spanning consensus structures (such as that used in *TR.6*) a careful assessment of false positive rates in such scenarios – such as, using a 'sliding-assessment' of specificity/sensitivity across different magnitudes of scale – might provide better insight into the 'optimal' search/shape space

3. **comparative assessment with other tools**. When compared to alternative methods such as *covariance models*, consensus structures inherently possess relative merits, such as the ability to and scalability in handling tertiary interactions, and also, the ability to directly manipulate the model constructed. In this sense, a direct comparison between the two might not be feasible, however, a basic comparison of performance – at the level of base-paired structures – might procure a baseline assessment on computational and search performance

4. **allowance for specifying highly-conserved sequences**. The main objective set out in this work was to investigate the potential of variable, topology-based search. However, a somewhat direct pitfall of this approach is the inability to reliably characterize a region of interest based on very simple topological structures[31]. In such cases where highly-conserved sequences provide for a much stronger signal than secondary structure alone (for example,

[31] a small hairpin structure inevitably yields a large number of hits, as may equally simple structures such as the sHP-3' SL

3' GC-SL in *TR.5*), it is desirable to allow for and investigate the complementary use of sequence-based constraints

5. **inclusion of thermodynamic calculations in scoring function**. In the same vein of the previous point, it is feasible to add-on arbitrary calculations to the scoring function, given that position/structure tracking is already in place in the current scoring scheme

6. **higher-order structural operators**. Although the current implementation of CSSD-based search was designed to support (distributed) high performance from the outset, any combinatorial approach is inherently limited by computational resources. Employing 'logical operators' on RNA (consensus) secondary structures might address 2 core issues: a) combining two or more separate consensa in one search might curb the shape space problem referred to above, and inherently decrease false-positive rates (as in the case for *TR.6*), and b) allow for more complex searches – where a 'signature' for a flaviviral 3' UTR might be defined by a suitable combination of structures and operators (e.g. *CSSD 1* and *CSSD 2* and *not CSSD 3*). Such operators could also include operand constraints such as for example the relative distance between two sub-structures, allowing for more expressive genomic signatures to be built in (e.g. *SLA* and *sHP-3' SL*)

7. **semi-assisted consensus structure build**. Whereas the search mechanism provided, in itself is robust and malleable to user intervention, the manual process of specifying a consensus structure using a multiple sequence alignment (or any suitable, alternative substrate) is a laborious and error-prone process. A possible extension to this work could recruit concepts from *evolutionary algorithms* (EA) in support of semi-assisted consensus building. Whereas brute-force techniques such as base-pair maximization are known to produce inadequate results, a semi-assisted approach would allow the user to employ an EA as a refinement tool at the point where an alignment-based approach would have already provided good candidate descriptors. Possibly improved CSSD candidates can then be automatically searched for in the neighborhood of the candidate consensus structure

A Prototype Implementation

A.1 Requirements

The core requirement for the prototype implemented in this work is based on the *consensus secondary structure descriptor* model described in **Section 2**. The model was defined in such a way as to follow the argumentation laid out in the preceding section, where the need for complex RNA topological models inclusive of tertiary interactions was highlighted. A well-formed consensus secondary structure C is a tuple of strings $(desc, var)$, where $desc$ is a non-empty string of one or more secondary structure symbols [‘<’, ‘>’, ‘(’, ‘)’, ‘_’, ‘−’, ‘:’, ‘{’, ‘}’, ‘[’, ‘]’] as defined in **Section 2**[32] to realize a specific secondary structure topology, and var is a variability specifier that may contain instances of members of the set [‘**1**’-‘**9**’, ‘**a**’-‘**p**’, ‘ ’], such that $|\, desc \,| = |\, var \,|$. By following the one-to-one positional correspondence between $desc$ and var, C thus denotes a structural descriptor with inbuilt variability, that is, one or more of the alphanumeric *positional variables* specified in var give rise (‘expanding $desc$’) to a finite set of unique structures of size $|\, C \,| = k$. Each resulting structure $c \in C$ is therefore a non-empty string of valid symbols as listed above, and each representing a valid RNA secondary structure descriptor.

For a given consensus structure C of size k, and a RNA sequence S of length n (indexed by nucleotide positions $1 \ldots n$, where $k, n \geq 1$), the solution discussed in this section attempts to answer the following query in an efficient and scalable fashion:

Given C and S, determine the set of nucleotide positions P in S at which one or more member structures of CSSD match;

where $P = \{\, p \mid S_p$ *is compatible with* $c \,\}$,

S_p *is the sub-sequence of* S *starting at nt position* p, *and* $1 \leq p \leq n$,

$c \in C$, *and*

$|\, S_p \,| = |\, c \,|$

In addition to the above formulation, the querying process may warrant further clarification by stating a few assumptions about how it is expected to operate in practice, and also by referring to specific details about the current implementation:

[32] though fixed in the current implementation, the symbol set members and semantics may be easily changed through code configuration, for example, to include a third pseudoknotted pair

query direction. A query is typically expressed (for convenience) in one direction only, as in the following way: given C, find matching secondary structures in S. The intent and scope of implementation, however, is to have queries execute in the opposite direction as well, that is, given a sequence S, search in a dataset of consensa for matching descriptors

sequences and consensus structures. Both sequences and consensus structure descriptors can be constructed to be of arbitrary length. As implemented in the current prototype, sequences include stop codons and terminal dinucleotides, and therefore the effective minimum length is 6 nt. Moreover, the current maximum sequence length is of 1000 nt, a limit set after current implementation hardware and system memory considerations were taken into account. Likewise, although a single-position CSSD is allowed, the resulting descriptor would correspond to a single unstructured nucleotide and would inherently match any given sequence. In view of this, CSSD limits are introduced after taking into account that a CSSD with meaningful topology would include at the very minimum a hairpin structure. The minimum helix size is therefore set at 2 base pairs (with a corresponding maximum limit of 14), whereas the apical loop length can range between 3 and 13 nucleotides – the effective minimum length for a CSSD descriptor therefore being 7. It should also be noted that although these minimum and maximum limits are enforced on the consensus structure descriptor $desc$, any positional variables assigned by the end-user to var may produce consensa that exceed these limits

compatibility between S_p and c. Given a CSSD member c, S_p is said to be compatible with c if the $5' \rightarrow 3'$ nucleotide sequence represented by S_p matches exactly the structural descriptor c such that any base-pair or pseudoknot contact requirement set by the descriptor is fulfilled. In the current implementation, only WC base-pairs and the G-U wobble pair are allowed, although pairing parameters can be readily modified using code constants. When scoring c against S_p under the current configuration, the scoring function requires an exact match, that is, only a binary score of 0 or 1 is issued. A score for a partial match – that is, a well-defined RNA secondary structure c' exists such that c' is a proper substring of c, and $S_{p'}$ is compatible with c', where $S_{p'}$ is a proper substring of S_p – may be produced if the appropriate compile-time directives are enabled and the system rebuilt. Also, given the ability to perform partial matching, it should be noted that arbitrary scoring functions can be easily coded in as extensions to the current prototype – for example, by superimposing thermodynamic calculations as matching progresses through a CSSD; or otherwise, allowing for conserved sequences to be matched as part of the descriptor

co-located and adjacent hits. The current implementation allows for multiple hits (corresponding to multiple, unique structures) to be returned at any matching position. In the command-line tools provided and in accordance with the objectives set at the outset, all hits are returned. Moreover, variability in the CSSD allows for two adjacent hits to be either overlapping, or possibly 'nested', such that a larger matching structure at position n is exactly one base-pair longer than the structure matching at position $n + 1$. The objectives initially set out for this project require that only base-pairing and PK constraints act as filters in the process of matching structures, however, it should be noted that post-processing of results may easily be added to filter out any undesired output

usability requirements. The querying tool is provided to the end-user in the form of a browser-based interface (as the main tool) and ancillary command line tools, all of which include the ability to distribute queries on a compute cluster[33]. Here, a high-level list of implemented features is provided, making reference to the prototype architecture shown in **Figure 17**. Brief algorithmic and other implementation details are provided in the subsequent section below:

- *web front-end*

  A web front-end ('*Structural RNA Homology Search*') allows the user to manage sequences, consensus structures, and computational resources (including booking of cluster resources), as well as execute queries and browse for query results. A brief overview of end-user functionality provided follows:

  consensus structures. The front-end allows for the definition and interactive building, storage and visualization of consensus structures. Interactivity is essential if a new (unknown) topology descriptor is to be built, wherein an initial structure is progressively refined by querying intermediate consensus structures against a known set of sequences (somewhat similar, conceptually, to a simultaneous *fold and align* approach). The in-browser visualization tool provided allows for representation of arbitrary secondary structures. Pseudoknot representation is currently limited to the first instance, with further extensions possible to allow for any number of tertiary interactions to be represented

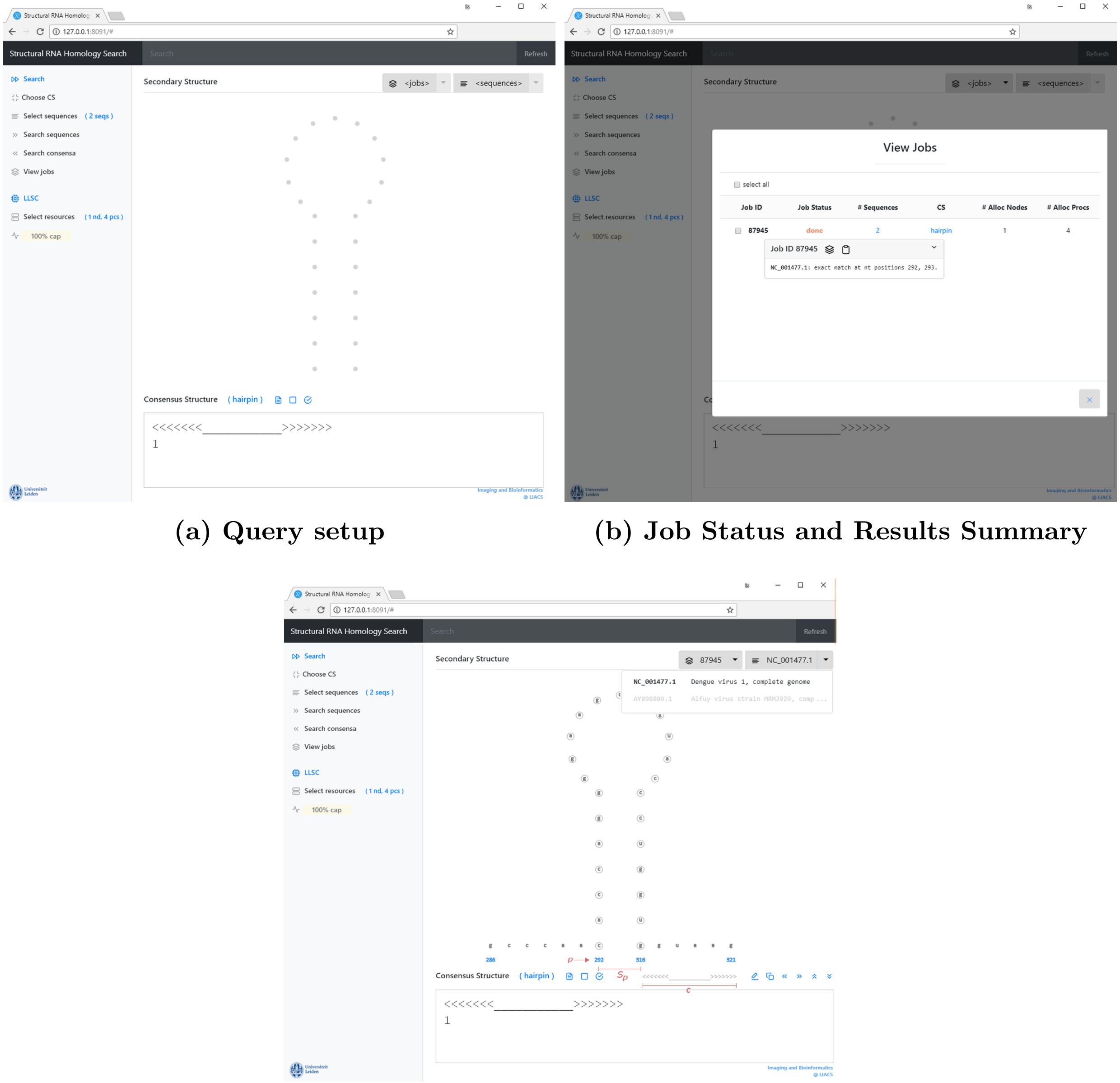
[33] LLSC, the *LIACS Life Science Cluster* at Leiden University, was used for both system development and testing

RNA sequences. New RNA sequences can be created, edited, and stored. Using existing sequences, a local or web-based BLAST[34] query can be used from within the UI to search and incorporate similar sequences. In the current implementation, the local BLAST version only supports the non-redundant nucleotide databases, with additional querying possible using the NCBI-provided RESTful APIs and command-line tools provided. (For future releases, it would be desirable to route local BLAST queries to a more comprehensively set-up, cluster-distributed BLAST installation.) Besides the nucleotide sequence itself, sequence properties include an *accession number*, an informative *text descriptor* of the nucleotide sequence, and a *group identifier* which can be used for effecting actions on a group of sequences

compute resources. Cluster nodes need to be selected prior to running a query. In the current implementation, whole compute nodes from a list of free nodes may be selected, allowing the end-user to map queries to appropriately dimensioned computing resources. Basic information about the compute resources used is made available including *job id* and *job execution time*. (It should be noted that as per the alogorithmic approach described below, excess resources are unutilized by the system, based on derived estimates of query computational requirement.)

query jobs and results. Using the web interface, query jobs can be set up by selecting one or more sequences, a consensus structure, and a number of compute nodes, following which the job is submitted to the LLSC for processing. Once complete, a summary of hits per sequence are returned with the option to browse individual results. If browsing mode is selected, sequences that yielded at least one hit can be viewed, whereby a subsequence (S_p) matching a specific structure (c) at position p is displayed. The user is allowed to browse through positions within the genome, however, in the current prototype only the first match at any given position is directly viewable in the UI. **Figure 16** illustrates the workflow, by providing three snapshots taken during a typical process starting from query setup through to result browsing

[34] https://blast.ncbi.nlm.nih.gov/


(a) Query setup

(b) Job Status and Results Summary

(c) Browse Results

View Jobs

☐ select all

	Job ID	Job Status	# Sequences	CS	# Alloc Nodes	# Alloc Procs
☐	87945	done	2	hairpin	1	4

Job ID 87945

```
NC_001477.1: exact match at nt positions 292, 293.
```

```
NC_001477.1    Dengue virus 1, complete genome
AY898809.1     Alfuy virus strain MRM3929, comp ...
```

Consensus Structure (hairpin)

```
<<<<<<<____________>>>>>>>
1
```

Figure 16 Example Web-based Workflow. In this example, 2 sequences are selected (Dengue virus, NC_001477.1 and Alfuy virus, AY898809.1), along with a simple hairpin structure with a positional variable of '1' at the stem base. The visualization in (**a**) shows the corresponding secondary structure, without taking into consideration the effect of positional variables. A single compute node is selected and the search query submitted. In (**b**), the job status is marked as complete and 2 matching positions (292, 293) returned for sequence NC_001477.1. Additional information is tracked per job, including job execution times, and sequence/CSSD submitted. If browsing mode is selected (refer to **c**), the matching sequence can be browsed, with the visualization elements changing to represent S_p, c, and p

- *back-end system*

  Most of the back-end environment and functionality is dedicated to executing search jobs submitted by the front-end and facilitating interaction with the cluster. Most operational scripts are therefore not intended to be used by non-technical users (refer to **Figure 17** for an architectural overview).

