# Conserved RNA Secondary Structures in Picornaviridae Genomes

CHRISTINA WITWER[†], SUSANNE RAUSCHER[†], IVO L. HOFACKER[†], AND PETER F. STADLER[†,‡,*]

[†]Institut für Theoretische Chemie, Universität Wien, Währingerstraße 17, A-1090 Wien, Austria

[‡]The Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

[*]Address for correspondence

**Abstract**

The family *Picornaviridae* contains important pathogens including, for example, Hepatitis A virus and Foot-and-Mouth Disease Virus. The genome of these viruses is a single messenger-active (+)-RNA of 7 200 to 8 500nts. Besides coding for the viral proteins, it also contains functionally important RNA secondary structures, among them an IRES region toward the 5'end.

This contribution provides a comprehensive computational survey of the complete genomic RNAs and a detailed comparative analysis of the conserved structural elements in seven of the currently nine genera in the family *Picornaviridae*. Compared with previous studies we find: (1) that only smaller sections of the IRES region than previously reported are conserved at single-base-pair resolution, and (2) that there is a number of significant structural elements in the coding region. Furthermore we identify potential CREs in four genera where this feature has not been reported so far.

**Keywords:** Picornaviridae, RNA secondary structure, IRES, CRE.

## 1. Introduction

The genomes of RNA viruses not only code for proteins, but often contain functionally active RNA structures that play a role during the various stages of the viral life cycle. Determining which parts of the huge RNA structure, formed by a viral genome, are functionally relevant, is a difficult task. In general, the secondary structures of such regions do not look significantly different from structures formed by random sequences. RNA secondary structures are however quite susceptible to point mutations:

**Table 1.** Complete genomic RNA of picornaviridae. We list the number $N$ of available sequences, the length $\ell_A$ of their alignment, their average pairwise sequence identity $\eta$, the location of the coding region in the alignment, and the mean pairwise sequence identity in the 5′-NTR. Only one complete sequence is known for both Erbovirus and Kobuvirus, which is not sufficient for the analysis presented here.

| Genus | $N$ | $\ell_A$ | $\eta$ | Coding region | $\eta(5')$ |
|---|---|---|---|---|---|
| Aphthovirus | 9 | 8231 | 0.791 | 1088-8124 | 0.543* |
| Cardiovirus | 6 | 8233 | 0.665 | 1088-8103 | 0.614 |
| Enterovirus | 29 | 7664 | 0.651 | 777-7548 | 0.765 |
| Hepatovirus | 10 | 7526 | 0.911 | 759-7449 | 0.901 |
| Parechovirus | 3 | 7391 | 0.774 | 716-7283 | 0.828 |
| Rhinovirus | 7 | 7296 | 0.687 | 652-7138 | 0.771 |
| Teschovirus† | 25 | 7135 | 0.893 | 433-7063 | 0.945 |

* From 7 of the 9 sequences since the 5'terminus is incomplete in the GeneBank entries.
† Teschovirus sequences do not include the S-fragment.

computer simulations [10, 44] showed that a small number of point mutations is very likely to cause large changes in the secondary structures: mutations in 10% of the sequence positions already leads almost surely 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 homology.

If selection acts to preserve a structural element then it must have some function. This observation led to the design of algorithms, briefly outlined in section 2 that reliably detect conserved RNA secondary structure elements in a small sample of related RNA sequences [15, 17]. Of course, it is not possible to determine the function of the conserved structure elements. Nevertheless, knowledge about their location can be used to guide, for instance, deletion studies [27].

The family *Picornaviridae* is currently divided into nine genera: Aphthovirus, Cardiovirus, Enterovirus, Hepatovirus, Rhinovirus, Parechovirus, Erbovirus, Kobuvirus, and Teschovirus [39]. These viruses are among the smallest ribonucleic acid-containing viruses known.

Although the members of the different genera share little sequence homology, they all have similar genomic structure and gene organization, see Fig. 1. The genome consists of a single strand messenger-active (+)-RNA of 7 200 to 8 500nts that is polyadenylated at the 3'-terminus and carries a small protein (virion protein, genome; VPg) covalently attached to its 5' end. The major part of the genomic RNA consists of a single large open reading frame coding a polyprotein. The 5' nontranslated region (5'NTR) is unusually long and contains multiple AUG triplets prior to the initiator of the viral translation. All Picornaviridae have an internal ribosomal entry site (IRES) instead of a 5'cap structure [14, 33].

A number of conserved RNA secondary structure elements have been described at least in some genera. Here we provide a comprehensive survey.

## 2. Methods

The algorithms `alidot` and `pfrali` for searching conserved secondary structure patterns in large RNAs are described in detail in [15, 17]. An `ANSI C` implementation is available from `http://www.tbi.univie.ac.at/RNA/`. The method requires an independent prediction 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. In this respect `alidot` and `pfrali` are similar to programs such as `construct` [26, 25] and `x2s` [21]. In contrast to efforts to simultaneously compute alignment and secondary structures e.g. [43, 3] the present approach emphasizes that the sequences may have common structural motifs but no single common structure. In this sense `alidot/pfrali` combines structure prediction and motif search [5].

A multiple sequence alignment is calculated using `CLUSTAL W` [46] and `Ralign` [45]. The latter program produces an amino acid sequence alignment for the coding parts of viral genomes which is translated back to the underlying RNA sequence and combined with RNA-alignments of the non-coding regions. The alignments are used without further modifications (except where stated explicitly).

RNA genomes are folded using McCaskill's partition function algorithm [29] as implemented in the `Vienna RNA Package` [16], based on the energy parameters published in [28].

Results are presented as conventional secondary structure drawing, as mountain plots, see e.g. Fig. 3(top), or dot plots, Fig. 7.
*Mountain plots:* A base pair $(i, j)$ is represented by a slab ranging from $i$ to $j$. The 5' and 3' sides of stems thus appear as up-hill and down-hill slopes, respectively, while plateaus indicated unpaired regions. Mountain plots are equivalent to the conventional drawing but have the advantage that (1) they can be compared more easily, and (2) it is easier to display additional information.
*Dot plots* are useful when structural alternatives have to be displayed. Each pair is shown as a small square at positions $(i, j)$ and $(j, i)$. The upper right and the lower left triangle can this be used to compare structures obtained by different methods.

Mountain plots and dot plots contain information about both sequence variation (color code) and thermodynamic likeliness of a base pair (indicated by the height of the slab and the size of the dot, respectively). Colors in the order red, ocher, green, cyan, blue, violet indicate 1 through 6 different types of base pairs. Pairs with one or two inconsistent mutation are shown in (two types of) pale colors.

In the conventional graphs paired positions with consistent mutations are indicated by circles around the varying position. Compensatory mutations thus are shown by circles around both pairing partners. Inconsistent mutants are indicated by gray instead of black lettering.

**2.1. Supplemental Material.** The complete data, consisting of the multiple alignments for each genus, the thermodynamic structure predictions for each sequence, the output of the `pfrali` program, and the secondary structure elements listed in fig. 1, are accessible through a web interface at `http://rna.tbi.univie.ac.at/virus/`.

