Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt
Technische Universität München

Viral mRNAs: evolution and structure-function relation

Michael Kiening

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften (Dr. rer. nat.)

genehmigten Dissertation.

Vorsitzender:

Prof. Dr. Fabian J. Theis

Prüfende der Dissertation:

1. Prof. Dr. Dmitrij Frishman
2. Ass.-Prof. Dr. Alexandre Goultiaev

Die Dissertation wurde am 01.07.2019 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 19.09.2019 angenommen.

Abstract

Viruses can cause enormous crop losses in agriculture and are responsible for severe infections, that can even be lethal. The continuous discovery of new viruses, and the lack of treatments for some of their most dangerous representatives like arenaviruses causing viral meningitis or HIV, highlights the importance enhancing our understanding of the viral replication cycle and its mechanisms. RNA structure has been shown to have a significant impact on the replication cycle of viruses.

Within the negative sense RNA viruses, the ambisense viruses are a special group of pathogens, that infect plants, animals and humans, leading to severe diseases like the hemorrhagic fever. Their segmented genome consists of at least one ambisense RNA segment that exhibits the ambisense expression strategy, i.e. it harbors two genes, which are oriented in opposite reading directions and are separated by a non-coding intergenic region. Within this intergenic region, some viruses contain a secondary structure element which is suspected to be involved in transcription termination. Within the first publication of this thesis we sought to give a thorough overview of conserved secondary structures in the intergenic regions of ambisense mRNAs, and give insight in their evolutionary relations.

While functional RNA secondary structures in non-coding regions have been studied excessively, there are only few examples of structures located in coding regions. Besides well characterized non-coding RNA structures, it has also been shown that structures in coding regions play an important role in regulatory processes and within the viral life cycle. The possible number of stable structural folds a sequence can adopt grows exponentially with increasing sequence length. This makes structure prediction of long molecules, such as mRNA sequences difficult. The second publication of this thesis covers the creation of a possibly complete census of conserved secondary structures in the coding regions of viral mRNAs. A structure and function prediction pipeline was applied to orthologous groups of viruses which were subsequently split the orthologous groups into structurally homogenous subgroups, which are called subVOGs. The resulting compilation of conserved RNA structures was embedded in our newly created web database RNASIV (RNA structures in viruses), providing access to the data for future research.

Zusammenfassung

Viren können für enorme Ernteausfälle verantwortlich sein und schwere Infektionen auslösen, welche tödlich enden können. Die ständige Entdeckung neuer Viren sowie der Mangel an Behandlungsmöglichkeiten für einige der gefährlichsten Vertreter wie die Meningitis auslösenden Arenaviren oder HIV zeigen die Relevanz unser Verständnis des viralen Replikationszyklus und dessen Mechanismen zu verbessern. Es wurde gezeigt, dass RNA Strukturen einen signifikanten Einfluss auf den viralen Replikationsmechanismus haben.

Innerhalb der negative sense RNA Viren, bilden ambisense Viren eine besondere Gruppe welche Pflanzen, Tiere und auch Menschen infiziert, und zu schweren Krankheiten, wie zum Beispiel dem Hämorrhagischem Fieber führen kann. Das segmentierte Genom beinhaltet mindestens eine RNA, welche nach der ambisense Strategie exprimiert wird. Das bedeutet, dass diese mRNA zwei Gene in entgegengesetzter Leserichtung enthält, welche von einer nicht kodierenden, intergenischen Region getrennt sind. Es wurde für einige dieser Viren gezeigt, dass innerhalb der intergenischen Region ein stabiles Sekundärstrukturelement lokalisiert ist, welches im Verdacht steht an der Transkriptionsterminierung beteiligt zu sein. Mit dieser ersten Publikation wurde versucht, sowohl einen vollständigen Überblick über alle Sekundärstrukturen in intergenischen Regionen von ambisense Viren zu geben, als auch deren evolutionäre Beziehungen zu klären.

Während RNA Strukturen in nicht kodierenden Bereichen bereits intensiv untersucht wurden, sind nur einige Beispiele von funktionellen Strukturen in kodierenden Regionen bekannt. Dennoch konnte man neben der Vielzahl von funktionellen nicht kodierenden Strukturen zeigen, dass auch Strukturen in kodierenden Bereichen an wichtigen regulatorischen Prozessen und dem viralen Replikationszyklus beteiligt sind. Die Zahl der möglichen strukturellen Faltungen die ein RNA Molekül annehmen kann, wächst exponentiell mit steigender Sequenzlänge. Dies erschwert die Strukturvorhersage von langen Molekülen wie mRNAs enorm. Mit der zweiten Publikation dieser Arbeit wird ein möglicherweise vollständiger Überblick über konservierte Sekundärstrukturen in viralen mRNAs gegeben. Unter Anwendung einer Strukturvorhersagepipeline wurden orthologe Gruppen von Viren in strukturell homogene Untergruppen, genannt subVOGs aufgeteilt.

Die daraus resultierende Sammlung konservierter RNA Strukturen wurde in die explizit dafür erstellte Webdatenbank RNASIV (RNA structures in viruses) eingepflegt, und steht somit für zukünftige Forschungszwecke zur Verfügung.

Acknowledgements

The years I was part of the bioinformatics group in Weihenstephan probably were the most challenging and at the same time life enhancing years in my life so far. Besides working on this thesis I have been involved in various scientific projects, teaching students the basics of bioinformatics, supervising bachelor and master thesis and many further administrative tasks. I attended conferences and travelled to collaborators and had the chance to meet various interesting people. Everyone who has gone through the publication process during a doctoral study or a similarly intense task knows how demanding this can be. I am very glad that I had and still have so many family members and friends that helped me to keep up the motivation during the whole time as a doctoral candidate.

First of all I want to thank my supervisor Prof. Dmitrij Frishman for leading me through the projects of my thesis. I really appreciate that he always has time for his students, whether in the lab, from Russia or Australia. He brought in his enormous scientific knowledge in the many fruitful discussions we had and accompanied my doctoral degree.

I also want to thank Ass.-Prof. Dr. Alexandre Goultiaev for spending his valuable time to be my second corrector, and Prof. Dr. Jan Baumbach who kindly agreed to be the chair of the examination committee.

I want to thank all of my former and present colleagues for the fruitful scientific discussions and of course also for the various off-topic meetings on serious as well as hilarious topics. Thanks to Léonie Corry, Martina Rüttger and Roswita Weinbrunn for keeping the department running, and taking so many tasks off the shoulders of PhD students and other employees. Thanks to Drazen Jalsovec for maintaining all of our soft- and hardware, installing rarely used software and solving dependency issues. Special thanks goes to Peter Hönigschmid for being the best friend one can imagine from school to graduation, in the office and free time, in Munich or Barcelona. My gratitude also goes Jan Zaucha, not only for all the proof reading of my writing, but also for the nice talks on scientific, educational and life regarding topics. I also want to thank Stephan Breimann for the interesting and educating talks. Thanks also to Usman Saed, Xeynub asrar, Evans Kataka, Alec Steep and all the other awesome people I had the pleasure to work with during my graduation time. I'm proud that I could be your colleague.

Last but not least, the most important persons in my life. I want to thank my whole family for accompanying me and giving me this tremendous support. I would have not finished this thesis without you. My parents Hubert and Brigitte for standing by me and supporting me during my whole life. My sister for always having an open ear and for being the best sister on earth. My grand uncle Hans for his support during graduation. And finally my beloved wife Claudia for supporting me in my decision to graduate and during the whole graduation process, motivating me, always being there for me, and for raising our beloved child Max and our soon expected unborn child Luis together with me, with so much love, passion and a smile on her face.

Publications

The two entries written in bold are part of this thesis.

- Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., ... & Pandey, G. (2013). A large-scale evaluation of computational protein function prediction. *Nature methods*, *10*(3), 221.
- Hamp, T., Kassner, R., Seemayer, S., Vicedo, E., Schaefer, C., Achten, D., ... & Heron, M. (2013, February). Homology-based inference sets the bar high for protein function prediction. In *BMC bioinformatics* (Vol. 14, No. 3, p. S7). BioMed Central.
- **Kiening, M., Weber, F., & Frishman, D. (2017). Conserved RNA structures in the intergenic regions of ambisense viruses. *Scientific reports*, *7*(1), 16625.**
- **Kiening, M., Ochsenreiter, R., Hellinger, H. J., Rattei, T., Hofacker, I., & Frishman, D. (2019). Conserved Secondary Structures in Viral mRNAs. *Viruses*, *11*(5), 401.**

Contents

List of Figures

List of Tables

1. Introduction

The following introduction provides information about the biological and methodological background as well as its implications for the approaches used in the published articles.

1.1. Motivation

1.1.1. Viruses

Viruses are small pathogenic particles that infect a host organism and hijack its cellular machinery responsible for protein production in order to reproduce. They essentially consist of single stranded (ss) or double stranded (ds) genetic material surrounded by a protein coat. Depending on their employed mechanism of mRNA production, they are categorized into DNA, RNA and reverse transcribing viruses. At a second level they are further distinguished according to the type of genetic material they harbor (Table 1).

DNA viruses	ds DNA viruses
	ss DNA viruses
RNA viruses	ds RNA viruses
	positive sense ss RNA viruses
	negative sense ss RNA viruses
	circular ss RNA viruses
Retro/Reverse transcribing viruses	Retro/Reverse transcribing viruses

Table 1: Top level of viral categorization.

Apart from being responsible for various severe diseases, they also pose a significant threat to human welfare by causing livestock and crop shortfalls. That makes viruses and the inhibition of their replication cycle an interesting and important field of research. A very effective way of mitigating their spread is vaccination with killed or attenuated viral particles, leading to an immunization of a potential host against an infection with a certain virus (Fiore, Bridges, & Cox, 2009). Nevertheless, vaccines against some of the most dangerous viruses, such as the West Nile virus, are still unavailable.

1.1.1.1. Viral entry into a host cell

The first requirement for a successful infection is the entry into the host's cell. Depending on whether the virus is enveloped by an additional outer membrane or not, the strategy for intrusion can be different (Cohen, 2016). The two possible methods are receptor mediated fusion or hijacking the endocytotic pathway. Viruses employing receptor mediated fusion, rely on binding to a protein, which is present on the surface of the host cell membrane. The binding leads to a conformational change of the viruses' fusion protein, facilitating the entry of the viral genetic material into the host cell. In the second path, hijacking the endocytotic pathway, the virus is encapsidated inside an endosome. A low PH level of the endosome then induces the fusion. Enveloped viruses can enter a cell employing either method, while viruses without an envelope usually enter the cell via the endocytotic pathway (Thorley, McKeating, & Rappoport, 2010).

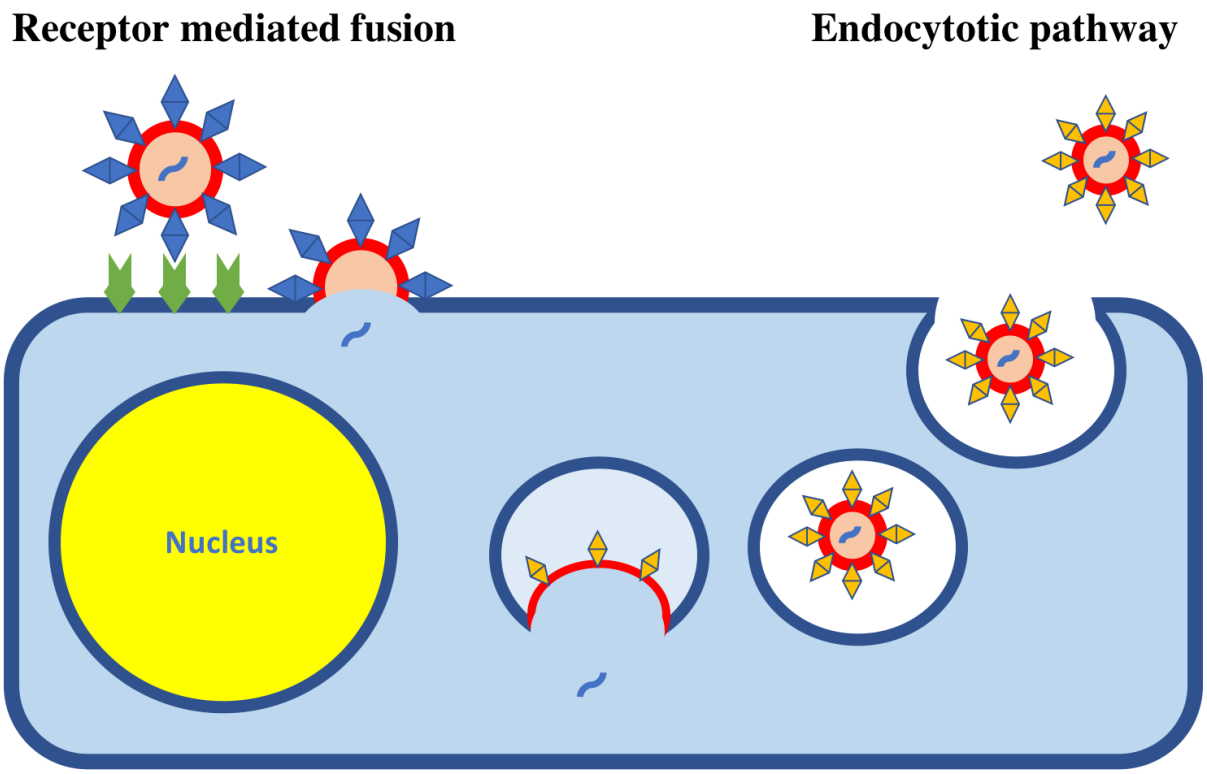


Figure 1: Visualization of the two main entry mechanisms of viruses into a host cell.

1.1.1.2. Viral exit strategies

Once the host cell has produced enough viral replicates and is no longer beneficial for the virus, the viral copies are released to infect further host cells and start a new replication cycle. In the following sections the three main mechanisms for exiting host cells are described briefly.

Budding

Budding means that a virus traverses the cell membrane, and gets enveloped into the hosts cell membrane in turn (Komarova, 2007). The budding mechanism is employed by enveloped viruses including the human immune deficiency (HIV) virus or Influenza (Pornillos, Garrus, & Sundquist, 2002). Budding does not result in immediate cell death, but as more viruses are budding through the membrane, it slowly dissolves and ultimately disintegrates. The released viruses can instantly infect other host cells.

Apoptosis

Apoptosis, the programmed cell death, is the last defence mechanism of a cell in case of a severe viral infection. The genetic material of the cell is fragmented, and the cell activates pathways that lead to its lysis, allowing it to be absorbed by macrophages. This mechanism is exploited primarily by non-enveloped viruses. Nevertheless, some enveloped viruses like HIV, use the process to either infect macrophages or to be transported into different host compartments (Stewart, Poon, Song, & Chen, 2000). In rare cases, apoptosis also leads to a quick dissolvement of the cell membrane, which immediately releases the viral replicates into the extracellular space (Bursch, Oberhammer, & Schulte-Hermann, 1992).

Exocytosis

This exit strategy is special, because it does not destroy the host cell. The virus hijacks the cell's transport system, and sends its progeny out of the cell in its own vesicles (Münz, 2017). Once the cell membrane is surpassed, the viral replicates are released in the extracellular space. This mechanism is mainly used by non-enveloped viruses, however, some enveloped viruses using this strategy exist (Olson & Grose, 1997).

1.1.2. Viral evolution

In evolutionary biology one of the most important tools are fossil records. Unfortunately, viruses do not leave such trails. To investigate viral evolution, we must rely on what we can learn from the extant viruses and their hosts. The fast evolution of viral genomes necessary to evade the constantly adapting immune systems of their hosts, and the large number of replicates each virus generation can have, make it difficult to rely on the same techniques used for other organisms to clarify their evolutionary relationships. At least the arms race of viruses and their hosts in adapting to each other resulted in a strong coevolutionary signal (Nasir, Kim, & Caetano-Anollés, 2012).

The origin of viruses is still an unsolved scientific question, for which three main theories have been proposed: i) virus first theory , ii) reduction hypothesis and iii) escape hypothesis (Nasir et al., 2012). The virus first theory claims that viruses existed before cellular life arose on earth, and even contributed to its emergence (Koonin, Senkevich, & Dolja, 2006; 2009). The fact that viruses can only replicate with a host organism questions this theory, and furthermore suggests that cellular organisms are essential for the emergence of viruses (Forterre, 2006). The reduction theory therefore suggests that viruses have evolved from parasitic organisms (Bândea, 1983). Finally, the escape hypothesis proposes that some genetic material of host organisms managed to escape the cellular control and acquired genes through horizontal gene transfer.

1.1.3. Taxonomy of viruses

The taxonomy of viruses is mainly handled by the International Committee on Taxonomy of Viruses (ICTV) (Lefkowitz et al., 2018). The viral taxonomy is defined in a hierarchical scheme, including naming conventions for each level in the hierarchy (Table 2). Only two levels within the hierarchy, the genus and subgenus, are mandatorily assigned to a virus, the other levels are optional. The taxonomic classification of viruses was published the first time in 1971 (MELNICK, 1971) for the first time and is updated very frequently in accordance with new findings. Therefore, it does not necessarily mirror the ground truth evolutionary context of viruses. Recent research showed that the relations between viruses are modelled more realistically by using protein sequence similarity rather than through taxonomic definitions (C. Chen et al., 2016). Furthermore it has been shown that RNA

structure can have an effect on the long term evolution of RNA sequences (Simmonds & Smith, 1999). Thus, RNA structure builds another layer of measurable relationship on top of the sequence conservation. Due to the high evolutionary pressure viruses are subjected to, it is believed that the evolutionary conservation of RNA structures preponderates RNA sequence conservation in many cases (Capriotti & Marti-Renom, 2010). In order to find new ways to stop viral infections from spreading, we need to enhance our understanding of the viral lifecycle and its regulatory mechanisms. Within this thesis we sought to investigate conserved potentially functional conserved secondary structures in viral mRNAs and give a possibly complete census.

Hierarchy level	Rank name	Name suffix	Mandatory / optional
0	Realm	-viria	Optional
1	Subrealm	-vira	Optional
2	Kingdom	-virae	Optional
3	Subkingdom	-virites	Optional
4	Phylum	-viricota	Optional
5	Subphylum	-viricotina	Optional
6	Class	-viricetes	Optional
7	Subclass	-viricetidae	Optional
8	Order	-virales	Optional
9	Suborder	-virineae	Optional
10	Family	-viridae	Optional
11	Subfamily	-virinae	Optional
12	Genus	-virus	Mandatory
13	Subgenus	-virus	Mandatory

Table 2: Hierarchical taxonomic classification scheme of the ICTV.

1.2. (m)RNA structure

1.2.1. (m)RNA structure and visualization

All forms of life make use of mRNA to encode proteins. The mRNA molecule consists of a sequence of the four ribonucleotides adenine (A), guanine (G), cytosine (C) and uracil (U). Complementary nucleotides (A-U and G-C) can form bonds within the chain and fold into complex secondary and tertiary structures. Additionally to the standard base pairs, non-canonical, less stable base pairs, e.g. G-U can be observed. The stability of an RNA molecule is measured by its free energy. The lower the free energy, the more stable the structure is. It is still very difficult to predict tertiary structures of RNAs and therefore, in most bioinformatic analyses only RNA secondary structure is analyzed. For most purposes

this is sufficient, because functional RNA secondary structures tend to be conserved through evolution, the folding energy of the molecule can be approximated sufficiently, and RNA function can be modeled successfully using secondary structure alone (Fontana, Konings, Stadler, & Schuster, 1993). The secondary structure can be split into seven distinct structural elements:

- **stacks** – stacked base pairs forming double helical regions
- **hairpin loop** – a series of unpaired bases that close a stack region
- **internal loop** – two series of unpaired bases, that connect two stack regions
- **bulge** – a single stretch of unpaired nucleotides that connect two stack regions
- **multiloop** – several stack regions (more than two) connected by several series of unpaired nucleotides
- **joints** – freely movable substructures, connected by a series of unpaired bases
- **free ends**

Depending on the desired level of detail, this definition can be altered into more general structural elements. An example is the RNAshape method, which provides five levels of decreasing complexity outlining how RNA secondary structures can be regarded (Giegerich, Voss, & Rehmsmeier, 2004). Looking at structures in a less detailed level makes it possible to align structures on different levels of abstraction and can sometimes help to identify similarities between large and complex structures in a more general fashion. The most common framework for visualizing RNA secondary structures is the dot bracket notation (Figure 2 a), where a dot indicates an unpaired base, and two matching brackets indicate a base pair. In Figure 2 provides further examples of common structure visualizations, i.e. mountain plots, circular and linear layouts and RNAshape. One additional visualization scheme that is worth mentioning here is FORNA (Kerpedjiev, Hammer, & Hofacker, 2015). The advantages of FORNA are, that the structures are drawn using force field calculations, and can be dynamically moved and bent by the user. In addition, it facilitates the visual comparison of elements, by allowing to load multiple structures into the same canvas.

1.2.2. Pseudoknots

Special cases of RNA secondary structure are the so-called pseudoknots. A pseudoknot consists of at least two helical regions, which are connected by loops or single stranded regions. In other words, two substructures of a structured region interact with each other. This phenomenon was first discovered in 1982 in the genome of the turnip yellow mosaic virus (Rietveld, Van Poelgeest, Pleij, Van Boom, & Bosch, 1982). An example for a pseudoknot structure from the literature is the H-type (Figure 2 g), which was found in the human telomerase RNA sequence (J.-L. Chen & Greider, 2005). Most RNA structure prediction algorithms use dynamic programming approaches and are therefore not able to predict pseudoknots, because the recursion equation cannot model base pairings between already defined substructures. There are special algorithms designed to predict pseudoknots at the cost of an enormous runtime increase compared to the conventional structure prediction algorithms. Due to these complications, pseudoknot prediction is as of yet not applicable to large scale RNA structure analyses.

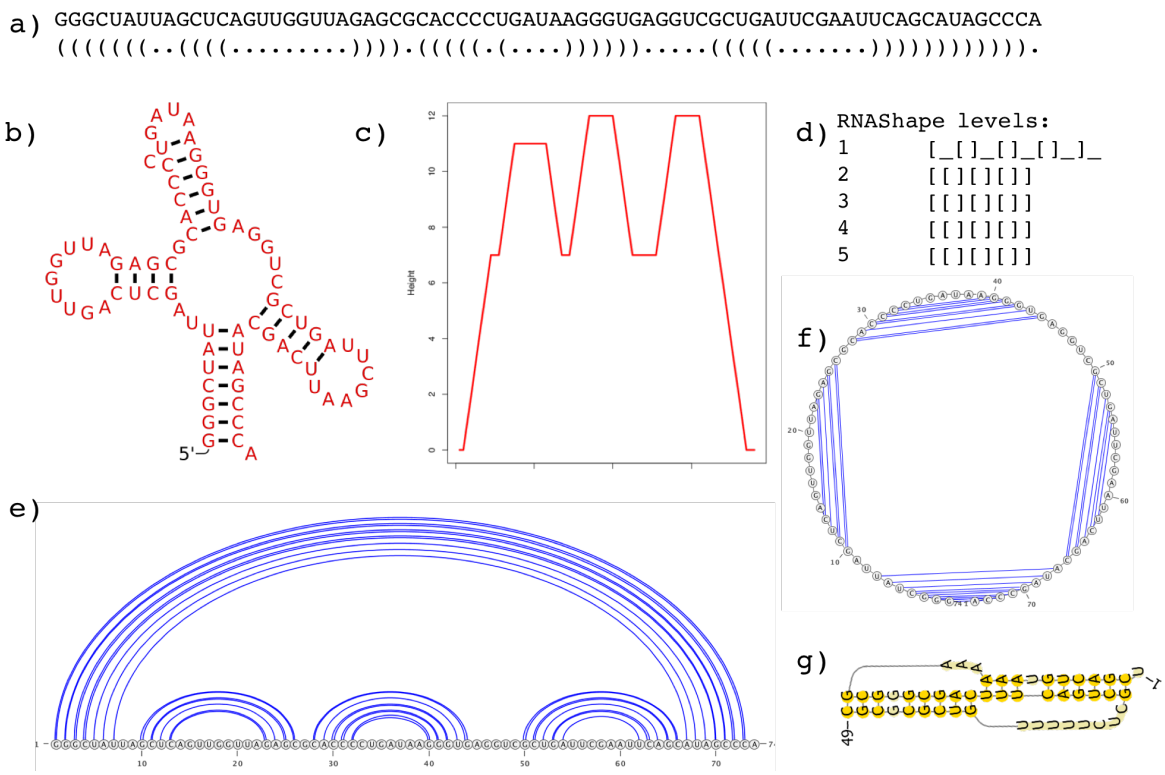


Figure 2: Different RNA structure visualizations:

a) RNA sequence with its minimum free energy (MFE) structure in dot bracket notation. Dots indicate unpaired residues, matching brackets indicate base pairs.

b) RNA diagram drawn with the R2R software package (Weinberg & Breaker, 2011).

c) mountain representation. The x-axis corresponds to sequence positions, and the y-axis to the number of base pairs enclosing the base at this position. Plateaus correspond to loops. The plot was created with the RNAfold webserver (Lorenz et al., 2011).

d) RNAshape representation in the five levels of abstraction, from the most detailed level 1 to the most abstract level 5 (Giegerich et al., 2004; Steffen, Voss, Rehmsmeier, Reeder, & Giegerich, 2006).

e) Linear graph representation. Blue lines indicate base pairs. The drawing was created with VARNA (Darty, Denise, & Ponty, 2009).

f) Circular graph representation. Blue lines indicate base pairs. The drawing was created with VARNA.

g) Pseudoknot structure of the human telomerase sequence, visualized with PseudoViewer2 (J.-L. Chen & Greider, 2005; Han & Byun, 2003).

1.2.3. Functions of RNA structure

Structured RNA can be involved in various functional processes. For example structures that change their conformation, when the matching ligand binds to the structure are called riboswitches. This structural alteration is for example used to dynamically regulate transcription in many cases. Riboswitches are common in prokaryotes, where about two

percent of the genes are regulated by this kind of structure (Sashital & Butcher, 2006). The thiamine pyrophosphate riboswitch was the first to be discovered in the human genome (Miranda-Ríos, 2007). Another interesting function that RNA structure can have is temperature sensing. RNA structures can be completely different depending on the temperature, a phenomenon used by some bacteria to control gene expression depending on the temperature of the environment (Johansson et al., 2002; Kortmann & Narberhaus, 2012). RNA structures can also perform catalytic functions, as in the case of ribozymes, which can for example perform self-cleavage or also cleave other transcripts (Vandivier, Anderson, Foley, & Gregory, 2016). Furthermore, RNA structure is essential for enabling translation, which only works due to the tRNAs' specific shape, which adopt a cloverleaf structure enabling aminoacylation (Bhaskaran, Rodriguez-Hernandez, & Perona, 2012). Another biological process that relies on RNA structure is microRNA silencing, which in some cases even leads to the digestion of mRNAs (Wienholds & Plasterk, 2005). An impressive example of structures located on mRNAs are internal ribosomal entry sites, which are used to initiate translation and are also frequently found in viruses (Mailliot & Martin, 2018). This list of functional RNA structures is by far not exhaustive and is likely to continue to grow with future research.