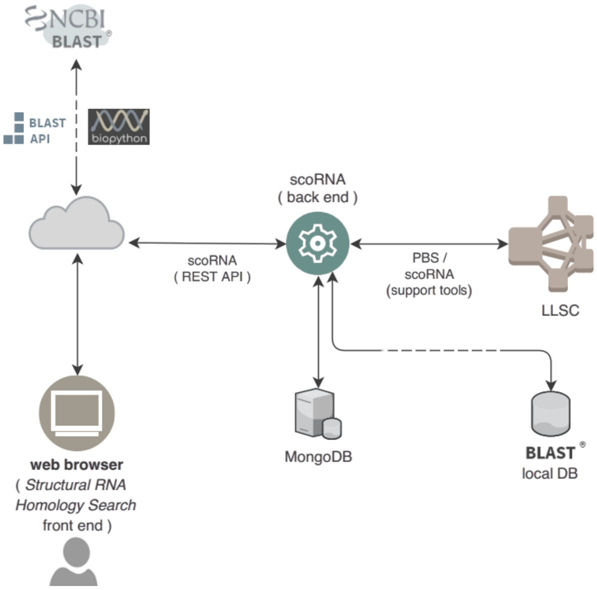


Figure 17 Prototype architecture

  Two command-line tools, however, are provided for potential use by an end-user: a) *scoRNA* is a tool for scoring RNA sequences against a consensus secondary structure. Supplied with a FASTA and a WUSS file (in CSSD format), the tool will list any hits (sequence#, nt position p, and matching structure c) for one or more sequences provided in the FASTA file. The tool also seamlessly works on a single machine as well as across a PBS-enabled cluster; and b) *run-mpi-batch* is a wrapper tool that automatically splits large jobs into manageable chunks that fit within the memory (and to a lesser extent, computational) constraints of the curren distributed environment

system requirements. The goal of this work, at the outset, was to test for feasibility of detecting signature structures in the 3' UTR of flaviviruses using only topological constraints (that is, descriptors for base-pairing and pseudoknot contacts), and in this respect no strict expectations were laid out on how the working prototype is to be implemented. In due course, the original scope was expanded to include a basic UI and ability to distribute queries over a compute cluster. The final prototype used to generate results detailed in **Section 3** was built (primarily) using the following tools and frameworks, with reference to the high-level architecture provided in **Figure 17** above. As with all libraries and frameworks adopted and mentioned in this section, the primary reasons for choosing the specific tools include interoperability and widespread adoption. In this regard, most components, frameworks, and tools used can be replaced with alternative solutions with relative ease. In some cases, such as with *MongoDB*, replacement does not involve any code or system changes other than replacement of the specific component itself:

- *front-end*

  The rationale for including a user front-end in the prototype was for ease of use of the search and scoring (back-end) system, and to allow visualization and interactive manipulation of consensus structures. With this goal in hand, the interface was built using *HTML/CSS*, *Javascript* and *jQuery* for cross-platform interoperability, while using *Bootstrap* for responsive design. Visualization and interactivity is provided through *D3.js*

- *back-end*

  The prototypical framework was built in *Python* (and launched using a command-line script *scoRNA.py*), and uses a number of standard packages: *Flask* as the main web framework and RESTful API provider, *Biopython* for sequence handling and BLAST support, and *PyMongo* as the database interface. *MongoDB* is used as a NoSQL repository for storage/retrieval of consensus structures and sequences. It should be noted that in this early prototype, search results are not stored but only persist for the duration of the user's browsing session. The cluster used for both prototype development and testing (the *LLSC*) is constituted of a 120-core, 22-node linux-based system currently running the *PBS* job scheduler. *scoRNA* interfaces to the cluster management environment using standard *PBS* job scripts, and also provides a number of Python command-line support scripts for interpreting search results. For performance reasons, the search algorithm itself is implemented in C. Two versions are made available, a single processor and a cluster-enabled prototype. For the latter, cluster message passage is done using MPI, along with standard C libraries for full interoperability[35]

- *front- and back-end communication*

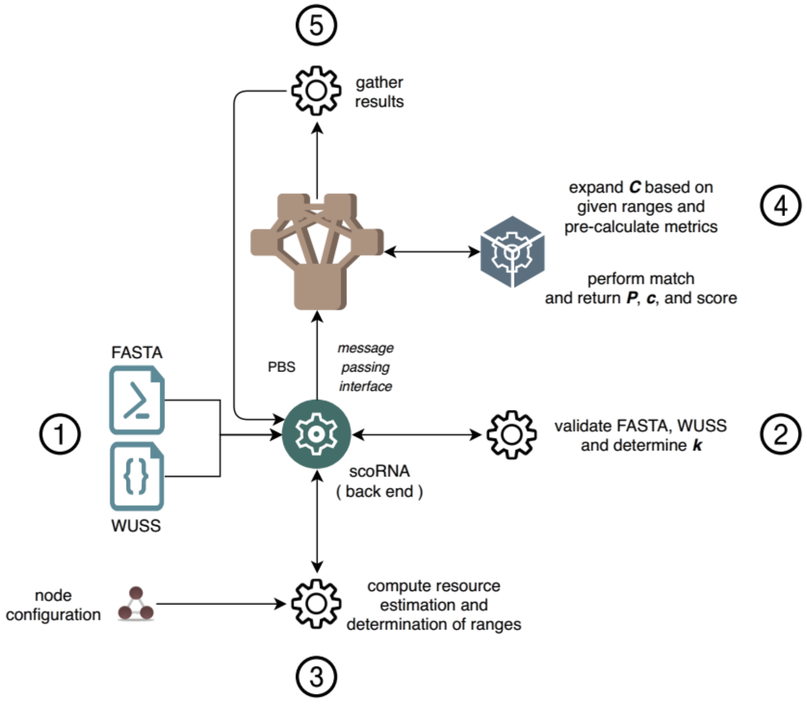
  A minimal RESTful interface has been adopted for interaction between user-interface and back-end environment, allowing for interoperability with multiple (possibly, additional) components in a loosely coupled manner. In the following, a few examples of the tail path of the base URL are given to demonstrate use of the API:

Path[36]	Request Methods Allowed	Action
/api/v1/cssd	GET, POST, DELETE	retrieve, create/update, or delete CSSD
/api/v1/sequences	GET, POST, DELETE	retrieve, create/update, or delete sequence
/api/v1/cluster-management/status	GET	get node status on cluster
/api/v1/search	POST, GET	submit search job or retrieve search results
/api/v1/search/BLAST	GET	search for similar sequences using BLAST

[35] both versions has been successfully tested on linux and (a quad-core) MS windows based system

A.2 Scoring algorithm and computational performance

Figure 18 provides an overview of the principal algorithmic steps involved in producing search results in response to the query stated in the previous section. In the following, the main steps highlighted (numbered 1 through 5) in the adjacent figure are briefly described.



5
gather results
expand C based on given ranges and pre-calculate metrics
4
perform match and return P, c, and score
FASTA
PBS
message passing interface
1
validate FASTA, WUSS and determine k
2
scoRNA (back end)
WUSS
node configuration
compute resource estimation and determination of ranges
3

Figure 18 Search Algorithm Worflow

In **step 1**, *FASTA* and *WUSS* formatted datasets are supplied to the search algorithm either as text files when using command-line tools, or as appropriately formatted objects (in *JSON* format) marshalled over the REST API at the front-end user interface. Though intimately related to the previous step, validation of both FASTA and WUSS data are performed separately (**step 2**), this being the main entry point to the search algorithm itself. Also in this step, the expansion factor k is estimated from any positional variables supplied in the given CSSD. It should be noted that whereas multiple FASTA-formatted sequences can be dealt with directly, the current prototype requires multiple invocations of the algorithm to search through multiple consensus structures – the latter not being an inherent limitation of the algorithmic approach adopted, but a limitation of the current implementation.

step 3 determines the computational resources (that is, the number of cluster nodes) required to distribute the query effectively over the cluster infrastructure. The input parameters being a) the number of sequences, b) the consensus structure tuple C, and c) the maximum number of nodes booked by the end user. The third parameter is useful as it allows the end-user to 'throttle' multiple jobs submitted concurrently to the cluster, based on job priority. The resource estimation

[36] for simplicity, addressed members of collections, for example, sequence accession numbers are not shown. Moreover, only the most frequently used URL snippets are provided

procedure is a relatively naive one, where firstly, the largest positional variables (in descending order) contributing to the size k of C are assessed to determine what is the best possible split across the number of nodes booked by the user. In a second iteration, the number of sequences submitted by the user are also assessed for distribution across the number of nodes. Letting the user-supplied number of nodes be n, then an estimate for the optimal number of nodes to be used o is produced, in the range $1 \leq o \leq n$. It should be noted that k is given priority, since, in the general case the order of magnitude of k is larger (typically, thousands to millions), when compared to the number of sequences provided (typically, hundreds to thousands).

step 4 involves farming out the search query across the designated number of nodes. On each node, the consensus structure C is first expanded according to the ranges determined in the previous step. This approach is preferred since, expanding C locally typically takes negligible amount of computation time, whereas expanding C centrally and distributing across nodes incurs a relatively expensive I/O penalty. More importantly, the way the CSSD is expanded is in a "**outwards-inwards**" fashion. This is particularly relevant as it allows the next step (that is, the actual matching step) to fail fast and avoid repeated, unnecessary matching. Moreover, scoring metrics are pre-calculated and stored during CSSD expansion for fast retrieval when partial scoring is applied.

Searching is performed by linearly scanning sequences (*only once, starting at each nucleotide position*) and matching against the current structure c, immediately failing to the next position or sequence upon a mismatch. Base-pair and tertiary interaction matching can be done efficiently since secondary and tertiary structure contacts are typically 'local', requiring storage of only a minimal amount of parsed information, when compared to the more voluminous shape space of consensus secondary structures.

step 5 gathers any partial results produced by worker nodes and updates the back-end system accordingly

Figures 19 and 20 highlight a selection of performance tests run using the above algorithm. The first figure provides runtime performance indicators (in milliseconds) for a number of sequences ranging between 1 and 1000. The CSSD descriptor used in this case was a simple hairpin structure, with a 4-base pair helix and an apical loop length of 5. The sequences used were a mixture of real sequences (same as those used in the results section) and random sequences, with the average sequence length also shown in the figure. In this case, a single-processor configuration was used. For the CSSD and sequence dataset sizes used in this test, the performance of the algorithm is approximately linear. In **Figure 20**, the performance of the algorithm is benchmarked against a sequence dataset ranging between 1 and 100 in size, and using multiple cluster node configuration ranging from a single-processor configuration to a selection of nodes having 64 cores in total. The CSSD used (below) evokes a shape space of ~8.600 secondary structures. As can be seen in the figure below, the algorithm performance reasonably well, providing a speed up factor of ~40 under the 64 processor configuration.

```
<<<<<-----<<<-<<<<-<_[[[[__>>>>>>>>>>>>>::]]]]
        5         123  1  4              5
```

