Conserved RNA Secondary Structures in Viral Genomes: A Survey

Ivo L. Hofacker[1], Peter F. Stadler[1,2], Roman Stocsits[1]

[1]Institut für Theoretische Chemie und Molekulare Strukturbiologie Universität Wien, Währingerstraße 17, A-1090 Wien, Austria
Phone: ++43 1 4277 52738; Fax: ++43 1 4277 52793;
Email: ivo@tbi.univie.ac.at.
[2]Bioinformatik, Institut für Informatik, Universität Leipzig, Kreuzstraße 7b, D-04103 Leipzig, Germany

The genomes of RNA viruses often carry conserved RNA structures that perform vital functions during the life cycle of the virus. Such structures can be detected using a combination of structure prediction and co-variation analysis. Here we present results from pilot studies on a variety of viral families performed during bioinformatics computer lab courses in past years.

1. Introduction

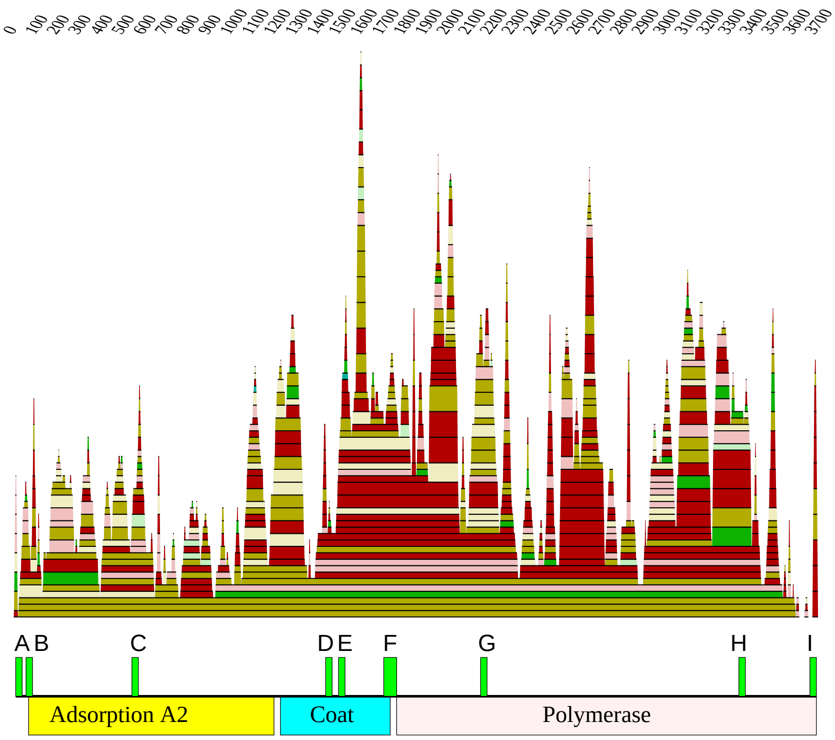
In addition to coding for proteins, the genomes of RNA viruses often contain functionally active RNA structures that are vital during the various stages of the viral life cycle. In many cases only a small part of the viral genome is functionally relevant at the level of RNA. Detecting these motifs is a difficult and tedious task since the secondary structures of such regions in general do not look significantly different from structures formed by random sequences.

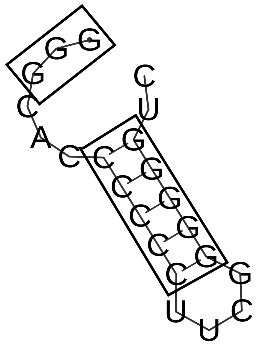
RNA secondary structures are however very fragile to accumulation of point mutations: computer simulations [11] showed that a small number of point mutations is very likely to cause large changes in the secondary structures: mutations in some 10% of the sequence positions already lead almost certainly to unrelated structures if the mutated positions are chosen randomly. Secondary structure elements that are consistently present in a group of sequences with less than, say 95%, average pairwise identity are therefore most likely the result of stabilizing selection, not a consequence of the high degree of sequence conservation.

This fact is exploited by the `alidot` algorithm for a systematic search of conserved secondary structure patterns in large RNA. In brief, independent predictions of the secondary structure for each of the sequences and a multiple sequence alignment that is obtained without any reference to the predicted secondary structures are combined to a list of homologous base pairs. This list is then sorted by means of hierarchical credibility criteria that explicitly take both thermodynamic information and information on sequence covariation into account. A detailed description of the method can be found in [6, 7], the programs are available from `http://www.tbi.univie.ac.at/RNA/` as part of the *Vienna RNA Package*.