1.2.4. RNA structure comparison

The fact that functional RNA structures tend to be evolutionary conserved makes it necessary to compare different structural elements with each other and measure their similarity (or the distance between them). There are various methods for RNA secondary structure comparison, e.g. tree edit distance, mountain metric, base pairs distance, or the structure conservation index (SCI). These diverse metrics use different approaches to calculate the structural similarity or distance. The base pair distance is one of the simplest approaches, as it just counts the number of base pairs that are not shared by the two structures. The discriminative power of this measure is limited (Agius, Bennett, & Zuker, 2010), especially for sequences of unequal length. A more complex approach is used by the mountain metric, which calculates the difference of two mountain plots of RNA structures (Hogeweg & Hesper, 1984). An even more advanced idea was used for the tree edit distance method, where the structures are represented as rooted, ordered trees, and the distance is expressed as the number of tree editing steps necessary to convert one structure

tree into another. Finally the SCI is based on the free energies of the structures; this might not be intuitive at first, because energies are compared instead of the structural sequences themselves, but the benefits of this approach become more apparent when we delve deeper into the details behind the measure: The average minimum free energy (MFE) of all structures is calculated, and divided by the energy of the consensus structure, into which all sequences can fold simultaneously. Comparing the MFEs of the sequences directly would not yield the desired result as two completely different structures still can have a similar MFE. However, comparing the average MFE to the energy of a consensus structure solves this problem, because the consensus structure can exhibit a similarly low MFE as the individual sequences, only if all the structures are similar. This measure has proven to be one of the most powerful metrics for structure comparison (Gruber, Bernhart, Hofacker, & Washietl, 2008).

1.2.5. RNA structure determination

Experimental approaches of RNA structure determination can be classified into two basic types: i) biophysical methods and ii) biochemical methods. While biophysical structure determination methods exploit physical traits of RNA molecules, biochemical approaches leverage the chemical properties of the nucleic acid bases by using specific reagents to infer the RNA structure. The reactivity of a nucleotide at a certain position with specific reagents gives information about the accessibility of the residue. In other words, a residue can only react with the reagent if it is not already paired to a different residue. Unfortunately, there are only a few experimentally verified RNA structures, and some of those are only available in a protein complex. Furthermore, RNA structures account only for a small fraction of entries in structural databases. The protein data bank (Berman et al., 2000) harbors about 141000 protein structures, but only about 11000 nucleic acid containing structures. Of these, only a small subset totaling 1384 structures contains RNA. The Rfam database (Griffiths-Jones, Bateman, Marshall, Khanna, & Eddy, 2003) a resource holding non-coding RNA families, where at least one member of the family is an experimentally verified RNA structure, contains only 3016 distinct families.

1.2.5.1. Biophysical methods

Biophysical methods represent the most accurate approaches for structure determination. The most prominent techniques are i) X-ray, ii) nuclear magnetic resonance spectroscopy (NMR) and iii) cryogenic electron microscopy (cryo-EM). These methods are described briefly in the following sections.

X-ray

The X-ray crystallography approach is the most accurate structure determination technique up until now. Unfortunately, it requires the molecule of interest to be crystallized, such that X-rays can be diffracted of a repetitive lattice comprised of multiple copies of the molecule of interest in order to yield accurate electron density maps. An example where the method allowed solving the RNA tertiary structure is the RNA tetraplex (Deng, Xiong, & Sundaralingam, 2001). It is still difficult to produce high quality crystals of RNA molecules, especially if their sequence is long Therefore, the number of X-ray structures determined for RNA sequences is still quite low (Ke & Doudna, 2004). Nevertheless, new approaches to enhance the crystallization process are being continuously developed (Golden & Kundrot, 2003; "Selecting New RNA Crystal Contacts," 2018; Shoffner, Wang, Podell, Cech, & Guo, 2018) and are expected to bring more high resolution RNA structures in the future. Due to technical difficulties, the high throughput analysis of RNAs using X-ray crystallography is not possible, and the number of solved structures using this approach is far too low for competitive bioinformatic analyses.

NMR

Nuclear magnetic resonance (NMR) enables researchers to study the dynamics and structure of molecules in solution with reasonable accuracy (Fürtig, Richter, Wöhnert, & Schwalbe, 2003). In 2003, half of all solved 3D nucleotide structures were obtained using the NMR method (Fürtig et al., 2003), demonstrating its importance. The method leverages the phenomenon of nuclear magnetic spin polarization when atomic nuclei with non-zero spins (atoms containing an odd number of protons or neutrons, whose spins do not cancel each other out) are placed within a magnetic field (Ashbrook, Griffin, & Johnston, 2018). Such nuclei can take one of two energy states (the lower state when the spin moments align with the external field or the excited state when they are anti-parallel). Upon excitation with a pulse of electromagnetic radiation matching the resonance frequency of the atoms, some

of the nuclei absorb energy and jump into the higher energy state. After some time, these nuclei relax down to the ground energy state, re-emitting energy in the process. These packets of energy have frequencies which are characteristic to the resonance frequency of the specific nucleus. The frequency and intensity of the emitted radiation is measured in order to determine the structure of the molecule being probed. The details of the structure are revealed thanks to the phenomenon of nuclear shielding i.e. shielding of the magnetic field that each nucleus experiences due to the induced magnetic fields in the electrons, which oppose the external magnetic field. If many such electrons are present in the vicinity of a nucleus, the magnitude of this shielding is greater and the exact amount of shielding is referred to as the chemical shift, defined as the difference between the expected resonance of a bare nucleus and the actual measured frequency of the nucleus shielded by the surrounding electrons (for example, electrons belonging to a carbon atom nearby). The method has been applied successfully to various RNA sequences, e.g. the 14-mer cUUCGg tetraloop hairpin RNA model system (Nozinovic, Fürtig, Jonker, Richter, & Schwalbe, 2010). The greatest benefit of this technique is, especially when compared to X-ray crystallography is that it captures multiple conformations of each molecule which is very informative when the molecule is unstable or undergoes conformational changes in order to perform its function. NMR is a promising method for determining RNA structures, but it is also not feasible for high-throughput determination of structures, especially with the goal to keep up with the quickly growing space of newly discovered RNA sequences, and the amount of time and cost a single NMR experiment entails. A further limiting factor for this method is the length of RNA sequences, as NMR cannot be used to analyze sequences of more than 1000 nucleotides, because of their slow rotational diffusion (Gopal, Zhou, Knobler, & Gelbart, 2012). The resolution of NMR structures is below that of X-ray experiments, but the fact that the molecules are studied in solution and their molecular motions can be captured renders this method equally important.

CryoEM

The cryoEM technique is used to analyze flash frozen RNA molecules and to obtain snapshots of their structural conformations out of the structural ensemble. The images are captured with an electron microscope (Gopal et al., 2012). The method produces images with considerably lower resolutions than NMR or X-ray experiments, but can still be used to infer the overall shape and the 3D size of molecules. For example, using cryoEM, it has

been shown that ensembles of large RNAs in solution are anisotropic (Gopal et al., 2012). The method can also indirectly be used to determine RNA secondary structure by recording 2D projections of RNA molecules and quantifying the observed branching patterns (Garmann et al., 2015).

1.2.5.2. Biochemical methods

Biochemical methods of structure probing use different chemical compounds that react with unpaired regions of folded molecules. Afterwards it is measured which residues have reacted with the compounds and which have not. In the following two sections, the main principles of two selected examples are explained briefly.

DMS (Dimethyl sulfate)

DMS is can be used to add a methyl adduct to adenine or cytosine residues (Ehresmann et al., 1987). This reaction is only possible for unpaired nucleotides, helix-terminating base pairs or base pairs located next to a GU wobble base pair. Using reverse transcription PCR (polymerase chain reaction) of the methylated RNA sequence the sequence stretches without methyl adducts can be amplified. The reverse transcriptase cannot continue transcription once it reaches a methylated residue. The resulting fragments can then be sequenced, and the paired or unpaired regions can then be mapped back to the sequence. The result of a DMS probing experiment is not a resolved RNA tertiary or secondary structure, but something analogous to an experimentally verified partition function of the sequence.

Carbodiimides

Carbodiimides are synthetic chemical compounds that can, similarly to DMS, form adducts with ribonucleases, which can then be detected using reverse transcription (Tijerina, Mohr, & Russell, 2007). Since the Carbodiimides reaction applies to G and U, while DMS reacts with A and C, the two compounds can be used as complementary analyses (Incarnato, Neri, Anselmi, & Oliviero, 2014).

SHAPE (Selective 2′-hydroxyl acylation analyzed by primer extension)

Similarly to DMS probing, the SHAPE method utilizes reagents that can react with the single stranded regions of a folded RNA molecule. But unlike the reagents of previous probing techniques, the SHAPE compounds react with all residues (Merino, Wilkinson,

Coughlan, & Weeks, 2005). The strength of the reaction gives insights into the flexibility of a residue, and thus the probability of being paired or unpaired. The sequence is extended with a complementary DNA primer, utilizing a reverse transcriptase. The fragments are then sequenced, and compared to an unmodified control sequence (Merino et al., 2005). Like other probing methods SHAPE also provides base pairing probabilities for all residues of a sequence. These probabilities can be used in RNA structure prediction methods to produce highly accurate secondary structures of RNAs (Deigan, Li, Mathews, & Weeks, 2009). The method has been successfully applied on whole genome sequences of e.g. the human immunodeficiency virus (Watts et al., 2009). With more and more probing-resolved structures emerging in the near future, it will soon be possible to bring RNA structure prediction to a new level for example by using e.g. machine learning techniques that predict the base paring probabilities of sequences based on large sets of prelabeled probing data.

1.2.6. RNA structure prediction

Since the discovery of the cloverleaf structure of transfer RNA molecules in 1965 various functional RNA structures were found (Holley, 1965). We are still at the beginning of uncovering the full diversity of RNA structures understating their importance for regulating processes within the cell. A well-known example are riboswitches, which alter their structural conformation when a certain stimulus is triggered, and as a consequence up- or down-regulate gene expression (Serganov & Patel, 2007). In non-coding regions, it has been shown that translation can be completely dependent on secondary structures (Gray & Hentze, 1994; Kozak, 2005). Also structures within coding regions of mRNAs have been proposed to alter translational processes for example by ribosomal stalling (Katz & Burge, 2003). The vast numbers and the length of nucleotide sequences makes it impossible to search for functional structures using experimental techniques only. Therefore, high-throughput structure prediction methods are crucial to conduct research in this field.

1.2.6.1. RNA structure prediction for single sequences

Most structure prediction methods nowadays rely on the minimization of the free energy using dynamic programming. For predicting structures of single sequences we use RNAfold from the ViennaRNA package (Lorenz et al., 2011). There are similar methods like mFold (Zuker & Stiegler, 1981) or Sfold (Ding & Lawrence, 2003) with analog

performance (Gardner & Giegerich, 2004). Personally we decided to prefer methods from the ViennaRNA package as long as they are competitive. RNAfold applies the McCaskill algorithm to calculate the partition function of a sequence (McCaskill, 1990). The partition function is a vector of base pairing probabilities for each nucleotide of a sequence. Incorporating this function into the dynamic programming algorithm enables the free energy minimization to be performed not only in reasonable time, but also gives the possibility to calculate suboptimal structures besides the MFE structure. The accuracy of energy minimizing algorithms for single sequences lies around 73%.

1.2.6.2. RNA structure prediction for multiple sequences

Predicting structures for single sequences is less accurate than using a set of sequences that potentially share structural features (Gardner & Giegerich, 2004). The diversity of sequences sharing the same structure gives additional information that can be leveraged to predict the consensus structure. This is due to the evolutionary pressure that acts on functional structures and preserves their shape. In a functional RNA structure, a mutation in the sequence that would disrupt it and compromise its functionality must be compensated by another mutation. These compensatory mutations are used in simultaneous folding algorithms, e.g. RNAalifold from the ViennaRNA package, to increase the accuracy of the predicted structures. Algorithms that simultaneously fold sets of RNA sequences into a common consensus structure take an alignment of the sequences as input. Most of them implement adapted versions of the Sankoff algorithm, which gives a solution to solving the three problems of i) aligning sequences ii) folding sequences and iii) reconstructing a phylogenetic tree of sequences, simultaneously in reasonable time using dynamic programming (Sankoff, 2006).

1.2.7. Function prediction of RNA secondary structures

All RNA molecules fold into secondary structures, but most of these structures are not functional. It becomes more and more important to distinguish random structures from functional elements and computational approaches are currently the only possibility for performing this task in a reasonable amount of cost and time is using computational approaches. RNAz from the ViennaRNA package is a widely used tool that performs function prediction for aligned RNA sequences with high accuracy. The main principle to

separate important structures from of the overwhelming mass of possible structures in RNAz is the calculation discriminating features regarding i) the thermodynamic stability and ii) the structural conservation. To efficiently model the thermodynamic stability of sequences, RNAz calculates z-scores. The z-score refers to the number of standard deviations the MFE of a structure deviates from a set of randomized sequences with the same length and base composition. The more negative the z-score, the more stable is the structure compared to the structures that are generated from random permutations of the sequence. The creation of sets of randomized sequences for each input alignment and the subsequent calculation of their MFEs is computationally very expensive. RNAz overcomes this problem by using a support vector regression to predict the z-score of the input sequences with high accuracy. The structural conservation is mainly modelled using the SCI, which compares the MFE of the sequences to the energy of their common consensus structure. The consensus structure is predicted using RNAalifold from the ViennaRNA package, which incorporates bonus energies for compensatory mutations. In the final prediction model, all features are combined in a support vector machine, and the input alignment is classified as functional RNA structure or not. Using this technique, we can predict functional RNA structure fragments in reasonable time and with high accuracy. Nevertheless, the length of RNA sequences is still a limiting factor, as the accuracy of structure prediction tools decreases with increasing sequence length. The main reason for this is that the number of possible secondary structures grows exponentially with increasing length. There are two common solutions to this problem: i) sliding windows and ii) RNAlalifold. The sliding window approach simply cuts an input alignment into overlapping slices of predefined size and uses them as input to the structure prediction software. RNAlalifold on the other hand scans a sequence alignment for local optimal substructures, by extending the current alignment window in both directions as long as the calculated MFE decreases. Both methods are used for screening large alignments, such as whole genome alignments, and are also applied within this thesis. The benefit of the RNAlalifold approach is that the boundaries of the potentially functional structures are better defined than using a randomly sliced window. The boundaries of structures are of special importance for the accuracy of structure prediction methods, because of the fast-growing structure space with increasing sequence length. To further increase the accuracy of structure or function prediction methods, the potentially functional structured regions

can be realigned using structure guided alignment methods like e.g. mLocARNA (Olivier et al., 2005; Will, Joshi, Hofacker, Stadler, & Backofen, 2012).

1.3. Research goals and investigation

The main goals of this research were to investigate structure-function relationships in conserved secondary structure elements of viral mRNAs and delineate their evolutionary impact. Most research has been done on structures in non-coding regions of genetic material, which leads to the question if the tools, that were developed specifically for these tasks can also be applied to coding regions. In a first publication we investigated conserved secondary structures in the intergenic regions of ambisense viruses and in turn compared the behavior of structure prediction methods in non-coding and coding regions. To further investigate the coding regions of viral mRNAs, the second publication deals with the creation of a thorough overview of conserved secondary structures in coding regions of viral mRNAs of all viral sequences within RefSeq (O'Leary et al., 2016), and the creation of a suitable web resource to provide the results for future research.

1.3.1. Ambisense viruses

Ambisense viruses belong to the negative sense ss RNA viruses and perform an unique expression strategy. They contain mRNA harboring two genes in opposite reading directions (Figure 3). These two genes are separated by a non-coding intergenic region, containing a potentially functional secondary structure element (López & Franze-Fernández, 2007).

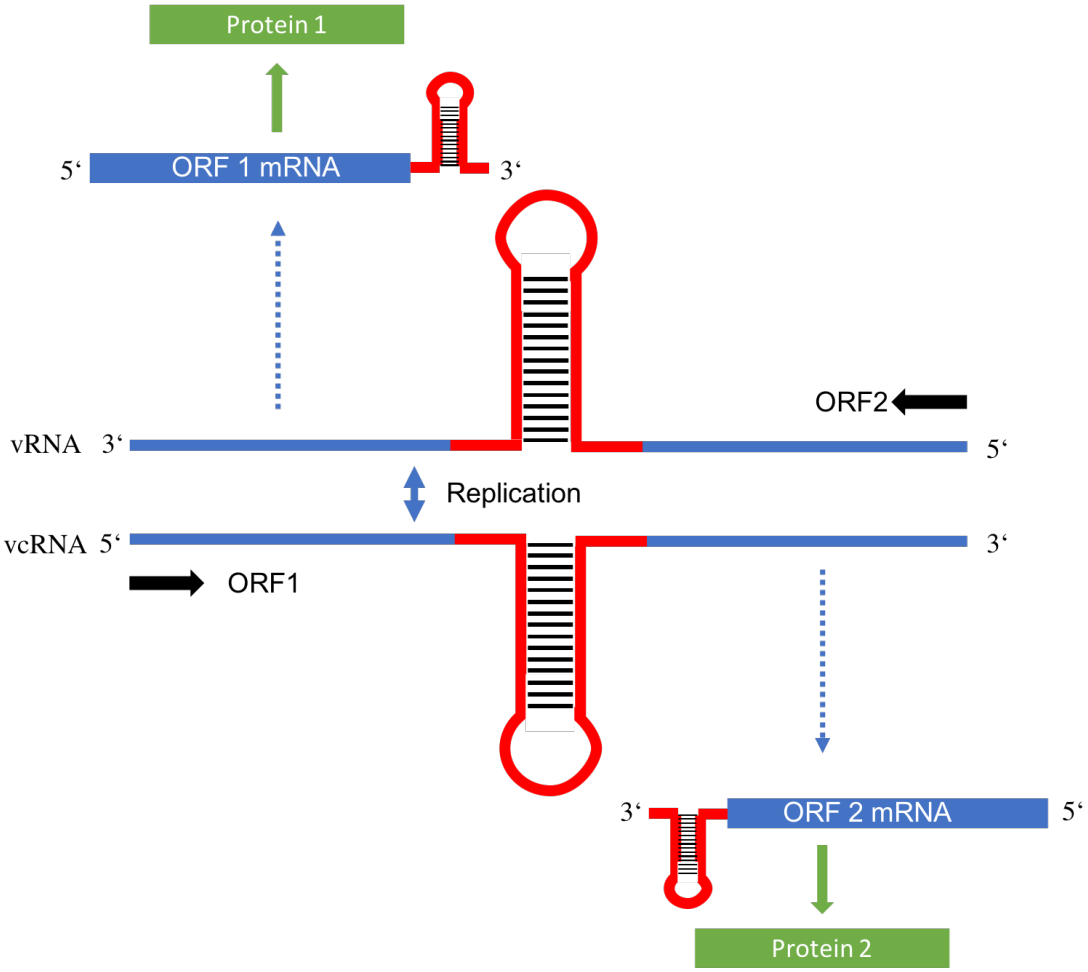


Figure 3: Schematic representation of the ambisense transcription strategy, as performed in e.g. arenaviruses.

The negative strand of the mRNA can be transcribed by the host cell immediately after the viral intrusion, since it is located on the inserted strand which is also called the viral strand (vRNA). Expression of the positive strand requires the intermediate replication step in which the viral complementary strand (vcRNA) is synthesized. In many cases the viruses encode their own RNA dependent RNA polymerases, to ensure transcription (Nguyen & Haenni, 2003). Ambisense viruses do contain a gene in the positive reading direction, but have more similarities with the negative sense ss RNA viruses than with the positive sense ss RNA viruses. Positive sense ss RNA viruses harbor mRNA that can be directly translated into a functional protein, while the ambisense RNA has to include a transcription step prior to translation (Nguyen & Haenni, 2003). The two genes are separated by a non-coding intergenic region, which was, in some cases, shown to comprise secondary structure

elements (López & Franze-Fernández, 2007). Furthermore it was shown that transcription termination takes place in the intergenic region (Brennan, Rezelj, & Elliott, 2017; López & Franze-Fernández, 2007), therefore it is suspected that the structural elements are involved in this process. Ambisense viruses infect various hosts and can cause severe damage to them. E.g. the tobacco mosaic virus and the tomato spotted wilt virus both infect plants and are responsible for enormous crop losses while arenaviruses infect humans and cause hemorrhagic fever. Within this thesis a pipeline that identifies conserved RNA secondary structures is applied to the intergenic regions of ambisense segments of all known virus genera performing this unique expression strategy. If the ambisense strategy involving a secondary structure element for transcription termination could be traced back to a common ancestor, the identified structures in the different viruses should be conserved between the species to some extend. Otherwise the results would point to a convergent evolution of the ambisense transcription strategy. Additionally, the study employed an approach to compare the usage of RNA structure and function prediction software, which is specifically trained with non-coding RNA sequences, for coding regions.

1.3.2. Virus orthologous groups

The virus orthologous groups (VOGs) are a collection of orthologous viral proteins, covering all viral peptides contained in the RefSeq database (O'Leary et al., 2016). The development was carried out by the CUBE institute at the university of Vienna, employing the COGsoft program (Kristensen et al., 2010). As many viruses make use of RNA structures in their replication cycle, and as these functional structures are potentially evolutionary conserved, these orthologous relations between viral proteins can be further investigated at the mRNA level. The fact that horizontal gene transfer is common in many viruses (Liu et al., 2010) suggests that it could be beneficial to look into the evolution of the individual components instead of trying to find a common ancestor for the whole virus. The VOGs provide information about the proteins of a virus, which can be transferred to the corresponding mRNAs. Therefore, each VOG represents group of proteins with their corresponding mRNA that potentially contains evolutionary conserved RNA secondary structures. These could give further insights into the evolutionary relationships between viruses. While RNA secondary structure has been investigated intensely in non-coding regions, only few analyses of the coding sequences can be found in the literature so far. A

thorough overview over conserved, potentially functional RNA structures in viral mRNAs provides targets for experimental verification in the first place, but also for the development of antiviral drugs in the long run, in case the former hypothesis is confirmed. Learning more about viruses, and especially RNA secondary structures and their functions will help us finding new ways to interrupt the viral replication cycle and stop viral epidemics from spreading. In this study we used the VOG data set as a source to create structurally homogenous subgroups of VOGs which we call subVOGs. We created an online database called RNASIV, where the subVOGs are accessible, and the identified conserved RNA secondary structures can be viewed and downloaded for future research.

2. Conserved RNA structures in the intergenic regions of ambisense viruses

The goal of this publication was to build a comprehensive overview over the conserved RNA secondary structures in the intergenic regions of all known ambisense viruses. The data was acquired using the NCBI taxonomy browser and redundancy reduced regarding the sequence of the intergenic regions. To get an overview over the structural potential of the intergenic coding regions we employed the RNAsurface algorithm to the sequences. The intergenic regions showed on average a higher structural potential than the coding regions, and amongst the non-coding regions, arenaviruses exhibited the highest potential to build stable secondary structures.

In order to identify conserved RNA structures within the intergenic regions, we employed the RNAclust algorithm. RNAclust clusters RNA sequences due to their predicted structural similarity. The predicted structural clusters were furthermore checked for potential functionality of the structures using RNAz and commonalities between different clusters using RNAshape. We found stable structures in the intergenic regions of arenaviruses, tospoviruses and a subset of phleboviruses, but the structures were not conserved between the species. This finding rather pins to a parallel evolution of the ambisense expression strategy, than to a common ancestor.

The similar results for coding and non-coding regions, when we investigated the relationship of sequence- and structure similarity in terms of the mean pairwise identity and the structure conservation index lead to the conclusion, that the structure prediction software, which was trained on non-coding sequences only, can also be applied to coding sequences.

The supplemental (Figures S1, S2, Table S1, S2, S3, S4) material can be accessed online with the published article. Prof. Friedemann Weber suggested the idea of the study. Dmitrij Frishman and Michael Kiening designed the study. Michael Kiening conducted the computational analyses. All authors interpreted the data and wrote the manuscript.

2.1. Abstract

Ambisense viruses are negative-sense single-stranded RNA viruses that use a unique expression strategy. Their genome contains at least one ambisense RNA segment that

carries two oppositely oriented reading frames separated by an intergenic region. It is believed that a structural RNA element within the intergenic region is involved in transcription termination. However, a general overview over the structural repertoire of ambisense intergenic regions is currently lacking. In this study we investigated the structural potential of the intergenic regions of all known ambisense viruses and compared their structural repertoire by structure-guided clustering. Intergenic regions of most ambisense viruses possess a high potential to build stable secondary structures and many viruses share common structural motifs in the intergenic regions of their ambisense segments. We demonstrate that (i) within the phylogenetic virus groups sets of conserved functional structures are present, but that (ii) between the groups conservation is low to non-existent. These results reflect a high degree of freedom to regulate ambisense transcription termination and also imply that the genetic strategy of having an ambisense RNA genome has evolved several times independently.

2.2. Introduction

Ambisense viruses comprise a subsection of the segmented negative-sense single-stranded RNA viruses. In contrast to the genome of the purely negative-sense RNA viruses, the ambisense genome contains at least one segment with an additional positive-sense reading frame (Nguyen & Haenni, 2003). The two parts are oppositely oriented and separated by a noncoding intergenic region (IGR), where transcription termination takes place (Emery & Bishop, 1987). The gene of the negative-sense reading frame is directly transcribed from the viral RNA (vRNA) that is delivered by the infecting particles. To express the gene of the positive-sense reading frame, however, the genome first has to be faithfully copied by the viral RNA-dependent RNA polymerase into the viral complementary RNA (vcRNA) which, in turn, serves as template for transcription. Ambisense viruses are found within the entire family *Arenaviridae*, within the genera Phlebovirus and Tenuivirus of the family *Phenuiviridae* (order Bunyavirales), and in the family *Tospoviridae* (order Bunyavirales). Arenaviruses and Phleboviruses infect animals and humans, while Tospoviruses and Tenuiviruses infect plants.

It has been shown for some of the ambisense viruses, e.g. Arenavirus or Phleboviruses that transcription termination takes place within the IGR (Albariño, Bird, & Nichol, 2007; Brennan et al., 2017; Ikegami, Won, Peters, & Makino, 2007; Lara, Billecocq, Leger, &

Bouloy, 2011; López & Franze-Fernández, 2007; Pinschewer, Perez, & la Torre, 2005) and it is suspected that a secondary structure element is involved in the termination process. Within the S segment of Arenaviruses Lassa virus and Mopeia virus this element is predicted to be a single- or a two-hairpin structure, respectively (Wilson & Clegg, 1991). Within the IGRs of viruses belonging to the family Tospoviridae, stable tetraloop structures were identified (Clabbers, Olsthoorn, & Gultyaev, 2014) and suggested to act in transcription termination (de Haan, Wagemakers, Peters, & Goldbach, 1990). Phleboviruses are thought to not contain any stable secondary structures in the IGR (Albariño et al., 2007).

So far, a general overview of the structural repertoire of ambisense IGRs is lacking. To fill this knowledge gap, we investigated the presence of conserved structural elements within the IGR of all known ambisense viruses. We demonstrate that the IGRs of most ambisense viruses have a high potential to build stable secondary structures, many of which are conserved within the phylogenetic groups, but not between them. Our findings imply that such structures are functional and that the ambisense coding strategy may have arisen several times independently during the evolution of segmented negative-strand RNA viruses.

2.3. Material and Method

2.3.1. Data set of ambisense RNA segments

GenBank (Clark, Karsch-Mizrachi, Lipman, Ostell, & Sayers, 2015) files containing genomic RNA sequences of the four known ambisense virus groups – Arenavirus (AV), Phlebovirus (PV), Tospovirus (TV), and Tenuivirus (TEV) - were collected using the NCBI taxonomy browser (NCBI Resource Coordinators, 2015). The sequences were filtered for segments fulfilling the typical criteria for the ambisense transcription strategy: two coding regions (CDSs) on opposite strands separated by an intervening noncoding IGR (Nguyen & Haenni, 2003). Subsequently, redundancy reduction was performed such that no two genomic segments shared 100% sequence identity in their IGRs. Sequences containing ambiguity codes corresponding to incompletely specified bases were excluded from consideration. In our final data set each virus segment was represented by its sequences in both 5'–3' and 3'-5' orientation, which are referred to as viral (v) and viral complementary (vc), respectively (Table 1).