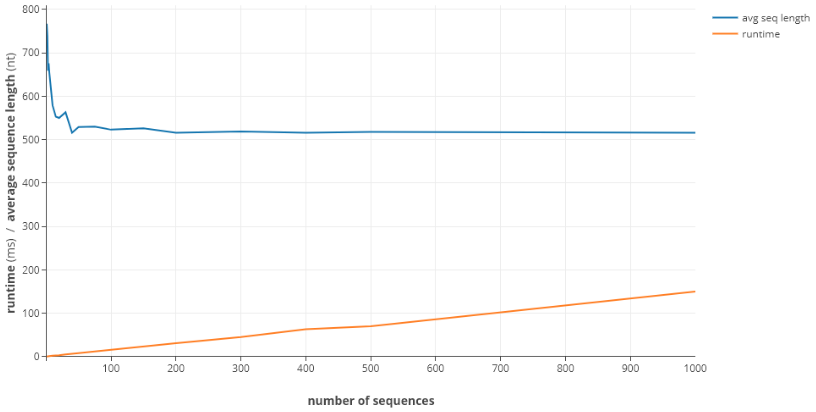


Figure 19 Single-processor performance against size of sequence dataset

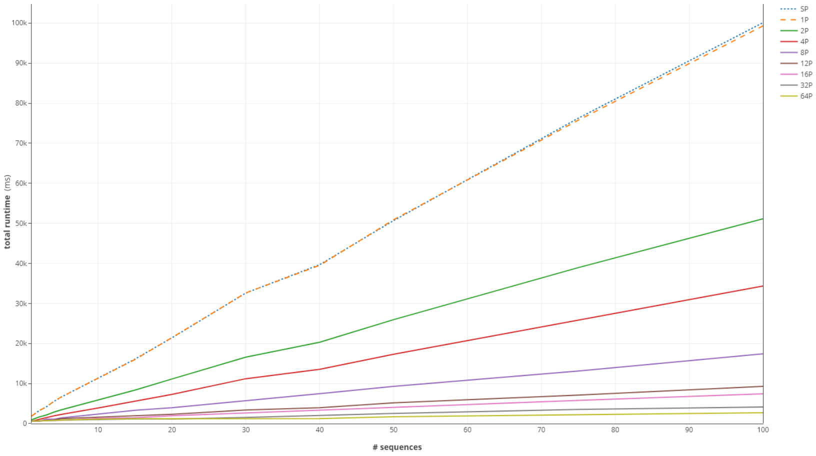


Figure 20 Performance against number of sequences and compute cores

A.3 Limitations and enhancements

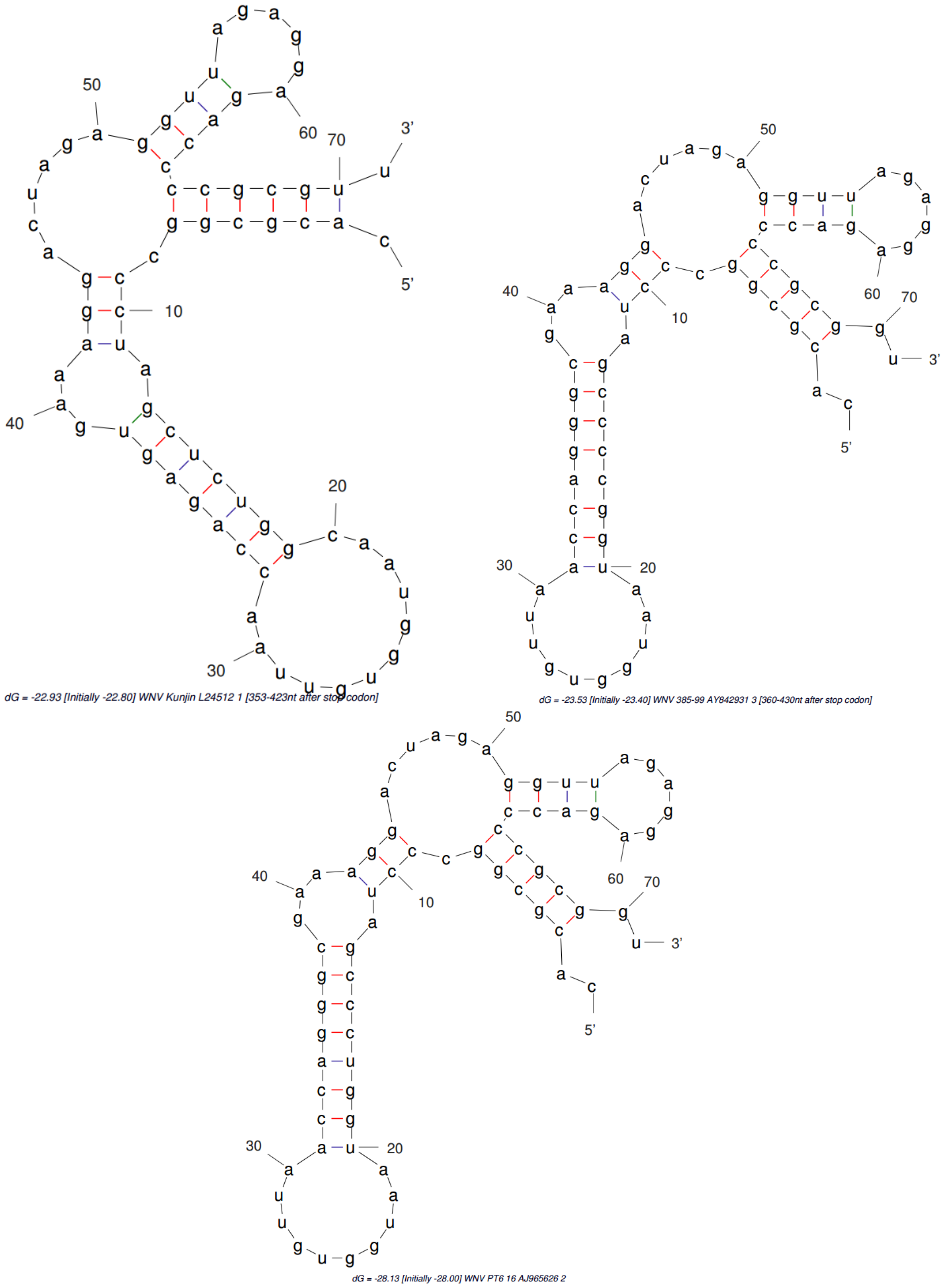
The following limitations in the design and use of the search algorithm have been identified, also with an eye to potential opportunities for future work:

1. usability testing, both in terms of the end-user tools and the range of viral genomes used is limited, and more fruitful considerations could be made if further testing is performed
2. as per original intent, the system developed runs under a 'proof of concept' premise. Further testing by end-users is necessary to quality check the prototype and obtain end-user feedback
3. degenerate nucleotide sequences are currently not supported and such sequences might require editing by the end-user from within the user interface
4. validation of sequences and consensus structures is done separately within the UI and the back-end. As an extension to the current prototype, the REST API (or alternative method) can be expanded to allow for centralized validation of system input
5. var positional variables are currently not visualized in the browser interface. Such functionality might assist the end-user in visualizing better the expanded consensus structure
6. cluster job status management currently lacks informative error reporting
7. in the current prototype, only the first (out of possibly more than one) hit is returned in the browser interface. Although the end-user can browse results manually using the back-end scripts, it is preferably to allow the user browse all applicable results at the front-end
8. as described above, large jobs (currently defined as a CSSD expanding into more than 2M structures) are split using back-end scripts. In a next iteration, such functionality could be integrated and automatically made available through the user interface
9. the current method of partitioning (expanded) consensus structures across the cluster does not make use of message passing when a search on a given node fails (i.e. a mismatch occurrence)
10. the current implementation uses standard C libraries, in support of cross-platform compatibility. Various specialized libraries, in particular, more efficient memory management libraries might be adopted to further boost performance
11. although the intent of this work is to provide a proof of concept for topology-based RNA modeling and search, a comparison with other tools might be warranted for a broader perspective on the utility of the method adopted

B Sequence/Structure Supplementary Information

B.1 Folded structures for sequences matched by CSSD A

For additional validation purposes, the matching sequences corresponding to the base-paired model for WNV 5' DB created in *TR.1*, were compared to folded structures of the same sequences using *mfold* (Zuker, 2003), as shown in **Figure 21** below.



dG = -22.93 [Initially -22.80] WNV Kunjin L24512 1 [353-423nt after stop codon]

dG = -23.53 [Initially -23.40] WNV 385-99 AY842931 3 [360-430nt after stop codon]

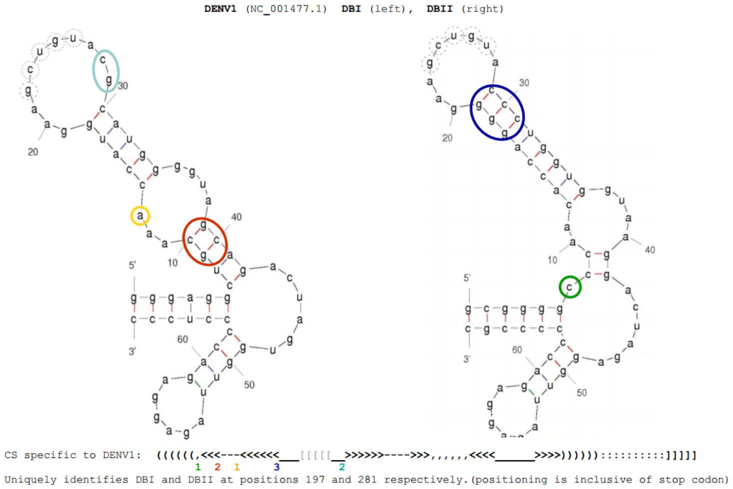
dG = -28.13 [Initially -28.00] WNV PT6 16 AJ965626 2

Figure 21 *mfold* structures of WNV Kunjin (top left), WNV 385-99 (top right), and WNV PT6.16 (bottom) matching the sequence/structure returned by *CSSD A*

B.2 Iterations required for CSSD B

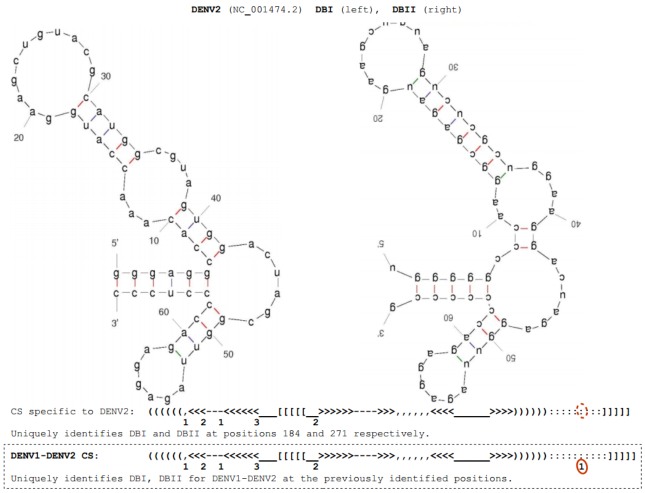
In producing *CSSD B*, the additional template sequences (DENV1-4, Japanese encephalitis, and Usutu) were sequentially used to 'widen' the initial consensus structure *CSSD A* to match up with the sequence (structural) variability required.

In each of **Figures 22 to 27**, new positional variables are introduced or existing ones modified (increased), to accommodate structural differences presented by the respective sequence. The colour scheme used for representing modified positional variables at each step is matched by same-colour circles overlayed on the *mfold*-structures, indicating which structural parts necessitated the respective change. 5' (3') DB structures are shown on the left (right).


DENV1 (NC_001477.1) **DBI** (left), **DBII** (right)

```
CS specific to DENV1:  ((((((,<<<---<<<<<<___[[[[[__>>>>>>---->>>,,,,,,<<<<______>>>>)))))):::::::::]]]]]
                              1  2  1     3        2
Uniquely identifies DBI and DBII at positions 197 and 281 respectively.(positioning is inclusive of stop codon)
```

Figure 22 Adding DENV1 to the initial *CSSD A*

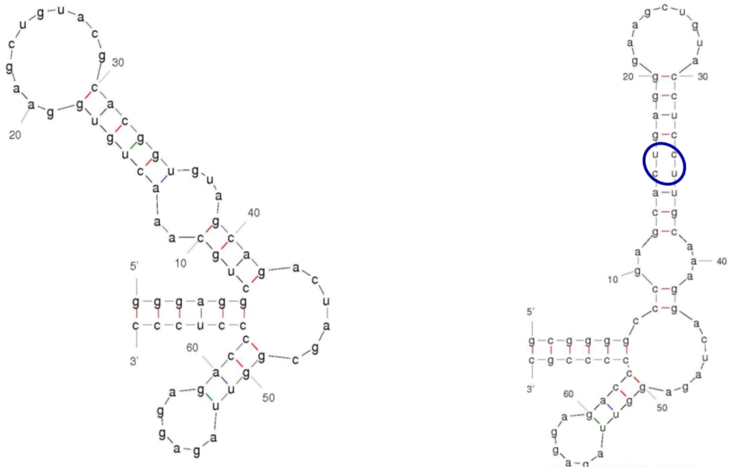

DENV2 (NC_001474.2) **DBI** (left), **DBII** (right)

```
CS specific to DENV2:  ((((((,<<<---<<<<<<___[[[[[__>>>>>>--->>>,,,,,,<<<<______>>>>)))))):::::::::]]]]]
                              1  2  1     3         2
Uniquely identifies DBI and DBII at positions 184 and 271 respectively.

DENV1-DENV2 CS:        ((((((,<<<---<<<<<<___[[[[[__>>>>>>--->>>,,,,,,<<<<______>>>>)))))):::::::::::]]]]]
                              1  2  1     3         2                                    1
Uniquely identifies DBI, DBII for DENV1-DENV2 at the previously identified positions.
```

Figure 23 Adding DENV2 to the *CSSD A* and DENV1

DENV3 (NC_001475.2) **DBI** (left), **DBII** (right)

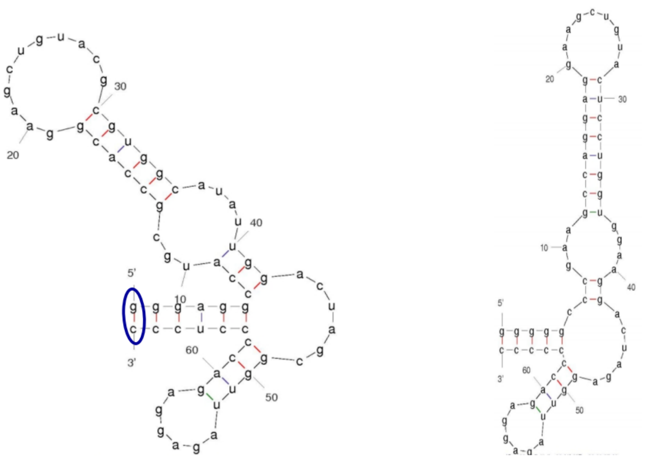


```
CS specific to DENV3:  ((((((,<<<--<<<-<<<<___[[[[[__>>>>->>>--->>>,,,,,,<<<<______>>>>))))))::::::::]]]]]
                             1  2     2   1         2    2                                       2
Uniquely identifies DBI and DBII at positions 176 and 260 respectively.
```

```
DENV1-DENV3 CS:        ((((((,<<<---<<<-<<<___[[[[[__>>>->>>---->>>,,,,,,<<<<______>>>>))))))::::::::]]]]]
                             1  2  1   2  3         2   2   1                                    3
Uniquely identifies DBI, DBII for DENV1-DENV3 at the previously identified positions.
```