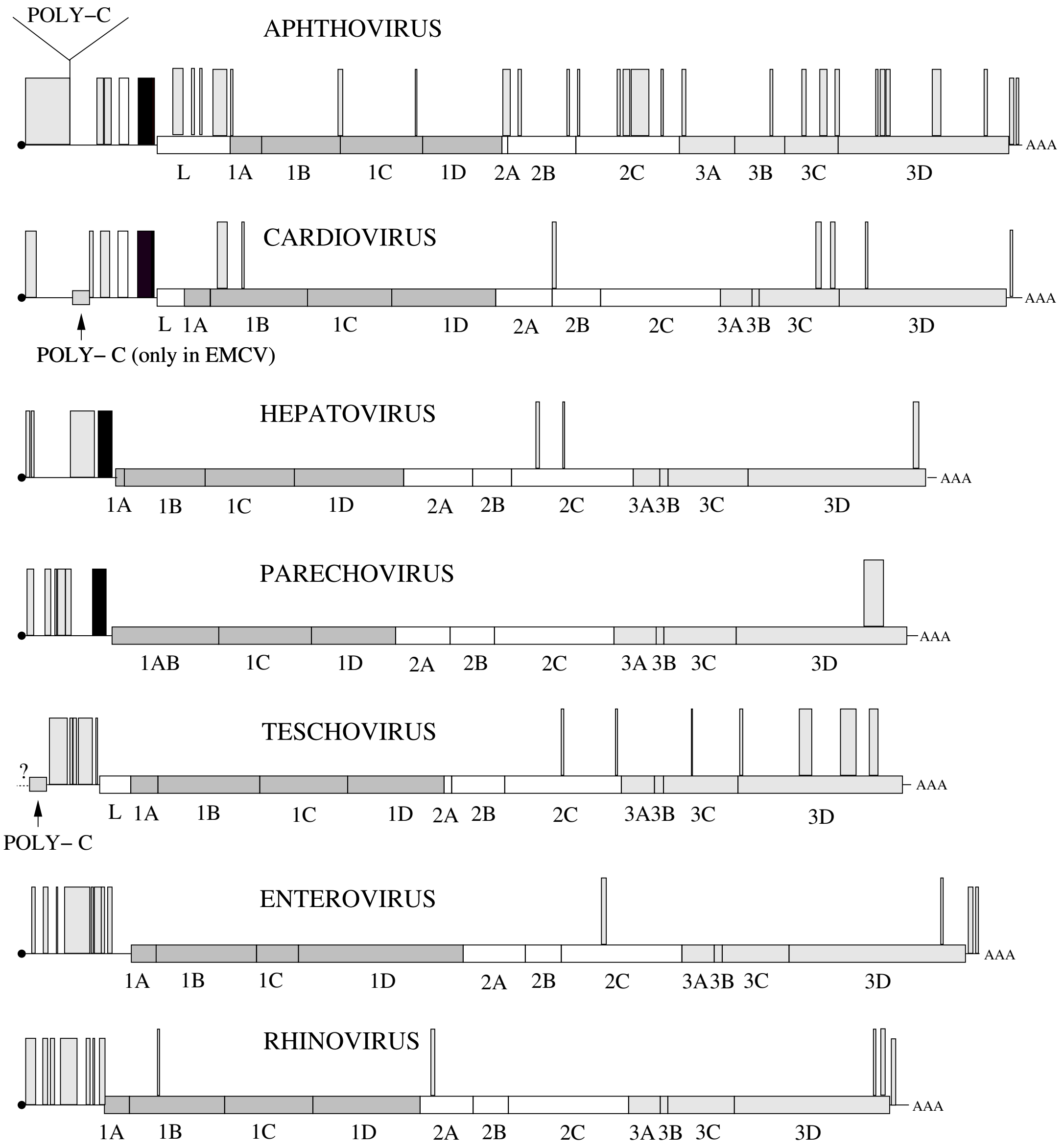



**Figure 1.** Overview of Picorna genomes. Putative conserved RNA elements are indicated above the diagrams of the reading frames.
The black boxes indicate the J,K element, the white box is the Ib element; for details see text. Proteins: leader protein L (only present in aphtho-, cardio- and teschovirus), capsid proteins 1A-1D, viral protease 2A, proteins involved in RNA synthesis 2B, 2C, unknown function 3A, VPg 3B, major viral protease 3C, RNA-dependent RNA-polymerase 3D; [42, 47, 41].

## 3. Results

The putative conserved structural features that have been identified for each Genus are summarized in Figure 1. The largest pieces of conserved structure are located in the 5'UTR. All groups with the exception of Rhinovirus show a hairpin motif close to the 3'-end. In addition, there is a substantial number of possibly conserved RNA structures within the ORF.

**3.1. The 5'-NTR.** The most prominent feature in the 5'-NTR is the IRES. The literature distinguished two or three types of IRES structures. Most common in the literature is the distinction between the RE-type found in Rhinovirus and Enterovirus, and the ACH type structure of Aphthovirus, Cardiovirus and Hepatovirus, see e.g. [36, 22, 35, 42]. Some authors distinguish three groups, type I (Aphthovirus, Cardiovirus), type II = RE, and type III (Hepatovirus), see e.g. [19, 49]. Figure 2 summarizes the results from our analysis which includes Teschovirus, Parechovirus, and Erbovirus for the first time.

Overall, we find that the IRES structure is less conserved even within the genera than expected. Fig. 2 indicates in color those features that are conserved within a group at the level of individual base pairs. Non-shaded parts of the structure are taken from folding for each genus the reference sequence listed in appendix A using the conserved base pairs as constraints.

We recover the close structural similarities of Rhinovirus and Enterovirus. A comparison of Aphthovirus, Cardiovirus, Hepatovirus and Parechovirus sets Hepatovirus apart from the other two groups, see also the details of the J and K elements in Fig. 3. The one available complete Erbovirus genome also shares the elements J and K with type I, see also [49]. Teschovirus forms a distinct fourth group of IRES structures that appears only vaguely related to the other groups.

**3.1.1.** *Cardiovirus and Aphthovirus.* The secondary structure of the 5'NTR of Cardiovirus is discussed in [36, 8, 22, 35]. Our results are very similar to the work of [35] on EMC virus. The main difference is that elements Ia and Ic, which flank element Ib, are not present in TME virus and therefore not conserved in the genus cardiovirus. In addition, the H element is longer in our data. In comparison to the earlier studies both our results and the structures in [35] have shorter conserved stem-loop regions.

The 5'NTR of the aphthovirus FMDV is discussed in [2, 36, 22]. There is only a single sequence for ERV-1 (equine rhino-virus I) which has only marginal sequence similarity with FMDV and hence was considered separately. Our data for FMDV are similar to the earlier studies. However, we find that the conserved parts of the stem-loop regions are significantly shorter than the ones reported in [36].

A comparison of cardiovirus and aphthovirus structures shows the following main differences: (1) The stem-loop structure A1 at the 5'end is much longer in FMDV compared to cardiovirus. (2) The D element in FMDV is enlarged at the expense of F.

The stem-loop structure H is very similar in aphthovirus and cardiovirus, the loop sequence `UCUUU` is strongly conserved in both genera. The stem contains many compensatory mutations that `Clustal W` failed to correctly align in this region. Manual improvement of the alignment shows that the Ib-elements as well as the M-elements of both genera can be superimposed and hence are structurally almost identical, Fig. 4 (left and right). In contrast, only the J-stem of the J,K feature is structurally (almost) identical in the two genera, Fig. 4 (middle), despite the fact that the topology of the J,K elements is conserved, Fig. 3.