Virus	Segment	Strand	[a]Data set name	Number of sequences	[b]Average GC-content	[b]Average length [bp]
Arenavirus	S	v	AV-S-v (-IGR)	97	0.44 (0.69)	3386 (78)
	S	vc	AV-S-vc (-IGR)			
	L	v	AV-L-v (-IGR)	61	0.40 (0.75)	7179 (124)
	L	vc	AV-L-vc (-IGR)			
Tospovirus	S	v	TV-S-v (-IGR)	70	0.34 (0.22)	3045 (681)
	S	vc	TV-S-vc (-IGR)			
	M	v	TV-M-v (-IGR)	77	0.35 (0.22)	4809 (325)
	M	vc	TV-M-vc (-IGR)			
Phlebovirus	S	v	PV-S-v (-IGR)	168	0.46 (0.51)	1769 (120)
	S	vc	PV-S-vc (-IGR)			
Tenuivirus	2	v	TEV-2-v (-IGR)	35	0.39 (0.36)	3555 (314)
	2	vc	TEV-2-vc (-IGR)			

	3	v	TEV-3-v (-IGR)	53	0.38 (0.26)	2511 (754)
	3	vc	TEV-3-vc (-IGR)			
	4	v	TEV-4-v (-IGR)	53	0.39 (0.35)	2218 (665)
	4	vc	TEV-4-vc (-IGR)			

Table 3: Ambisense virus data sets.

[a]*Data set names are composed of three abbreviations: virus name-segment name–strand. If only the intergenic regions are used for a specific analysis, -IGR is added to the data set name.*

[b]*Values in parantheses refer to IRs only.*

2.3.2. Evaluating the structural potential of RNA sequences

To evaluate the structural potential within the IGRs and CDSs of ambisense viruses we employed the *RNAsurface* algorithm (Soldatov, Vinogradova, & Mironov, 2014). *RNAsurface* converts the values of minimal free energy (MFE) for a given RNA sequence into a z-score calculated as:

$$z = \frac{E-\mu}{\sigma}, \quad (1)$$

where E denotes the MFE while μ and σ are the average and standard deviation of the energy distribution of random sequences with comparable nucleotide composition and length. *RNAsurface* uses z-score evaluation to reconstruct the structural potential surface of an RNA sequence, which can be visualized by a two dimensional heat map (see Figure 4). The x-axis corresponds to sequence positions while the y-axis corresponds to segment length. Each point and its color on the heat map represent a substring of the sequence and its structural potential. z-scores for all possible substrings of length between a certain minimal and maximal window size (W_{min}, W_{max}) are calculated. Locally optimal segments are visualized as peaks within the structural potential surface. A segment is regarded as locally

optimal if small changes of its boundaries lead to worse z-scores. We used the default values of W_{min} and W_{max} of 30 and 200 nucleotides, respectively.

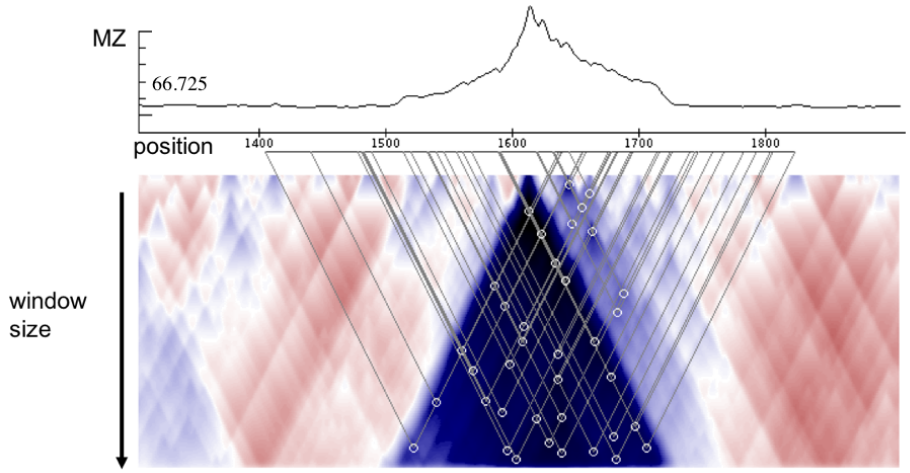


Figure 4: Fragment of the RNAsurface heatmap and the MZ plot for the Arenavirus segment S (GeneBank ID: AB261991), including the IR (positions 1579–1642). The line plot in the upper half of the picture shows the MZ values (y-axis) for each sequence position (x-axis). The heatmap below represents the surface of the structural potential, where the triangles represent locally optimal segments. Each point on the surface corresponds to a RNA subsequence. Colors represent the significance of the secondary structure predicted for the subsequence at each point. Blue corresponds to highly structured regions while red corresponds to unstructured regions. Circles in the heatmap denote local optimal segments.

The structural potential surface defined in this fashion can be reduced to a one-dimensional curve reflecting the structuredness of a sequence segment. For each individual sequence position i the maximum squared z-score (MZ) among all negative z-scores of sequences of length between W_{min} and W_{max} covering this position is computed.

RNAsurface was applied to each strand of CDS and IGR RNA sequences separately. To make the resulting MZ curves comparable between sequences of different length n, sequence positions p were converted to relative positions p_{rel} on the percentage scale, with the leftmost and rightmost positions of each sequence corresponding to 0% and 100% for CDS1, 100% and 200% for the IGR, and 200% and 300% for CDS2 respectively, according to the following equations:

$$p < IR_{start} \rightarrow p_{rel} = \frac{p}{IR_{start} - 1},$$

$$p \epsilon IR \rightarrow p_{rel} = \left(\frac{p - IR_{start}}{IR_{start} - IR_{end}} * 100\right) + 100, \quad (2)$$

$$p > IR_{end} \rightarrow p_{rel} = \left(\frac{p - IR_{end}}{n - IR_{end}} * 100\right) + 200,$$

with IR_{start} and IR_{end} being the first and last residue within the IGR respectively.

2.3.3. Clustering and identification of evolutionary conserved RNA structures in the intergenic regions of ambisense segments

In addition to the analysis of the structural potential described above we delineated conserved structural motifs in all 16 IGR of ambisense viruses listed in Table 1 by performing motif clustering using *RNAclust* (Will, Reiche, Hofacker, Stadler, & Backofen, 2007). Applying this technique to CDS regions would be computationally prohibitive because they are too long to build structure guided alignments in reasonable time. In a first step *RNAclust* employs *RNAfold* from the Vienna package (Hofacker & Stadler, 2006) to calculate for each input sequence the base pairing matrix, which contains the probability for each base to be paired with each other base of the sequence. These matrices are then used to calculate pairwise alignments between all input sequences using *LocARNA* (Will et al., 2012). Subsequently, a hierarchical tree, based on the pairwise alignment scores is derived, using the WPGMA (weighted pair group method with averaging) (Lemey, Salemi, & Vandamme, 2009) method. Each node in the tree corresponds to a possible cluster of sequences sharing a structural motif. For each node a multiple sequence alignment (MSA) and a consensus secondary structure are computed using *mLocARNA* and *RNAalifold* (Hofacker, 2007), respectively. *(m)LocARNA* performs a variation of the Sankoff Fold and Align algorithm and thus produces structure guided (multiple) sequence alignments (Lemey et al., 2009; Will et al., 2012).

To obtain the final motif clusters on the tree we employed *RNAsoup*, a script implementing an adapted version of the Duda and Hart (Duda, Hart, & Stork, 2012) rule to find the optimal motif clusters. Briefly, the null hypothesis is that the sequences i to n belonging to an internal node C form one cluster and should not be further split into separate clusters corresponding to the child nodes of C, C1 and C2. The squared error for this hypothesis

(Je(1)) is calculated as a sum of differences between the MFE of each individual sequence (Ei) and the consensus of all sequences belonging to C (Econs) (Kaczkowski et al., 2009):

$$Je(1) = \sum_{i=1}^{n} (E_i - E_{cons})^2 \quad (3)$$

The squared error for the opposite hypothesis, namely that the cluster C should be split into separate clusters C1 and C2, is calculated as:

$$Je(2) = \sum_{j=1}^{2} \sum_{i=1}^{n_j} (E_i - E_{cons})^2 \quad (4)$$

The null hypothesis is rejected if:

$$\frac{Je(1)}{Je(2)} < 1 - \frac{2}{\pi} k \sqrt{\frac{2 - \frac{16}{\pi^2}}{n}} \quad (5)$$

where k is a user-defined parameter, with larger k values resulting in larger clusters and vice versa. The procedure results in an hierarchical tree containing all input sequences, which are then divided into clusters. Since altering the levels of k leads to a change in the cluster size, each value of k represents a possible clustering. Based on the analysis of Rfam sequence families (Will et al., 2007) it was previously shown that k values between 0.8 and 1.2 result in the best structural consistency of RNA sequences within each family. We discuss the optimal choice of the k value for our study in the Results section. For the visualization of the structures FORNA (Kerpedjiev et al., 2015) was used.

2.3.4. Potential functionality of structures

The clusters predicted by the *RNAclust* pipeline were further checked for functionality using *RNAz* (Gruber, Findeiß, Washietl, Hofacker, & Stadler, 2010). For each set of aligned RNA sequences secondary structure prediction was affected by *RNAz*, which implements the 'Align, then Fold' strategy. The *RNAz* method uses the *RNAfold* algorithm from the

Vienna package to calculate secondary structures and the corresponding MFE for each individual RNA sequence in the alignment. In addition, for each aligned sequence set *RNAz* calculates a consensus secondary structure and its MFE using the *RNAalifold* algorithm from the Vienna package. Subsequently *RNAz* calculates three measures of structure conservation: i) the MFE z-score for each individual sequence, ii) the mean MFE z-score amongst all sequences, and iii) the structure conservation index (SCI) of the entire alignment. The SCI is calculated as the average MFE value of the sequences contained in the input MSA, divided by the MFE value of the consensus structure. Similar to the *RNAsurface* described above, an *RNAz* z-score describes the number of standard deviations by which the MFE of a given sequence deviates from the MFEs of a set of randomized sequences with the same length and base composition. However, since MFE calculations for a large set of randomized sequences are computationally prohibitive, *RNAz* predicts z-score values for each sequence by a support vector regression that estimates the mean and standard deviation of random MFEs dependent on the nucleotide composition of the given sequence. Negative and positive z-scores indicate structures that are more and less stable, respectively, than would be expected by chance (Gruber et al., 2010). The mean z-score is calculated as the sum of all individual z-scores divided by the number of sequences. *RNAz* assumes that conserved and thermodynamically stable structures are functional, in which case it outputs 'RNA', otherwise it outputs 'OTHER'. For this purpose a p-value, called class probability, is calculated. p-values greater or smaller than 0.5 trigger the prediction of the 'RNA' or 'OTHER' class, respectively. The default classifier of *RNAz* is trained using MSAs that are created without any structure information. Predicting conserved structures based on alignments derived from *LocARNA* would result in an over-prediction of functional RNAs, as *LocARNA* alignments are already guided by structures. To avoid such over-prediction the option '-l' for *LocARNA* type alignments was used, which enables a different training model in *RNAz* that is optimized on structure guided alignments.

2.3.5. Deducing recurring structural motifs using shape abstraction

The clustering procedure described above joins together sequences that display global structural similarity, as judged by the SCI calculated over the entire sequence length. In order to find local structural motifs shared by ambisense viruses we converted the consensus structures of each cluster into the abstract shape notation (Giegerich et al., 2004).

Shapes are defined at five levels of abstraction – from the most realistic to the most abstract ones:

Level 1: nesting pattern for all loop types and all unpaired regions.

Level 2: nesting pattern for all loop types and unpaired regions in external loop and multiloop.

Level 3: nesting pattern for all loop types, but no unpaired regions.

Level 4: helix nesting pattern and unpaired regions in external loop and multiloop.

Level 5: helix nesting pattern and no unpaired regions.

In order to infer the overall structural similarity between RNA molecules we used the most abstract level 5, which strongly compresses structural diversity. In addition level 3, which provides the best trade-off between accuracy and abstraction, was used to assess specific differences between the consensus structures.

To convert the dot-bracket notation of *RNAalifold* into shape notation the tool RNAshape (Steffen et al., 2006) was used. Only shapes of the same type can be compared with each other. For each data set we locally aligned the consensus shapes of all clusters against each other using the dynamic programming tool water from the EMBOSS toolkit ("EMBOSS: The European Molecular Biology Open Software Suite," 2000). As water accepts only nucleotide sequences as input, we replaced the shape notation with nucleotides ('['='A', ']'='T', '_'=G), and used a simple scoring scheme (match +1, mismatch 0, gap -1). The nucleotide alignments were then translated back into the shape notation.

2.3.6. Investigation of the relationship between sequence and structure conservation in the intergenic and coding regions

To investigate sequence-structure relationships in ambisense virus segments we calculated two measures of evolutionary conservation using *RNAz*: SCI and mean pairwise identity (mPID). In order to find out whether the relationship between mPID and SCI is different if pure sequence alignments, rather than structure guided alignments, are employed we also conducted the same analysis using the multiple sequence alignment method *clustalOmega* (Sievers & Higgins, 2014). To derive the SCI and mPID values for the three regions (CDS1, CDS2, IGR) in each cluster we first multiply aligned sequences and then split the MSAs

into overlapping windows using the helper perl script rnaz*Window.pl*, which is part of the *RNAz* package. As the full-length CDS are too long to produce structure guided alignments in reasonable time, we realigned the alignment windows initially produced with *clustalOmega* using *mLocARNA*. In a last step the alignment windows generated by both methods were passed to *RNAz* and correlation coefficients between mPID and SCI for each region were calculated separately.

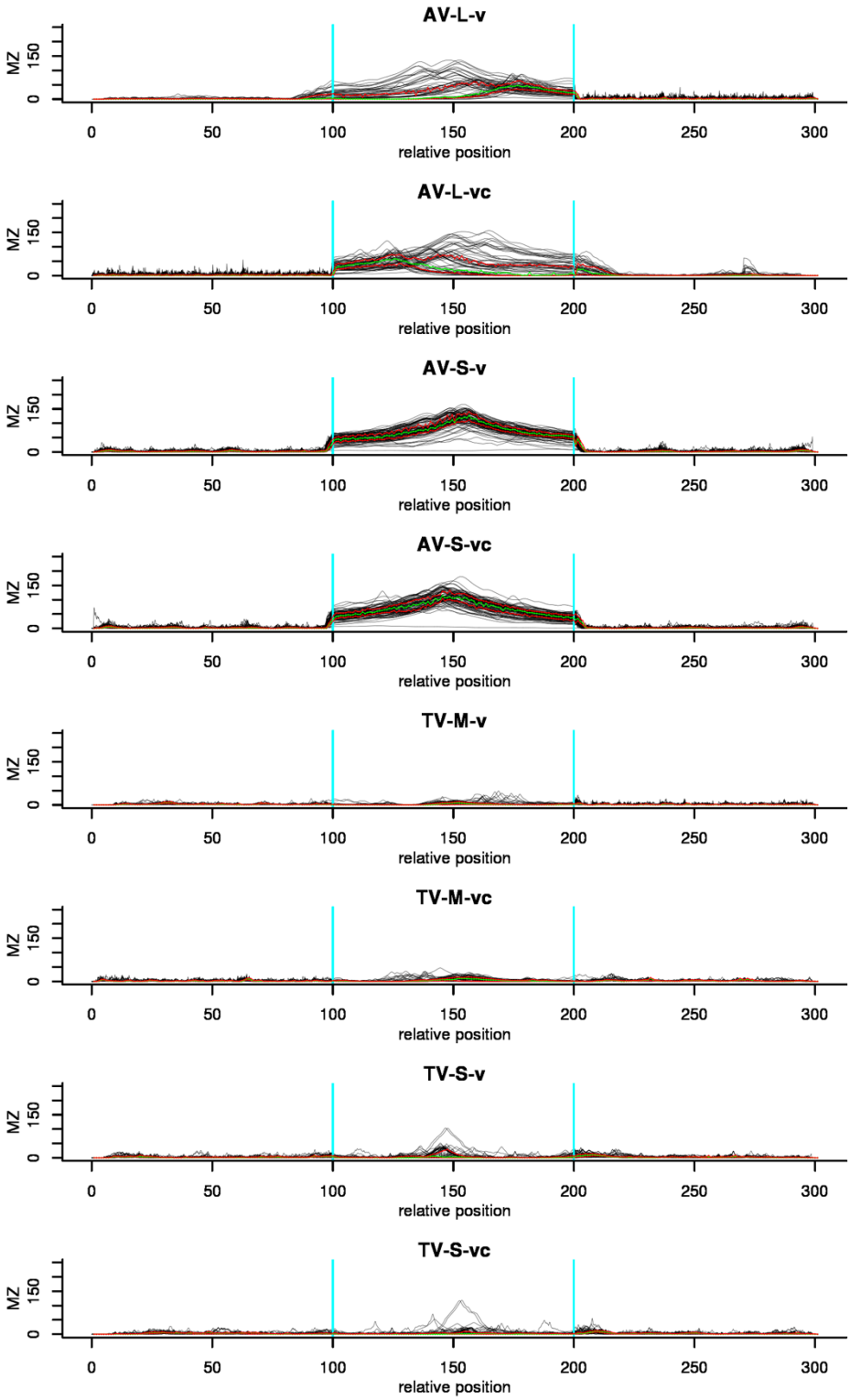
Data Availability

The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

2.4. Results and Discussion

2.4.1. Highest structural potential within IGR in most sequences

For each of the data sets listed in Table 1 we compared the structural potential of the IGRs to that of the CDSs by calculating MZ values for each sequence position using *RNAsurface* (Figures 5 and 6). Sequence positions were converted to a percentage scale for each part of the sequences (CDS1, IGR, CDS2), as explained in *Methods*. The largest contrast in structure potential between CDSs and IGR was identified in the Arenavirus data sets. The median MZ values within the IGRs are between 20 for AV-L-v and 67 for AV-S-v, while in the CDSs they range between 0.2 and 2. In TV and TEV the difference in structural potential between CDSs and IGR is less pronounced, although the highest MZ peaks are still located within the IGR (see boxplots in Figure 6). Within the TV sequences and the sequences of TEV segment 4 the higher structuredness of the IGR is still quite apparent, while in TEV segments 2 and 3 the difference between CDS and IGR is very small. The CDSs of PV appear to contain potentially structured regions, while their IGRs can be subdivided into three groups: i) those with no structural potential within IGRs at all (24% and 43% for v and vc, respectively), ii) those containing local optimal segments, whose MZ peak is however lower than that of the CDSs (25% and 45% for v and vc, respectively), and iii) those containing the highest MZ peak in the entire sequence, including CDSs (51% and 12% for v and vc, respectively). Thus, our analysis shows that there is high structural potential within the IGSRs of most of the ambisense virus sequences.


AV-L-v
AV-L-vc
AV-S-v
AV-S-vc
TV-M-v
TV-M-vc
TV-S-v
TV-S-vc
MZ
0
150
0
50
100
150
200
250
300
relative position

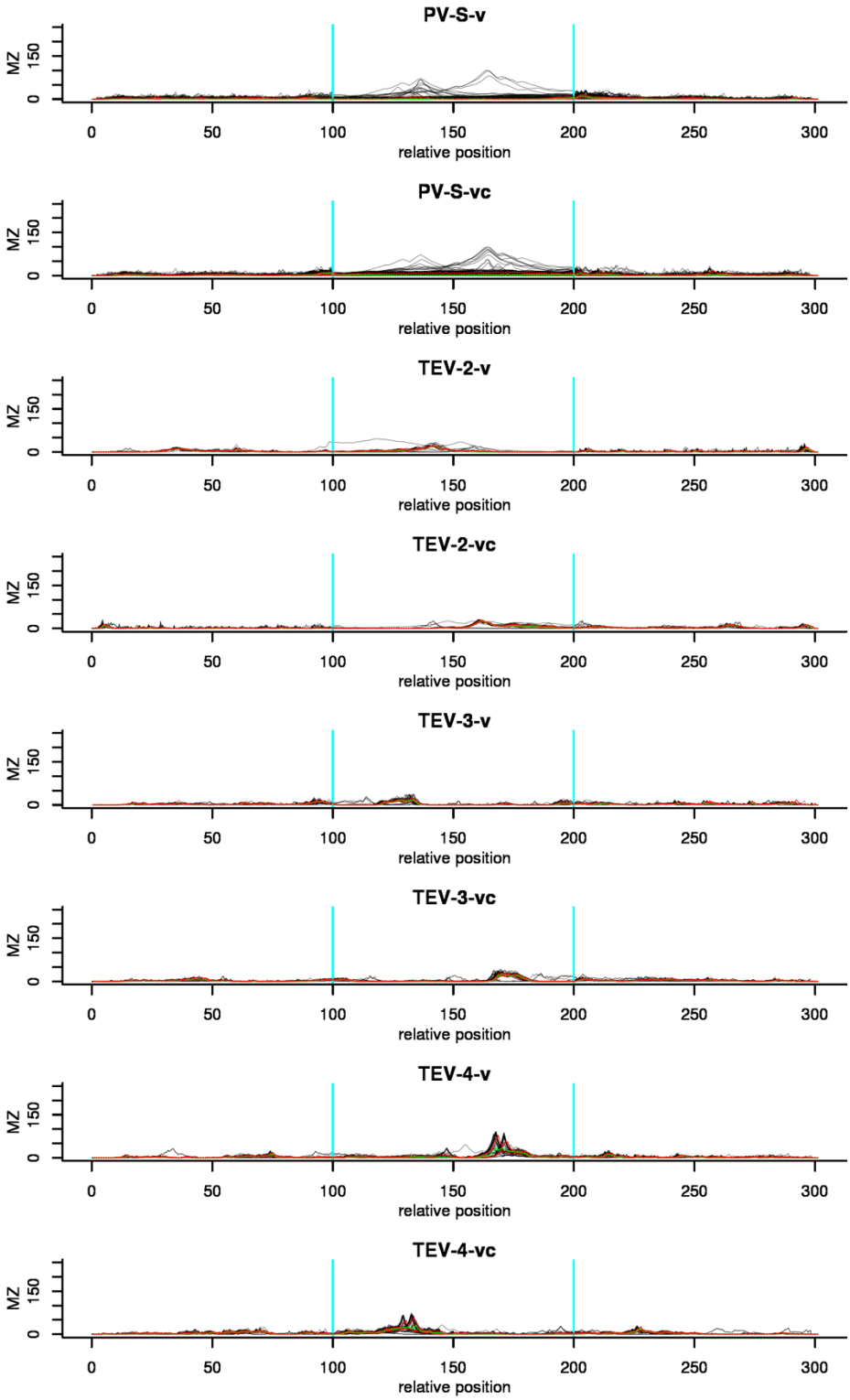


Figure 5: Structural potential plots for all data sets listed in Table 1. The x-axis corresponds to the relative sequence position, with the ranges 0–100%, 100–200%, and 200–300% corresponding to CDS1, IR, and CDS2, respectively. Black lines indicate maximum squared z-score values (MZ) of individual sequences derived from RNAsurface. Green and red lines correspond to the median and 25%/75% quartiles, respectively. Vertical blue lines highlight the borders of the IR. High MZ values indicate potentially structured regions.

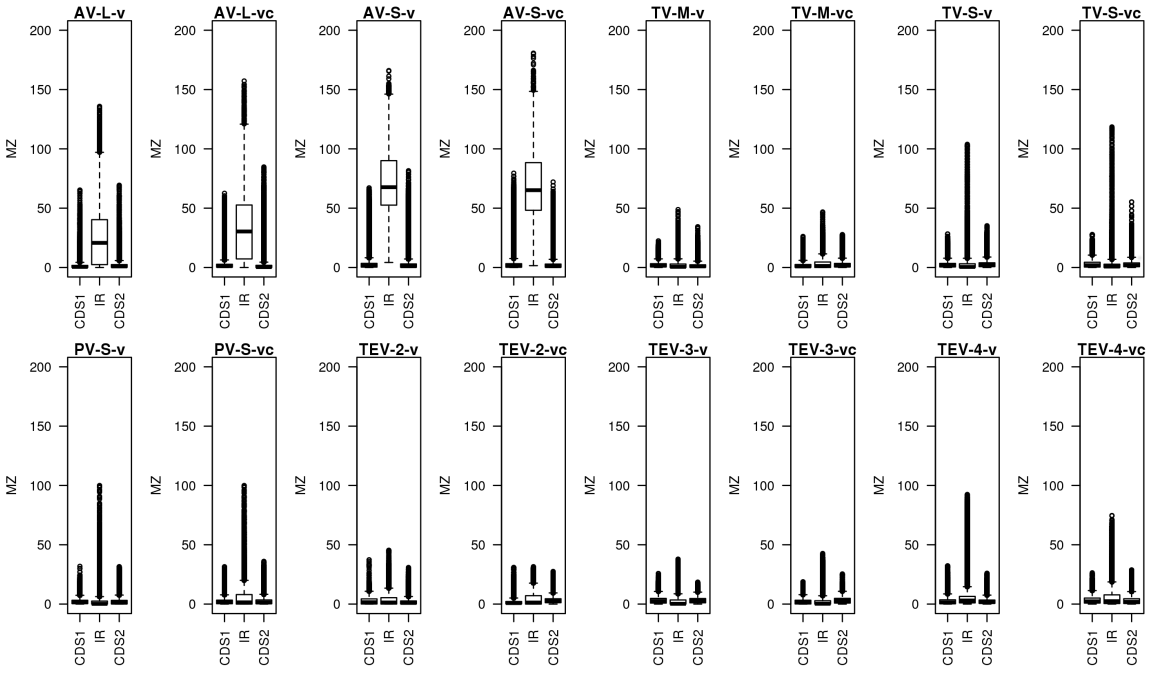


Figure 6: Boxplots showing the distribution of MZ values in each data set for CDS1, IR, and CDS2. MZ values reflect the potential of a sequence to form stable secondary structures.

2.4.2. Structural Clustering

Using the *RNAclust* pipeline, hierarchical trees and multiple structure guided sequence alignments were calculated for each data set. The trees, and a table showing all sequences in their corresponding clusters can be found in the supplementary material (Figures S1, S2 and Tables S1A, S1B). The full clustering results including all alignments are available upon request. Sequence clusters potentially sharing common structural motifs were delineated from trees using a certain cutoff value of the parameter k (see *Methods*).

The consensus structure of each cluster is described in terms of the structural features it contains (Table S2). We define the following five structural features (Figure 7):

1. Stem: a series of consecutive base pairs that can include bulges and/or internal loops.
2. Loop: a series of unpaired residues followed by one base pair. Smaller loops are named according to the amount of unpaired residues in the loop, e.g. triloop (three unpaired residues), tetraloop (four unpaired residues) or pentaloop (five unpaired residues).

3. Internal loop: two or more unpaired residues on both sides of a stem.
4. Bulge: a series of unpaired residues on one side of a stem.
5. Multiloop (Mloop): a region where at least three stems come together. A stem can be followed by either unpaired residues or by another stem.