Figure 24 **Adding DENV3 to the *CSSD A* and DENV1-2**

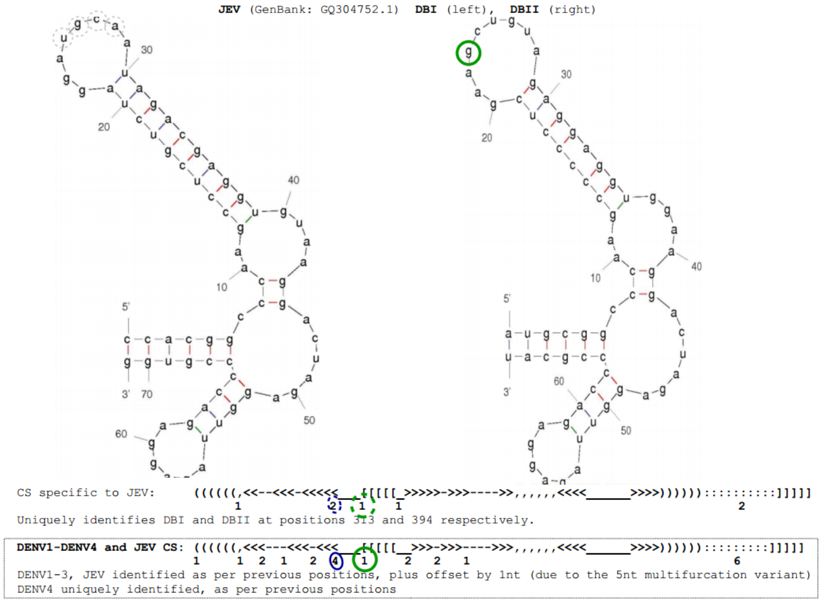
DENV4 (NC_002640.1) **DBI** (left), **DBII** (right)



```
CS specific to DENV4:  ((((((,<<<---<<<<<<<___[[[[[_>>>>>>>---->>>,,,,,,<<<<______>>>>))))))::::::::]]]]]
                       1     1  2         2        3                                            5
Uniquely identifies DBI and DBII at positions 114 and 205 respectively.
```

```
DENV1-DENV4 CS:        ((((((,<<<---<<<-<<<___[[[[[__>>>->>>---->>>,,,,,,<<<<______>>>>))))))::::::::]]]]]
                       1     1  2  1   2  3         2   2   1                                    6
DENV4 uniquely identified. For DENV1-3, as before, plus offset by 1nt (due to the 5nt multifurcation variant)
```

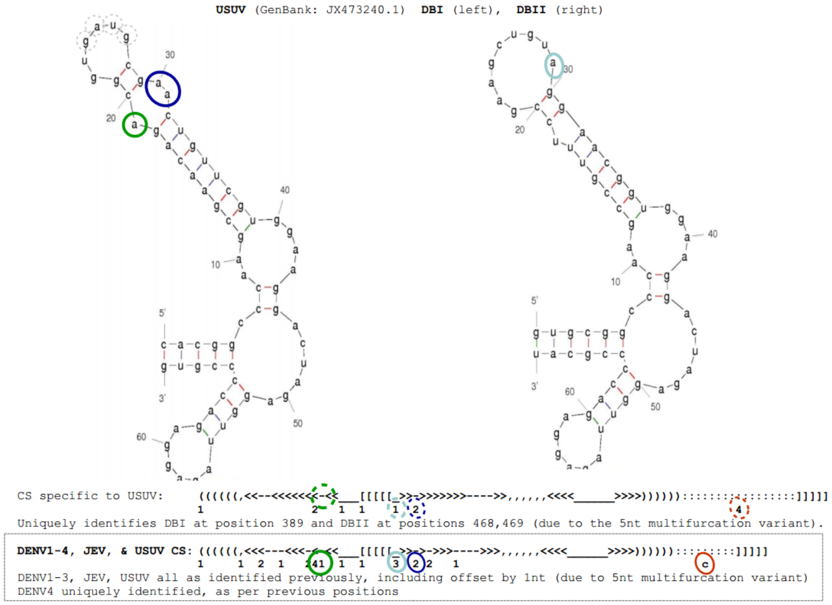
Figure 25 **Adding DENV4 to the *CSSD A* and DENV1-3**



```
CS specific to JEV:      ((((((,<<--<<<-<<<<<___[[[[[_>>>>>->>>---->>,,,,,,<<<<______>>>>)))))):::::::::]]]]]
                               1              2   1   1                                              2
Uniquely identifies DBI and DBII at positions 313 and 394 respectively.
```

```
DENV1-DENV4 and JEV CS:  ((((((,<<<---<<<-<<<___[[[[[__>>>->>>---->>>,,,,,,<<<<______>>>>)))))):::::::::]]]]]
                         1     1  2  1   2   4   1      2   2   1                                    6
DENV1-3, JEV identified as per previous positions, plus offset by 1nt (due to the 5nt multifurcation variant)
DENV4 uniquely identified, as per previous positions
```

Figure 26 Adding Japanese encephalitis to the *CSSD A* and DENV1-4



```
CS specific to USUV:     ((((((,<<--<<<<<<<-<<___[[[[[_>>->>>>>>>---->>,,,,,,<<<<______>>>>)))))):::::::::::::::]]]]]
                         1                  2   1  1     1 2                                             4
Uniquely identifies DBI at position 389 and DBII at positions 468,469 (due to the 5nt multifurcation variant).
```

```
DENV1-4, JEV, & USUV CS: ((((((,<<<---<<<-<-<<___[[[[[_>>->->>>---->>>,,,,,,<<<<______>>>>)))))):::::::::]]]]]
                         1     1  2  1   241  1  1     3 2 2   1                                   c
DENV1-3, JEV, USUV all as identified previously, including offset by 1nt (due to 5nt multifurcation variant)
DENV4 uniquely identified, as per previous positions
```

Figure 27 Adding Usutu to the *CSSD A* and DENV1-4, Japenese encephalitis

B.3 5'-proximal nucleotide frequency distribution for all sequences used

Figure 28 below reveals the uncharacteristic distribution of the four RNA nucleotides for Usutu and WNV$_{KUN}$, relative to the other 42 sequences used in test runs 1 to 6 in **Section 3**. The former two sequences are underscored along the x-axis, where it is evident that these sequences have relatively higher a-g-u content and are, correspondingly, c-poor. This suggests that the initial (5' proximal) regions of these sequences are able to conform to a variety of secondary structure descriptors, thus allowing for a higher false positive hit rate. Indeed, this could be the principal reason why under many test scenarios Usutu and Kunjin strains yield a number of false positives in their initial 5' proximal region (refer to, for example, **Table 4** for *TR.4*).

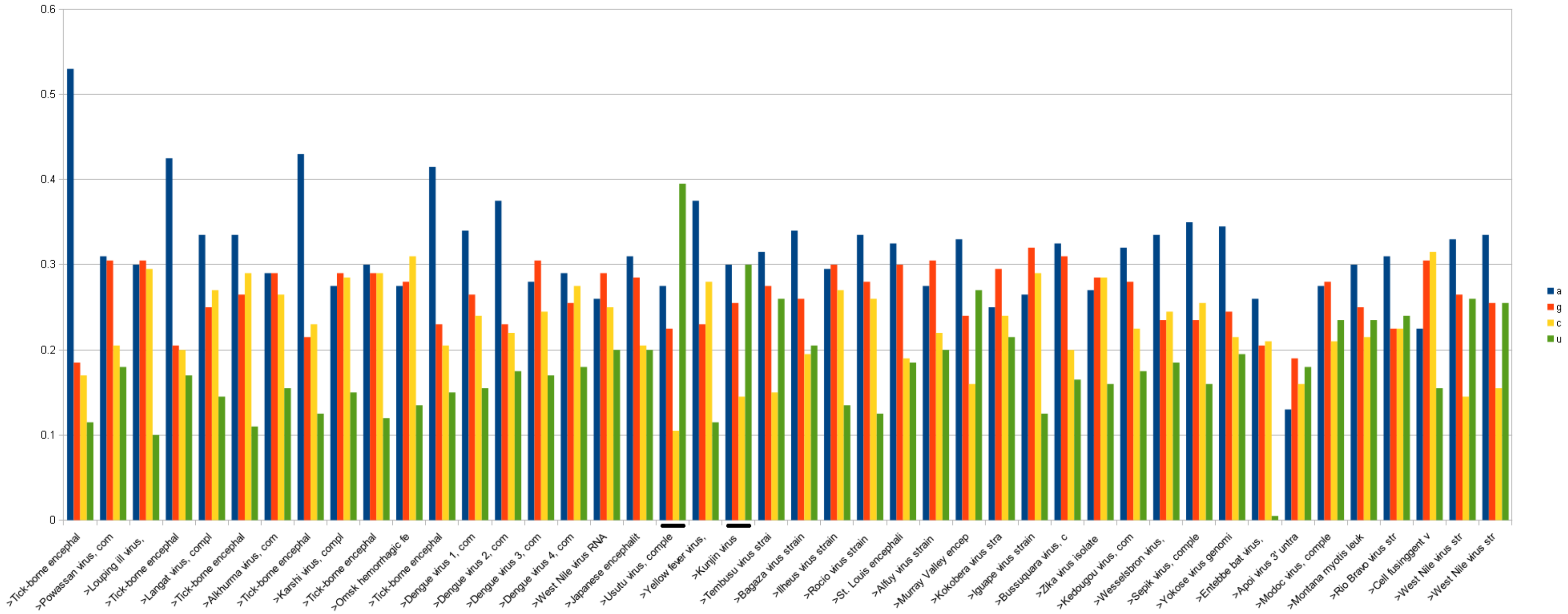
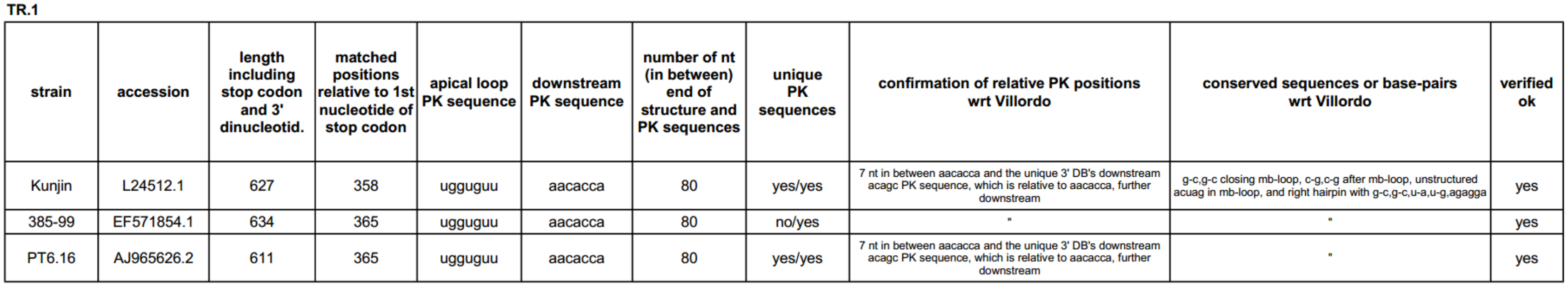


Figure 28 Frequency distribution of nucleotides in the first 200 (5'-proximal) nucleotides

B.4 Extended Test Run 1 to 6 Results

Figures 29 to **36** below provide extended information in relation to the test run results reported in **Section 3**. The figures include the basic data provided in summary format in that section, in addition to the following information:

- all upstream/downstream sequences involved in the pseudoknots incorporated into the model being tested (columns 5 and 6);
- whether the PK sequences are unique within the test sequence's 3' UTR (column 8);
- confirmation of the relative distances between structures (and their downstream PK sequences) and conserved sequences/base pairing information as provided by Villordo et al. (columns 9 and 10); and
- an verification assessment (column 11) based on all the above information for the respective hit(s) listed. In cases where multiple hits are reported at a particular nt position, the verification data is with respect to the one highlighted in bold. In a few isolated cases, the verification status is tagged as 'maybe', in which case, all conserved sequences and base-pairing information is verified but the sequences involved in PK interactions could not be directly verified (in Villordo or in the literature, generally)



TR.1

strain	accession	length including stop codon and 3' dinucleotid.	matched positions relative to 1st nucleotide of stop codon	apical loop PK sequence	downstream PK sequence	number of nt (in between) end of structure and PK sequences	unique PK sequences	confirmation of relative PK positions wrt Villordo	conserved sequences or base-pairs wrt Villordo	verified ok
Kunjin	L24512.1	627	358	ugguguu	aacacca	80	yes/yes	7 nt in between aacacca and the unique 3' DB's downstream acagc PK sequence, which is relative to aacacca, further downstream	g-c,g-c closing mb-loop, c-g,c-g after mb-loop, unstructured acuag in mb-loop, and right hairpin with g-c,g-c,u-a,u-g,agagga	yes
385-99	EF571854.1	634	365	ugguguu	aacacca	80	no/yes	"	"	yes
PT6.16	AJ965626.2	611	365	ugguguu	aacacca	80	yes/yes	7 nt in between aacacca and the unique 3' DB's downstream acagc PK sequence, which is relative to aacacca, further downstream	"	yes