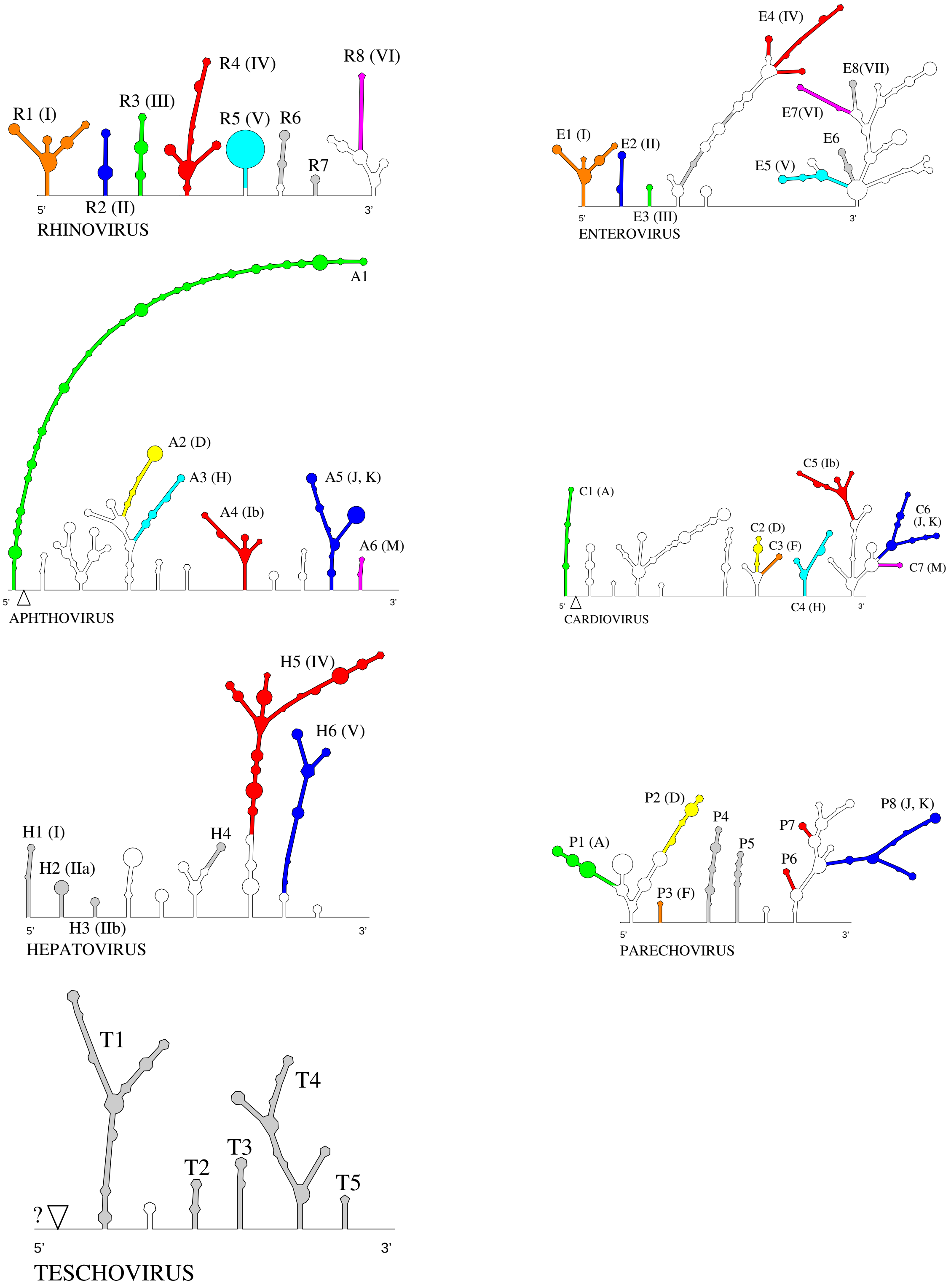



**Figure 2.** Schematic illustration of the 5'-NTR's of Picornaviridae. The minimum energy structure of one sequence of the respective virus genus is represented. Colored backgrounds mark regions which are conserved within all investigated sequences of that genus. The labels in brackets correspond to the notation of the 'classic' model of the IRES [42]. The sequence positions of the structure elements in a reference structure are listed in appendix A. △ denotes a poly-**C** region, the ? indicates the missing data for the 5'end of the teschovirus sequences.

**3.1.2.** *Parechovirus.* Until recently parechoviruses Echovirus 22 and Echovirus 23 were classified as members of the genus Enterovirus. The secondary structure of their 5'NTR is described in [34, 12]. As our analysis is based on only 3 sequences we might still overestimate the conserved parts, in particular in regions P1, P4, and P6 where the sequence is highly conserved. Ghazi [12] finds the same structure for Parechovirus and Cardiovirus. Our results agree in part with this analysis. In particular: Our element P1 corresponds to A in [12], but is shorter. P2 corresponds to D, P3 corresponds to F. Our element P4 is located in region of Ghazi's H. The sequence is rather conserved in this region and both Ghazi's H and our P4 have comparable thermodynamic stability, with a pairing probability of approximately $p = 1/3$ for each of the two alternatives. Both variants have little similarity with the H-element in cardiovirus and aphthovirus. The element P5 has been reported before. P8 contains the J and K elements, where J is identical to Ghazi's structures, our prediction for the K-element shows minor differences with previous studies. Our elements P6 and P7 are part of Ghazi's clover-leaf like motif I. The clover-leaf structure is thermodynamically feasible ($p \approx 1/4$), but appears to have significant structural variability in this genus so that it is not detected as a conserved feature. The analysis in [34], which used only two sequences and Zuker's `mfold` program, agrees in part with our consensus structure.

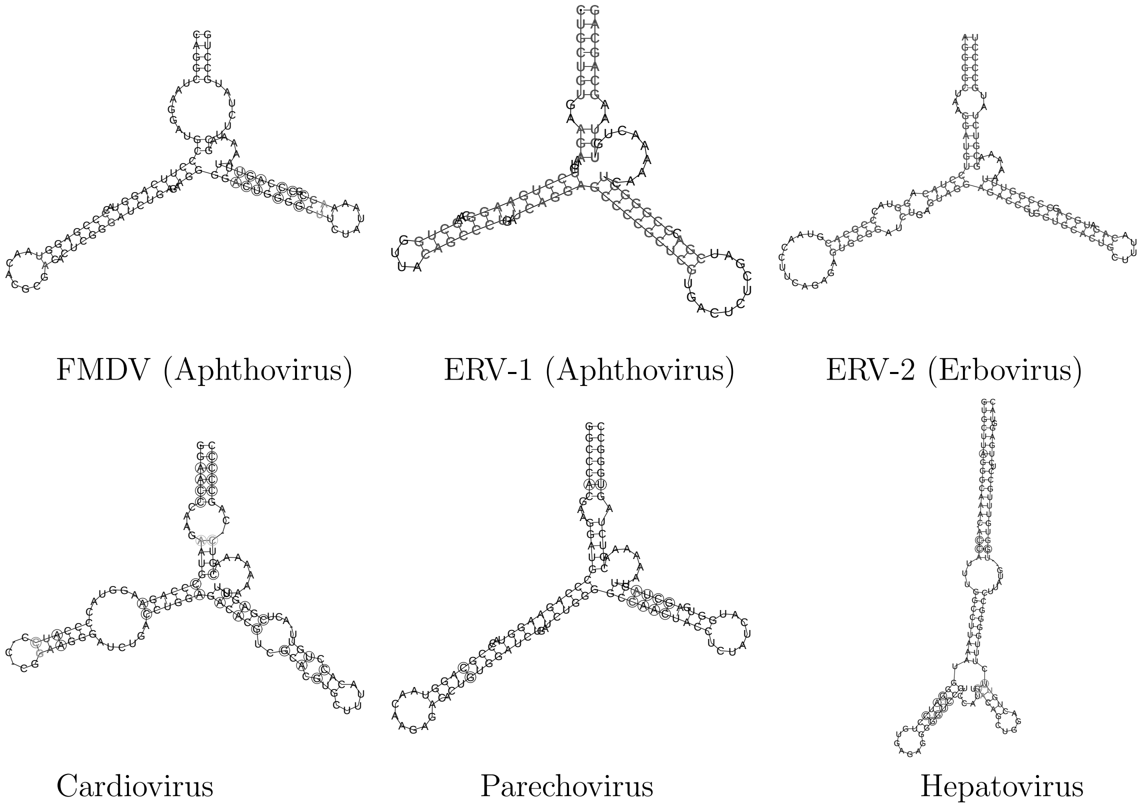



**Figure 3.** Consensus structures of the J and K elements. There appears to be some variation within aphthovirus: The sizes of the loops varies between different strains, FMDV and ERV-1, in this example. Hepatovirus shows an analogous structure.