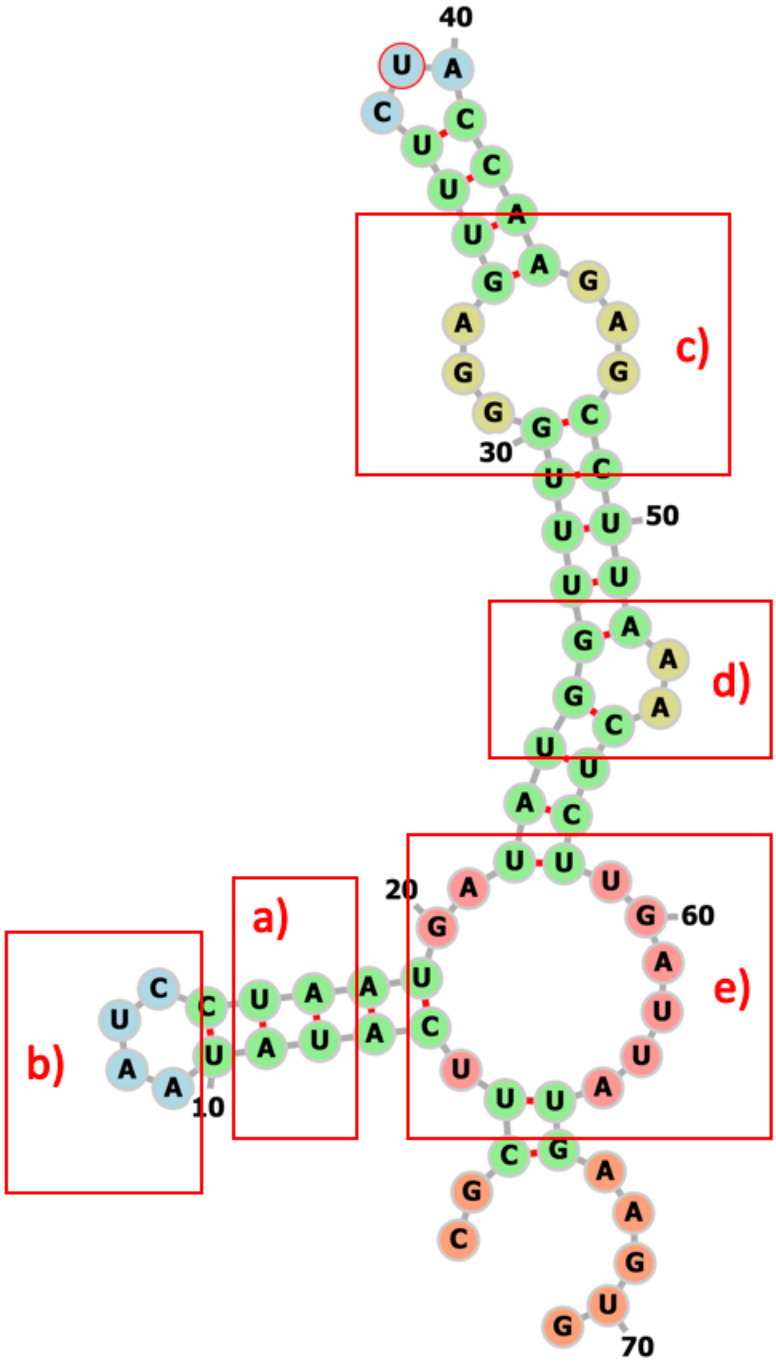


Figure 7: Example structure visualizing the structural features used in our analysis: (a) stem, (b) loop, (c) internal loop, (d) bulge, (e) multi loop.

We identified a large number of structural motifs in the clustered IGRs of ambisense viruses (Tables S2 and S3). The clusters are characterized by the mPID of the underlying structure guided MSA, the SCI, the *RNAz* prediction of functionality ('RNA' or 'OTHER'), and by structural features contained in the consensus structure, if available. We paid special attention to the hairpin loop motifs, which are believed to play a role in transcription termination of ambisense RNAs (Nguyen & Haenni, 2003). Using the recommended values of k (between 0.80 and 1.20) we were not able to reproduce previously reported structural features of TV sequences (e.g. y-shaped tetraloops (Clabbers et al., 2014)), which only became visible in our clusters at k=0.6. Based on this limited benchmark we used the value of k=0.6 in this study. A lower value of k results in more fine-grained clusters, which is acceptable as no wrong structures are accumulated in a cluster. The size of clusters and their number vary drastically between data sets (Figure S2).

2.4.3. Conserved stem-loop structures in Arenaviruses

In general, clustering results echo those obtained by analysing the structural potential of IGRs. Arenavirus sequences show the strongest structural conservation among all data sets, as evidenced by the consistently high SCI values in the AV-S-v/vc-IGR data sets (Table S2). Almost all the clusters containing more than one sequence (either viral and viral complementary) are predicted by *RNAz* to contain functional RNAs (Table S2 and Figure S2). The only structural features present in the consensus structures of the S segment are stems and loops (Table

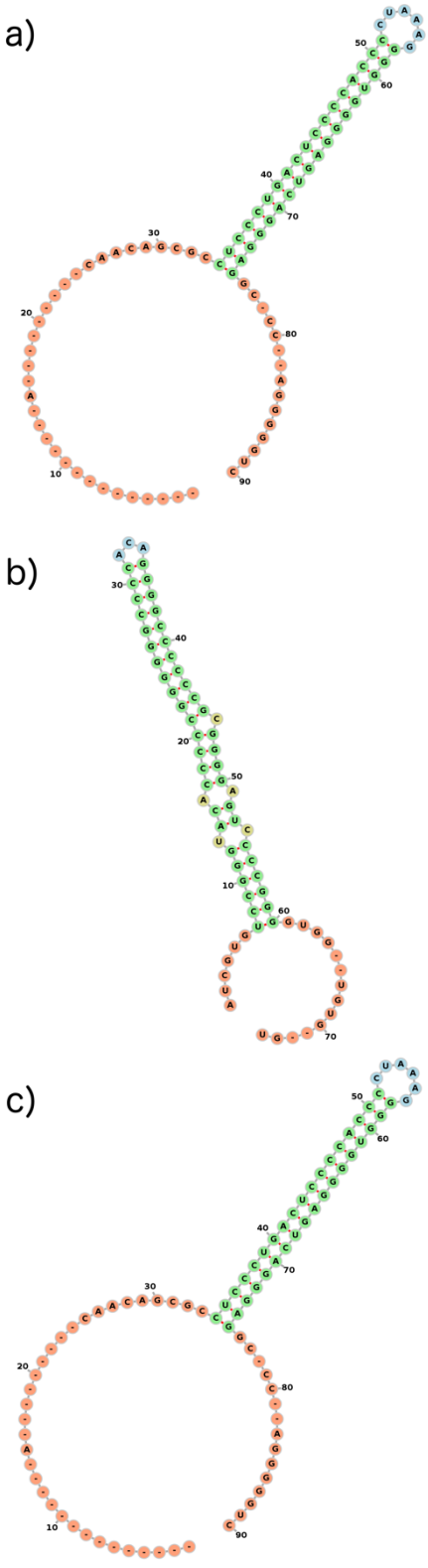


Figure 8: Consensus secondary structures, visualized with FORNA: (a) AV-S-v-IR cluster 3 (b) AV-L-v-IR cluster 55 (c) TV-M-v-IR cluster 11.

S3). The stem size varies between 5 and 18 base pairs, and the loop size between 3 and 7 unpaired bases. The biggest cluster in the AV-S-v-IGR set (node ID 3 in the hierarchical tree) consists of 56 sequences. The only consensus stem-loop structure for this cluster, predicted with *RNAalifold*, contains the hexaloop sequence 'CCUAAAGG' (Figure 8a), with the first and the last bases forming a base pair. The mPID of the alignment is 0.65 while the SCI is 0.94 (Table S2). The high SCI value indicates that the structures of the individual sequences within the cluster are energetically very close to the consensus structure, and thus are strongly conserved. The cluster contains 18 species (Tables S1A, S1B), of which four have been previously reported to contain the single stem-loop feature: Pichinde (Auperin, Galinski, & Bishop, 1984), Lymphotic choriomeningitis virus (LCV) (Romanowski & Bishop, 1985), Lassa virus (Auperin, Sasso, & McCormick, 1986) and Lujo virus (Briese et al., 2009). Thus, our analysis proposes 14 further species to fall into this structural class. The smaller clusters have more than one stem-loop structure. Clusters 155, 138 and 180 consist of 13, 9 and 5 sequences, respectively, and have two stem-loops. This structural feature was already described for the Tacaribe virus (López & Franze-Fernández, 2007), which did not fit into any cluster in our analysis. All sequences but one, the Morogoro virus, that share the two stem-loop feature belong to the group of new world Arenaviruses. Cluster 119 consists of 10 sequences, all belonging to the species Junin virus, and has three stem-loops, two triloops and one pentaloop. The two triloops found in the Junin virus sequences are already described in the literature (Ghiringhelli, Rivera-Pomar, Lozano, Grau, & Romanowski, 1991), while the pentaloop has not been mentioned before. All Machupo virus sequences form one cluster (138) and also share two highly conserved triloops. All but one found consensus structures consist of loops with GC closing pairs and thus are very stable. The AV-S-vc-IGR set is subdivided into a larger number of smaller clusters, but shows a similar distribution of stem-loops, also with very high SCI values.

The IGR structures of the L segment of AV have received much less attention in the literature so far. The distinction between the New World and the Old World Arenaviruses becomes even more apparent within the L segment, as all clusters only contain sequences from one of these groups. In contrast to S segment, which only harbors stems and loops, some consensus structures of the L segment contain two further structural features, internal loops and Mloops (Table S3). The stem size varies between 3 and 27 base pairs, the loop

size between 3 and more than 10 unpaired residues, and the internal loop size between 2 and 9 unpaired residues. The biggest cluster (node ID 55 with 10 sequences), containing only Old World Arenaviruses of the species Lassa virus, Mopeia virus, Mobala virus and Luna virus Arenaviruses also adopts the single stem-loop feature with an additional internal loop (Figure 8b). The consensus structure of this cluster has quite a high SCI value of 0.74 and the underlying MSA has a mPID of 0.57 (Table S2). The representatives of the New World Arenaviruses, Machupo virus, Tacaribe virus, Sabia virus and Chapere virus share one conserved triloop, while all other members of this group have at least two stem-loops. The clusters comprising LCV (92, 97, 106) tend to contain more than one stem-loop, connected through an Mloop (Table S3, structures 10, 11). The stems sometimes also contain internal loops. Only one cluster containing LCV sequences (113) shows a different consensus with only one stem-loop.

As expected, Arenavirus sequences belonging to the same species tend to appear in the same cluster (this also applies to all other ambisense virus genera). Our analysis thus indicates that structures are more conserved within species than between them. It is also remarkable that all clusters in the AV-S-v/vc-IGR data set, except the cluster 180 in the v set and the cluster 69 in the vc set, as well as all clusters in the AV-L-v-IGR data set, and all clusters in the AV-L-vc-IGR data set, except clusters 8 and 16, are predicted to contain functional RNAs by *RNAz* (Tables S2, S3). This finding is in line with the previous studies that show that a stable hairpin in the IGR of Arenaviruses is required for transcription termination (López & Franze-Fernández, 2007).

2.4.4. Conserved motifs within Tospovirus segment M

The TV-M-v/vc-IGRs contain diverse structural features, with stem sizes varying between 4 and 44 base pairs and loops containing between 3 and 10 unpaired nucleotides (Table S3). All clusters of the v set contain at least one multiloop with 3 base pairs, internal loops between 2 and 11 nucleotides, as well as occasional bulges of size 1 to 3 nucleotides. The four tomato spotted wilt virus (TSWV) clusters (5, 31, 71, 90) share two strongly conserved y-shaped tetraloops: ‘UGAAAA’ and ‘CCGAAG’ (Table S3, structure 16). Clusters 71 and 90 additionally have a pentaloop (‘UGACAAG’) and a decaloop (‘UAAUCUGACUAA’) in common, while cluster 5 contains 2 additional triloops (‘CAAUG’, ‘UCAAA’) that are not conserved in other clusters. The presence of y-shaped tetraloops in the IGR of the TV-

M segment was reported previously (Clabbers et al., 2014). Cluster 110 also contains three tetraloops ('AAAUAU', 'CUUAGG', 'UUUU-A'), of which two adopt the y-shape (Figure 8c). Cluster 115 is the only cluster where no consensus structure could be found at all. The vc sequences show the same picture. All TSWV clusters (49, 56, 64, 84) share the tetraloop 'UUUUCA', (Table S3, structures 18, 20, 22) with the exception of the cluster 6, in which this tetraloop exhibits the sequence 'UACAAA'. The y-shape is not always present in the vc sequences, where some of the structures tend to build a longer stem loop (Table S3, structure 22). The TSWV clusters also contain a number of partially conserved internal loops as well as triloops, tetraloops and several larger loops up to a septaloop.

The IGRs of the TV-M segment, with a mean length of 324 nucleotides, are quite long compared to the size of the local structural elements, that range in size between 40 and 80 nucleotides. For this reason the overall conservation signal in terms of the global SCI tends to be weak.

2.4.5. Phleboviruses

Based on the structural elements contained in the IGR, PV sequences can be separated into several groups. The first group (clusters 6, 296, 299, 130, 135, 138, 146, 151, 156, 165, 200, and 282) contains only one stem-loop and in some cases internal loops (Table S3, structure 23). The stem size within this group varies between 3 and 15 base pairs (Table S3). The loops typically contain between 4 and 15 unpaired residues, although larger loops with more than 20 unpaired residues also exist. The second group (clusters 289, 319, 271, 227, 127, 230, and 214) possesses consensus structures containing several stem-loops, with the stem length between 3 and 40 base pairs, the loop size ranging between 3 and more than 20 unpaired nucleotides, the internal loops containing between 2 and over 20 unpaired nucleotides, and some bulges of size 1 or 2 (Table S3, structure 25). The third group (clusters 221, 235, 279, 286, 306, 309, and 332) is characterized by the presence of an additional multiloop with 3 to 6 base pairs (Table S3, structure 24). In all three groups we find clusters predicted to contain functional RNAs by *RNAz* (see Methods) as well as clusters predicted to not contain functional RNAs (Tables S2, S3). Finally, the fourth group, comprising all other clusters, does not contain any structural elements at all. We were unable to find any published indication that transcription termination by phleboviral IGRs depends on a specific secondary structure element. However, for Rift Valley fever virus it

was shown that the UTR of the L segment (which is not ambisense) forms a functional stem-loop structure that takes part in transcription termination (Ikegami et al., 2007). Our analysis indicates that at least some Phleboviruses contain similar local structural elements in the IGR that could serve as transcription regulating elements.

2.4.6. Tenuiviruses

Sequences from the TEV-2-v/vc-IGR data sets possess the lowest structural potential compared to all other data sets. The structural potential of a sequence is assessed based on the z-scores of its MFE structures and thus reflects the significance of these MFE structures (Soldatov et al., 2014). Low structural potential in a set of sequences may imply that the MFE structures are random and thus unlikely to be evolutionarily conserved. Indeed, the *RNAclust* analysis revealed that although the IGRs are very similar at the sequence level, the structural conservation in terms of the SCI is weak for the v sequences and even lower for the vc sequences. The biggest cluster of the v set (node ID 4), making up 70% of the data set, has a mPID of 96% and a SCI of 78%. In the vc set, all sequences are clustered into one cluster (0) sharing a mPID of 93% and a SCI of 25%. All consensus structures of both sets are predicted to be non-functional by *RNAz*. We were thus unable to detect any structural conservation in the IGRs of these sequences.

Similarly, the sequences of the TEV-4-v/vc-IGR data sets were grouped into a single cluster and showed almost no structural conservation, although they share a mPID of 81% and 87% in the v and vc set respectively, and both alignments were classified to contain a functional RNA by *RNAz*.

2.4.7. The relation between the sequence and structural similarity in the intergenic and coding regions of ambisense viruses

Previous research suggests that there is a certain relationship between mPID and structural conservation in RNA sequences. For non-coding RNAs, such as small nuclear RNAs, the correlation between sequence identity and structural similarity quickly increases as the sequence similarity approaches ~60% and saturates between 60-100% (Capriotti & Marti-Renom, 2010), while in mRNAs this correlation only exists at sequence identity levels between 85% and 100% (Chursov et al., 2012). We investigated the relation between SCI and mPID in ambisense virus IGR and CDS based both on pure multiple sequence

alignments computed by *ClustalOmega* and on structure guided alignments created with *mLocarna*. The SCI and mPID values were calculated for MSAs produced by both alignment algorithms for all nodes in the *RNAclust* hierarchical trees (at k=0.6). Since the SCI and mPID values follow a normal distribution according to the Kolmogorov-Smirnov test (Lilliefors, 1967), we computed Pearson correlation coefficients between them and retained only significant correlations (p-value <0.05; two sided t-test; Figure 9). mPID values calculated both from sequence and structure-based alignments are strongly positively correlated with SCI, in the intergenic as well as in the coding regions (Figures 9 and 10; Table 4). The correlation is generally stronger for higher mPID values, which is in line with our previous results obtained for yeast mRNAs (Chursov et al., 2012).

Region	**Pearson correlation coefficient between SCI and mPID**	
	Sequence alignments (MAFFT)	Structure-guided alignments (mLocARNA)
CDS1	0.85	0.89
CDS2	0.86	0.89
IGR	0.34	0.89

Table 4: Correlation between mPID and SCI across all data sets.

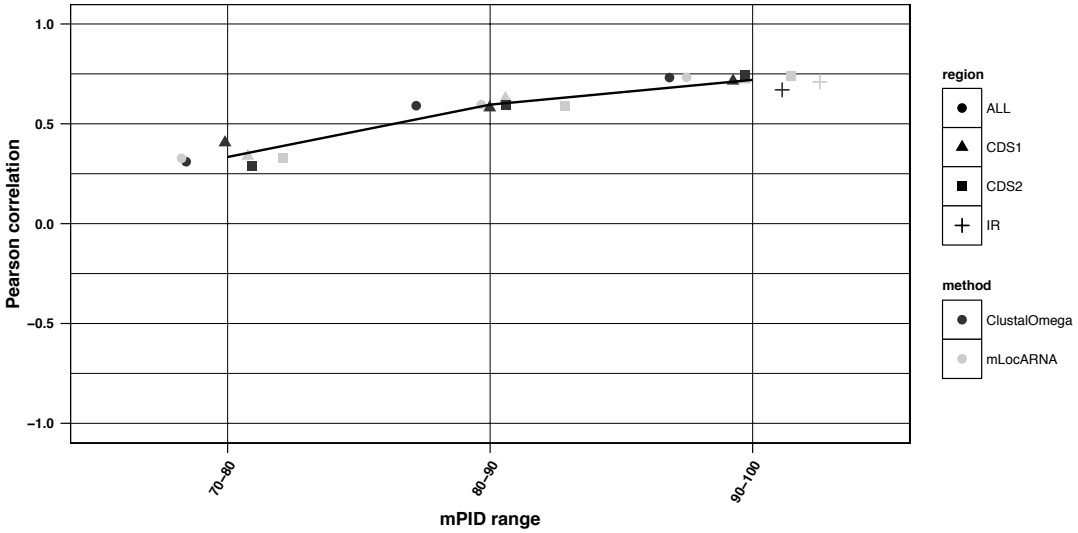
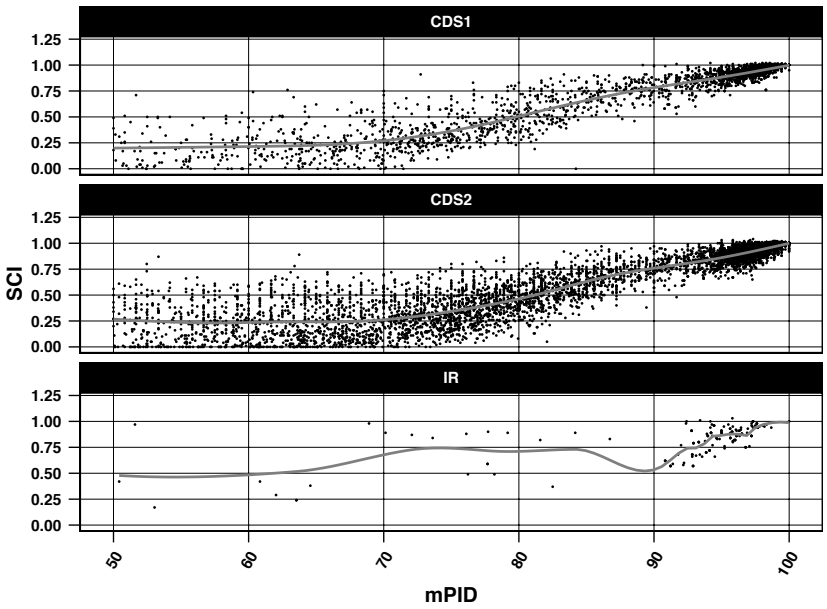


Figure 9: Correlation of mPID and SCI within different mPID threshold ranges. The shape of the data points refers to the three regions: triangles – CDS1, squares – CDS2, pluses – IR; circles represent all three regions taken together. Colours correspond to the alignment method used: grey – mLocARNA, black ClustalOmega. Only significant (confidence interval of 0.95, P-value<0.05) correlations are shown. For better visibility of overlapping points a horizontal jitter function was applied. The black line shows the regression performed on all data points, and the grey shaded area the standard error. Correlations for the mPID values below 50% were not significant and are thus not shown.

a)



b)

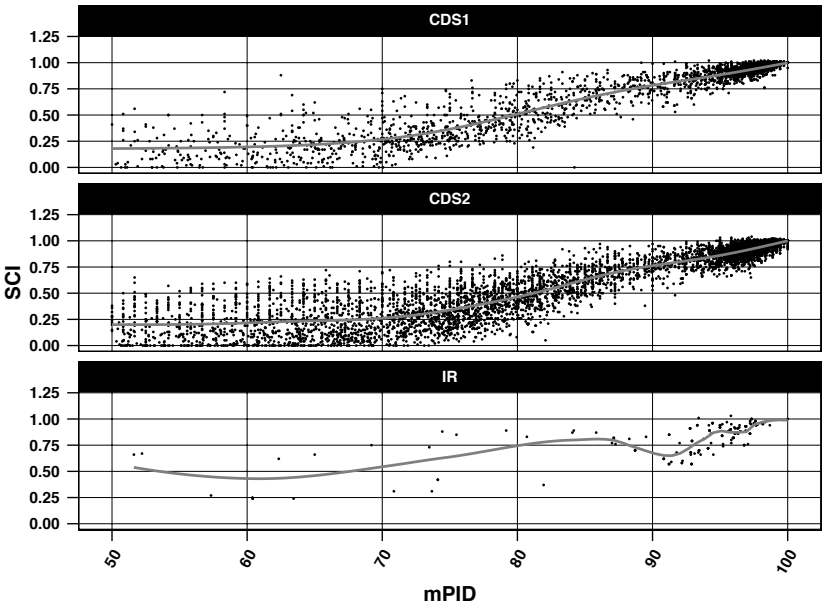


Figure 10: Comparison of sequence similarity in terms of the mPID on the x-axis and structure similarity in terms of the SCI on the y-axis, for the CDSs and IRs of all data sets for (a) Multiple sequence alignments generated with ClustalOmega. (b) Structure guided multiple sequence alignments generated with mLocaRNA. Each point corresponds to an alignment that covers a window of 120 nucleotides either of CDS1, CDS2 or the IR.

2.4.8. Inference of shape motifs

To find conserved local motifs in the IGRs of ambisense viruses, we converted the consensus secondary structures of RNAs in each cluster into shapes at the five levels of abstraction described in *Methods*. At each abstraction level shapes were locally aligned using the program *water* from the EMBOSS toolkit. To assess the differences between the consensus structures we chose layer 3 as the best tradeoff between the accuracy and abstraction, because it depicts all nesting patterns for all loop types, but no unpaired regions. To define the maximum motif the structures have in common, the most abstract level 5 was used. All shape consensus structures are listed in Table S4.

Within the AV-S-v/vc-IGR data sets the consensus structures already showed a high overall similarity between different clusters. Expectedly, the shape analysis led to qualitatively similar results: the consensus structures of the largest clusters, that make up 58% and 40% of the v and vc data set, respectively, contain a single stem-loop ([] in shape notation). 28% of the v set and 31% of the vc set share an additional stem-loop ([][]), and sometimes also internal loops (e.g. [][[]]). Only 10% of both data sets contain a third stem-loop ([][][]), occasionally with internal loops. Likewise, the results for the AV-L-v/vc-IGR data sets are as follows. 32% of the v and vc sets share one stem-loop (containing internal loops), 25% of both sets share two stem-loops (containing internal loops), three stem-loops are harbored by 15% of the data sets, and 10% of the v sequences as well as 16% of the vc sequences harbor four stem-loops.

The TV-M-v/vc-IGR data set showed a weak structural conservation based on the *RNAclust* analysis, but it is known from previous research that the sequences contain conserved structural elements (Clabbers et al., 2014). Interestingly, this conservation is visible in the shape analysis, with the exception of cluster 115 (25% of the sequences) in the v set and cluster 122 (12% of the sequences) in the vs set, which did not produce any consensus structure. Clusters 90 and 71 from the v set show the exact same shape, consisting of four stem-loops, including the y-shaped loops already described in literature, and cluster 5 only differs from those by an additional internal loop. These three clusters make up ~43% of the sequences. Clusters 31 and 110 (30%) harbor only three stem-loops, but also share the y-shaped loops. The level 5 shape motif visible from the local alignments of all clusters of the v set, except 115 and 110, is [[[][]][]], while in the consensus structure of cluster 110

the third stem-loop is situated before the y-shaped loop: [[][[][]]]. Within the vc set, the structures are more diverse, but also share a conserved core. Cluster 64 (13%) consists of a single long stem-loop containing several internal loops and bulges, while cluster 111 (6%) harbors two loops. The majority of sequences share three stem-loops (31%). The remaining 30% of the sequences harbor four or more loop structures. The y-shaped motif is only visible in clusters 84 ([[][][[][]][]]) and 147 ([][[[][]][][]]), which account for 21% of the sequences; in both cases the y-shaped loops are part of a multiloop.

In line with the *RNAclust* analysis several groups become apparent within the Phlebovirus data sets (PV-S-v/vc-IGR):

- 11% of the v set and 21% of the vc set do not contain any structure.
- 58% of the v set as well as 29% of the vc sequences harbor one stem-loop ([]) and in some cases also internal loops ([[]])
- Two stem-loops are harbored by 13% and 44% of the v and vc sequences, respectively.
- Within the viral sequences, 4% harbor three stem-loops.
- Finally, two groups of vc sequences, accounting for 10% and 5% of the data, respectively, are characterized by the presence of a multiloop and three ([[][][]]) to four ([[][][][]])stem-loops.

In contrast to what is currently known, the vast majority of the Phlebovirus sequences harbors at least one single stem-loop with a varying number of internal loops within the stems on the v strand and two stem-loops on the vc strand. Only a small proportion of sequences does not show any structural feature.

2.4.9. Conclusions

Segmented negative-strand RNA viruses are in principle only capable of expressing one gene per segment (Ferron, Weber, la Torre, & Reguera, 2017). To expand the coding capacity, they express additional genes either via polyprotein expression, insertion of ORFs with a shifted reading frame, or by the ambisense strategy (Elliott & Brennan, 2014; Wuerth & Weber, 2016). The IGR is an inherent feature of the ambisense RNA viruses, pathogens of major medical and economical importance. Here, we present the first comprehensive comparison of the predicted IGR structures among and between the arenaviruses, tospoviruses, phleboviruses and tenuiviruses.

In order to obtain a general overview of the structural repertoire of the ambisense RNA segments, we explored their potential to build stable secondary structures by clustering the sequences globally based on structural similarity, predicting consensus structures for each cluster, and identifying local structural motifs using shape abstraction. The shape analysis showed that the consensus structures of the clusters can be further compressed due to their similar shapes, reducing the structural repertoire of the IGR to three to four recurring patterns within each data set. Arenaviruses showed the highest structural potential within their IGRs, and a very low degree of structuredness in the two CDSs. The structural potential of the Arenavirus IGRs is also clearly the strongest among all ambisense viruses. The most prominent feature of Arenavirus IGRs is the presence of at least one stem-loop structure, which is presumed to be a transcription termination signal. This structure has been previously discovered in four of the Arenavirus species, and our analysis also suggests that it is conserved in 14 further species. Approximately 60% of the sequences contain only one stem-loop structure (shape []). Further 30% of the IGRs contain two stem-loop structures (shape [][]) in 10 species, including the Tacaribe virus, where they have been previously described (López & Franze-Fernández, 2007). Finally, 10% of the IGRs, all belonging to the species Junin virus, contain three stem-loop structures (shape [][][]) – an arrangement that has not been mentioned in literature so far. According to previous research this species also belongs to the group, which contains two stem-loops (Ghiringhelli et al., 1991). Interestingly, only the sequences belonging to the "New world Arenaviruses" harbor more than one stem-loop. Almost all structures were classified as functional, which reinforces the assumption that they may play a role in transcription termination.