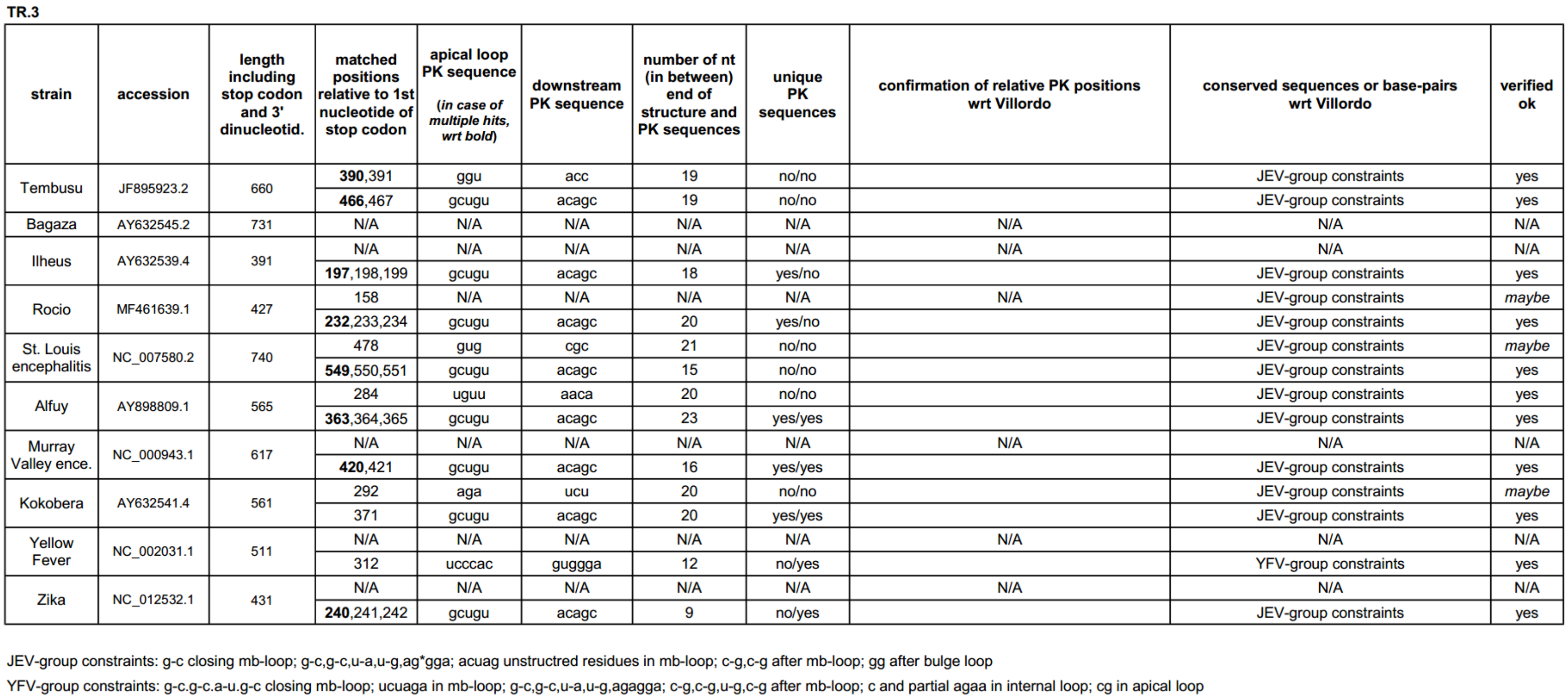
Figure 29 Extended TR.1 results

TR.2

strain	accession	length including stop codon and 3' dinucleotid.	matched positions relative to 1st nucleotide of stop codon	apical loop PK sequence (*in case of multiple hits, wrt bold*)	downstream PK sequence	number of nt (in between) end of structure and PK sequences	unique PK sequences	confirmation of relative PK positions wrt Villordo	conserved sequences or base-pairs wrt Villordo	verified ok
Kunjin	L24512.1	627	**357**,358	ugguguu	aacacca	79	yes/yes	7 nt in between aacacca and the unique 3' DB's downstream acagc PK sequence, which is relative to aacacca, further downstream	DENV/JEV-group constraints	yes
			434,435	gcugu	acagc	19	yes/yes	7 nt in between acagc and the unique 5' DB's downstream aacacca PK sequence, which is relative to acagc, further upstream	DENV/JEV-group constraints	yes
385-99	EF571854.1	634	365	ugguguu	aacacca	80	no/yes	7 nt in between aacacca and the unique 3' DB's downstream acagc PK sequence, which is relative to aacacca, further downstream	DENV/JEV-group constraints	yes
			441,442	gcugu	acagc	19	yes/yes	7 nt in between acagc and the unique 5' DB's downstream aacacca PK sequence, which is relative to acagc, further upstream	DENV/JEV-group constraints	yes
PT6.16	AJ965626.2	611	365	ugguguu	aacacca	80	yes/yes	7 nt in between aacacca and the unique 3' DB's downstream acagc PK sequence, which is relative to aacacca, further downstream	DENV/JEV-group constraints	yes
			441,442	gcugu	acagc	19	yes/no	7 nt in between acagc and the unique 5' DB's downstream aacacca PK sequence, which is relative to acagc, further upstream	DENV/JEV-group constraints	yes
DENV1	NC 001477.1	465	**197**,198	gcugu	gcagc	10	no/no	1 nt in between downstream gcagc and stem base of 3' DB	DENV/JEV-group constraints	yes
			281,282	gcugu	acagc	10	no/no	10 nt in between end of structure and downstream PK	DENV/JEV-group constraints	yes
DENV2	NC 001474.2	454	**184**,185	gcugu	gcagc	9	no/no	5 nt in between downstream gcagc and stem base of 3' DB	DENV/JEV-group constraints	yes
			271,272	gcugu	acagc	9	no/yes	9 nt in between end of structure and downstream PK	DENV/JEV-group constraints	yes
DENV3	NC_001475.2	443	**176**,177	gcugu	gcagc	10	no/yes	1 nt in between downstream gcagc and stem base of 3' DB	DENV/JEV-group constraints	yes
			260,261	gcugu	acagc	8	no/yes	8 nt in between end of structure and downstream PK	DENV/JEV-group constraints	yes
DENV4	NC_002640.1	387	**114**,115	gcugu	gcagc	14	no/yes	4 nt in between downstream gcagc and stem base of 3' DB	DENV/JEV-group constraints	yes
			205	gcugu	acagc	9	no/yes	9 nt in between end of structure and downstream PK	DENV/JEV-group constraints	yes
Japanese encephalitis	GQ304752.1	585	**313**,314	ugca	ugcg	11	no/no	downstream ugcg starts at 2nd base pair of 3' DB stem	DENV/JEV-group constraints	yes
			394,395	gcugu	acagc	9	yes/yes	9 nt in between end of structure and downstream PK	DENV/JEV-group constraints	yes
Usutu	NC_006551.1	669	389	gaug	cguu	20	no/no	downstream cguu lies below apical loop of 3' DB	DENV/JEV-group constraints	*maybe*
			468,469	gcug*u*	*a*cagc	16	yes/yes	16 nt in between end of structure and downstream PK	DENV/JEV-group constraints	yes

DENV/JEV-group constraints: g-c,g-c closing mb-loop, c-g,c-g after mb-loop, acuag in mb-loop, right hairpin with g-c,g-c,u-a,u-g,agugga

Figure 30 Extended TR.2 results



TR.3

strain	accession	length including stop codon and 3' dinucleotid.	matched positions relative to 1st nucleotide of stop codon	apical loop PK sequence (*in case of multiple hits, wrt bold*)	downstream PK sequence	number of nt (in between) end of structure and PK sequences	unique PK sequences	confirmation of relative PK positions wrt Villordo	conserved sequences or base-pairs wrt Villordo	verified ok
Tembusu	JF895923.2	660	**390**,391	ggu	acc	19	no/no		JEV-group constraints	yes
			466,467	gcugu	acagc	19	no/no		JEV-group constraints	yes
Bagaza	AY632545.2	731	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Ilheus	AY632539.4	391	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
			197,198,199	gcugu	acagc	18	yes/no		JEV-group constraints	yes
Rocio	MF461639.1	427	158	N/A	N/A	N/A	N/A	N/A	JEV-group constraints	*maybe*
			232,233,234	gcugu	acagc	20	yes/no		JEV-group constraints	yes
St. Louis encephalitis	NC_007580.2	740	478	gug	cgc	21	no/no		JEV-group constraints	*maybe*
			549,550,551	gcugu	acagc	15	no/no		JEV-group constraints	yes
Alfuy	AY898809.1	565	284	uguu	aaca	20	no/no		JEV-group constraints	yes
			363,364,365	gcugu	acagc	23	yes/yes		JEV-group constraints	yes
Murray Valley ence.	NC_000943.1	617	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
			420,421	gcugu	acagc	16	yes/yes		JEV-group constraints	yes
Kokobera	AY632541.4	561	292	aga	ucu	20	no/no		JEV-group constraints	*maybe*
			371	gcugu	acagc	20	yes/yes		JEV-group constraints	yes
Yellow Fever	NC_002031.1	511	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
			312	ucccac	guggga	12	no/yes		YFV-group constraints	yes
Zika	NC_012532.1	431	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
			240,241,242	gcugu	acagc	9	no/yes		JEV-group constraints	yes

JEV-group constraints: g-c closing mb-loop; g-c,g-c,u-a,u-g,ag*gga; acuag unstructred residues in mb-loop; c-g,c-g after mb-loop; gg after bulge loop

YFV-group constraints: g-c.g-c.a-u.g-c closing mb-loop; ucuaga in mb-loop; g-c,g-c,u-a,u-g,agagga; c-g,c-g,u-g,c-g after mb-loop; c and partial agaa in internal loop; cg in apical loop