Detailed analysis of the genomic secondary features are available for picornaviridae [16], flaviviridae [14], hepadnaviridae [13, 8]. In this contribution we report on preliminary surveys of a variety of virus families.

A compilation of conserved cis-acting **sequences** curated by Henry V. Huang can be found at `http://www.microbiology.wustl.edu/dept/fac/huang/ccas/`.





	clustal	code2aln
A	—	1-150
B	80-110	80-110
C	540-575	545-615
D	1480-1495	1427-1451
E	1520-1590	1560-1640
F	1785-1855	—
G	—	2250-2310
H	3405-3435	3470-3540
I	—	3690-3717

Figure 1. The Hogeweg mountain plot on the l.h.s. gives an overview of the potentially conserved secondary structures of the complete RNA genome of Levivirus. The underlying alignment is obtained using `code2aln` [12]. In the genome chart below credible RNA structures and positions of the open reading frames are indicated. Position numbers of these elements (compiled in the small table on the r.h.s., refer to the alignment available at `http://rna.tbi.univie.ac.at/virus/`. The 5'-end structure A is homologous to the 5'-terminal RNA replicase recognition site of the allolevivirus Q_β where it is required for the recognition of the template and its amplification. Element F could correspond to the coat protein binding site that is located just upstream of the polymerase gene in Q_β.

2. Leviviridae

Leviviridae are small ssRNA phages without envelop and tail. The replication cycle includes no DNA stage. The total genome length is 3466 up to 4276 nucleotides depending on type of strain. Most Levivirus species have four (partly) overlapping genes, some exceptions exist which contain only three open reading frames. Currently, two structural virion proteins have been found and identified.

We have investigated 8 sequences of the Levivirus genus. Currently, there are 5 types of strains represented in the `GenBank` data base as complete genomes: The Enterobacteria phages `MS2`, `KU1`, `GA`, `fr`, and `AP205`. The last one was eliminated from the data set analyzed here because of large changes in genomic structure.

At the 5'-terminal end of the Levivirus sequences we detect a short GC-rich hairpin (tetraloop) which follows to an unpaired GGG element, see Figure 1. This is probably the analogon to the recognition signal site for the RNA replicase in Alloleviviruses, a stem-loop-structure that is well known and defined in Q_β. The Q_β replicase amplifies RNA templates autocatalytically with high efficiency. This recognition element, consisting of a hairpin and a very short unpaired region, at the 5'-terminus, is essential for function [1].

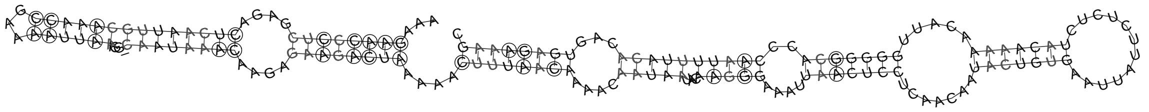


Figure 2. Common RNA structure of Reston Ebola and Marburgvirus in the region from 13100-13270 of their `clustalw` alignment. The signal was detected by `qrna` then re-analyzed using `alidot`. Circles mark sequence variation supporting the structure.

3. Negative-Stranded RNA Viruses

The latest version of the Virus Taxonomy [2] classifies the families Paramyxoviridae, Rhabdoviridae, Filoviridae, and Bornaviridae as members of the order Mononegavirales. The families Orthomyxoviridae, Bunyaviridae and Arenaviridae remain unassigned.

Paramyxoviridae have genome sizes of about 15-16kb. The only unambiguous RNA structure is a panhandle structure that is highly conserved. In a addition, there are few weak signals for specific small RNA features in the genera Respirovirus and Morbillivirus that deserve a more detailed analysis.

Rhabdoviridae have a genome size of about 12kb. No unambiguous signals have been detected. In the genus rhabdovirus conserved gene-end and start sequences have been reported [15] which might be associated also with conserved structural features.

Filoviridae. Complete sequences are available only for Reston and Zaire Ebola and for the Marburg virus. These three sequences are insufficient for the `alidot` approach. The program `qrna` [10] was therefore used to identify possible conserved RNA features in a pairwise alignment. An example is shown in Fig. 2.

Bunyaviridae exhibit panhandle structures that are confirmed by the existence of compensatory mutations. Neither the genus Hantavirus nor the genus Bunyavirus (which has a genome consisting of three distinct chromosomes) exhibit convincing additional RNA secondary structure motifs in either the $(+)$ or the $(-)$-strand although there are some small hairpins containing a few consistent and compensatory mutations are reported in [5].

Orthomyxoviridae have genomes that contain a moderate number of small segments (8 in the case of the influenza virus) of small segments. As for most negative stranded RNA viruses our computations confirm the well-known panhandle structures [3]. A hairpin loop structure at both the 5' and the 3' end (that competes with the panhandle and hence does not show in our data) is required for efficient endonuclease activity of influenza virus RNA polymerase, an activity that is required for the cap-snatching activity of primers from host pre-mRNA [9].