In contrast to Arenaviruses, Tospovirus sequences showed only a moderate structural potential in the IGR, although it is still higher than in the CDS regions. We found conserved y-shaped tetraloops and triloops in the IGRs of the M segment on the viral strand, with the y-shaped loops containing a varying amount of internal loops: i) [[[[[[]]]]]], ii) [[[][[]]]], and iii) [[[[[[]]]][]]]. Sequences of the viral complementary strand tend to fold into a long single stem-loop ([]), while the y shape is only present in 21% of the vc sequences. 48% of the sequences harbor one to seven additional stem-loops. All consensus structures of the Tospovirus clusters, viral as well as viral complementary, were predicted to be functional.

The Phlebovirus IGRs have so far not been thought to contain any functionally relevant structural features (Albariño et al., 2007), which led us to expect a low structural potential in their IGRs. This is indeed true for most of the analysed Phlebovirus sequences. Nevertheless, a subset of Phlebovirus sequences, comprising approximately 76% and 57% of the vRNA and vcRNA data set respectively, showed a high potential to build stable secondary structures in the IGR. For 51% and 12% of the v and vc sequences, respectively, the potential is higher than the structural potential of the two CDS. *RNAclust* analysis confirmed that these structures are highly conserved within the corresponding clusters. Only 11% of the sequences did not produce any consensus structure. Shape analysis revealed that the most prominent feature on the vRNA strand is a single stem-loop with a varying amount of internal loops, and on the vcRNA strand two stem-loops with internal loops. The found structures are predicted to be functional. Tenuiviruses appear to be the only ambisense viruses that lack conserved structures in the IGRs, in spite of the high level of sequence similarity. Still the consensus structures of segment four were predicted to be functional.

The structure conservation index was used to identify conserved structures in the IGRs of ambisense viruses and to measure their reliability. As this measure is dependent on the sequence similarity of the underlying alignment (Gruber et al., 2010), we investigated the relation between sequence and structure conservation in terms of the mPID and the SCI respectively both for pure sequence and structure guided MSAs. The analysis showed that SCI and mPID are highly correlated both in the intergenic and in the coding regions, and the correlation is stronger for higher mPID values. Similarity of sequence-structure relationships between the coding and non-coding regions suggests that *RNAz* and other methods for predicting stable and conserved RNA secondary structures, which are usually trained on non-coding regions, can also be applied to coding sequences.

Genomic RNA of ambisense RNA viruses is largely encapsidated by N proteins (Reguera, Cusack, & Kolakofsky, 2014). In non-ambisense bunyaviruses, transcription is believed to be terminated by 3' UTR structures formed in either the genome template or the nascent mRNA (Barr, 2007). Similarly, whether the IGR-located transcription termination signal of ambisense bunyaviruses acts on the mRNA level, or whether it already forms on the nucleocapsids, is currently unknown.

To summarize, we detected certain levels of structural conservation (or the absence of any predicted structures) within the ambisense virus groups, but not between them. This may indicate that transcriptional termination *per se* could be achieved by many different types of IGRs. Other, group-specific, IGR functions may exist that constrain the structural freedom despite the fact that the IGR can be a target of antiviral host defences (Moy et al., 2014).

In any case, our results indicate that the ambisense gene expression strategy has evolved several times independently as a means to alleviate the one-gene-per-segment restriction of the segmented negative-strand RNA viruses. These findings may improve our understanding of negative-strand RNA virus gene regulation and evolution and prompt further research into the structure-function relationship of the IGR.

3. Conserved Secondary Structures in Viral mRNAs

In this second publication, we sought to give a comprehensive overview over conserved structures on viral mRNAs of all viruses which are part of RefSeq. We used the virus orthologous groups (VOGs) data set as starting point. The VOGs were provided by the CUBE institute of the university of Vienna and represent orthologous groups, containing all proteins with an entry in RefSeq.

We calculated distance trees for the mRNA sequences belonging to VOG proteins and subsequently predicted structural conservation and functionality of the identified structures for the inner nodes of the trees. The nodes with the maximum number of sequences still exhibiting conserved structures were annotated as structurally homogenous subgroups of VOGs (subVOG). Using this procedure, we could resemble structures that were already described in the literature and propose new viral sequences which contain them as well. We furthermore investigated the relationship of mRNA structure and protein function, and found trends previously described for cellular organisms, as well as newly described virus specific relationships. An online resource RNASIV was created, where the subVOGs can be accessed and downloaded for future research.

The supplemental (Figures S1, S2, S3, Table S1) material can be accessed online with the published article. Michael Kiening designed the study, the methodology, implemented the software, interpreted the results and wrote the manuscript. Roman Ochsenreiter designed the study and interpreted the results. The data set was provided by Prof. Thomas Rattei and Hans-Jörg Hellinger. Ivo Hofacker, Thomas Rattei and Dmitrij Frishman had the idea to the study, designed the study, supervised the project and wrote the manuscript.

3.1. Abstract

RNA secondary structure in untranslated and protein coding regions has been shown to play an important role in regulatory processes and the viral replication cycle. While structures in non-coding regions have been investigated extensively, a thorough overview of the structural repertoire of protein coding mRNAs, especially for viruses, is lacking. Secondary structure prediction of large molecules, such as long mRNAs remains a challenging task, as the contingent of structures a sequence can theoretically fold into grows

exponentially with sequence length. We applied a structure prediction pipeline to Viral Orthologous Groups that first identifies the local boundaries of potentially structured regions and subsequently predicts their functional importance. Using this procedure, the orthologous groups were split into structurally homogenous subgroups, which we call subVOGs. This is the first compilation of potentially functional conserved RNA structures in viral coding regions, covering the complete RefSeq viral database. We were able to recover structural elements from previous studies and discovered a variety of novel structured regions. The subVOGs are available through our web resource RNASIV (RNA structure in viruses).

3.2. Introduction

Secondary structures formed in single-stranded mRNA molecules through complementary self-interactions, both in the untranslated (UTR) and coding (CDS) regions of mRNAs, have been implicated in a variety of regulatory functions (Bevilacqua & Blose, 2008). For example, riboswitches modulate gene expression through conformational changes in response to various stimuli (Serganov & Patel, 2007). Translation initiation, elongation, and termination as well as translation efficiency depend on higher order mRNA secondary structures in non-coding regions (Gray & Hentze, 1994; Kozak, 2005). CDS hairpins have also been suggested to play a role in the regulation of translation (Katz & Burge, 2003), in particular by causing ribosomal stalling and modulating translational efficiency (Mortimer, Kidwell, & Doudna, 2014). The relationship between mRNA structure in the CDS and gene expression has been demonstrated both computationally and experimentally (Carlini, Chen, & Stephan, 2001; Duan et al., 2003; Ilyinskii et al., 2009; Kudla, Murray, Tollervey, & Plotkin, 2009; Nackley et al., 2006). In particular, reduced mRNA stability near the start codon has been observed in a wide range of species, probably as a mechanism to facilitate ribosome binding or start codon recognition by initiator-tRNA (Gu, Zhou, & Wilke, 2010). Structured elements within CDS directly influence mRNA abundance (Del Campo, Bartholomäus, Fedyunin, & Ignatova, 2015). Computational studies show that native mRNAs have lower folding energies and are thus more stable than codon-randomized ones (Katz & Burge, 2003). The three mRNA functional domains—5'UTR, CDS, and 3'UTR—form largely independent folding units, while base pairing across domain borders is rare (Shabalina, Ogurtsov, & Spiridonov, 2006). The ability of viruses to persist in their host in

a genus-specific manner is influenced by the interplay between local structural motifs and genome-scale ordered RNA structures (GORS) (Simmonds, Tuplin, & Evans, 2004), which impose additional restraints on the RNA sequence space. Evolutionarily conserved local secondary structures have been identified in CDSs (Meyer & Miklós, 2005) and shown to be functional (Olivier et al., 2005). An indirect indication of the global importance of RNA structures in the coding regions comes from the recent study of Fricke et al. who identified selection favoring specific pairing patterns between synonymous codons within RNA hairpins (Fricke, Gerst, Ibrahim, Niepmann, & Marz, 2019).

Increasing evidence suggests that secondary structural elements in the CDSs of viral RNAs also constitute a previously underappreciated, evolutionarily conserved level of functional organization of viruses. A large number of conserved secondary structural motifs were computationally identified in the Flavivirus genomes (Fricke et al., 2015; Pirakitikulr, Kohlway, Lindenbach, & Pyle, 2016; Thurner, Witwer, Hofacker, & Stadler, 2004), predicted to restrain sequence variability (Simmonds & Smith, 1999) and experimentally shown to regulate important biological processes, such as replication and infection (Pirakitikulr et al., 2016). Multiple secondary structures were described in the coding regions of the (+) sense RNA of the Influenza A virus (Moss, Priore, & Turner, 2011). Another example is a secondary structural element within the coding region of the Dengue virus type 2, which is essential for its replication (Clyde & Harris, 2006). More recently, using a comparative genomics approach, Goz and Tuller identified a large number of potentially functionally important regions in the coding regions of Dengue viruses, in which the RNA folding strength is conserved independently of sequence conservation and compositional bias (Goz & Tuller, 2015). Specific regions in the HIV structural genes were reported to be under strong selection for stable secondary structures (Goz & Tuller, 2016). Recent research shows that mechanisms of translational control by RNA structures can be shared between viruses and cellular organisms (Díez & Jungfleisch, 2017).

Given the important role played by RNA structures in shaping the evolutionary dynamics of viruses and modulating their interaction with the host, a large-scale investigation of RNA motifs in viruses would be warranted. However, there are two major challenges that need to be addressed before embarking on such an investigation. First, accurate structure prediction for long RNA molecules, such as mRNAs, is generally out of reach for the existing computational methods. Second, conserved stem-loop structures can only be

derived from a collection of high-quality alignments of orthologous viral transcripts, which are difficult to obtain, given the rapid pace of viral evolution and the ensuing poor sequence conservation, even between closely related species.

Here, we propose a computational approach to explore the RNA structurome of the viral coding regions, in which local structure predictions are applied to VOG (Viral Orthologous Groups, http://vogdb.org), the first comprehensive collection of orthologous groups derived for all viral proteins contained in the RefSeq (Pruitt, Tatusova, Brown, & Maglott, 2012) database. We utilize RNALalifold (Lorenz et al., 2011) to scan long input sequences for locally optimal secondary structures. The identified structural boundaries are more accurate than those derived from using a sliding window of fixed length. Functional importance of structured regions is assessed by RNAz (Gruber et al., 2010). We present a novel database, RNASIV (RNA structure in viruses; http://rnasiv.bio.wzw.tum.de), which contains the largest currently available collection of predicted RNA structures in viruses. It provides access to 201,708 viral mRNA sequences clustered into 42,293 structurally homogenous groups and is intended to become a useful tool for exploring structure–function relationships in virus families.

3.3. Materials and Methods

3.3.1. Viral Orthologous Groups (VOGs)

All genome sequences and their annotations were retrieved from the RefSeq viral database release 79 (O'Leary et al., 2016) and grouped into phages and non-phages, based on the available taxonomic information. Assemblies containing inconsistently annotated or completely unannotated polyproteins were identified based on the manually curated information provided by ViralZone (Hulo et al., 2011) and excluded from consideration. Phage and non-phage protein sequences were clustered into phage and non-phage preVOGs, using the NCBI's COG software package with all default settings.

For all phage and non-phage preVOGs, multiple sequence alignments were constructed with Clustal Omega v1.2.4 (Sievers & Higgins, 2014) and used to build HMM-profiles using HMMer 3 (Eddy, 2009). The profiles were subsequently aligned against each other, using HHalign from the HHsuite toolkit (Remmert, Biegert, Hauser, & Söding, 2011). The number of aligned HMM columns was used as an alignment score. All scores for alignments with HHalign probability >85, HHalign *e*-Value $<10^{-5}$, and more than 70% of aligned columns

between the query and the match HMM were stored as an all-against-all matrix. This matrix was clustered into 21,200 VOGs, using the MCL (Markov Clustering) method (Enright, Van Dongen, & Ouzounis, 2002). Based on the manual inspection of the homogeneity of the protein function descriptions in the resulting clusters, we selected the inflation value of 2.0 for the MCL clustering. For all VOG member proteins, we determined the closest homolog in the UniProt database (UniProt Consortium, 2015) from BLAST (Altschul, Gish, Miller, Myers, & Lipman, 1990) hits with E-values better than 10^{-5} and a minimal query coverage of 90%. Functional descriptions of VOGs were automatically derived based on the most frequent protein description found in the UniProt entries or, if not available, in the RefSeq annotation (O'Leary et al., 2016). The complete VOG dataset, which was used in this study, and supplementary files are available for download at http://vogdb.org.

3.3.2. Mapping VOG Sequences to Specific Hosts

We used Virus-Host DB (Mihara et al., 2016) to assign host information to VOG proteins. For 20757 VOGs, we were able to map all contained sequences to a specific host, while 428 VOGs contain proteins from at least one viral species for which we could not find host annotation. Most VOGs include viruses infecting hosts from only one domain of life, i.e., bacteria (~72%), eukaryotes (~22%), or archaea (4%), while only 2% of VOGs are taxonomically mixed (Figure 11). Only 12 VOGs contain viruses that infect hosts from all three domains of life. The VOG sizes range from 15 proteins of 12 distinct species, up to 265 proteins belonging to 261 different species (on average, 104 proteins from 95 different species). These VOGs mostly harbor highly conserved core enzymes of double-stranded DNA viruses, such as kinases, ligases, methylases, helicases, hydrolases, and synthases (Kazlauskas, Krupovic, & Venclovas, 2016). The other two VOGs additionally contain proteins from viruses belonging to the order of Caudovirales, which belong to the bacteriophages, which are not classified as double-stranded DNA viruses, according to the NCBI taxonomy. We excluded from consideration 15 VOGs containing satellite viruses infecting other viruses.

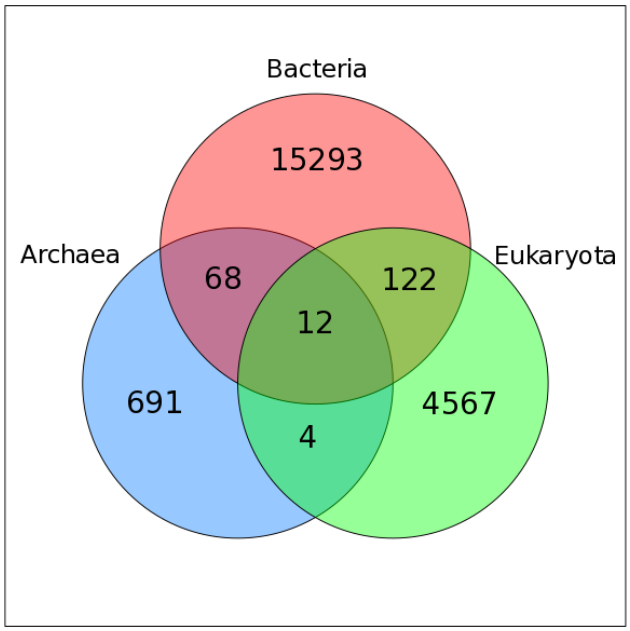


Figure 11: Venn diagram showing the taxonomy of the host organisms within all viral orthologous groups (VOGs). Only those VOGs are included for which host annotation for all viruses is available in the Virus-Host DB.

3.3.3. Distance Trees of VOG Proteins

Expectedly, we found that RNA structure conservation within VOGs decreases with increasing VOG size. Most VOGs (66%) consist of at least three sequences (size distribution shown in Figure 12) and can therefore potentially be split into smaller groups containing structures that are not conserved across the entire VOG. We therefore utilized distance trees derived by the neighbor-joining algorithm (Saitou & Nei, 1987) to identify structurally homogeneous subsets of VOGs (subVOGs). All-against-all pairwise alignments of protein sequences were calculated using Clustal Omega and then converted to the nucleotide alphabet. The distance matrices were derived from pairwise sequence identity values, and the trees were created from the matrices using neighbor joining, as implemented in the BioPerl toolkit (Stajich et al., 2002). The inner nodes of the sequence trees represent possible subVOG candidates, potentially containing structurally homogenous sequences.

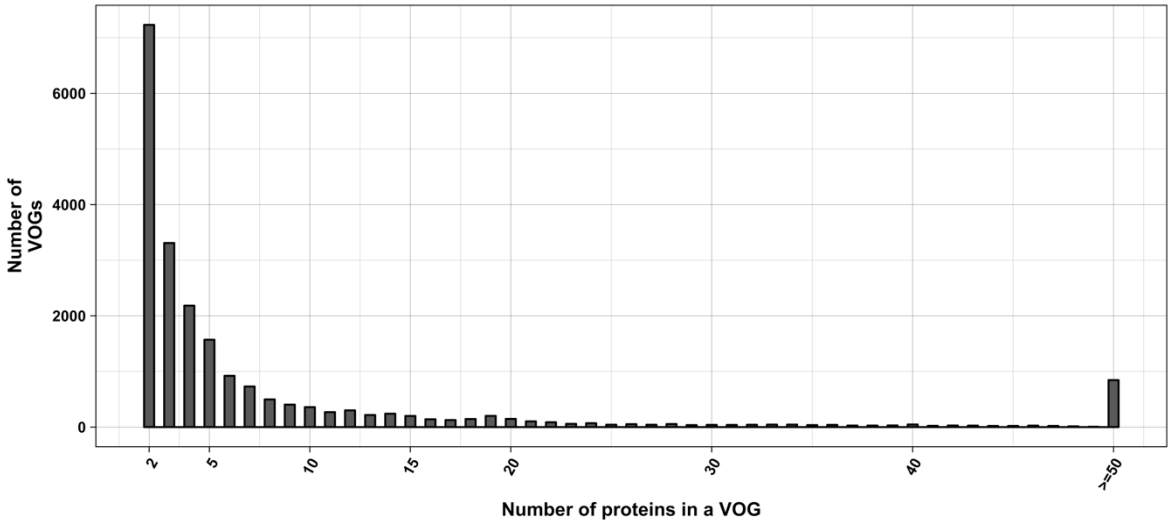


Figure 12: Distribution of VOG sizes.

3.3.4. Structure Prediction and subVOG Assignment

In order to assess the amount of structural RNA conservation present in subVOG candidates, multiple sequence alignments (MSAs) of proteins were calculated for each inner node of the distance trees and converted to the nucleotide alphabet. The RefSeq nucleotide and protein sequences were obtained from the VOGDB. We then employed RNALalifold from the ViennaRNA package (Lorenz et al., 2011), with default parameters, to determine the boundaries of locally stable structures within each MSA, and realigned these local regions using mLocARNA (Will et al., 2012). MLocARNA produces structure-guided multiple sequence alignments, using an adapted version of the Sankoff algorithm. The significance and conservation of the found structures was assessed with RNAz (Gruber et al., 2010). This procedure is simpler and arguably more accurate than the usual approach of applying RNAz to the entire MSA within a sliding window. RNAz classifies fragments of an MSA pre-selected by RNALalifold as containing or not containing a functional RNA secondary structural element. Realignment with mLocARNA significantly increases the precision of RNAz (Gruber et al., 2010). As no sequence of a potential subVOG can be regarded as a reference sequence, the option “no reference” was used for the subsequent RNAz analysis. The RNAz method uses the RNAfold algorithm from the ViennaRNA package to calculate secondary structures and the corresponding minimum free energy (MFE) for each individual RNA sequence in the alignment. In addition, for each aligned sequence set, RNAz calculates a consensus secondary structure and its MFE using the

RNAalifold algorithm. RNAz assumes that conserved and thermodynamically stable structures are functional, in which case it outputs "RNA". Otherwise, it outputs "OTHER". For this purpose, a class probability value, combining all information on an input alignment is calculated. We used a stringent threshold of 0.9 (default 0.5) for the class probability value, which is recommended for finding high confidence structures (Gruber et al., 2010). Subsequently, the trees were scanned for subtrees containing at least one conserved structural element, that is, predicted to be functional, and the largest subtrees were designated as structurally homogenous subVOGs. We found that sequences that are only distantly related according to the neighbor-joining tree may still share conserved RNA structures. In order to account for structure-level relationships between sequences, we built covariance models for all conserved structures found within subVOGs, using the tool cmbuild from the infernal package (Cui, Lu, Wang, Jing-Yan Wang, & Gao, 2016), and used them to search against all sequences in the entire VOG database.

3.3.5. mRNA Stability

Following Tuller et al. (Tuller et al., 2011) and Faure et al. (Faure, Ogurtsov, Shabalina, & Koonin, 2016), we employed RNAfold to calculate the folding energy of the most and the least stable 30-nucleotide segment of mRNAs (ΔGmin and ΔGmax, respectively), as well as the average folding energy of all possible 30 nucleotide segments (ΔGmean). Faure et al. investigated the effect of mRNA stability on the translation rate and protein folding. During translation, the ribosome sequentially unfolds parts of the mRNA. These parts are typically 30 nucleotides long, which explains the choice of segment length in Faure et al. As this procedure does not take into account the actual boundaries of local structures, but rather limits all structures to the size of 30 nucleotides, we additionally calculated the three energy values for all local optimal structures found with RNALfold.

3.3.6. mRNA Structures and Protein Function

We investigated the relationship between protein function, described in terms of gene ontology (GO) annotation (Ashburner et al., 2000), and mRNA structures. Instead of using the global folding energy for classifying mRNAs as highly or lowly structured (Vandivier et al., 2013), we considered structural coverage—the portion of an mRNA covered by functional and conserved structures. GO terms for all VOG proteins were downloaded using QuickGO (Binns et al., 2009), where available. Based on the Evidence & Conclusion

Ontology (ECO) evidence codes (Giglio et al., 2019), two separate datasets were created: (i) Proteins annotated by manually or experimentally derived GO terms (ECO evidence codes: ECO:0000352, ECO:0000269), and (ii) proteins annotated by GO terms with any evidence codes. To find out whether mRNAs of proteins with certain functions tend to harbor more or fewer structures, we pooled together functionally similar GO terms with the average structural coverage of their corresponding mRNAs, using Revigo (Supek, Bošnjak, Škunca, & Šmuc, 2011). Revigo uses a semantic similarity measure to group similar GO terms together, which results in a concise list of distinct functions. To perform this analysis, we calculated the average structural coverage of all subVOG mRNAs with available GO annotation. For the experimental dataset we allowed a coverage value to be associated with a GO term if more than 50% of the sequences in a particular subVOG were annotated with this term. Within the dataset based on all evidence codes, we only allowed GO terms shared by all sequences of a subVOG. We only used mRNAs that were clustered into a subVOG. For sequences that were not part of any subVOG, we did not find conserved structures, although this does not necessarily mean that the mRNA did not contain functional structures. The distributions of standard deviations of the structural coverage values were compared within the actual and randomly generated Revigo clusters. Randomization was performed 1000 times by preserving the size of the clusters and filling them with randomly chosen GO terms.

3.4. Results

3.4.1. Overview of the Study

A graphical overview of the study is given in Figure 13. In a first step, we created distance trees for all protein sequences contained in each VOG, using the neighbor joining method, as described in Materials and Methods. All sequences of the inner nodes of each tree, representing potential subVOGs, were multiply aligned, converted to the nucleotide alphabet and processed with RNALalifold to obtain all potentially conserved local optimal structures. Each part of the alignment covering a potential structure was then realigned with the structure-guided alignment method mLocARNA and checked for functionality using RNAz. The use of structure-guided alignments as input for RNAz improves the performance, compared to pure sequence-based alignments (Gruber et al., 2010). The tree nodes containing the most sequences that yielded conserved structures were taken as final

subVOGs. For all obtained subVOG structures, we computed covariance models that could be used to search for similar structures in future research.

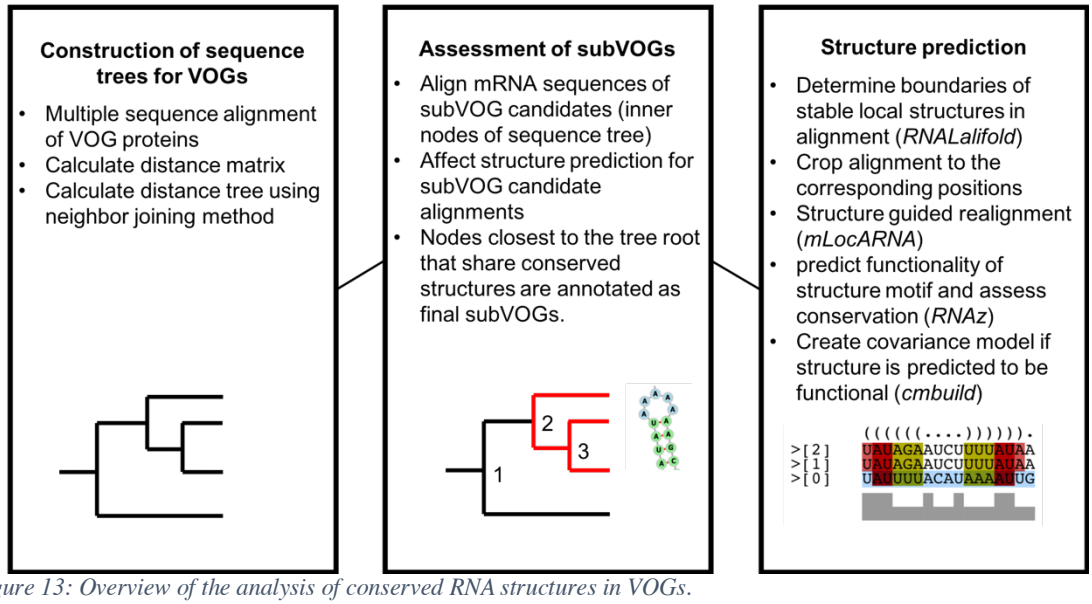


Figure 13: Overview of the analysis of conserved RNA structures in VOGs.

3.4.2. Structure Conservation in VOGs

The current release of the VOG database, derived from the RefSeq release 77, contains 21,200 VOGs, composed of 251,796 proteins from 6252 phages and eukaryotic viruses (Figure S1). Protein sequences in each VOG were aligned by Clustal Omega, converted to the nucleotide alphabet, and used as input for RNA structure prediction by RNALalifold. As seen in Figure 14, the number of local optimal structures conserved within entire VOGs decreases quickly with the number of aligned sequences, which may in part be the consequence of poor multiple alignment quality in large sets of sequences. Indeed, we found that proteins in smaller VOGs tend to be more closely related (Figure S2). To exclude structures found due to sequence conservation only, the potential functionality of structures was verified with RNAz. However, even those VOGs that only consist of a few sequences do not always contain conserved structures. There are 7232 VOGs with exactly two sequences, and for 1237 of these, we could not find any conserved structures. The remaining 5995 VOGs of size two had an average structural coverage of approximately 25% (Figure 15a). Out of the 13,968 VOGs with more than two sequences, 7238 VOGs were predicted to contain RNA structures conserved across the entire VOG, with an average structural coverage of approximately 18% (Figure 15b). These contain between 3

and 96 sequences, with an average of 6. On average, VOGs contain sequences from three different genera, mostly belonging to the same taxonomic family and thus also to the same order (Figure 16a–c). The 25 most diverse VOGs contain sequences from three different orders and up to 19 taxonomic families. On average, a VOG contains mRNAs from viruses that infect hosts from four different genera, belonging to three different taxonomic families and two orders. The VOG with the highest host diversity corresponds to 209 different host genera from 114 families and 64 orders.