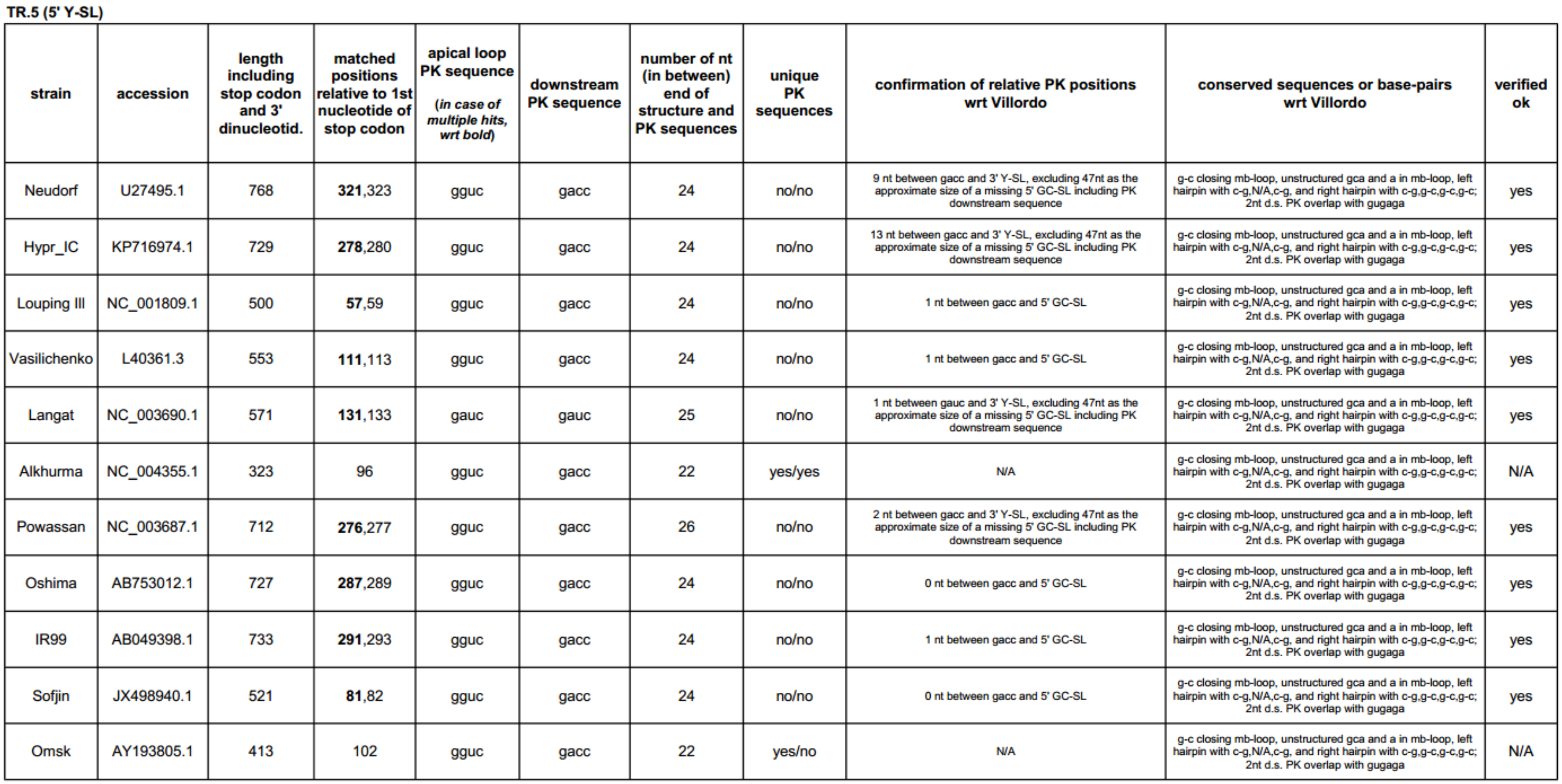
Figure 31 Extended TR.3 results

TR.4

strain	accession	length including stop codon and 3' dinucleotid.	matched positions relative to 1st nucleotide of stop codon	apical loop PK sequence	downstream PK sequence	number of nt (in between) end of structure and PK sequences	unique PK sequences	confirmation of relative PK positions wrt Villordo	conserved sequences or base-pairs wrt Villordo	verified ok
Kunjin	L24512.1	627	106	guugagu	acucaac	2	yes/yes		JEV-group constraints	yes
			266	gcg	ugc	6	no/no		JEV-group constraints	yes
385-99	EF571854.1	634	113	guugagu	acucaac	2	yes/yes		JEV-group constraints	yes
			273	gcg	ugc	6	no/no		JEV-group constraints	yes
PT6.16	AJ965626.2	609	113	guugagu	acucaac	2	yes/yes		JEV-group constraints	yes
			273	gcg	ugc	6	no/no		JEV-group constraints	yes
DENV1	NC 001477.1	465	57	cgg	ccg	6	no/no		DENV-group constraints	yes
			130	gagc	gcuc	2	no/yes		DENV-group constraints	yes
DENV2	NC 001474.2	454	37	cgg	ccg	6	no/no		DENV-group constraints	yes
			110	gagu	gcuc	2	yes/yes		DENV-group constraints	yes
DENV3	NC_001475.2	443	33	cgg	ccg	6	no/no		DENV-group constraints	yes
			109	gagc	gcuc	2	no/yes		DENV-group constraints	yes
DENV4	NC_002640.1	387	37	gag	cuc	3	no/no		DENV-group constraints with no unstructured c above mb-loop	yes
			N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Japanese enceph.	GQ304752.1	585	70	acuggg	cucagu	3	no/no		JEV-group constraints	*maybe*
			230	gcu	agc	3	no/no		JEV-group constraints	yes
Usutu	NC_006551.1	668	150	guugagu	acucaac	2	yes/yes		JEV-group constraints	yes
			309	gcc	ggc	3	no/no		JEV-group constraints	yes
Tembusu	JF895923.2	660	141	guugga	uccaac	3	yes/yes		JEV-group constraints	yes
			299	gcg	cgc	11	no/no		JEV-group constraints with g-c i.o. a-u below mb-loop and u-g instead of c-g above mb-loop	*maybe*
Bagaza	AY632545.2	731	286	guuggau	guccaac	2	yes/yes		JEV-group constraints	yes
			448	gca	ugc	10	no/no		JEV-group constraints	yes
Ilheus	AY632539.4	391	34	gcuug	caagc	3	yes/no		JEV-group constraints	yes
			N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Rocio	MF461639.1	427	70	gcaug	caugc	3	yes/yes		JEV-group constraints	yes
			N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
St. Louis encephalitis	NC_007580.2	740	233	gucaggu	accuggc	2	yes/yes		JEV-group constraints	yes
			N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Alfuy	AY898809.1	565	44	guuggu	accaac	2	yes/yes		JEV-group constraints	yes
			204	gcc	ggc	3	no/no		JEV-group constraints	yes
Murray Valley ence.	NC_000943.1	617	100	gguuggu	accaacc	2	yes/yes		JEV-group constraints	yes
			262	gcc	ggc	3	no/no		JEV-group constraints	yes
Kokobera	AY632541.4	561	41	ccg	cgg	4	no/no		JEV-group constraints	*maybe*
			203	ggu	acu	5	no/no		JEV-group constraints	*maybe*
Yellow Fever	NC_002031.1	511	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
			N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Zika	NC_012532.1	431	23	uuugggg	ccccagg	6	yes/yes		JEV-group constraints	yes
			107	cgg	cug	18	no/no		JEV-group constraints with u-a i.o. c-g above mb-loop	*maybe*

DENV-group constraints:	g-c,g-c,a-u,c-g below mb-loop; 5' gu and 3' agc..cc unstructured nt at dangling ends; g-c,c-g,c,a-u above mb-loop; c unstructured in mb-loop
JEV-group constraints:	g-c,a-u,c-g below mb-loop; gu unstructured nt at 5' dangling end; g-c,c-g,a-u,c-g above mb-loop; c unstructured in mb-loop
YFV-group constraints:	c-g,c-g,g-c,a-u,c-g below mb-loop; gu unstructured nt at 5' dangling end; g-c,c-g,c,a-u above mb-loop; c unstructured in mb-loop; u-g,c-g,g-c closing apical loop

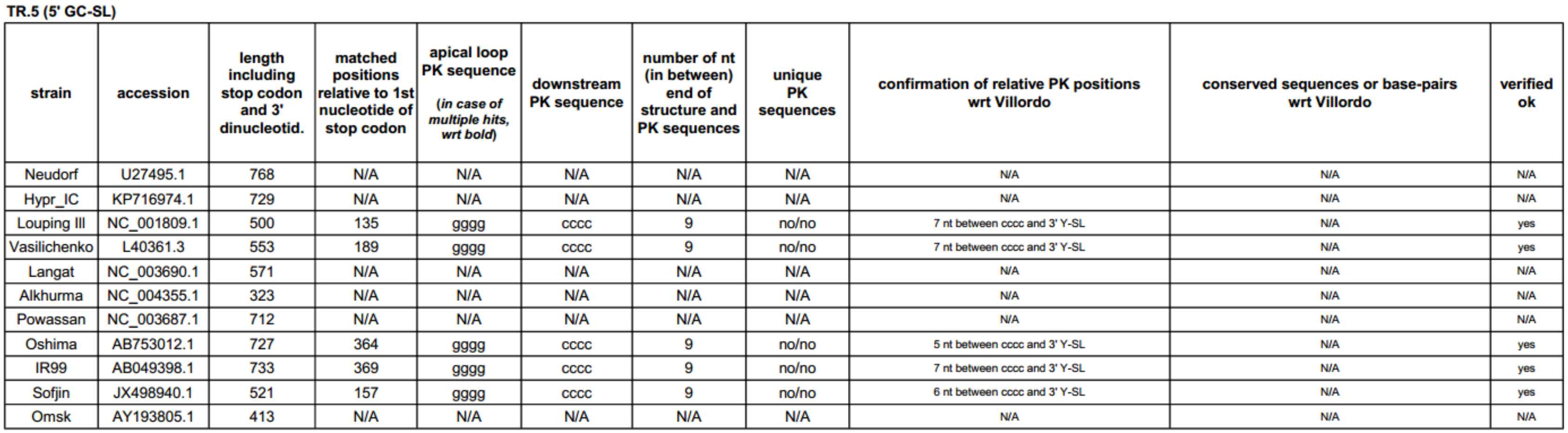
Figure 32 Extended TR.4 results



TR.5 (5' Y-SL)

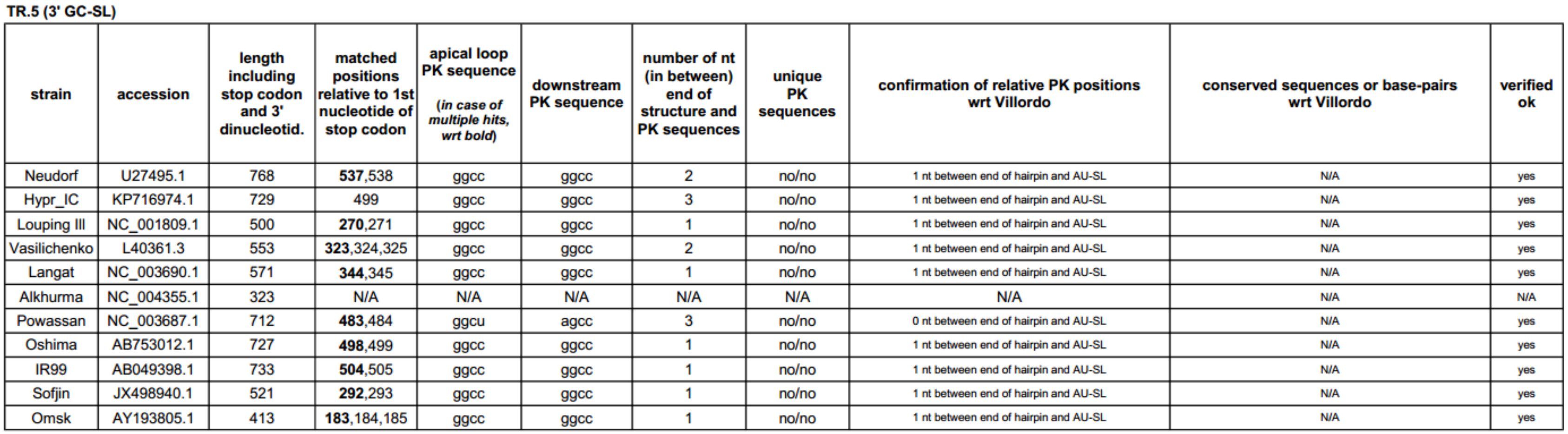
strain	accession	length including stop codon and 3' dinucleotid.	matched positions relative to 1st nucleotide of stop codon	apical loop PK sequence (*in case of multiple hits, wrt bold*)	downstream PK sequence	number of nt (in between) end of structure and PK sequences	unique PK sequences	confirmation of relative PK positions wrt Villordo	conserved sequences or base-pairs wrt Villordo	verified ok
Neudorf	U27495.1	768	**321**,323	gguc	gacc	24	no/no	9 nt between gacc and 3' Y-SL, excluding 47nt as the approximate size of a missing 5' GC-SL including PK downstream sequence	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Hypr_IC	KP716974.1	729	**278**,280	gguc	gacc	24	no/no	13 nt between gacc and 3' Y-SL, excluding 47nt as the approximate size of a missing 5' GC-SL including PK downstream sequence	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Louping III	NC_001809.1	500	**57**,59	gguc	gacc	24	no/no	1 nt between gacc and 5' GC-SL	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Vasilichenko	L40361.3	553	**111**,113	gguc	gacc	24	no/no	1 nt between gacc and 5' GC-SL	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Langat	NC_003690.1	571	**131**,133	gauc	gauc	25	no/no	1 nt between gauc and 3' Y-SL, excluding 47nt as the approximate size of a missing 5' GC-SL including PK downstream sequence	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Alkhurma	NC_004355.1	323	96	gguc	gacc	22	yes/yes	N/A	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	N/A
Powassan	NC_003687.1	712	**276**,277	gguc	gacc	26	no/no	2 nt between gacc and 3' Y-SL, excluding 47nt as the approximate size of a missing 5' GC-SL including PK downstream sequence	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Oshima	AB753012.1	727	**287**,289	gguc	gacc	24	no/no	0 nt between gacc and 5' GC-SL	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
IR99	AB049398.1	733	**291**,293	gguc	gacc	24	no/no	1 nt between gacc and 5' GC-SL	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Sofjin	JX498940.1	521	**81**,82	gguc	gacc	24	no/no	0 nt between gacc and 5' GC-SL	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Omsk	AY193805.1	413	102	gguc	gacc	22	yes/no	N/A	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	N/A

Figure 33 Extended TR.5 results (5' Y-SL)



TR.5 (5' GC-SL)

strain	accession	length including stop codon and 3' dinucleotid.	matched positions relative to 1st nucleotide of stop codon	apical loop PK sequence (*in case of multiple hits, wrt bold*)	downstream PK sequence	number of nt (in between) end of structure and PK sequences	unique PK sequences	confirmation of relative PK positions wrt Villordo	conserved sequences or base-pairs wrt Villordo	verified ok
Neudorf	U27495.1	768	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Hypr_IC	KP716974.1	729	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Louping III	NC_001809.1	500	135	gggg	cccc	9	no/no	7 nt between cccc and 3' Y-SL	N/A	yes
Vasilichenko	L40361.3	553	189	gggg	cccc	9	no/no	7 nt between cccc and 3' Y-SL	N/A	yes
Langat	NC_003690.1	571	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Alkhurma	NC_004355.1	323	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Powassan	NC_003687.1	712	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Oshima	AB753012.1	727	364	gggg	cccc	9	no/no	5 nt between cccc and 3' Y-SL	N/A	yes
IR99	AB049398.1	733	369	gggg	cccc	9	no/no	7 nt between cccc and 3' Y-SL	N/A	yes
Sofjin	JX498940.1	521	157	gggg	cccc	9	no/no	6 nt between cccc and 3' Y-SL	N/A	yes
Omsk	AY193805.1	413	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A



TR.5 (3' GC-SL)