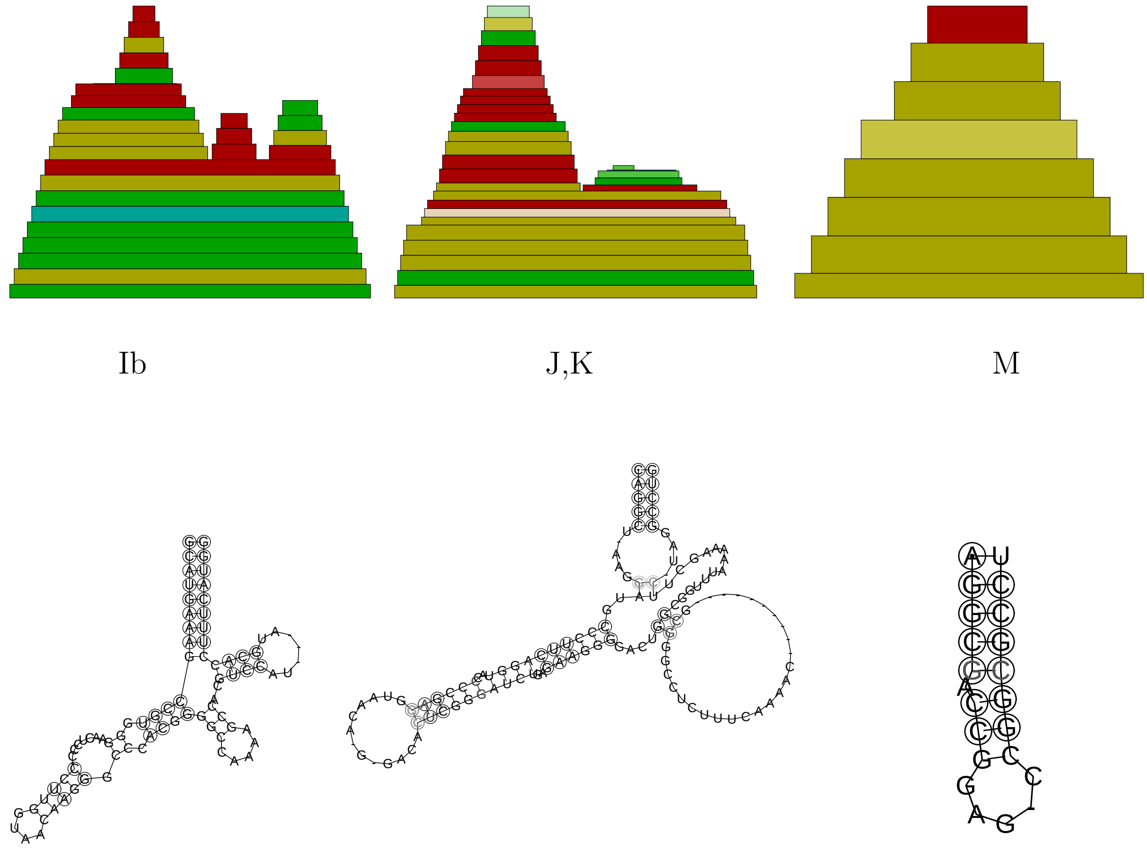



**Figure 4.** Common features in the IRES of Aphthovirus and Cardiovirus. Alignment manually improved.

**3.1.3.** *Hepatovirus.* The secondary structure of hepatovirus RNA is considered in [1]. The sequences in our data set have about 90% pairwise identity, hence we have only a small number of compensatory mutations to verify structural features predicted based on the thermodynamic rules.

Elements I and II are identical to Brown's structure; the sequences are completely conserved in this regions, i.e., there are no co-variations to verify the thermodynamic prediction in the region. Domain III is not present in our data. We find high structural flexibility here. It is noted in [1] already that "the structure of domain III was poorly defined." The only possibly significant structure in this region is a stem-loop with a completely conserved sequence around position 300, which does not appear in Brown's prediction.

Stem IV is significantly shorted in our analysis and the multiloop of the cloverleaf structure is slightly different. The deletion studies reported in [1] indicate that that domain IV is critically involved in formation of an HAV IRES element. Compared to the earlier study we find a larger element V at the expense of part of IV.

The conserved secondary structure elements of hepatovirus cannot be compared directly to that of aphtho- and cardiovirus. But there is a structural analogy of the