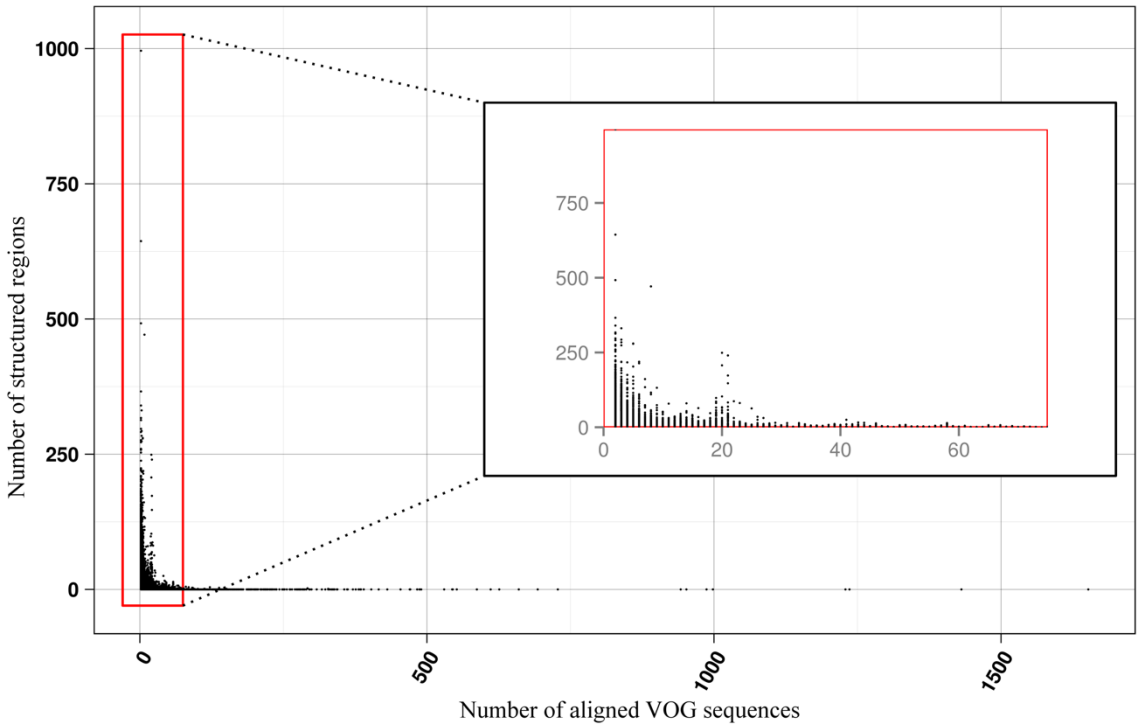
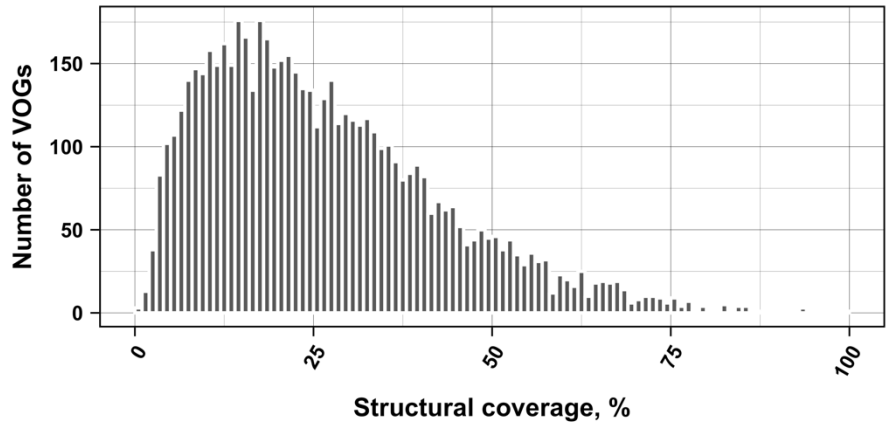
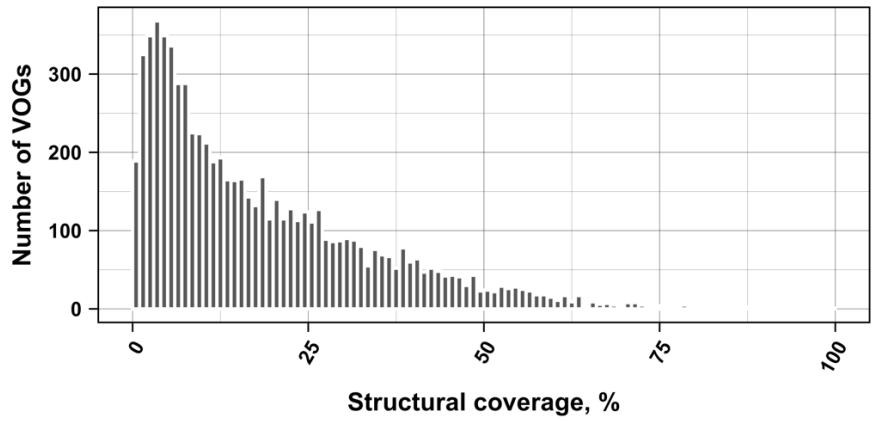


Figure 14: Number of local RNA structures as a function of VOG size.

a)



b)



c)

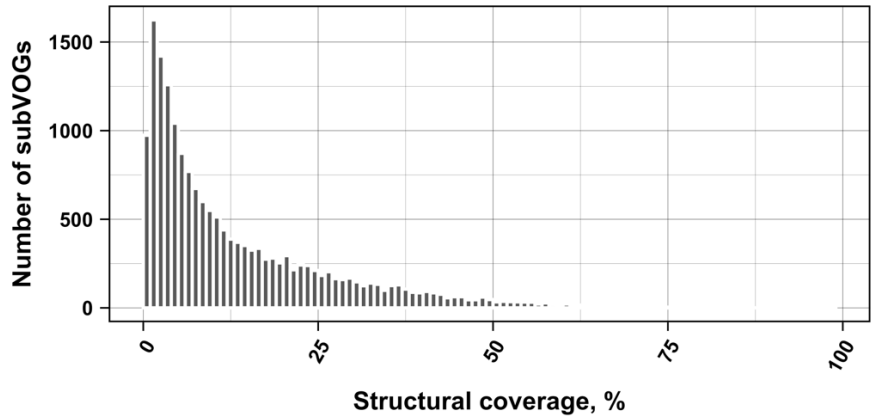


Figure 15: Coverage of VOG alignments by local optimal RNA structures. (a) VOGs with two sequences. (b) VOGs with more than two sequences, in which structures are conserved across all sequences. (c) subVOGs. VOGs that did not contain conserved structures, even after splitting into subVOGs, are not shown.

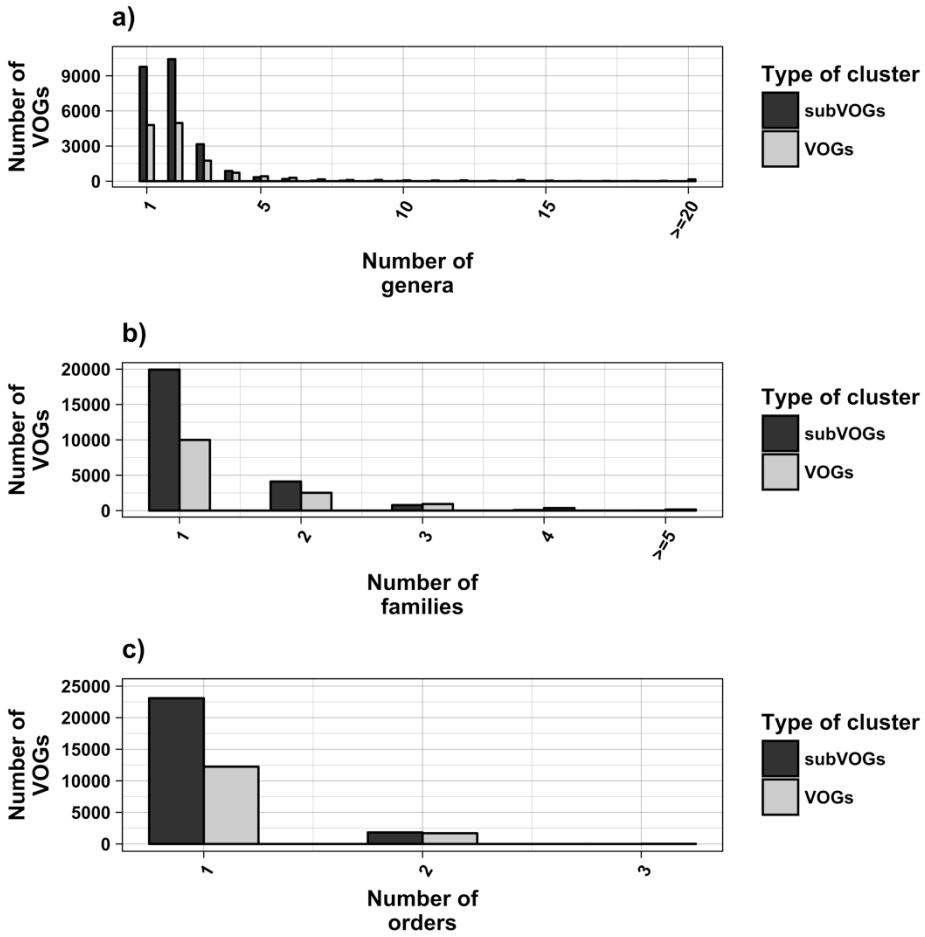


Figure 16: Taxonomic distribution of proteins in VOGs (with more than two sequences) and subVOGs.

3.4.3. Structure Conservation in subVOGs

We attempted to subdivide 6730 VOGs with more than two sequences and without conserved structures into structurally homogeneous subsets, which we call subVOGs, using phylogenetic trees derived by the neighbor-joining method. This procedure resulted in 17,678 subVOGs with an average structural coverage of approximately 13% (Figure 15c). The average number of genera per subVOG is 2 and the most diverse of them contains sequences from three orders and 14 families. A subVOG contains on average sequences that infect two different host genera, and the most diverse subVOG infects hosts of 42 different genera, belonging to 33 families and 20 different orders (Figure 17a–c). Thus, unsurprisingly, subVOGs, which constitute subsets of full VOGs with increased structural homogeneity, exhibit a reduced taxonomic spread, both of the viruses they contain and their hosts. A large fraction of subVOGs (63%) contains sequences from more than one genus and 21% contain sequences from more than one family. The structural coverage of

subVOGs, i.e., the fraction of alignment positions that are located within conserved RNA structures, decreases with increasing taxonomic diversity of the viruses and their hosts (Figure 18). An example that demonstrates the reduction of taxonomic spread between a VOG and its corresponding subVOGs is given in Figure 19. Here, the VOG 00052, which contains 20 proteins from 12 different virus species belonging to 4 different taxonomic families, was split into four structurally homogenous subVOGs. Two of the four subVOGs consist of mRNAs belonging to the genus Avipoxvirus from the family Poxviridae, the third subVOG contains sequences from the family Mimiviridae, and the fourth subVOG consists of two mRNAs belonging to viruses from two different taxonomic families, the Ascoviridae and the Iridoviridae. For two mRNAs, we could not find structures conserved in any of the other VOG members and they are therefore not part of any subVOG.

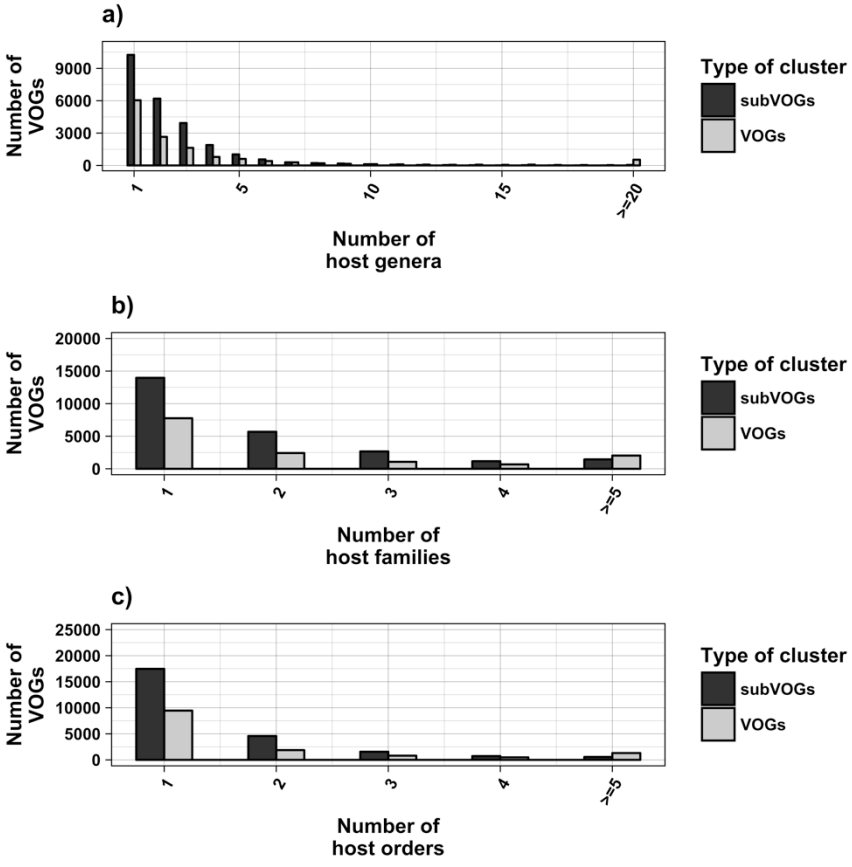


Figure 17: Taxonomic distribution of hosts in VOGs (with more than two sequences) and subVOGs.

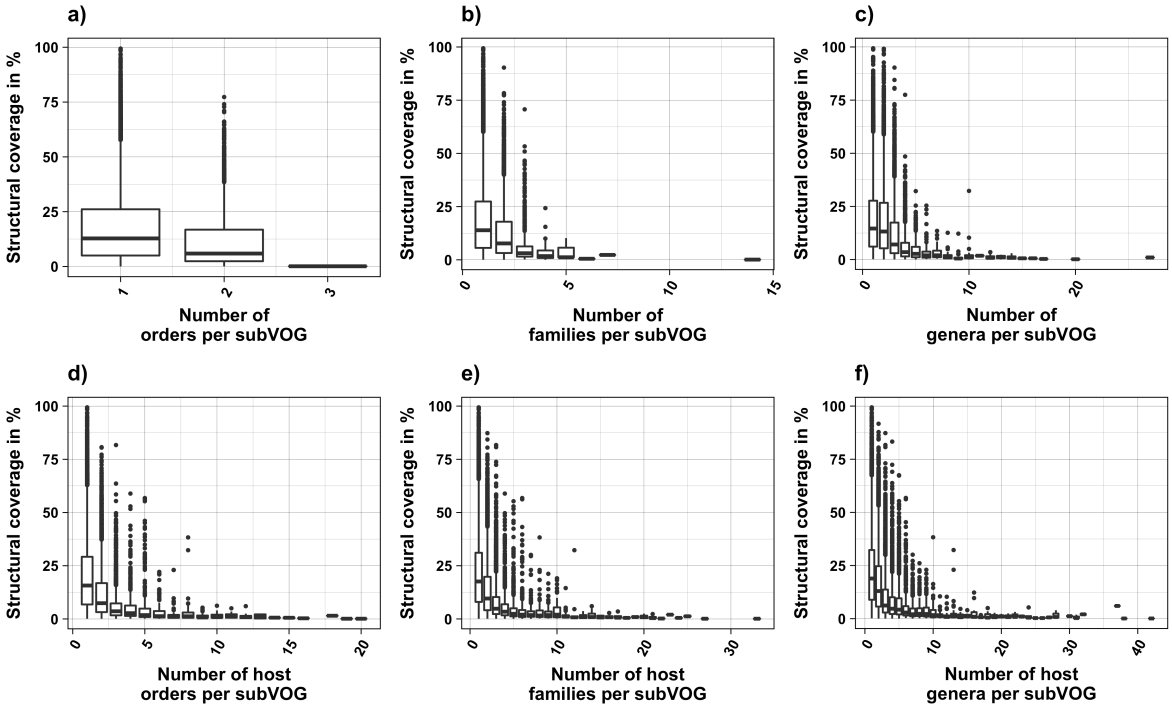


Figure 18: Structural coverage as a function of the taxonomic variety of subVOGs and their host organisms.

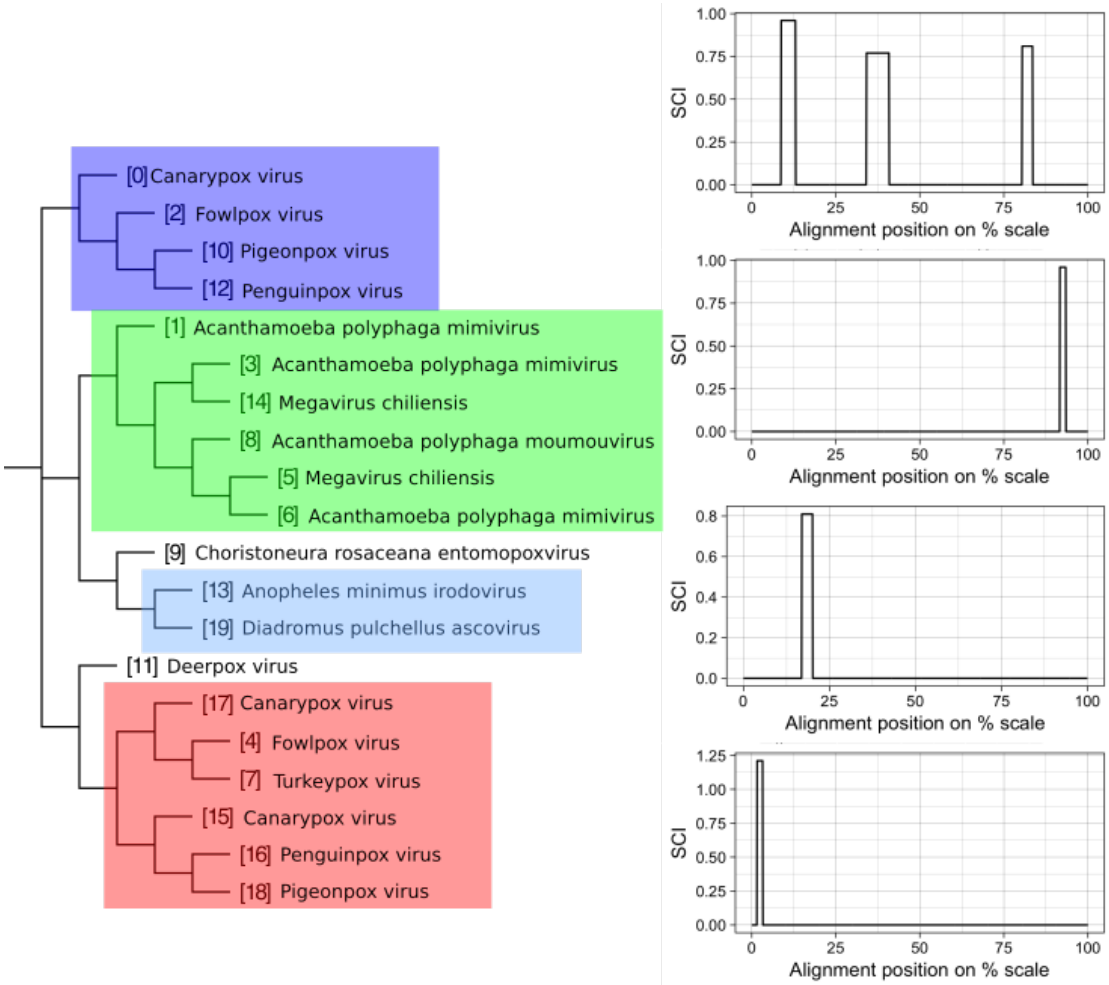


Figure 19: Example of a VOG split into structurally homogenous subVOGs. Shown is the VOG 00052 containing 20 mRNAs, encoding for Kila-N domain proteins, from 12 virus species. On the left, the neighbor-joining tree based on the pairwise sequence identity between the protein sequences is shown. Colored boxes indicate subVOGs, within which conserved structures were predicted. The tree nodes outside colored boxes did not yield any conserved structures. On the right, the structure conservation index (SCI) (black line for each subVOG alignment) is plotted against the alignment position on the percentage scale. Plots are ordered according to the subVOG position in the tree.

As an example, Figure 20 shows the subVOG 1 of VOG11160, which contains two mRNAs encoding the matrix protein 1 from the Influenza A virus (H3N2) and the Influenza B virus. There are three RNA structural motifs described in the literature for the Influenza A mRNA. Nucleotides 105 to 192 form either a multibranch structure, according to Moss et al. (Moss et al., 2011) and Jiang et al. (Jiang, Nogales, Baker, Martinez-Sobrido, & Turner, 2016), or a double hairpin structure, proposed by Jiang et al. (Jiang et al., 2016). Two consecutive stem-loop structures are formed from position 682 to 744, according to Moss et al. (Moss et al., 2011). Despite the sequences' dissimilarity between Influenza A and B, both motifs are partly conserved, according to our RNAz analysis of the corresponding subVOG (Figure 20). Our analysis supports the second hairpin loop from the double hairpin

structure, described by Jiang et al. (Figure 20a–c). From the second motif, proposed by Moss et al., we also found that the second hairpin structure was partly conserved (Figure 20d–e). The consensus structure of the first motif has a high structure conservation index (SCI) of 0.78, although the part of the alignment covering the structure has a low pairwise identity of 29%. The second motif has an SCI of 0.58 and a pairwise identity of 32%. Our analysis also revealed three further conserved stem-loop structures—position 346 to 369, 438 to 483, and 654 to 674, with SCIs and mPIDs of 0.81 and 29%, 0.66 and 48%, and 0.65 and 33%, respectively.

A recent study of secondary structures in alphaviruses by Kutchko et al. revealed that Sindbis virus mRNAs harbor many functional structures, but they are poorly conserved in the closely related Venezuelan equine encephalitis virus (Kutchko et al., 2018). The corresponding subVOG containing mRNAs coding for the non-structural protein 1 includes orthologous mRNAs from 12 further alphaviruses. We identified three short structures that are conserved in all of the contained species and overlap with the functional structures described by Kutchko et al., while all other structures reported by Kutchko et al. are indeed poorly conserved in further Alphavirus species.

An example of a subVOG in which structures are conserved across mRNAs from different taxonomic families is given in Figure 21. Shown is a subVOG containing proteins from two mosaic viruses (Maracuja mosaic virus, Tobacco mosaic virus), the Bell pepper mottle virus, and the Odontoglossum ringspot virus (Figure 21a,b). The proteins are classified as replicases and RNA polymerases. The subVOG contains overall 15 locally conserved structured regions. Figure 21 shows the region covering alignment positions 4766 to 4815. The alignment covering this structure has an mPID of 72% and the structures are conserved with an SCI of 0.9.

Overall, we subdivided 21,200 VOGs containing, on average, 11 proteins (233,380 in total) into a total of 42,293 subVOGs, containing, on average, five mRNAs (201,708 in total) and three structured regions (147,087 in total). The VOGs with more than two sequences that had to be split up contain, on average, four subVOGs.

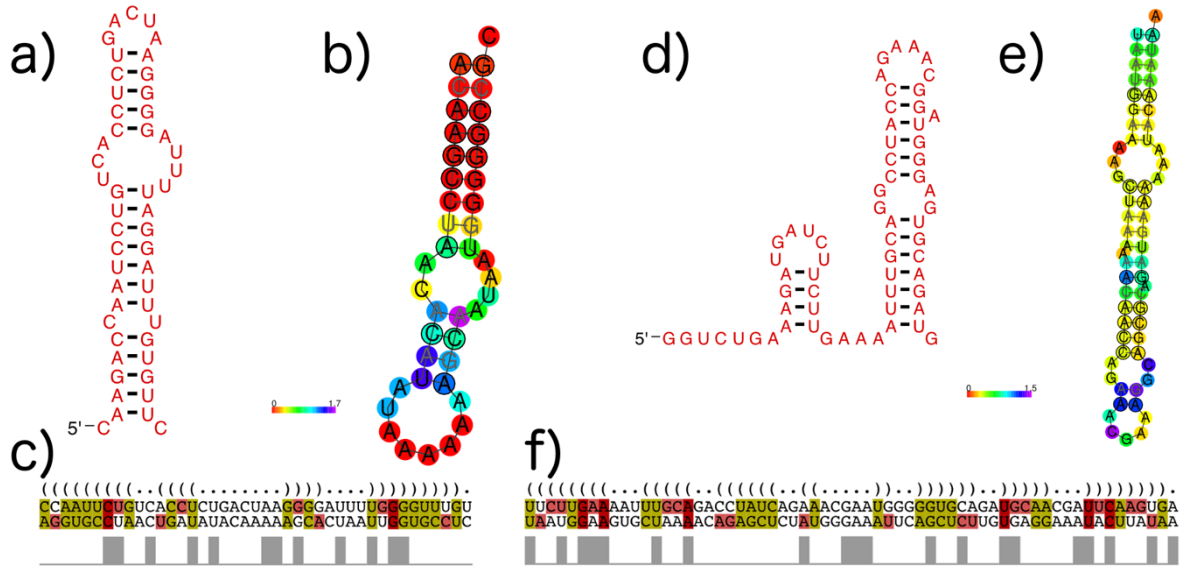


Figure 20: Structures found in Influenza A and B mRNAs encoding the matrix protein (VOG11160). Colors in MSA pictures encode compensatory mutations supporting the consensus structure. Red marks pairs with no sequence variation; ochre, green, turquoise, blue, and violet mark pairs with 2, 3, 4, 5, and 6 different types of pairs, respectively. (a) The second of the two consecutive stem loops of the structure proposed by Jiang et al. [52], covering positions 147–192, visualized with R2R [54]; (b) The predicted conserved consensus structure for nucleotides 148–188 supports the second hairpin loop of the model of Jiang et al., shown in (a). Colors encode the positional entropy; (c) Structure-guided alignment and dot bracket structure notation for the consensus structure shown in (a). The upper sequence corresponds to Influenza A and the lower sequence to Influenza B; (d) Shown are two consecutive hairpin loops for nucleotide positions 682 to 744, proposed by Moss et al. [23], visualized with R2R; (e) The predicted conserved structure for nucleotides 697–758 partly supports the model shown in (e). Colors encode the positional entropy; (f) Structure-guided alignment and dot bracket notation for the consensus structure shown in (e). The upper sequence corresponds to Influenza A and the lower sequence to Influenza B.

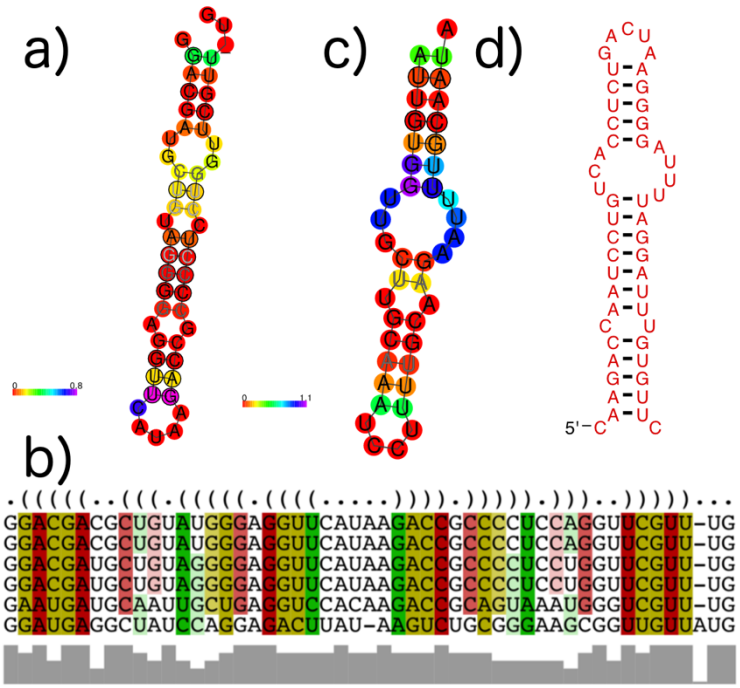


Figure 21: Example structures that were identified within subVOGs. (a) Structural annotation of the subVOG 30, belonging to VOG00029, which contains six mRNAs encoding a replicase protein of different Tobamovirus species. Consensus structure visualized by RNAalifold. Colors encode the positional entropy; (b) Structure-guided MSA and consensus structure in dot bracket notation corresponding to consensus structure shown in (a). Colors encode compensatory mutations supporting the consensus structure. Red marks pairs with no sequence variation; ochre, green, turquoise, blue, and violet mark pairs with 2, 3, 4, 5, and 6 different types of pairs, respectively; (c) Consensus structure of subVOG 64 from VOG00003, which contains four mRNAs coding for a p28-like protein of different alphabaculoviruses; (d) Structure found in a Heliothis virescens ascovirus 3e, by covariance model search of the structure shown in (c), using cmsearch in the entire sequence space of all VOGs.