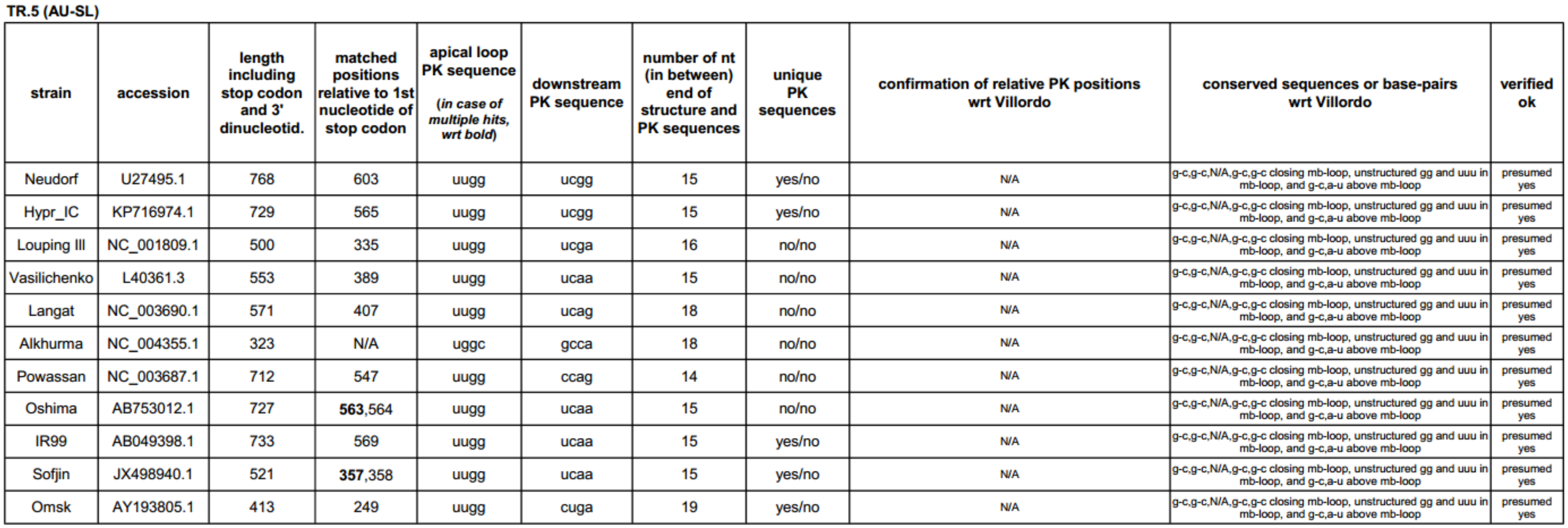
strain	accession	length including stop codon and 3' dinucleotid.	matched positions relative to 1st nucleotide of stop codon	apical loop PK sequence (*in case of multiple hits, wrt bold*)	downstream PK sequence	number of nt (in between) end of structure and PK sequences	unique PK sequences	confirmation of relative PK positions wrt Villordo	conserved sequences or base-pairs wrt Villordo	verified ok
Neudorf	U27495.1	768	**537**,538	ggcc	ggcc	2	no/no	1 nt between end of hairpin and AU-SL	N/A	yes
Hypr_IC	KP716974.1	729	499	ggcc	ggcc	3	no/no	1 nt between end of hairpin and AU-SL	N/A	yes
Louping III	NC_001809.1	500	**270**,271	ggcc	ggcc	1	no/no	1 nt between end of hairpin and AU-SL	N/A	yes
Vasilichenko	L40361.3	553	**323**,324,325	ggcc	ggcc	2	no/no	1 nt between end of hairpin and AU-SL	N/A	yes
Langat	NC_003690.1	571	**344**,345	ggcc	ggcc	1	no/no	1 nt between end of hairpin and AU-SL	N/A	yes
Alkhurma	NC_004355.1	323	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Powassan	NC_003687.1	712	**483**,484	ggcu	agcc	3	no/no	0 nt between end of hairpin and AU-SL	N/A	yes
Oshima	AB753012.1	727	**498**,499	ggcc	ggcc	1	no/no	1 nt between end of hairpin and AU-SL	N/A	yes
IR99	AB049398.1	733	**504**,505	ggcc	ggcc	1	no/no	1 nt between end of hairpin and AU-SL	N/A	yes
Sofjin	JX498940.1	521	**292**,293	ggcc	ggcc	1	no/no	1 nt between end of hairpin and AU-SL	N/A	yes
Omsk	AY193805.1	413	**183**,184,185	ggcc	ggcc	1	no/no	1 nt between end of hairpin and AU-SL	N/A	yes

Figure 34 Extended TR.5 results (5' and 3' G-SL

TR.5 (3' Y-SL)

strain	accession	length including stop codon and 3' dinucleotid.	matched positions relative to 1st nucleotide of stop codon	apical loop PK sequence *(in case of multiple hits, wrt bold)*	downstream PK sequence	number of nt (in between) end of structure and PK sequences	unique PK sequences	confirmation of relative PK positions wrt Villordo	conserved sequences or base-pairs wrt Villordo	verified ok
Neudorf	U27495.1	768	**456**,457	gguc	gacc	23	no/no	0 nt in between gacc and the 3' GC-SL	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Hypr_IC	KP716974.1	729	**417**,418	gguc	gacc	23	no/no	1 nt in between gacc and the 3' GC-SL	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Louping III	NC_001809.1	500	**189**,190	gguc	gacc	22	no/no	0 nt in between gacc and the 3' GC-SL	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Vasilichenko	L40361.3	553	**243**,244	gguc	gacc	22	no/no	-1 nt in between gacc and the 3' GC-SL	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Langat	NC_003690.1	571	**262**,263	gguc	gacc	23	no/no	0 nt in between gacc and the 3' GC-SL	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Alkhurma	NC_004355.1	323	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
Powassan	NC_003687.1	712	410	gguc	gacc	23	no/no	0 nt in between gacc and the 3' GC-SL	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Oshima	AB753012.1	727	**416**,417,418	gguc	gacc	23	no/no	0 nt in between gacc and the 3' GC-SL	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,u-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
IR99	AB049398.1	733	**423**,424	gguc	gacc	22	no/no	0 nt in between gacc and the 3' GC-SL	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,c-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Sofjin	JX498940.1	521	**210**,211,212	gguc	gacc	22	no/no	-1 nt in between gacc and the 3' GC-SL	g-c closing mb-loop, unstructured gca and a in mb-loop, left hairpin with c-g,N/A,u-g, and right hairpin with c-g,g-c,g-c,g-c; 2nt d.s. PK overlap with gugaga	yes
Omsk	AY193805.1	413	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Figure 35 Extended TR.5 results (3' Y-SL)



TR.5 (AU-SL)

strain	accession	length including stop codon and 3' dinucleotid.	matched positions relative to 1st nucleotide of stop codon	apical loop PK sequence (*in case of multiple hits, wrt bold*)	downstream PK sequence	number of nt (in between) end of structure and PK sequences	unique PK sequences	confirmation of relative PK positions wrt Villordo	conserved sequences or base-pairs wrt Villordo	verified ok
Neudorf	U27495.1	768	603	uugg	ucgg	15	yes/no	N/A	g-c,g-c,N/A,g-c,g-c closing mb-loop, unstructured gg and uuu in mb-loop, and g-c,a-u above mb-loop	presumed yes
Hypr_IC	KP716974.1	729	565	uugg	ucgg	15	yes/no	N/A	g-c,g-c,N/A,g-c,g-c closing mb-loop, unstructured gg and uuu in mb-loop, and g-c,a-u above mb-loop	presumed yes
Louping III	NC_001809.1	500	335	uugg	ucga	16	no/no	N/A	g-c,g-c,N/A,g-c,g-c closing mb-loop, unstructured gg and uuu in mb-loop, and g-c,a-u above mb-loop	presumed yes
Vasilichenko	L40361.3	553	389	uugg	ucaa	15	no/no	N/A	g-c,g-c,N/A,g-c,g-c closing mb-loop, unstructured gg and uuu in mb-loop, and g-c,a-u above mb-loop	presumed yes
Langat	NC_003690.1	571	407	uugg	ucag	18	no/no	N/A	g-c,g-c,N/A,g-c,g-c closing mb-loop, unstructured gg and uuu in mb-loop, and g-c,a-u above mb-loop	presumed yes
Alkhurma	NC_004355.1	323	N/A	uggc	gcca	18	no/no	N/A	g-c,g-c,N/A,g-c,g-c closing mb-loop, unstructured gg and uuu in mb-loop, and g-c,a-u above mb-loop	presumed yes
Powassan	NC_003687.1	712	547	uugg	ccag	14	no/no	N/A	g-c,g-c,N/A,g-c,g-c closing mb-loop, unstructured gg and uuu in mb-loop, and g-c,a-u above mb-loop	presumed yes
Oshima	AB753012.1	727	**563**,564	uugg	ucaa	15	no/no	N/A	g-c,g-c,N/A,g-c,g-c closing mb-loop, unstructured gg and uuu in mb-loop, and g-c,a-u above mb-loop	presumed yes
IR99	AB049398.1	733	569	uugg	ucaa	15	yes/no	N/A	g-c,g-c,N/A,g-c,g-c closing mb-loop, unstructured gg and uuu in mb-loop, and g-c,a-u above mb-loop	presumed yes
Sofjin	JX498940.1	521	**357**,358	uugg	ucaa	15	yes/no	N/A	g-c,g-c,N/A,g-c,g-c closing mb-loop, unstructured gg and uuu in mb-loop, and g-c,a-u above mb-loop	presumed yes
Omsk	AY193805.1	413	249	uugg	cuga	19	yes/no	N/A	g-c,g-c,N/A,g-c,g-c closing mb-loop, unstructured gg and uuu in mb-loop, and g-c,a-u above mb-loop	presumed yes

Figure 36 Extended TR.5 results (AU-SL)

B.5 Approximate stacking energy for Test Run 3 to 5 hits

Figures 37 through **39** display approximate base stacking energies for all hits returned in test runs 3 through 5, respectively. In each of the plots below, template sequence true and false positive hits are represented by solid blue and red plot markers, respectively, whereas test sequence true and false positive hits are marked by hollow blue and red plot markers. Sequence names run along the x-axis, with stacking energy shown in the y-axis. The three figures provided give an early indication that suitable *energy thresholding* parameters may be adopted to separate true hits from false positives, across the different test scenarios.