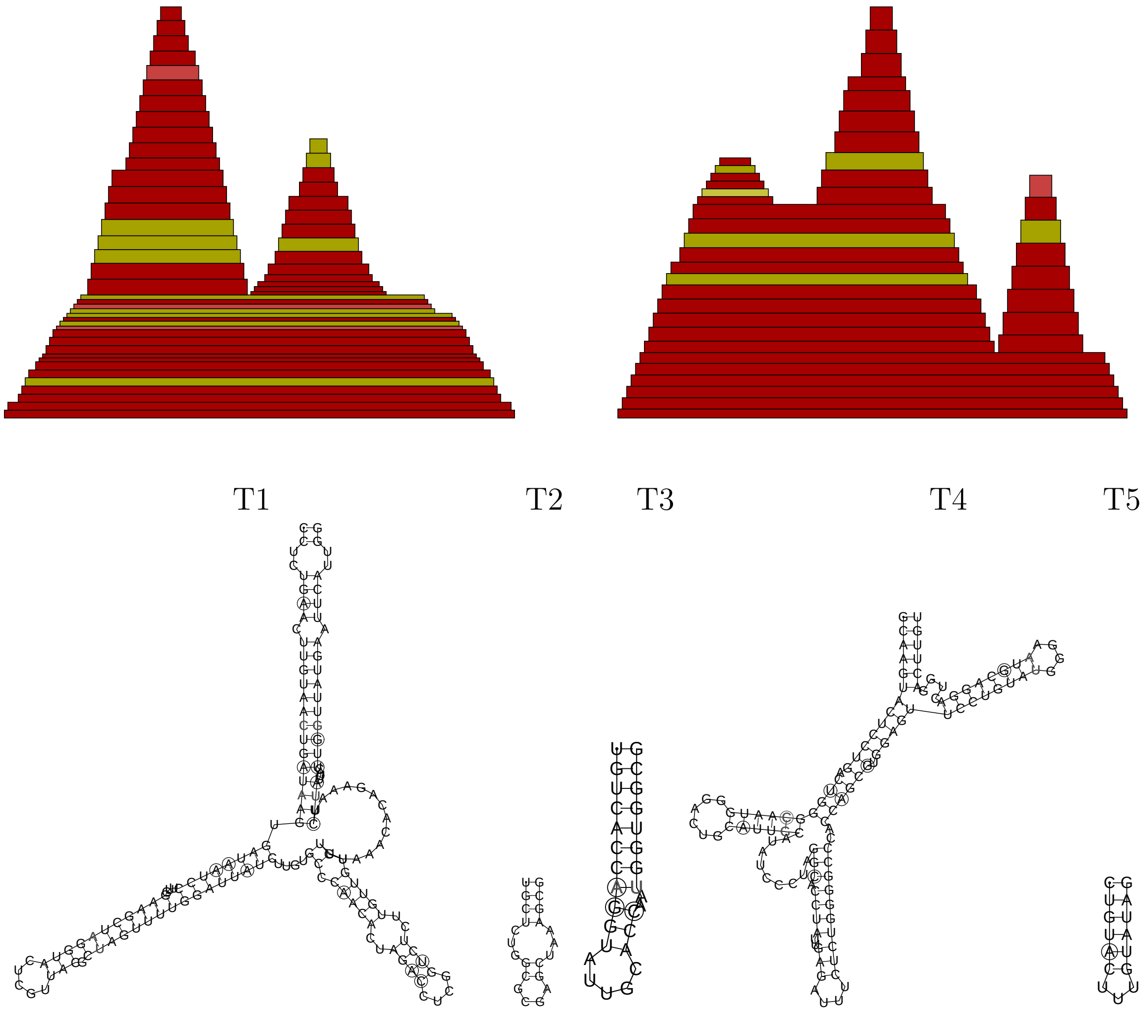



**Figure 5.** Conserved Structures in the 5'NTR of Teschovirus. Mountain plots are given only for T1 and T4.

cloverleaf structure (Ib in cardiovirus/aphthovirus and IV in Hepatovirus) and the branching stem-loop (J,K in cardiovirus/aphthovirus and V in hepatovirus).

**3.1.4.** *Teschovirus.* The sequences of the teschovirus 5'NTR are not known completely. The presence of an oligo-C stretch was demonstrated for F65 [6]. The nucleotide sequence of the 5'NTR up to this C tract could be determined successfully only in 3 of the 25 assayed strains (Talfan, Bozen, and Vir-1626/89) [50]. Hence we report no structure before the oligo-C-region.

The secondary structure of the 5'NTR of teschovirus has not been studied previously. Only element T4 shows some similarities with element V in hepatovirus. The other conserved structures do not have obvious similarities with conserved elements in other Picorna genera. The conserved elements T1-T5 are shown in Fig. 5.

**3.1.5.** *Enterovirus and Rhinovirus.* The secondary structure of IRES of Enterovirus and Rhinovirus has been the subject of a larger number of studies, see e.g. [37, 42, 51]. We recover elements I through VII in Enterovirus and I-VI in Rhinovirus, some of them with a slightly shorter stem. In addition we found the stem-loop structures E6 and R6, and R7, respectively, see Fig. 2. E6 and R7 are homologous structures, the sequence is absolutely conserved here. Elements R5 (=V) and R7 can be detected unambiguously only in HRV-A.

An attempt to extract the common structures of enterovirus and rhinovirus for a common multiple alignment yielded only a fraction of the structures found in each genus separately. In part this is due to small differences in the structural elements and in part the lack of structures can be attributed to the poor quality of the alignment.

**3.1.6.** *Other Picornaviridae.* According to [49] ERV-1 (=ERAV, aphthovirus) and ERV-2 (=ERBV, erbovirus) have a type I IRES structure. The similarity to FMDV is insufficient for a good alignment of ERV-2 with the FMDV sequences. The computed minimum energy structure shows an identifiable J,K-element. The one complete sequence of kobuvirus does not exhibit any features that can be matched unambiguously with the conserved structural elements of the other genera.

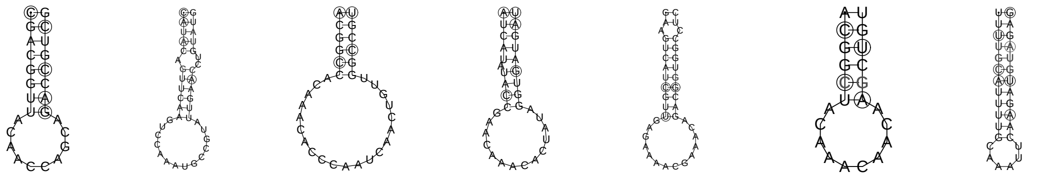
### 3.2. Coding Region.

**3.2.1.** *Cis-acting Replication Element (CRE).* A cis-acting replication element (CRE) within the coding region of several picornaviruses has been described in a number of different picornaviruses. The function of the CRE probably involves the initiation of the synthesis of the negative-sense strand template RNA during virus replication [13]. The CRE has been identified in HRV14 in region 1B of the genome [30], in Cardiovirus in region 1B [24] and in Poliovirus in region 2C [13].

Although located within a protein-coding segment of the genome , the CRE function is independent of its translation. Thus, this segment of the viral RNA has dual functions, both encoding the VP1 capsid protein and participating directly in the replication of the viral genome. The existence of the computer-predicted structure was confirmed by mutational analysis [30, 24]. Furthermore, the activity of the CRE is not position dependent [13].

In cardiovirus we recover the CRE in the 1B region which encodes the capsid protein VP2. For EMCV (excluding Mengo-Virus) and Theilovirus our structure agrees with the one reported in [24]. In [35] a different structure is given for Mengo-Virus. We find that the Mengo-Virus CRE-structures agree with the consensus of the other species. In enterovirus we recover the CRE in 2C as described in [13], in HRV-B the element is found in 1B as described in [30].

We find putative CRE elements in the 2C region of aphthovirus and teschovirus, and in the 2A region of HRV-A. There are three conserved elements in the coding region of hepatovirus. The most likely candidate for a CRE is located in region 2C. For parechovirus we were not able to identify a putative CRE. The locations of the (putative) CREs are summarized in table 2.



Aphthovirus Enterovirus Cardiovirus HRV-A HRV-B Teschovirus Hepatovirus

```
Aphto     ~~~~CGAC-GGUU------ACA-CCAAGCA------GACCGUCG~~~~~
Entero    CAUACAGU-UCAAG--------UCCAAAU-GCCGUAUUGAACCUGUAUG
Cardio    ~~~~~ACG-GCCA---CAAACACCCAAUCAACUGU-UGGCCGU~~~~~~
HRV-A     ~~~AUCAUAUACCGAACAAACA---------CUAUAGGUGAUGAU~~~~
HRV-B     GAAGUCAU-CGUUGAGAAAACG---AAACA------GACGGUGGCCUC~
Tescho    ~~~~~~AC-GGCU--ACAAACA-----ACA------AGCUGU~~~~~~~
Hepato    UUUUGCAU-UUUG---CAAA--------------UUCAAGAUGUAGAG~
          ~~~(((((-((((.......................)))))))))~~~~
          1.......10........20........30........40.........
```

**Figure 6.** Known and putative CREs in picornaviridae: secondary structures (top) and sequences (below). Nucleotides in the loops are highlighted to emphasize the AC-rich composition.