3.4.4. subVOG Covariance Models

We built covariance models for all structures found within subVOGs and, using cmsearch, found that in many cases, structures are conserved between different subVOGs and even between different VOGs. In most cases, this was due to a shared sequence domain. For example, the subVOG 64 from VOG00003 harbors four mRNA sequences from different nucleopolyhedroviruses, belonging to the family Baculoviridae. This subVOG was predicted to contain four conserved structures. One of these structures is a highly conserved stem-loop structure (Figure 21c). This structure can also be found in an mRNA of Heliothis virescens ascovirus 3e, belonging to the family Ascoviridae, which is part of VOG01276 (Figure 21d). The two structures are highly conserved with an SCI close to 1, although they are part of different VOGs and belong to mRNAs of different virus families. The alignment of the corresponding proteins revealed that these sequences share a common domain, but the sequence similarity is below the inclusion threshold of the VOG pipeline (Figure S3).

3.4.5. mRNA Stability and Length

It was shown for a number of eukaryotic and prokaryotic organisms that longer mRNAs exhibit more stable RNA structures, which allows for more efficient control of co-translational protein folding (Faure et al., 2016; Tuller et al., 2011). In our dataset of viral mRNA sequences, we also found a correlation between the free energy of the most stable 30-nucleotide segment of an mRNA (ΔGmin) and mRNA length (Pearson correlation coefficient –0.27; from here on referred to as Pearson's *r*), but no correlation between the average energy of all possible 30-nucleotide windows (ΔGmean) and mRNA length (Table 5, Figure 22a). We additionally calculated the free energy of the most and least stable local optimal segment found by RNALalifold as well as the mean energy of all found RNALalifold segments, and obtained Pearson's r values of –0.25, –0.07, and 0.29 respectively. The Pearson's r of folding energy and GC content lies between –0.5 for ΔGmax and –0.94 for ΔGmean (Table 5, Figure 22b). The number of bases that are within functional structures is positively correlated with the alignment length of subVOGS (Pearson's r 0.40, *p*-value < 2.2^{-16}), while this correlation becomes negative when considering the percentage of bases within structures (structural coverage) instead of the absolute value (Pearson's r –0.27, *p*-value < 2.2^{-16}) (Figure 23). In other words, longer mRNAs harbor more or longer structured regions, but at the same time, the percentage of positions in functional structures decreases with increasing length. The only explanation for this effect that we can think of is that there is a certain number of structured elements needed for regulatory functions, which is largely independent of the mRNA length. As expected (see Figure 18), there is a weak but significant negative correlation (Pearson's r –0.23, *p*-value < 2.2^{-16}) between structural coverage and the number of sequences in the MSA, with more taxonomically diverse alignments containing fewer conserved structures.

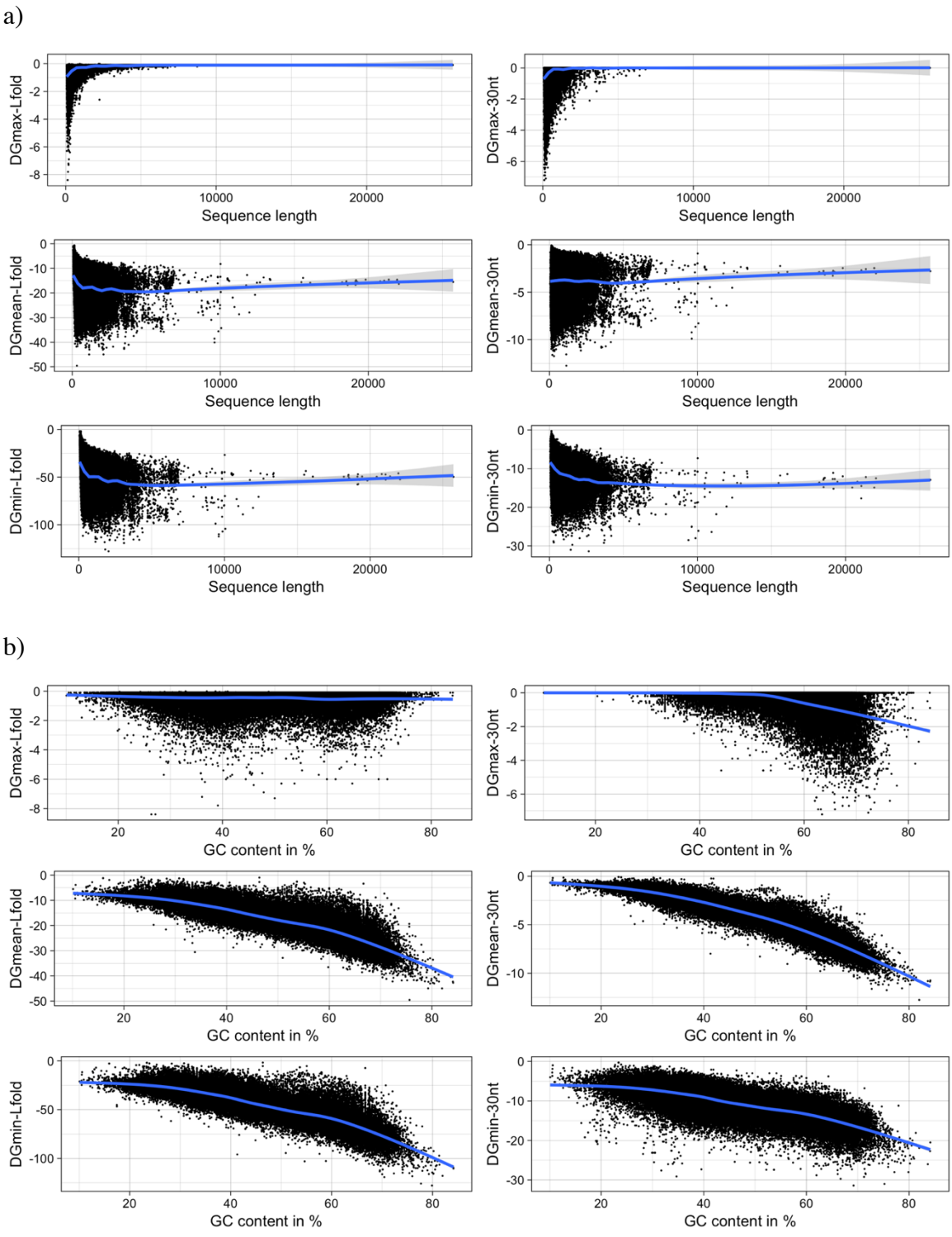


Figure 22: MRNA folding energy as a function of (a) sequence length and (b) GC-content. DGmin: Minimum folding energy of either all possible 30-nucleotide windows of a sequence or all found local optimal structures using RNALfold. DGmean and DGmax: Mean and maximum of all windows, respectively.

Type of ΔG	Pearson correlation coefficient	
	ΔG vs. sequence length	ΔG vs. GC-content
ΔGmin	−0.27 (<2.2^{-16})	−0.73 (<2.2^{-16})
ΔGmean	0.004 (0.1655)	−0.94 (<2.2^{-16})
ΔGmax	0.17 (<2.2^{-16})	−0.50 (<2.2^{-16})
ΔGmin (RNALfold)	−0.24 (<2.2^{-16})	−0.86 (<2.2^{-16})
ΔGmean (RNALfold)	−0.16 (<2.2^{-16})	−0.86 (<2.2^{-16})
ΔGmax (RNALfold)	0.29 (<2.2^{-16})	−0.07 (<2.2^{-16})

Table 5: Pearson correlation between alignment length or GC-content and the minimum (ΔGmin), maximum (ΔGmax), or mean (ΔGmean) folding energy of either all possible 30-nucleotide long-sequence windows or all local optimal structures found with RNALfold, of all mRNAs in our data set. P-values are given in parentheses.

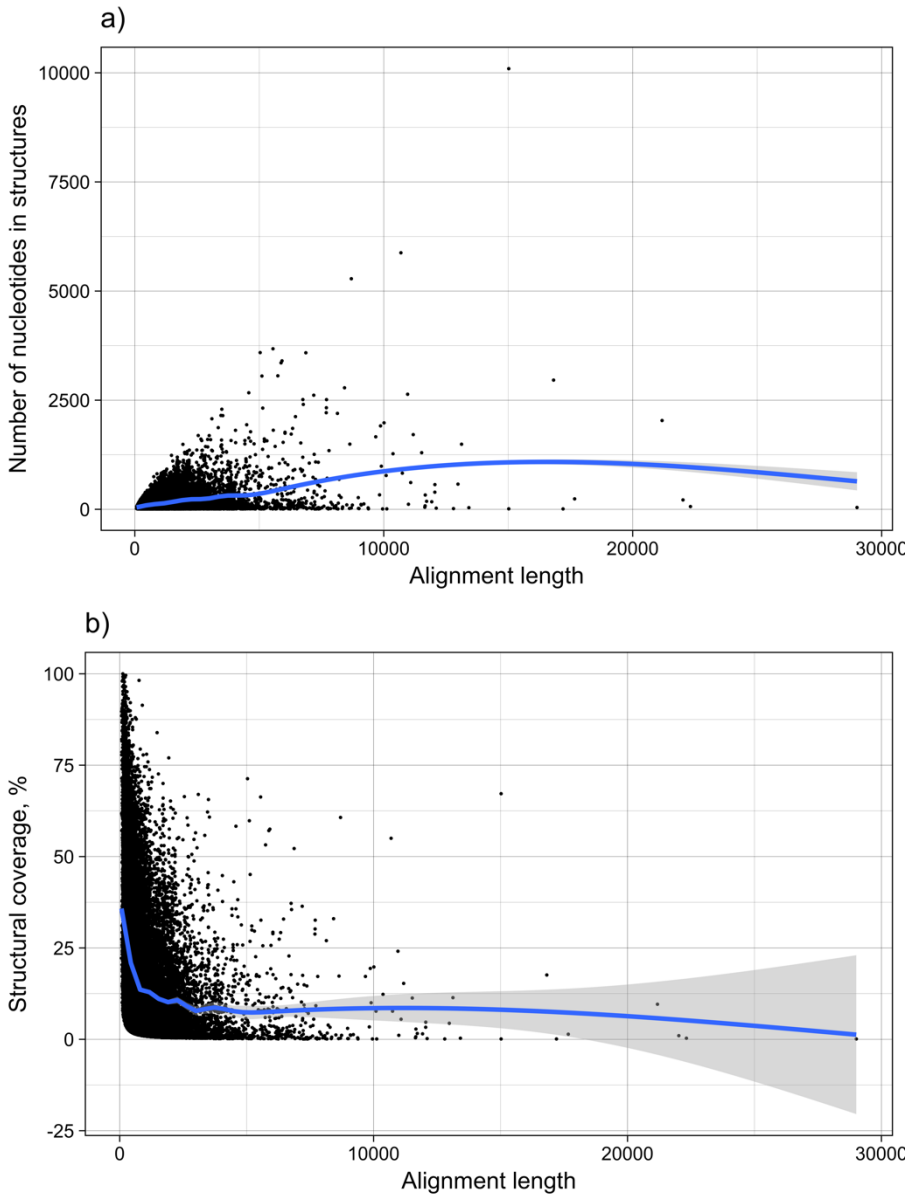


Figure 23: MRNA structure as a function of length. The graph shows the dependence of (a) the number of nucleotides within structures predicted to be functional, and (b) the structural coverage of the mRNAs in %, from the total length of mRNAs. Each point corresponds to one subVOG.

3.4.6. mRNA Structures and Protein Function

We analyzed the relationship between protein function and mRNA structure in viral subVOGs by comparing RNA structural coverage with gene ontology (GO) annotation. Using the QuickGO database, we identified a total of 814 VOG proteins that are manually or experimentally annotated (according to ECO evidence codes, as described in Materials and Methods) with GO terms, of which 727 are part of a subVOG, and thus harbor conserved structures according to our analysis. (For the sake of completeness, we also performed the same analysis for all GO annotated proteins, without regard for the annotation evidence codes, see Table S2). For each individual GO term, we only considered the structural coverage of mRNA sequences if that term was assigned to more than 50% of the proteins in a given subVOG. This resulted in 106 GO terms from the biological process sub-ontology and 17 terms from the molecular function sub-ontology. Note that no GO terms from the cellular component sub-ontology satisfied the criteria explained above.

Using Revigo, we derived 70 functionally similar groups of GO terms, with 57 belonging to the biological process ontology and 13 to the function sub-ontology (Table S1). The resulting GO term groups were subdivided into three categories, according to the average structural coverage of the corresponding subVOGs: Low structural coverage (up to 10%), medium structural coverage (up to 20%), and high structural coverage (more than 20%). We found that the standard deviation of the structural coverage values within the Revigo clusters was significantly smaller (Wilcoxon test p-value 1.068^{-10}), compared to randomized clusters (Figure 24). In other words, our findings suggest that mRNAs encoding the proteins with coherent functions tend to exhibit a similar structural coverage.

These findings are in line with the previous study by Vandivier et al. who found that transcripts in *Arabidopsis thaliana* with similar levels of secondary structure in their untranslated and coding regions tend to encode functionally similar proteins (Vandivier et al., 2013). Likewise, Wang et al. also identified GO terms associated with highly or lowly folded mRNAs in yeast (X. Wang, Li, & Gutenkunst, 2017). Four of the GO terms associated with highly structured mRNAs, according to Wang et al. (regulation of translation, posttranscriptional regulation of gene expression, regulation of cellular protein metabolic process, and cellular nitrogen compound biosynthetic process), correspond to highly structured viral mRNAs in our data. At the same time, none of the GO terms

corresponding to lowly structured yeast mRNAs according to Wang et al. were enriched in our results. On the other hand, Fan Li et al. found that *Arabidopsis thaliana* mRNAs related to "regulation of transcription" were structurally unstable (F. Li et al., 2012), while we found that mRNAs encoding the proteins related to "viral transcription" do harbor conserved RNA structures. We also found virus-specific trends not previously observed for cellular proteins, such as the high structure of viral mRNAs coding for proteins that regulate replication and transcription, suppression by viruses of host translation, or modulation by viruses of host process (Table S1). It has been reported that mRNA folding strength influences the efficiency of gene expression and that mRNAs encoding abundant proteins generally tend to be more structured (Zur & Tuller, 2012). In the future, once RNA-seq data for a sufficient number of viral genes becomes available, it will be interesting to investigate whether functional coherence between mRNAs with similar structural coverage is, at least in part, caused by similar expression levels.

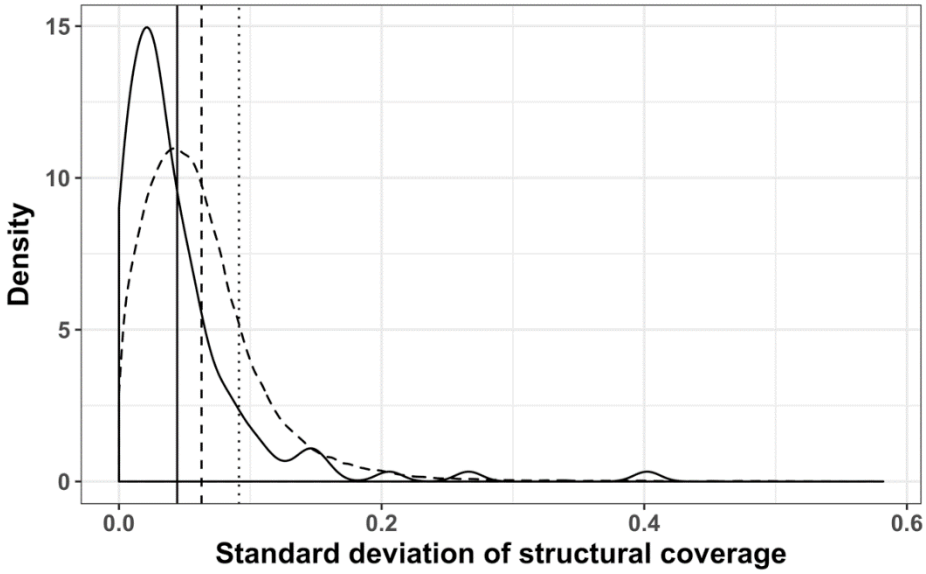


Figure 24: Distribution of standard deviations of mRNA structural coverage, mapped to GO-terms: Clustered with Revigo (solid line); randomized Revigo clusters (dashed line); not clustered (dotted line); vertical lines represent the mean of the corresponding dataset.

3.4.7. subVOG Online Resource

Structurally homogenous subVOGs can be accessed online (http://rnasiv.bio.wzw.tum.de) through two entry points: "Browse by VOG" and "Browse by taxonomy". The first option is a list of all VOGs, together with the consensus description of their constituent proteins. The list can be filtered with a keyword search and links to the corresponding subVOGs of

each VOG are provided. The second option is an expandable taxonomic tree, based on the NCBI taxonomy (Federhen, 2012), which allows navigation to the viral species of interest. For each species, mRNA sequences are provided, if available, interlinked to the corresponding subVOGs. Tree nodes containing only mRNAs that are not part of any subVOG are colored grey. Each subVOG contains at least two sequences that share at least one structural element predicted to be functional. If a species of interest is not contained in the subVOG database, the taxonomy tree makes it possible to find the taxonomically closest species. Web pages describing individual subVOGs contain four parts:

i) General information, i.e., number of mRNAs in the subVOG, the number of proteins and species in the parent VOG, as well as a consensus functional description;

ii) Information on conserved structures among the subVOG sequences. A plot outlining the SCI for each column of the subVOG MSA gives a brief overview over the structure of the subVOG members. Also provided is a table that shows a list of all structures found, including the corresponding values of SCI, mPID, and the GC content. The consensus structure can also be visualized by Forna, and a covariance model is provided, which can be used to search for similar structures. Additionally, the RNAz results for each individual structured region can be accessed, including structure visualization, dot plots, and the local structure-guided alignments;

iii) The global MSA for the subVOG sequences. Alignment columns colored in blue correspond to the structured regions described in the previous section. The alignment is visualized with the javascript library MSAviewer (Yachdav et al., 2016), which is based on Jalview (Waterhouse, Procter, Martin, Clamp, & Barton, 2009);

iv) The list of subVOG members, including protein names, descriptions, and taxonomy. For each protein, a link to the RefSEQ entry is provided, as well as the amino acid and nucleotide sequences. The leftmost column of the list contains a checkbox for each subVOG member, which can be used to build a subset of members and analyze the RNA structures shared by these.

3.5. Discussion

In this work we set out to create a possibly complete census of conserved RNA secondary structures in the coding regions of viruses and to shed light on their biological role. Using

sequence comparison and structure prediction methods, we derived structurally homogenous groups of viral mRNAs from subsets of viral orthologous groups (VOGs), which we call subVOGs. We identified a total of 147,087 conserved structures in 42,293 subVOGs, which we make accessible through our database RNASIV (RNA Structures in Viruses). On average, subVOGs contain three structured regions and their structural homogeneity decreases with increasing taxonomic diversity of the viruses and their hosts. We found that 63% of all subVOGs contain mRNAs from at least two genera and 21% from more than one taxonomic family. In line with the previous studies on cellular organisms, we confirm that, in viruses, longer mRNAs tend to contain more stable structures. However, the number of structures grows only slowly with length, which implies that there is a certain minimum amount of structures required to maintain regulatory functions and control protein folding. MRNAs annotated with similar GO terms tend to have a similar structural coverage, hinting at possible commonalities in the regulatory mechanisms of functionally related proteins. It is hoped that RNASIV will be a useful resource for exploring the structure–function relationships in viral mRNAs.

4. References

Agius, P., Bennett, K. P., & Zuker, M. (2010). Comparing RNA secondary structures using a relaxed base-pair score. *RNA (New York, N.Y.)*, *16*(5), 865–878. http://doi.org/10.1261/rna.903510

Albariño, C. G., Bird, B. H., & Nichol, S. T. (2007). A shared transcription termination signal on negative and ambisense RNA genome segments of Rift Valley fever, sandfly fever Sicilian, and Toscana viruses. *Journal of Virology*, *81*(10), 5246–5256. http://doi.org/10.1128/JVI.02778-06

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. http://doi.org/10.1016/S0022-2836(05)80360-2

Ashbrook, S. E., Griffin, J. M., & Johnston, K. E. (2018). Recent Advances in Solid-State Nuclear Magnetic Resonance Spectroscopy. *Annual Review of Analytical Chemistry (Palo Alto, Calif.)*, *11*(1), 485–508. http://doi.org/10.1146/annurev-anchem-061417-125852

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genetics*, *25*(1), 25–29. http://doi.org/10.1038/75556

Auperin, D. D., Galinski, M., & Bishop, D. H. (1984). The sequences of the N protein gene and intergenic region of the S RNA of pichinde arenavirus. *Virology*, *134*(1), 208–219.

Auperin, D. D., Sasso, D. R., & McCormick, J. B. (1986). Nucleotide sequence of the glycoprotein gene and intergenic region of the Lassa virus S genome RNA. *Virology*, *154*(1), 155–167.

Barr, J. N. (2007). Bunyavirus mRNA synthesis is coupled to translation to prevent premature transcription termination. *RNA (New York, N.Y.)*, *13*(5), 731–736. http://doi.org/10.1261/rna.436607

Bândea, C. I. (1983). A new theory on the origin and the nature of viruses. *Journal of Theoretical Biology*, *105*(4), 591–602.

Bevilacqua, P. C., & Blose, J. M. (2008). Structures, kinetics, thermodynamics, and biological functions of RNA hairpins. *Annual Review of Physical Chemistry*, *59*(1), 79–103. http://doi.org/10.1146/annurev.physchem.59.032607.093743

Bhaskaran, H., Rodriguez-Hernandez, A., & Perona, J. J. (2012). Kinetics of tRNA folding monitored by aminoacylation. *RNA (New York, N.Y.)*, *18*(3), 569–580. http://doi.org/10.1261/rna.030080.111

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., & Apweiler, R. (2009). QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics (Oxford, England)*, *25*(22), 3045–3046. http://doi.org/10.1093/bioinformatics/btp536

Brennan, B., Rezelj, V. V., & Elliott, R. M. (2017). Mapping of Transcription Termination within the S Segment of SFTS Phlebovirus Facilitated Generation of NSs Deletant Viruses. *Journal of Virology*, *91*(16), e00743–17. http://doi.org/10.1128/JVI.00743-17

Briese, T., Paweska, J. T., McMullan, L. K., Hutchison, S. K., Street, C., Palacios, G., et al. (2009). Genetic detection and characterization of Lujo virus, a new hemorrhagic fever-associated arenavirus from southern Africa. *PLoS Pathogens*, *5*(5), e1000455. http://doi.org/10.1371/journal.ppat.1000455

Bursch, W., Oberhammer, F., & Schulte-Hermann, R. (1992). Cell death by apoptosis and its protective role against disease. *Trends in Pharmacological Sciences*, *13*(6), 245–251.

Capriotti, E., & Marti-Renom, M. A. (2010). Quantifying the relationship between sequence and three-dimensional structure conservation in RNA. *BMC Bioinformatics*, *11*(1), 322. http://doi.org/10.1186/1471-2105-11-322

Carlini, D. B., Chen, Y., & Stephan, W. (2001). The relationship between third-codon position nucleotide content, codon bias, mRNA secondary structure and gene expression in the drosophilid alcohol dehydrogenase genes Adh and Adhr. *Genetics*, *159*(2), 623–633.