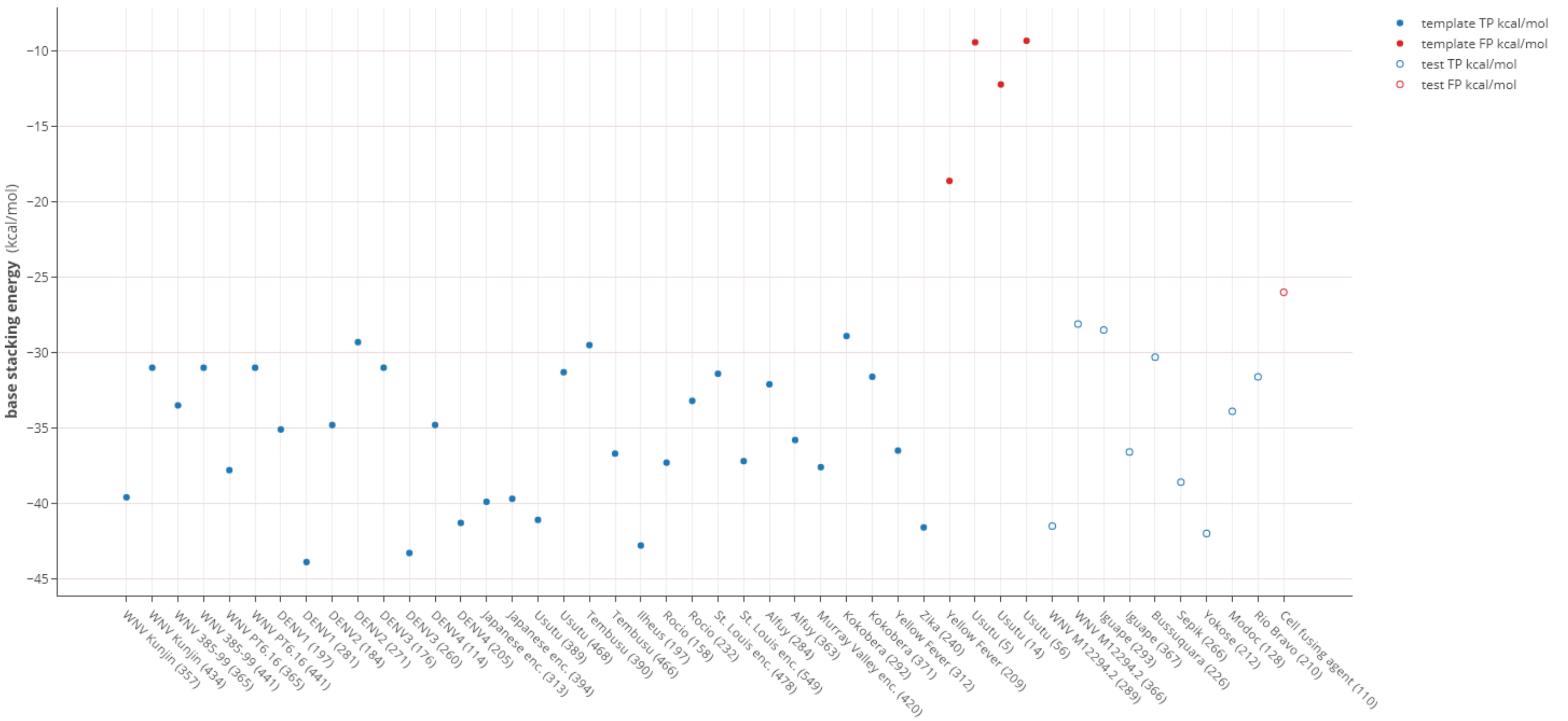


Figure 37 Base stacking energy for Test Run 3

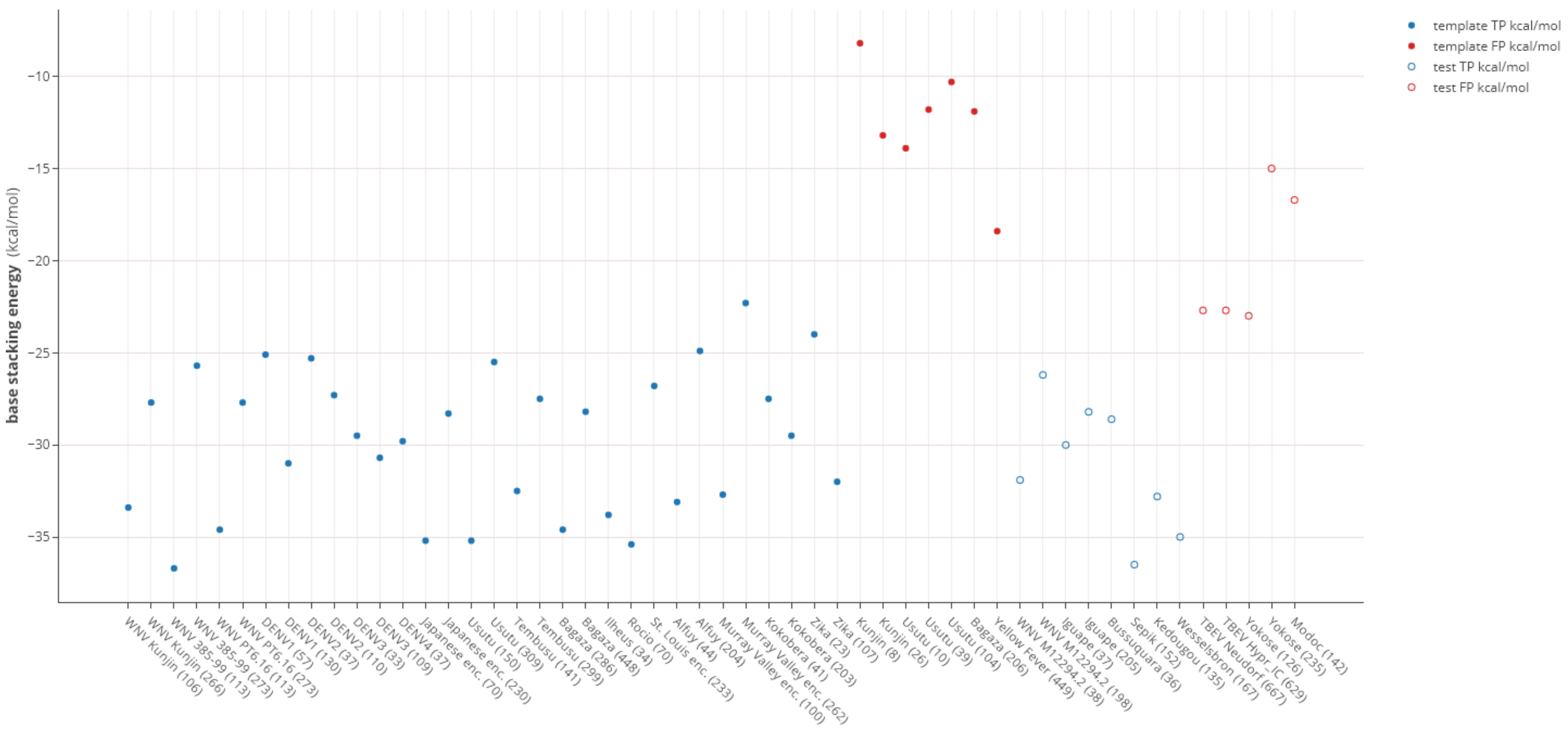


Figure 38 Base stacking energy for Test Run 4

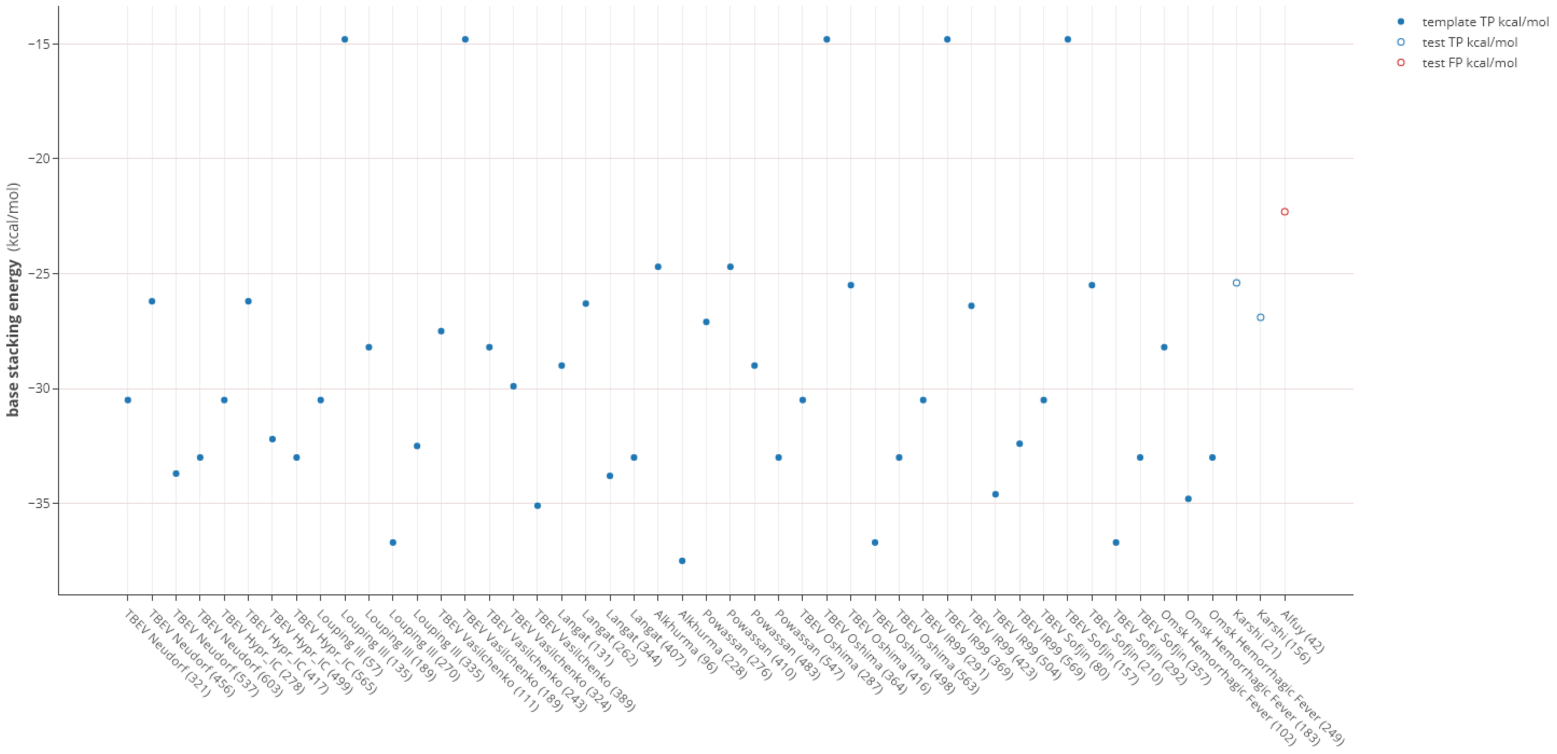


Figure 39 Base stacking energy for Test Run 5

References

Alvarez DE, Lodeiro MF, Luduena SJ, Pietrasanta LI & Gamarnik AV (2005). Long-range RNA-RNA interactions circularize the dengue virus genome. *J. Virol.* 79(11):6631–6643.

Bennett S et al. (2009). Epidemic dynamics revealed in dengue evolution. *Mol. Biol. Evol.* 27(4):811–818.

Bidet K & Garcia-Blanco MA (2014). Flaviviral RNAs: weapons and targets in the war between virus and host. *Biochem. J.* 462(2):215–230.

Brinton MA & Dispoto JH (1988). Sequence and secondary structure analysis of the 5'-terminal region of flavivirus genome RNA. *Virology.* 162(2):290–299.

Chapman EG et al. (2014). The Structural Basis of Pathogenic Subgenomic Flavivirus RNA (sfRNA) Production. *Science.* 344(6181):307–310.

Clyde K, Barrera J & Harris E (2008). The capsid-coding region hairpin element (cHP) is a critical determinant of dengue virus and West Nile virus RNA synthesis. *Virology.* 379(2):314–323.

Clyde K & Harris E (2006). RNA secondary structure in the coding region of dengue virus type 2 directs translation start codon selection and is required for viral replication. *J. Virol.* 80(5):2170–2182.

Davis WG, Basu M, Elrod EJ, Germann MW & Brinton MA (2013). Identification of cis-acting nucleotides and a structural feature in West Nile virus 3'-terminus RNA that facilitate viral minus strand RNA synthesis. *J. Virol.* 87(13):7622–7636.

Dong H et al. (2014). Flavivirus RNA methylation. *J. Gen. Virol.* 95(4):763–778.

Fernández-Sanlés A, Ríos-Marco P, Romero-López C & Berzal-Herranz A (2017). Functional Information Stored in the Conserved Structural RNA Domains of Flavivirus Genomes. *Front. Microbiol.* 8:546.

Filomatori CV et al. (2006). A 5' RNA element promotes dengue virus RNA synthesis on a circular genome. *Genes & development.* 20(16):2238–2249.

Filomatori CV et al. (2017). Dengue virus genomic variation associated with mosquito adaptation defines the pattern of viral non-coding RNAs and fitness in Hum. Cells. *PLoS Pathog.* 13(3):e1006265.

Funk A et al. (2010). RNA structures required for production of subgenomic flavivirus RNA. *J. Virol.* 84(21):11407–11417.

Gebhard LG, Filomatori CV & Gamarnik AV (2011). Functional RNA elements in the dengue virus genome. *Viruses.* 3(9):1739–1756.

Georg S, Thomas M & Paul A (n.d.). ARE-mRNA degradation requires the 5'-3' decay pathway. *EMBO Rep*. 7(1):72–77.

Giegerich R, Voß B & Rehmsmeier M (2004). Abstract shapes of RNA. *Nucleic Acids Res*. 32(16):4843–4851.

Göertz G et al. (2016). Noncoding subgenomic flavivirus RNA is processed by the mosquito RNA interference machinery and determines West Nile virus transmission by Culex pipiens mosquitoes. *J. Virol*. 90(22):10145–10159.

Gritsun T & Gould E (2006). Origin and evolution of 3' UTR of flaviviruses: long direct repeats as a basis for the formation of secondary structures and their significance for virus transmission. *Advances in Virus Res*. 69:203–248.

Gritsun T et al. (1997). Complete sequence of two tick-borne flaviviruses isolated from Siberia and the UK: analysis and significance of the 5' and 3'-UTRs. *Virus Res*. 49(1):27–39.

Jones CI, Zabolotskaya MV & Newbury SF (2012). The 5'-3' exoribonuclease XRN1/Pacman and its functions in cellular processes and development. *WIRES: RNA*. 3(4):455–468.

Kieft JS, Rabe JL & Chapman EG (2015). New hypotheses derived from the structure of a flaviviral Xrn1-resistant RNA: Conservation, folding, and host adaptation. *RNA Biol*. 12(11):1169–1177.

Liu Y, Liu H, Zou J, Zhang B & Yuan Z (2014). Dengue virus subgenomic RNA induces apoptosis through the Bcl-2-mediated PI3k/Akt signaling pathway. *Virology*. 448:15–25.

Lobo F et al. (2009). Virus-Host Coevolution: Common Patterns of Nucleotide Motif Usage in Flaviviridae and Their Hosts. *PloS one*. 4:e6282.

Lodeiro MF, Filomatori CV & Gamarnik AV (2009). Structural and functional studies of the promoter element for dengue virus RNA replication. *J. Virol*. 83(2):993–1008.

MacFadden A et al. (2018). Mechanism and structural diversity of exoribonuclease-resistant RNA structures in flaviviral RNAs. *Nat. Commun*. 9(1):119.

Mathews DH (2004). Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*. 10(8):1178–1190.

Mathews DH, Moss WN & Turner DH (2010). Folding and finding RNA secondary structure. *Cold Spring Harbor Perspect. Biol*. 2(12):a003665.

Mohan PM & Padmanabhan R (1991). Detection of stable secondary structure at the 3' terminus of dengue virus type 2 RNA. *Gene*. 108(2):185–191.

Moon SL et al. (2012). A noncoding RNA produced by arthropod-borne flaviviruses inhibits the cellular exoribonuclease XRN1 and alters host mRNA stability. *RNA*. 18(11):2029–2040.

Moon SL et al. (2015). XRN1 stalling in the 5' UTR of Hepatitis C virus and Bovine Viral Diarrhea virus is associated with dysregulated host mRNA stability. *PLoS Pathog*. 11(3):e1004708.

Nawrocki EP, Kolbe DL & Eddy SR (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics*. 25(10):1335–1337.

Ng WC, Soto-Acosta R, Bradrick SS, Garcia-Blanco MA & Ooi EE (2017). The 5' and 3' Untranslated Regions of the Flaviviral Genome. *Viruses*. 9(6):137.

Pijlman GP et al. (2008). A Highly Structured, Nuclease-Resistant, Noncoding RNA Produced by Flaviviruses Is Required for Pathogenicity. *Cell Host & Microbe*. 4(6):579–591.

Polacek C, Friebe P & Harris E (2009). Poly (A)-binding protein binds to the non-polyadenylated 3' untranslated region of dengue virus and modulates translation efficiency. *J. Gen. Virol*. 90(3):687–692.

Roby JA, Pijlman GP, Wilusz J & Khromykh AA (2014). Noncoding subgenomic flavivirus RNA: multiple functions in West Nile virus pathogenesis and modulation of host responses. *Viruses*. 6(2):404–427.

Schuessler A et al. (2012). West Nile virus noncoding subgenomic RNA contributes to viral evasion of the type I interferon-mediated antiviral response. *J. Virol*. 86(10):5708–5718.

Schuster P (1995). How to search for RNA structures theoretical concepts in evolutionary biotechnology. *J. Biotechnol*. 41(2-3):239–257.

Silva PA, Pereira CF, Dalebout TJ, Spaan WJ & Bredenbeek PJ (2010). An RNA pseudoknot is required for production of yellow fever virus subgenomic RNA by the host nuclease XRN1. *J. Virol*. 84(21):11395–11406.

Simmonds P et al. (2017). ICTV virus taxonomy profile: Flaviviridae. *J. Gen. Virol*. 98(1):2–3.

Villordo SM, Carballeda JM, Filomatori CV & Gamarnik AV (2016). RNA Structure Duplications and Flavivirus Host Adaptation. *Trends Microbiol*. 24(4):270–283.

Villordo SM, Filomatori CV, Sánchez-Vargas I, Blair CD & Gamarnik AV (2015). Dengue virus RNA structure specialization facilitates host adaptation. *PLoS Pathog*. 11(1):e1004604.

Vlachakis D, Koumandou VL & Kossida S (2013). A holistic evolutionary and structural study of *flaviviridae* provides insights into the function and inhibition of HCV helicase. *PeerJ*. 1:e74.

Wengler G & Castle E (1986). Analysis of structural properties which possibly are characteristic for the 3'-terminal sequence of the genome RNA of flaviviruses. *J. Gen. Virol*. 67(6):1183–1188.

Zuker M (2003). Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* 31(13):3406–3415.

Zuker M & Sankoff D (1984). RNA secondary structures and their prediction. *Bull. Math. Biol.* 46(4):591–621.