**Table 2.** Position of (putative) CRE elements.

| Genus | Acc.No | Gene. | Position | Remark |
|---|---|---|---|---|
| Aphthovirus | AJ007347 | 2C | 4834-4859 | putative |
| Enterovirus | V01150 | 2C | 4456-4494 | as in [13] |
| Cardiovirus | M81861 | 1B | 1308-1340 | as in [24] |
| Hepatovirus | K02990 | 2C | 4187-4245 | possible |
| Teschovirus | AF231769 | 2C | 4228-4249 | putative |
| Rhinovirus-A | M16248 | 2A | 3325-3357 | putative |
| Rhinovirus-B | K02121 | 1B | 1727-1764 | as in [30] |
| Parechovirus | ? | | | |

The loop of the CRE is relatively large in all genera and contains predominantly A and C. We note that an alignment of region 2C of all aphthovirus and teschovirus sequences shows that the CRE element is conserved between the two genera.

**3.2.2.** *Other Conserved Elements.* There appears to be no structural feature in the coding region that is shared among all picorna genera besides the CRE. On the other hand we find a number of structures that are conserved within a genus, see Fig. 1. The are 5 such structures in cardiovirus, a single feature in the 3D region of parechovirus, 6 in teschovirus, 3 in hepatovirus, 2 in enterovirus, 25 (!) in aphthovirus, and 1 in rhinovirus. A complete list is provided in electronic form, see "supplemental material". We do not have an explanation for the large number of conserved elements in genus aphthovirus in comparison with all other picornaviridae.

The two species HRV-A and HRV-B in the genus rhinovirus show significant differences at the level of their secondary structures. In HRV-A we find 4 conserved elements inside the coding region, only one of which is also conserved between HRV-A + HRV-B. These differences are emphasized by the fact that the CRE is located in different parts of the genome in these to species, see Tab. 2.

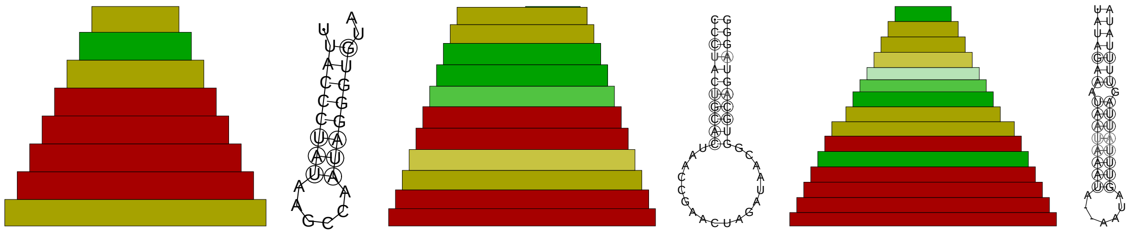


**Figure 7.** The most prominent features in the 3'NTR. Left: Cardiovirus region I; Middle: Region II of Enterovirus cluster 3; Right: hairpin structure of rhinovirus.

**3.3. 3' Non-Translated Region.** Recent deletion studies [4, 9] show that the 3'NTR, which ends in a poly-A region of variables length in all genera, is important for RNA synthesis.

**3.3.1.** *Cardiovirus.* Duque and Palmenberg [9] report three conserved stem-loop motifs (I, II, III) in the 3'NTR. Deletion studies by these authors indicated that deletion of III is lethal, deletion of II resulted in marginal RNA synthesis activity but failure of transfection with genomic RNA, while stem I was found to be dispensable for viral growth. Surprisingly, we find stem I with large thermodynamic stability and a significant number of compensatory mutations, while regions II and III do not form a conserved structure in our data set. The structures reported in [9] are consistent with all sequences of our data-set but are not thermodynamically favorably in any one of them. Neither II nor III can be recovered by considering the species Theilovirus and EMCV separately.

The secondary structures reported in [4] are completely different from both our results and the structures reported in [9]. Cui and Porter [4] suggest that a U-rich stretch, essentially region III of [9], interacts with the poly-A tail.

**3.3.2.** *Enterovirus.* There is ample literature on the structure of the 3'NTR of enterovirus, e.g. [38, 40, 51, 32, 48, 31]. None of the reported structures is conserved within the entire genus. Following the previous studies we have therefore split the 29 available genomes into three clusters because there are not enough sequences available for the fourth cluster described in [51].

Cluster 1 consists of Poliovirus and Human Enterovirus C, cluster 2 contains Human Enterovirus A [18]. 3'NTR sequences are highly conserved within each of these two clusters. While we find the published structures in our data, their equilibrium probabilities are small and they appear as one of a number of thermodynamically feasible alternatives.

Cluster 3 contains Human Enterovirus B. We find domains I and II as reported in [38, 51]. It is interesting to note that the structure of domain II is supported by a substantial number of compensatory mutations, Fig. 8.

**3.3.3.** *Other Genera.* The 3'NTR of aphthovirus apparently has not been considered before. We find two hairpin structures, one of which has an almost conserved GAAA sequence motif in the loop. In one of the nine sequences we find GCAAA instead.

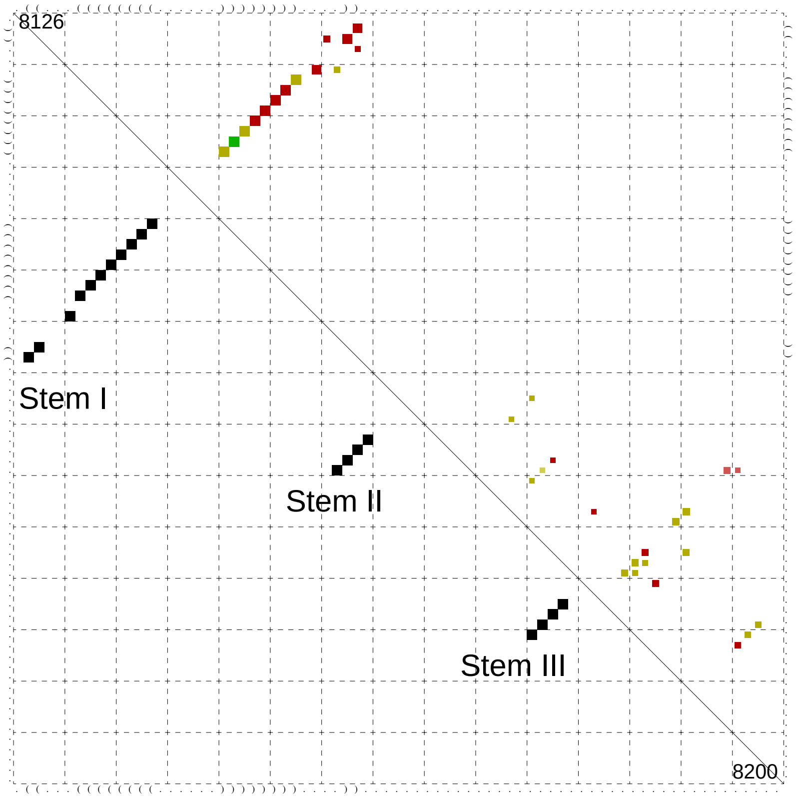



**Figure 8.** Dot plot of the 3'NTR of Enterovirus. Upper triangle: Output of `pfrali` showing only stem I. Its prediction is well supported by a number of consistent mutations and the absence of inconsistent mutations. The size of the squares is proportional to $\log p$ here. The signals in the area of stems II and III have probabilities of only a few percent and hence are not significant. The lower triangles shows the structures reported in [9] for comparison.

A hairpin motif, which was already reported in [38] is detected unambiguously in rhinovirus, Fig. 8. The structure is conserved between HRV-A and HRV-B.

In Parechovirus, Hepatovirus, and Teschovirus there are no significant conserved secondary structures in the 3'NTR. In particular, we could not confirm the published minimum energy structures for individual teschovirus [6, 50] and hepatovirus [40] sequences as conserved features.