Chen, C., Huang, H., Mazumder, R., Natale, D. A., McGarvey, P. B., Zhang, J., et al. (2016). Computational clustering for viral reference proteomes. *Bioinformatics (Oxford, England)*, *32*(13), 2041–2043. http://doi.org/10.1093/bioinformatics/btw110

Chen, J.-L., & Greider, C. W. (2005). Functional analysis of the pseudoknot structure in human telomerase RNA. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(23), 8080–5– discussion 8077–9. http://doi.org/10.1073/pnas.0502259102

Chursov, A., Walter, M. C., Schmidt, T., Mironov, A., Shneider, A., & Frishman, D. (2012). Sequence-structure relationships in yeast mRNAs. *Nucleic Acids Research*, *40*(3), 956–962. http://doi.org/10.1093/nar/gkr790

Clabbers, M. T. B., Olsthoorn, R. C. L., & Gultyaev, A. P. (2014). Tospovirus ambisense genomic RNA segments use almost complete repertoire of stable tetraloops in the intergenic region. *Bioinformatics (Oxford, England)*, *30*(13), 1800–1804. http://doi.org/10.1093/bioinformatics/btu122

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2015). GenBank. *Nucleic Acids Research*, gkv1276. http://doi.org/10.1093/nar/gkv1276

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. *Journal of Virology*, *80*(5), 2170–2182. http://doi.org/10.1128/JVI.80.5.2170-2182.2006

Cohen, F. S. (2016). How Viruses Invade Cells. *Biophysical Journal*, *110*(5), 1028–1032. http://doi.org/10.1016/j.bpj.2016.02.006

Cui, X., Lu, Z., Wang, S., Jing-Yan Wang, J., & Gao, X. (2016). CMsearch: simultaneous exploration of protein sequence space and structure space improves not only protein homology detection but also protein structure prediction. *Bioinformatics (Oxford, England)*, *32*(12), i332–i340. http://doi.org/10.1093/bioinformatics/btw271

Darty, K., Denise, A., & Ponty, Y. (2009). VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics (Oxford, England)*, *25*(15), 1974–1975. http://doi.org/10.1093/bioinformatics/btp250

de Haan, P., Wagemakers, L., Peters, D., & Goldbach, R. (1990). The S RNA segment of tomato spotted wilt virus has an ambisense character. *The Journal of General Virology*, *71 (Pt 5)*(5), 1001–1007. http://doi.org/10.1099/0022-1317-71-5-1001

Deigan, K. E., Li, T. W., Mathews, D. H., & Weeks, K. M. (2009). Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(1), 97–102. http://doi.org/10.1073/pnas.0806929106

Del Campo, C., Bartholomäus, A., Fedyunin, I., & Ignatova, Z. (2015). Secondary Structure across the Bacterial Transcriptome Reveals Versatile Roles in mRNA Regulation and Function. *PLoS Genetics*, *11*(10), e1005613. http://doi.org/10.1371/journal.pgen.1005613

Deng, J., Xiong, Y., & Sundaralingam, M. (2001). X-ray analysis of an RNA tetraplex (UGGGGU)(4) with divalent Sr(2+) ions at subatomic resolution (0.61 A). *Proceedings of the National Academy of Sciences of the United*

States of America, *98*(24), 13665–13670. http://doi.org/10.1073/pnas.241374798
Ding, Y., & Lawrence, C. E. (2003). A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, *31*(24), 7280–7301. http://doi.org/10.1093/nar/gkg938
Díez, J., & Jungfleisch, J. (2017). Translation control: Learning from viruses, again. *RNA Biology*, *14*(7), 835–837. http://doi.org/10.1080/15476286.2017.1325068
Duan, J., Wainwright, M. S., Comeron, J. M., Saitou, N., Sanders, A. R., Gelernter, J., & Gejman, P. V. (2003). Synonymous mutations in the human dopamine receptor D2 (DRD2) affect mRNA stability and synthesis of the receptor. *Human Molecular Genetics*, *12*(3), 205–216.
Duda, R. O., Hart, P. E., & Stork, D. G. (2012). Pattern Classification. John Wiley & Sons.
Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics*, *23*(1), 205–211.
Ehresmann, C., Baudin, F., Mougel, M., Romby, P., Ebel, J. P., & Ehresmann, B. (1987). Probing the structure of RNAs in solution. *Nucleic Acids Research*, *15*(22), 9109–9128. http://doi.org/10.1093/nar/15.22.9109
Elliott, R. M., & Brennan, B. (2014). Emerging phleboviruses. *Current Opinion in Virology*, *5*, 50–57. http://doi.org/10.1016/j.coviro.2014.01.011
EMBOSS: The European Molecular Biology Open Software Suite. (2000). EMBOSS: The European Molecular Biology Open Software Suite, 1–2.
Emery, V. C., & Bishop, D. H. (1987). Characterization of Punta Toro S mRNA species and identification of an inverted complementary sequence in the intergenic region of Punta Toro phlebovirus ambisense S RNA that is involved in mRNA transcription termination. *Virology*, *156*(1), 1–11.
Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, *30*(7), 1575–1584.
Faure, G., Ogurtsov, A. Y., Shabalina, S. A., & Koonin, E. V. (2016). Role of mRNA structure in the control of protein folding. *Nucleic Acids Research*, *44*(22), 10898–10911. http://doi.org/10.1093/nar/gkw671
Federhen, S. (2012). The NCBI Taxonomy database. *Nucleic Acids Research*, *40*(Database issue), D136–43. http://doi.org/10.1093/nar/gkr1178
Ferron, F., Weber, F., la Torre, de, J. C., & Reguera, J. (2017). Transcription and replication mechanisms of Bunyaviridae and Arenaviridae L proteins. *Virus Research*. http://doi.org/10.1016/j.virusres.2017.01.018
Fiore, A. E., Bridges, C. B., & Cox, N. J. (2009). Seasonal influenza vaccines. *Current Topics in Microbiology and Immunology*, *333*(Suppl), 43–82. http://doi.org/10.1007/978-3-540-92165-3_3

Fontana, W., Konings, D. A., Stadler, P. F., & Schuster, P. (1993). Statistics of RNA secondary structures. *Biopolymers*, *33*(9), 1389–1404. http://doi.org/10.1002/bip.360330909

Forterre, P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. *Virus Research*, *117*(1), 5–16. http://doi.org/10.1016/j.virusres.2006.01.010

Fricke, M., Dünnes, N., Zayas, M., Bartenschlager, R., Niepmann, M., & Marz, M. (2015). Conserved RNA secondary structures and long-range interactions in hepatitis C viruses. *RNA (New York, N.Y.)*, *21*(7), 1219–1232. http://doi.org/10.1261/rna.049338.114

Fricke, M., Gerst, R., Ibrahim, B., Niepmann, M., & Marz, M. (2019). Global importance of RNA secondary structures in protein-coding sequences. *Bioinformatics (Oxford, England)*, *35*(4), 579–583. http://doi.org/10.1093/bioinformatics/bty678

Fürtig, B., Richter, C., Wöhnert, J., & Schwalbe, H. (2003). NMR spectroscopy of RNA. *Chembiochem : a European Journal of Chemical Biology*, *4*(10), 936–962. http://doi.org/10.1002/cbic.200300700

Gardner, P. P., & Giegerich, R. (2004). A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, *5*(1), 140. http://doi.org/10.1186/1471-2105-5-140

Garmann, R. F., Gopal, A., Athavale, S. S., Knobler, C. M., Gelbart, W. M., & Harvey, S. C. (2015). Visualizing the global secondary structure of a viral RNA genome with cryo-electron microscopy. *RNA (New York, N.Y.)*, *21*(5), 877–886. http://doi.org/10.1261/rna.047506.114

Ghiringhelli, P. D., Rivera-Pomar, R. V., Lozano, M. E., Grau, O., & Romanowski, V. (1991). Molecular organization of Junin virus S RNA: complete nucleotide sequence, relationship with other members of the Arenaviridae and unusual secondary structures. *The Journal of General Virology*, *72 (Pt 9)*(9), 2129–2141. http://doi.org/10.1099/0022-1317-72-9-2129

Giegerich, R., Voss, B., & Rehmsmeier, M. (2004). Abstract shapes of RNA. *Nucleic Acids Research*, *32*(16), 4843–4851. http://doi.org/10.1093/nar/gkh779

Giglio, M., Tauber, R., Nadendla, S., Munro, J., Olley, D., Ball, S., et al. (2019). ECO, the Evidence & Conclusion Ontology: community standard for evidence information. *Nucleic Acids Research*, *47*(D1), D1186–D1194. http://doi.org/10.1093/nar/gky1036

Golden, B. L., & Kundrot, C. E. (2003). RNA crystallization. *Journal of Structural Biology*, *142*(1), 98–107.

Gopal, A., Zhou, Z. H., Knobler, C. M., & Gelbart, W. M. (2012). Visualizing large RNA molecules in solution. *RNA (New York, N.Y.)*, *18*(2), 284–299. http://doi.org/10.1261/rna.027557.111

Goz, E., & Tuller, T. (2015). Widespread signatures of local mRNA folding structure selection in four Dengue virus serotypes. *BMC Genomics*, *16 Suppl 10*(S10), S4. http://doi.org/10.1186/1471-2164-16-S10-S4

Goz, E., & Tuller, T. (2016). Evidence of a Direct Evolutionary Selection for Strong Folding and Mutational Robustness Within HIV Coding Regions. *Journal of Computational Biology : a Journal of Computational Molecular Cell Biology*, *23*(8), 641–650. http://doi.org/10.1089/cmb.2016.0052

Gray, N. K., & Hentze, M. W. (1994). Regulation of protein synthesis by mRNA structure. *Molecular Biology Reports*, *19*(3), 195–200. http://doi.org/10.1007/BF00986961

Gruber, A. R., Bernhart, S. H., Hofacker, I. L., & Washietl, S. (2008). Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, *9*(1), 122. http://doi.org/10.1186/1471-2105-9-122

Gruber, A. R., Findeiß, S., Washietl, S., Hofacker, I. L., & Stadler, P. F. (2010). RNAz 2.0: improved noncoding RNA detection. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 69–79. http://doi.org/10.1142/9789814295291_0009

Gu, W., Zhou, T., & Wilke, C. O. (2010). A universal trend of reduced mRNA stability near the translation-initiation site in prokaryotes and eukaryotes. *PLoS Computational Biology*, *6*(2), e1000664. http://doi.org/10.1371/journal.pcbi.1000664

Han, K., & Byun, Y. (2003). PSEUDOVIEWER2: Visualization of RNA pseudoknots of any type. *Nucleic Acids Research*, *31*(13), 3432–3440. http://doi.org/10.1093/nar/gkg539

Hofacker, I. L. (2007). RNA consensus structure prediction with RNAalifold. *Methods in Molecular Biology (Clifton, N.J.)*, *395*, 527–544.

Hofacker, I. L., & Stadler, P. F. (2006). Memory efficient folding algorithms for circular RNA secondary structures. *Bioinformatics (Oxford, England)*, *22*(10), 1172–1176. http://doi.org/10.1093/bioinformatics/btl023

Hogeweg, P., & Hesper, B. (1984). Energy directed folding of RNA sequences. *Nucleic Acids Research*, *12*(1 Pt 1), 67–74. http://doi.org/10.1093/nar/12.1part1.67

Holley, R. W. (1965). Structure of an Alanine Transfer Ribonucleic Acid. *Jama*, *194*(8), 868–871. http://doi.org/10.1001/jama.1965.03090210032009

Hulo, C., de Castro, E., Masson, P., Bougueleret, L., Bairoch, A., Xenarios, I., & Le Mercier, P. (2011). ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Research*, *39*(Database issue), D576–82. http://doi.org/10.1093/nar/gkq901

Ikegami, T., Won, S., Peters, C. J., & Makino, S. (2007). Characterization of Rift Valley fever virus transcriptional terminations. *Journal of Virology*, *81*(16), 8421–8438. http://doi.org/10.1128/JVI.02641-06

Ilyinskii, P. O., Schmidt, T., Lukashev, D., Meriin, A. B., Thoidis, G., Frishman, D., & Shneider, A. M. (2009). Importance of mRNA secondary structural elements for the expression of influenza virus genes. *Omics : a Journal of Integrative Biology*, *13*(5), 421–430. http://doi.org/10.1089/omi.2009.0036

Incarnato, D., Neri, F., Anselmi, F., & Oliviero, S. (2014). Genome-wide profiling of mouse RNA secondary structures reveals key features of the mammalian

transcriptome. *Genome Biology*, *15*(10), 491. http://doi.org/10.1186/s13059-014-0491-2

Jiang, T., Nogales, A., Baker, S. F., Martinez-Sobrido, L., & Turner, D. H. (2016). Mutations Designed by Ensemble Defect to Misfold Conserved RNA Structures of Influenza A Segments 7 and 8 Affect Splicing and Attenuate Viral Replication in Cell Culture. *PloS One*, *11*(6), e0156906. http://doi.org/10.1371/journal.pone.0156906

Johansson, J., Mandin, P., Renzoni, A., Chiaruttini, C., Springer, M., & Cossart, P. (2002). An RNA thermosensor controls expression of virulence genes in Listeria monocytogenes. *Cell*, *110*(5), 551–561.

Kaczkowski, B., Torarinsson, E., Reiche, K., Havgaard, J. H., Stadler, P. F., & Gorodkin, J. (2009). Structural profiles of human miRNA families from pairwise clustering. *Bioinformatics (Oxford, England)*, *25*(3), 291–294. http://doi.org/10.1093/bioinformatics/btn628

Katz, L., & Burge, C. B. (2003). Widespread selection for local RNA secondary structure in coding regions of bacterial genes. *Genome Research*, *13*(9), 2042–2051. http://doi.org/10.1101/gr.1257503

Kazlauskas, D., Krupovic, M., & Venclovas, Č. (2016). The logic of DNA replication in double-stranded DNA viruses: insights from global analysis of viral genomes. *Nucleic Acids Research*, *44*(10), 4551–4564. http://doi.org/10.1093/nar/gkw322

Ke, A., & Doudna, J. A. (2004). Crystallization of RNA and RNA-protein complexes. *Methods (San Diego, Calif.)*, *34*(3), 408–414. http://doi.org/10.1016/j.ymeth.2004.03.027

Kerpedjiev, P., Hammer, S., & Hofacker, I. L. (2015). Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics (Oxford, England)*, *31*(20), 3377–3379. http://doi.org/10.1093/bioinformatics/btv372

Komarova, N. L. (2007). Viral reproductive strategies: How can lytic viruses be evolutionarily competitive? *Journal of Theoretical Biology*, *249*(4), 766–784. http://doi.org/10.1016/j.jtbi.2007.09.013

Koonin, E. V., Senkevich, T. G., & Dolja, V. V. (2006). The ancient Virus World and evolution of cells. *Biology Direct*, *1*(1), 29. http://doi.org/10.1186/1745-6150-1-29

Koonin, E. V., Senkevich, T. G., & Dolja, V. V. (2009). Compelling reasons why viruses are relevant for the origin of cells. *Nature Reviews. Microbiology*, *7*(8), 615–author reply 615. http://doi.org/10.1038/nrmicro2108-c5

Kortmann, J., & Narberhaus, F. (2012). Bacterial RNA thermometers: molecular zippers and switches. *Nature Reviews. Microbiology*, *10*(4), 255–265. http://doi.org/10.1038/nrmicro2730

Kozak, M. (2005). Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, *361*, 13–37. http://doi.org/10.1016/j.gene.2005.06.037

Kristensen, D. M., Kannan, L., Coleman, M. K., Wolf, Y. I., Sorokin, A., Koonin, E. V., & Mushegian, A. (2010). A low-polynomial algorithm for assembling

clusters of orthologous groups from intergenomic symmetric best matches. *Bioinformatics (Oxford, England)*, *26*(12), 1481–1487. http://doi.org/10.1093/bioinformatics/btq229

Kudla, G., Murray, A. W., Tollervey, D., & Plotkin, J. B. (2009). Coding-sequence determinants of gene expression in Escherichia coli. *Science (New York, N.Y.)*, *324*(5924), 255–258. http://doi.org/10.1126/science.1170160

Kutchko, K. M., Madden, E. A., Morrison, C., Plante, K. S., Sanders, W., Vincent, H. A., et al. (2018). Structural divergence creates new functional features in alphavirus genomes. *Nucleic Acids Research*, *46*(7), 3657–3670. http://doi.org/10.1093/nar/gky012

Lara, E., Billecocq, A., Leger, P., & Bouloy, M. (2011). Characterization of wild-type and alternate transcription termination signals in the Rift Valley fever virus genome. *Journal of Virology*, *85*(23), 12134–12145. http://doi.org/10.1128/JVI.05322-11

Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., Orton, R. J., Siddell, S. G., & Smith, D. B. (2018). Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Research*, *46*(D1), D708–D717. http://doi.org/10.1093/nar/gkx932

Lemey, P., Salemi, M., & Vandamme, A.-M. (2009). The Phylogenetic Handbook. Cambridge University Press.

Li, F., Zheng, Q., Vandivier, L. E., Willmann, M. R., Chen, Y., & Gregory, B. D. (2012). Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. *The Plant Cell*, *24*(11), 4346–4359. http://doi.org/10.1105/tpc.112.104232

Lilliefors, H. W. (1967). ON THE KOLMOGOROV-SMIRNOV TEST FOR NORMALITY WITH MEAN AND VARIANCE UNKNOWN. *Journal of the American Statistical Association*, 1–5.

Liu, H., Fu, Y., Jiang, D., Li, G., Xie, J., Cheng, J., et al. (2010). Widespread horizontal gene transfer from double-stranded RNA viruses to eukaryotic nuclear genomes. *Journal of Virology*, *84*(22), 11876–11887. http://doi.org/10.1128/JVI.00955-10

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for Molecular Biology : AMB*, *6*(1), 26. http://doi.org/10.1186/1748-7188-6-26

López, N., & Franze-Fernández, M. T. (2007). A single stem-loop structure in Tacaribe arenavirus intergenic region is essential for transcription termination but is not required for a correct initiation of transcription and replication. *Virus Research*, *124*(1-2), 237–244. http://doi.org/10.1016/j.virusres.2006.10.007

Mailliot, J., & Martin, F. (2018). Viral internal ribosomal entry sites: four classes for one goal. *Wiley Interdisciplinary Reviews. RNA*, *9*(2), e1458. http://doi.org/10.1002/wrna.1458

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, *29*(6-7), 1105–1119. http://doi.org/10.1002/bip.360290621

MELNICK, J. L. (1971). Classification and nomenclature of viruses, 1971. *Prog Med Virology*, *13*, 462–484.
Merino, E. J., Wilkinson, K. A., Coughlan, J. L., & Weeks, K. M. (2005). RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *Journal of the American Chemical Society*, *127*(12), 4223–4231. http://doi.org/10.1021/ja043822v
Meyer, I. M., & Miklós, I. (2005). Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Research*, *33*(19), 6338–6348. http://doi.org/10.1093/nar/gki923
Mihara, T., Nishimura, Y., Shimizu, Y., Nishiyama, H., Yoshikawa, G., Uehara, H., et al. (2016). Linking Virus Genomes with Host Taxonomy. *Viruses*, *8*(3), 66. http://doi.org/10.3390/v8030066
Miranda-Ríos, J. (2007). The THI-box riboswitch, or how RNA binds thiamin pyrophosphate. *Structure (London, England : 1993)*, *15*(3), 259–265. http://doi.org/10.1016/j.str.2007.02.001
Mortimer, S. A., Kidwell, M. A., & Doudna, J. A. (2014). Insights into RNA structure and function from genome-wide studies. *Nature Reviews. Genetics*, *15*(7), 469–479. http://doi.org/10.1038/nrg3681
Moss, W. N., Priore, S. F., & Turner, D. H. (2011). Identification of potential conserved RNA secondary structure throughout influenza A coding regions. *RNA (New York, N.Y.)*, *17*(6), 991–1011. http://doi.org/10.1261/rna.2619511
Moy, R. H., Cole, B. S., Yasunaga, A., Gold, B., Shankarling, G., Varble, A., et al. (2014). Stem-loop recognition by DDX17 facilitates miRNA processing and antiviral defense. *Cell*, *158*(4), 764–777. http://doi.org/10.1016/j.cell.2014.06.023
Münz, C. (2017). The Autophagic Machinery in Viral Exocytosis. *Frontiers in Microbiology*, *8*, 269. http://doi.org/10.3389/fmicb.2017.00269
Nackley, A. G., Shabalina, S. A., Tchivileva, I. E., Satterfield, K., Korchynskyi, O., Makarov, S. S., et al. (2006). Human catechol-O-methyltransferase haplotypes modulate protein expression by altering mRNA secondary structure. *Science (New York, N.Y.)*, *314*(5807), 1930–1933. http://doi.org/10.1126/science.1131262
Nasir, A., Kim, K. M., & Caetano-Anollés, G. (2012). Viral evolution: Primordial cellular origins and late adaptation to parasitism. *Mobile Genetic Elements*, *2*(5), 247–252. http://doi.org/10.4161/mge.22797
NCBI Resource Coordinators. (2015). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, gkv1290. http://doi.org/10.1093/nar/gkv1290
Nguyen, M., & Haenni, A.-L. (2003). Expression strategies of ambisense viruses. *Virus Research*, *93*(2), 141–150.
Nozinovic, S., Fürtig, B., Jonker, H. R. A., Richter, C., & Schwalbe, H. (2010). High-resolution NMR structure of an RNA model system: the 14-mer

cUUCGg tetraloop hairpin RNA. *Nucleic Acids Research*, *38*(2), 683–694. http://doi.org/10.1093/nar/gkp956

O'Leary, N. A., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research*, *44*(D1), D733–45. http://doi.org/10.1093/nar/gkv1189

Olivier, C., Poirier, G., Gendron, P., Boisgontier, A., Major, F., & Chartrand, P. (2005). Identification of a conserved RNA motif essential for She2p recognition and mRNA localization to the yeast bud. *Molecular and Cellular Biology*, *25*(11), 4752–4766. http://doi.org/10.1128/MCB.25.11.4752-4766.2005

Olson, J. K., & Grose, C. (1997). Endocytosis and recycling of varicella-zoster virus Fc receptor glycoprotein gE: internalization mediated by a YXXL motif in the cytoplasmic tail. *Journal of Virology*, *71*(5), 4042–4054.

Pinschewer, D. D., Perez, M., & la Torre, de, J. C. (2005). Dual role of the lymphocytic choriomeningitis virus intergenic region in transcription termination and virus propagation. *Journal of Virology*, *79*(7), 4519–4526. http://doi.org/10.1128/JVI.79.7.4519-4526.2005

Pirakitikulr, N., Kohlway, A., Lindenbach, B. D., & Pyle, A. M. (2016). The Coding Region of the HCV Genome Contains a Network of Regulatory RNA Structures. *Molecular Cell*, *62*(1), 111–120. http://doi.org/10.1016/j.molcel.2016.01.024

Pornillos, O., Garrus, J. E., & Sundquist, W. I. (2002). Mechanisms of enveloped RNA virus budding. *Trends in Cell Biology*, *12*(12), 569–579.

Pruitt, K. D., Tatusova, T., Brown, G. R., & Maglott, D. R. (2012). NCBI Reference Sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research*, *40*(Database issue), D130–5. http://doi.org/10.1093/nar/gkr1079

Reguera, J., Cusack, S., & Kolakofsky, D. (2014). Segmented negative strand RNA virus nucleoprotein structure. *Current Opinion in Virology*, *5*, 7–15. http://doi.org/10.1016/j.coviro.2014.01.003

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*, *9*(2), 173–175. http://doi.org/10.1038/nmeth.1818

Rietveld, K., Van Poelgeest, R., Pleij, C. W., Van Boom, J. H., & Bosch, L. (1982). The tRNA-like structure at the 3' terminus of turnip yellow mosaic virus RNA. Differences and similarities with canonical tRNA. *Nucleic Acids Research*, *10*(6), 1929–1946. http://doi.org/10.1093/nar/10.6.1929

Romanowski, V., & Bishop, D. H. (1985). Conserved sequences and coding of two strains of lymphocytic choriomeningitis virus (WE and ARM) and Pichinde arenavirus. *Virus Research*, *2*(1), 35–51.

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425.

Sankoff, D. (2006). Simultaneous Solution of the RNA Folding, Alignment and Protosequence Problems. *SIAM Journal on Applied Mathematics*, *45*(5), 810–825. http://doi.org/10.1137/0145048

Sashital, D. G., & Butcher, S. E. (2006). Flipping off the riboswitch: RNA structures that control gene expression. *ACS Chemical Biology*, *1*(6), 341–345. http://doi.org/10.1021/cb6002465

Selecting New RNA Crystal Contacts. (2018). Selecting New RNA Crystal Contacts. *Structure (London, England : 1993)*, *26*(9), 1166–1167. http://doi.org/10.1016/j.str.2018.08.009

Serganov, A., & Patel, D. J. (2007). Ribozymes, riboswitches and beyond: regulation of gene expression without proteins. *Nature Reviews. Genetics*, *8*(10), 776–790. http://doi.org/10.1038/nrg2172

Shabalina, S. A., Ogurtsov, A. Y., & Spiridonov, N. A. (2006). A periodic pattern of mRNA secondary structure created by the genetic code. *Nucleic Acids Research*, *34*(8), 2428–2437. http://doi.org/10.1093/nar/gkl287

Shoffner, G. M., Wang, R., Podell, E., Cech, T. R., & Guo, F. (2018). In Crystallo Selection to Establish New RNA Crystal Contacts. *Structure (London, England : 1993)*, *26*(9), 1275–1283.e3. http://doi.org/10.1016/j.str.2018.05.005

Sievers, F., & Higgins, D. G. (2014). Clustal omega. *Current Protocols in Bioinformatics*, *48*, 3.13.1–16. http://doi.org/10.1002/0471250953.bi0313s48

Simmonds, P., & Smith, D. B. (1999). Structural constraints on RNA virus evolution. *Journal of Virology*, *73*(7), 5787–5794.

Simmonds, P., Tuplin, A., & Evans, D. J. (2004). Detection of genome-scale ordered RNA structure (GORS) in genomes of positive-stranded RNA viruses: Implications for virus evolution and host persistence. *RNA (New York, N.Y.)*, *10*(9), 1337–1351. http://doi.org/10.1261/rna.7640104

Soldatov, R. A., Vinogradova, S. V., & Mironov, A. A. (2014). RNASurface: fast and accurate detection of locally optimal potentially structured RNA segments. *Bioinformatics (Oxford, England)*, *30*(4), 457–463. http://doi.org/10.1093/bioinformatics/btt701

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Research*, *12*(10), 1611–1618. http://doi.org/10.1101/gr.361602

Steffen, P., Voss, B., Rehmsmeier, M., Reeder, J., & Giegerich, R. (2006). RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics (Oxford, England)*, *22*(4), 500–503. http://doi.org/10.1093/bioinformatics/btk010

Stewart, S. A., Poon, B., Song, J. Y., & Chen, I. S. (2000). Human immunodeficiency virus type 1 vpr induces apoptosis through caspase activation. *Journal of Virology*, *74*(7), 3105–3111. http://doi.org/10.1128/jvi.74.7.3105-3111.2000

Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS One*, *6*(7), e21800. http://doi.org/10.1371/journal.pone.0021800

Thorley, J. A., McKeating, J. A., & Rappoport, J. Z. (2010). Mechanisms of viral entry: sneaking in the front door. *Protoplasma*, *244*(1-4), 15–24. http://doi.org/10.1007/s00709-010-0152-6

Thurner, C., Witwer, C., Hofacker, I. L., & Stadler, P. F. (2004). Conserved RNA secondary structures in Flaviviridae genomes. *The Journal of General Virology*, *85*(Pt 5), 1113–1124. http://doi.org/10.1099/vir.0.19462-0

Tijerina, P., Mohr, S., & Russell, R. (2007). DMS footprinting of structured RNAs and RNA-protein complexes. *Nature Protocols*, *2*(10), 2608–2623. http://doi.org/10.1038/nprot.2007.380

Tuller, T., Veksler-Lublinsky, I., Gazit, N., Kupiec, M., Ruppin, E., & Ziv-Ukelson, M. (2011). Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biology*, *12*(11), R110. http://doi.org/10.1186/gb-2011-12-11-r110

UniProt Consortium. (2015). UniProt: a hub for protein information. *Nucleic Acids Research*, *43*(Database issue), D204–12. http://doi.org/10.1093/nar/gku989

Vandivier, L. E., Anderson, S. J., Foley, S. W., & Gregory, B. D. (2016). The Conservation and Function of RNA Secondary Structure in Plants. *Annual Review of Plant Biology*, *67*(1), 463–488. http://doi.org/10.1146/annurev-arplant-043015-111754

Vandivier, L., Li, F., Zheng, Q., Willmann, M., Chen, Y., & Gregory, B. (2013). Arabidopsis mRNA secondary structure correlates with protein function and domains. *Plant Signaling & Behavior*, *8*(6), e24301. http://doi.org/10.4161/psb.24301

Wang, X., Li, P., & Gutenkunst, R. N. (2017). Systematic Effects Of mRNA Secondary Structure On Gene Expression And Molecular Function In Budding Yeast. http://doi.org/10.1101/138792

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics (Oxford, England)*, *25*(9), 1189–1191. http://doi.org/10.1093/bioinformatics/btp033

Watts, J. M., Dang, K. K., Gorelick, R. J., Leonard, C. W., Bess, J. W., Swanstrom, R., et al. (2009). Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature*, *460*(7256), 711–716. http://doi.org/10.1038/nature08237

Weinberg, Z., & Breaker, R. R. (2011). R2R--software to speed the depiction of aesthetic consensus RNA secondary structures. *BMC Bioinformatics*, *12*(1), 3. http://doi.org/10.1186/1471-2105-12-3

Wienholds, E., & Plasterk, R. H. A. (2005). MicroRNA function in animal development. *FEBS Letters*, *579*(26), 5911–5922. http://doi.org/10.1016/j.febslet.2005.07.070

Will, S., Joshi, T., Hofacker, I. L., Stadler, P. F., & Backofen, R. (2012). LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. *RNA (New York, N.Y.)*, *18*(5), 900–914. http://doi.org/10.1261/rna.029041.111

Will, S., Reiche, K., Hofacker, I. L., Stadler, P. F., & Backofen, R. (2007). Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Computational Biology*, *3*(4), e65. http://doi.org/10.1371/journal.pcbi.0030065

Wilson, S. M., & Clegg, J. C. (1991). Sequence analysis of the S RNA of the African arenavirus Mopeia: an unusual secondary structure feature in the intergenic region. *Virology*, *180*(2), 543–552.

Wuerth, J. D., & Weber, F. (2016). Phleboviruses and the Type I Interferon Response. *Viruses*, *8*(6), 174. http://doi.org/10.3390/v8060174

Yachdav, G., Wilzbach, S., Rauscher, B., Sheridan, R., Sillitoe, I., Procter, J., et al. (2016). MSAViewer: interactive JavaScript visualization of multiple sequence alignments. *Bioinformatics (Oxford, England)*, *32*(22), 3501–3503. http://doi.org/10.1093/bioinformatics/btw474

Zuker, M., & Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, *9*(1), 133–148. http://doi.org/10.1093/nar/9.1.133

Zur, H., & Tuller, T. (2012). Strong association between mRNA folding strength and protein abundance in S. cerevisiae. *EMBO Reports*, *13*(3), 272–277. http://doi.org/10.1038/embor.2011.262