Arenaviridae also exhibit panhandle structures that have been repeatedly reported in the literature. Conserved intergenic regions containing poly-U stretches are reminiscent of Rhabdoviridae. We did not find evidence that these sequences form conserved secondary structures.

Panhandle structures appear to be the only common RNA secondary structure features of negative stranded RNA viruses despite at least weak indications for additional conserved features in some groups.

4. Positive-Stranded RNA Viruses

The survey of positive-stranded RNA Viruses is far from complete at this point. Apart from the families picornaviridae [16] and flaviviridae [14] which have been studied in detail, preliminary data are available for three additional unrelated groups.

Dicistroviridae are an only recently described family of arthodpod viruses. The genome contains a 5'-untranslated region of 500-800nt followed by two ORFs of around 5500nt and 2600nt separated by an untranslated intergenic region (IGR) of approximately 190nt. This di-cistronic structure of the genome is special among RNA viruses. The IGR between the replicase and the four capsid proteins contains an IRES region while the 5'-end contains another IRES motif despite the fact that translation is fMet initiated. Preliminary data indicate that the IGR-IRES is structurally well conserved.

Comoviridae A number of conserved elements with a larger number of compensatory mutations can be found within the three genera Comovirus, Fabavirus, and Nepovirus. A hairpin structure with a well-conserved stem and a relatively large A/U-rich loop region (pos 1257-1271 of the nepovirus alignment) is reminiscent of the cis-acting replication element (CRE) of picornaviridae. The close similarity in the life cycle of picornaviridae and comoviridae lends credibility to this conjecture.

Potyviridae have genomes of almost 10000nt consisting of a single ORF, coding for a polyprotein, and short UTRs at the 5' and 3' ends. A number of interesting features were found, especially the 3'UTR seems to contain several well-conserved structures.

5. Retroviridae

Retroviridae, in contrast, exhibit a large number of functional RNA secondary structures. In the case of the lentivirus HIV some of them are very well characterized both experimentally and computationally, such as the TAR hairpin at the 5' end, the gag/pol frameshift hairpin (also present in Mammalian Type B Retrovirus) and the RRE structure.

Our pilot study yield indications for functional RNA structures throughout the family. There seems to be little similarity in the motifs used by different genera, however. A comparison between different genera is difficult because the low sequence homology does not permit reliable alignments. A much more detailed analysis will be necessary to determine whether some secondary motifs are shared between all Retroviridae.

6. Discussion

By combining thermodynamic structure prediction with co-variation analysis conserved RNA secondary structures in viral RNA genomes can be reliably detected. The analysis works best if at least 5 sequences with pairwise identities of around 80% are available. For many types of viruses such sequence data are now available, and we have begun a systematic survey of viral RNA structures, the results of which are made available in an on-line database at `http://rna.tbi.univie.ac.at/virus/`.

All RNA viruses studied so far contain at least some functional RNA structures. These conserved structures occur particularly often in the 5'- and 3'-UTRs, but even within coding regions conserved structures are not rare.

To our surprise, the bibliographic search for experimental evidence of and further information on "RNA secondary structures" in a given group of virus — a seemingly rather straightforward task — turned out to be more tedious than the work on the actual sequence and structure data. We have therefore developed the `litsift` tool [4] to make bibliographic search in context of the Atlas of Viral Secondary Structures more effective. To this end we use classifiers trained on sample corpora in a system that filters and ranks search results from bibliographic databases such as Pubmed. Preliminary results indicate that a classifier trained on one virus group can indeed be applied successfully to search the literature on other virus groups.