## 4. Discussion

Structural genomics, the systematic determination of all macro-molecular structures represented in a genome, is at present focused almost exclusively on proteins. Over the past two decades it has become clear, however, that a variety of RNA molecules have important, and sometimes essential, biological functions beyond their roles as rRNAs, tRNAs, or mRNAs. To comprehensively understand the biology of a cell, it will ultimately be necessary to know the identity of all encoded RNAs, the molecules with which they interact and the molecular structures of these complexes [7]. Viral RNA genomes, because of their small size and the strong selection that acts upon them, are an ideal proving ground for techniques that aim at identifying functional RNA *structures*.

A combination of structure prediction based on the thermodynamic rules of the "standard energy model" for nucleic acid secondary structures and the evaluation of consistent and compensatory mutations can be employed for scanning complete viral genomes for functional RNA structure motifs. This contribution reports a detailed, comprehensive survey of such structural features for those six (out of nine) genera of the family of picornaviridae for which sufficient sequence information is currently available: Aphthovirus, Cardiovirus, Enterovirus, Hepatovirus, Parechovirus, Rhinovirus, and Teschovirus.

The 5'-region of a number of these viruses has been studied previously because of the particular interest in the IRES region. Our automatic approach confirms many of the patterns identified previously based on smaller data sets. However, we find that in many cases the parts of these features that are conserved base-pair by base-pair are significantly smaller. This conclusion is mainly based on the fact that some sequences that are now contained in the database simply cannot form parts of the structures that have have previously been reported as conserved. The same conclusion can be drawn for the 3'NTR.

On the other hand, there is a large number of secondary structure elements that have not been described before, most importantly within the coding region. Most notably, we have been able to identify likely or at least possible candidates for the CRE region in Aphthovirus, Hepatovirus, Rhinovirus-A and Teschovirus, apart from recovering the known locations of the CRE in Enterovirus, Cardiovirus, and Rhinovirus-B. Only for Parechovirus we did not find a significant signal.

The approach used here goes beyond search software such as `RNAMOT` [11] in that it does not require any *a priori* knowledge of the functional structure motifs and it goes beyond searches for regions that are thermodynamically especially stable or well-defined [20] in that it returns a specific prediction for a structure if *and only if* there is sufficient evidence for structural conservation. The results collected here (and in the supplementary material available on-line) could be used to refine descriptors e.g. for the CRE that can then be used for structure-specific scans in other RNAs.

**Acknowledgments.** This work is supported by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Project Nos. P-12591-INF (SR) and P-13545-MAT (CW).

## Appendix

**Appendix A: Figure 2.** The schematic drawings in Fig. 2 are obtained from a typical strain with the strictly conserved structural features indicated by shadings. Here we give the reference sequences, alternative nomenclature where available, and the exact sequence positions of the outermost base pair of each of the indicated elements.

Aphthovirus: FMDV, strain C3Arg85 (Acc.No. AJ007347) [42]:
A1 2-367, A2 (D) 587-640, A3 (H) 648-703, A4 (Ib) 769-846, A5 (J,K) 924-1033, A6 (N) 1037-1058.

Cardiovirus: TMEV, strain DA (Acc.No. M20301) [8, 42, 35]:
C1 (A) 1-86, C2 (D) 524-554, C3 (F) 580-602, C4 (H) 610-680, C5 (Ib) 749-831, C6 (J,K) 909-1020, C7 (M) 1023-1042.

Parechovirus: HPeV-1, strain Harris (Acc.No. S45208 L00675) [12]:
P1 (A) 14-67, P2 (D) 157-205, P3 (F) 239-253, P4 261-325, P5 327-373, P6 416-431, P7 452-464, P8 (J,K) 550-661. P6 and P7 are part of Ib.

Teschovirus: PTV-11, strain Dresden (Acc.No. AF296096), no S-fragment:
T1 19-166, T2 187-208, T3 212-242, T4 257-372, T5 401-415.

Hepatovirus: HAV, strain MBB (Acc.No. M20273) [1]:
H1 (I) 5-37, H2 (II) 49-72, H3 (IV) 349-545, H4 (V) 577-688.

Enterovirus: Coxackievirus B, strain 1 Japan (Acc.No. M16560), [23, 42, 51]:
E1 (I) 2-86, E2 (II) 127-165, E3 (III) 200-215, E4 (IV) 240-443, E5 (V) 477-534, E6 535-559, E7 (VI) 583-622, E8 (VII) 625-641.

Rhinovirus: strain HRV89 (Acc.No. M16248, A10937), [23, 42, 51]:
R1 (I) 3-85, R2 (II) 128-166, R3 (III) 183-229, R4 (IV) 272-405, R5 (V) 422 462, R6 479-511, R7 536-548, R8 (VI) 582-624.

**Appendix B: Access Codes.** Aphthovirus/FMDV: AF18915, AJ133359, AF154271, AJ007347, X00871, X00429, X74812, M10975, AJ251473, M14409, M14408, L11360, Y18531, X74811, X83209;
Aphthovirus/Equine rhinitis A virus: X96870;
Cardiovirus: L22089, M81861, M22457, K01410, M16020, M20562, M20301, M80887;
Hepatovirus: X75214, AB020569, AB020567, AB020565, AB020564, D00924, X83302, M20273, K02990, M59808;
Parechovirus: AJ005695, AF055846, L02971;
Teschovirus: AJ011380, AF23176, AF231768, AF296104, AF296100, AF296102, AF296087, AF296107, AF296108, AF296109, AF296088, AF296089, AF296111, AF296112, AF296113, AF296090, AF296091, AF296115, AF296117, AF296092, AF296093, AF296118, AF296094, AF296119, AF296096;
Rhinovirus: M16248, D00239, L24917, X02316, K02121, L05355, U60874, RV-0007*, RV-0002*;
Enterovirus: V01150, X00595, D00625, K01392, X04468, U05876, AF177911, AF176044, U22521, U22522, D00627, M16560, AF085363, U57056, M16572, X05690, S76772, X67706, AF083069, U16283, X92886, X84981, X80059, X79047, AF162711, D00435, D00538, D90457;
Erbovirus: X96871;
Kobuvirus: AB010145.

The sequences indicated by * are not in `genbank`. These data were obtained from [18].

## References


[1] E. A. Brown, S. P. Day, R. Jansen, and S. M. Lemon. The 5' nontranslated region of hepatitis a virus RNA: Secondary structure and elements required for translation in vitro. *J. Virol.*, 65:5828–5838, 1991.

[2] B. E. Clarke, A. L. Brown, K. M. Currey, S. E. Newton, D. J. Rowlands, and A. R. Carroll. Potential secondary and tertiary structure in the genomic RNA of foot and mouth disease virus. *Nucleic Acids Res.*, 15:7067–7079, 1987.

[3] J. Corodkin, L. J. Heyer, and G. D. Stormo. Finding common sequences and structure motifs in a set of RNA molecules. In T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, C. Sander, and A. Valencia, editors, *Proceedings of the ISMB-97*, pages 120–123, Menlo Park, CA, 1997. AAAI Press.

[4] T. Cui and A. G. Porter. Localization of the binding site for encephalomyocarditis virus RNA polymerase in the 3'-noncoding region of the viral RNA. *Nucleic Acids Res.*, 23:377–382, 1995.

[5] T. Dandekar and M. W. Hentze. Finding the hairpin in the haystack: searching for RNA motifs. *Trends. Genet.*, 11:45–50, 1995.

[6] M. Doherty, D. Todd, N. McFerran, and E. M. Hoey. Sequence analysis of a porcine enterovirus serotype 1 isolate: relationships with other picornaviruses. *J. Gen. Virol.*, 80:1929–1941, 1999.