Acknowledgments. This work is supported by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project Nos. P-13545-MAT and P-15893, and the German *DFG* Bioinformatics Initiative. Pilot studies on individual virus groups were conducted as part of bioinformatics computer lab courses in Vienna and Leipzig: Winter 2000: Leviviridae (Erwin Gaubitzer, Philipp Hohensinner); Winter 2001: Retroviridae (Peter Andorfer, Veronika Bayer, Alexander Biedermann, Matthias Binder, Claudia Blaukopf, Ulrich Braunschweig, Claudia Fried, Stefan Fringer, Sebastian Glatt, David Grote, Michaela Gruber, Jakob Haglmüller, Leonhard Heinz, Astrid Hrdina, Alexander Kassler, Vedhu Krystufek, Christian Lahsnig, Beate Lichtenberger, Christian Martschitsch, Georg Mitterer, Michael Müller, Gregor Obernosterer, Florian Pauler, Paul Perco, Thomas Perlot, Sonja Prohaska, Markus Reschke, Miriam Satler, Manfred Schifrer, Yvonne Schindlegger, Ulrike Seifert, Sabine Stampfl, Stefanie Strobach, Andrea Tanzer, Thomas Taylor, Siegfried Ussar, Stefan Washietl, Wolfgang Winkler); Winter 2002: Mononegalovirales (Philipp Adaktylos, Katja Balazs, Florian Baumgart, Stefan Bruckner, Christoph Buchmann, Valerie Diederichs, Karin Ebner, Jörg Ettenauer, Wolfgang Fischl, Andrea Förster, Andreas Gruber, Susanne Haider, Philipp Hainzl, Jennifer Hetzl, Hans-Peter Kantner, Martina Knapp, Stefan Kraus, Roman Ferdinand Kreindl, Dominik Muggenhumer, Murat Özcelik, Jürgen Pfatschbacher, Petra Pokorny, Josef Ruppert, Johann Schmidt, Simone Schopper, Andreas Verhounig, Michael Wittinger) Spring 2003: Dicistroviridae (Christine Körner, Torsten Glomb); Potyviridae (Franziska Heinze, Sonja Karam, Jörg Lehmann, Markus Rohrschneider); Comoviridae (Jan Buck, Alexander Groß, Jana Hertel, Manuela Lindemeter, Peter Menzel, Alexander Schubert, Stefan Seemann).

References

[1] C. K. Biebricher and R. Luce. Sequence analysis of RNA species synthesized by Qbeta replicase without template. *Biochemistry*, 32:4848–4854, 1993.

[2] C. Büchen-Osmond. The universal virus database ICTVdB. *Computing Science Eng.*, 5:16–25, 2003.

[3] H.-K. Cheong, C. Cheong, Y.-S. Lee, B. L. Seong, and B.-S. Choi. Structure of influenza virus panhandle RNA studied by NMR spectroscopy and molecular modeling. *Nucl. Acids Res.*, 27:1392–1397, 2003.

[4] L. C. Faulstich, P. F. Stadler, C. Thurner, and C. Witwer. `litsift`: Automated text categorization in bibliographic search. In *Data Mining and Text Mining for Bioinformatics (ECML/PKDD 2003)*, 2003. submitted.

[5] M. Fekete. *Scanning RNA Virus Genomes for Functional Secondary Structures*. PhD thesis, University of Vienna, 2000.

[6] I. L. Hofacker, M. Fekete, C. Flamm, M. A. Huynen, S. Rauscher, P. E. Stolorz, and P. F. Stadler. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucl. Acids Res.*, 26:3825–3836, 1998.

[7] I. L. Hofacker and P. F. Stadler. Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comp. & Chem.*, 23:401–414, 1999.

[8] K. Kidd-Ljunggren, M. Zuker, I. L. Hofacker, and A. H. Kidd. The hepatitis B virus pregenome: prediction of RNA structure and implications for the emergence of deletions. *Intervirology*, 43:154–64, 2000.

[9] M. B. Leahy, H. C. Dobbyn, and e. G. GBrownlee. Hairpin loop structure in the 3' arm of the influenza A virus virion RNA promoter is required for endonuclease activity. *J. Virology*, 75:7042–7049, 2001.

[10] E. Rivas and S. R. Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2(8):19 pages, 2001.

[11] P. Schuster, W. Fontana, P. F. Stadler, and I. L. Hofacker. From sequences to shapes and back: A case study in RNA secondary structures. *Proc. Royal Soc. London B*, 255:279–284, 1994.

[12] R. Stocsits. *Nucleic Acid Sequence Alignments of Partly Coding Regions*. PhD thesis, University of Vienna, 2003.

[13] R. Stocsits, I. L. Hofacker, and P. F. Stadler. Conserved secondary structures in hepatitis B virus RNA. In *Computer Science in Biology*, pages 73–79, Bielefeld, D, 1999. Univ. Bielefeld. Proceedings of the GCB'99, Hannover, D.

[14] C. Thurner, C. Witwer, I. Hofacker, and P. F. Stadler. Conserved RNA secondary structures in Flaviviridae genomes. *J. Gen. Virol.*, 2003. submitted.

[15] G. Wertz, S. Whelan, A. LeGrone, and B. L.A. Extent of terminal complementarity modulates the balance between transcription and replication of vesicular stomatitis virus rna. *Proc. Natl. Acad. Sci. USA*, 91:8587–8591, 1994.

[16] C. Witwer, S. Rauscher, I. L. Hofacker, and P. F. Stadler. Conserved RNA secondary structures in picornaviridae genomes. *Nucl. Acids Res.*, 29:5079–5089, 2001.