[7] J. A. Doudna. Structural genomics of RNA. *Nature Struct. Biol.*, 7:954–956, 2000.

[8] G. M. Duke, M. A. Hoffman, and A. C. Palmenberg. Sequence and structural elements that contribute to efficient encephalomyocarditis virus RNA translation. *J. Virol.*, 66:1602–1609, 1992.

[9] H. Duque and A. C. Palmenberg. Phenotypic characterization of three phylogenetically conserved stem-loop motifs in the mengovirus 3' untranslated region. *J. Virol*, 75:3111–3120, 2001.

[10] W. Fontana, D. A. M. Konings, P. F. Stadler, and P. Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33:1389–1404, 1993.

[11] D. Gautheret, F. Major, and R. Cedergren. Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA. *Comput. Appl. Biosci.*, 6:325–331, 1990.

[12] F. Ghazi, P. J. Hughes, T. Hyypiae, and G. Stanway. Molecular analysis of human parechovirus type 2 (formerly echovirus 23). *J. Gen. Virol.*, 79:2642–2650, 1998.

[13] I. Goodfellow, Y. Chaudhry, A. Richardson, J. Meredith, J. W. Almond, W. Barclay, and D. J. Evans. Identification of a cis-acting replication element within the poliovirus coding region. *J. Virol.*, 74:4590–4600, 2000.

[14] M. J. Hewlett, J. K. Rose, and D. Baltimore. 5' terminal structure of poliovirus polyribosomal RNA is pUp. *Proc. Natl. Acad. Sci. USA*, 73:327–330, 1976.

[15] 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.

[16] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

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

[18] Institute for Animal Health, UK. The picornavirus home page. `http://www.iah.bbsrc.ac.uk/virus/Picornaviridae/picornavirus.htm`.

[19] R. J. Jackson, M. T. Howell, and A. Kaminski. The novel mechanism of initiation of picornavirus RNA translation. *Trends in Biochemical Sciences*, 15:477–483, 1990.

[20] A. B. Jacobson and M. Zuker. Structural analysis by energy dot plot of large mRNA. *J. Mol. Biol.*, 233:261–269, 1993.

[21] V. Juan and C. Wilson. RNA secondary structure prediction based on free energy and phylogenetic analysis. *J. Mol. Biol.*, 289(4):935–947, 1999.

[22] S. Y. Le, J. H. Chen, N. Sonenberg, and J. V. Maizel. Conserved tertiary structural elements in the 5' nontranslated region of cardiovirus, aphthovirus and hepatitis a virus RNAs. *Nucleic Acids Res.*, 21:2445–2451, 1993.

[23] S. Y. Le and M. Zuker. Common structures of the 5' non-coding RNA in enteroviruses and rhinoviruses. *J. Mol. Biol.*, 216:729–741, 1990.

[24] P.-E. Lobert, N. Escriou, J. Ruelle, and T. Michiels. A coding RNA sequence acts as a replication signal in cardioviruses. *Proc. Natl. Acad. Sci. USA*, 96:11560–11565, 1999.

[25] R. Lück, S. Graf, and G. Steger. `ConStruct`: a tool for thermodynamic controlled prediction of conserved secondary structure. *Nucl. Acids. Res.*, 27:4208–4217, 1999.

[26] R. Lück, G. Steger, and D. Riesner. Thermodynamic prediction of conserved secondary structure: Application to the RRE element of HIV, the tRNA-like element of CMV, and the mRNA of prion protein. *J. Mol. Biol.*, 258:813–826, 1996.

[27] C. W. Mandl, H. Holzmann, T. Meixner, S. Rauscher, P. F. Stadler, S. L. Allison, and F. X. Heinz. Spontaneous and engineered deletions in the 3'-noncoding region of tick-borne encephalitis virus: Construction of highly attenuated mutants of flavivirus. *J. Virology*, 72:2132–2140, 1998.

[28] D. Mathews, J. Sabina, M. Zucker, and H. Turner. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[29] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.

[30] K. McKnight and S. M. Lemon. The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication. *RNA*, 4:1569–1584, 1998.

[31] J. M. Meredith, J. B. Rohll, J. W. Almond, and D. J. Evans. Similar interactions of the poliovirus and rhinovirus 3d polymerases with the 3' untranslated region of rhinovirus 14. *J. Virol.*, 73:9952–9958, 1999.

[32] M. H. Mirmomeni, P. J. Hughes, and G. Stanway. An tertiary structure in the 3' untranslated region of enteroviruses is necessary for efficient replication. *J. Virol.*, 71:2363–2370, 1997.

[33] A. Nomoto, Y. F. Lee, and E. Wimmer. The 5'-end of poliovirus mRNA is not capped with m7g(5')pppg(5')np. *Proc. Natl. Acad. Sci. USA*, 73:375–380, 1976.

[34] M. S. Oberste, K. Maher, and M. A. Pallansch. Complete sequence of echovirus 23 and its relationship to echovirus 22 and other human enteroviruses. *Virus Res.*, 56:217–223, 1998.

[35] A. C. Palmenberg and J. Sgro. Topological organization of picornaviral genomes: Statistical prediction of RNA structural signals. *Seminars in Virology*, 8:231–241, 1997.

[36] E. V. Pilipenko, V. M. Blinov, B. K. Chernov, T. M. Dmitrieva, and V. I. Agol. Conservation of the secondary structure elements of the 5'-untranslated region of cardio- and aphthovirus RNAs. *Nucleic Acids Res.*, 17:5701–5711, 1989.

[37] E. V. Pilipenko, V. M. Blinov, L. I. Romanova, A. N. Sinyakow, S. V. Maslova, and V. I. Agol. Conserved structural domains in thr 5'-untranslated region of picornaviral genomes: an analysis of the segment contolling translation and neurovirulence. *Virology*, 168:201–209, 1989.

[38] E. V. Pilipenko, S. V. Maslova, A. Sinyakov, and V. I. Agol. Towards identifaction of cis-acting elements involved in the replication of enterovirus RNAs - a proposal for the existence of tRNA-like terminal structures. *Nucleic Acids Res.*, 20:1739–1745, 1992.

[39] C. Pringle. Virus taxonomy at the XIth international congress of virology, Sydney, Australia. *Arch. Virol.*, 144:2065–2070, 1999.

[40] J. B. Rohll, D. H. Moon, D. J. Evans, and J. W. Almond. The 3' untranslated region of picornavirus RNA: Features required for efficient genome replication. *J. Virol.*, 69:7835–7844, 1995.

[41] D. J. Rowlands. Foot-and-mouth disease viruses (picornaviridae). In R. Webster and A. Granoff, editors, *Encyclopedia of Virology*, pages 586–575. Academic Press, 2nd edition, 1999.

[42] R. R. Rueckert. Picornaviridae: The viruses and their replication. In N. Fields, D. Knipe, and P. Howley, editors, *Virology*, volume 1, pages 609–654. Lippincott-Raven Publishers, Philadelphia, New York, third edition, 1996.

[43] D. Sankoff. Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. *SIAM J. Appl. Math.*, 45:810–825, 1985.

[44] 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 Society London B*, 255:279–284, 1994.

[45] 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.

[46] J. D. Thompson, D. G. Higgs, and T. J. Gibson. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucl. Acids Res.*, 22:4673–4680, 1994.

[47] L. M. Vance, N. Moscufo, M. Chow, and B. A. Heinz. Poliovirus 2C region functions during encapsidation of viral RNA. *J. Virol.*, 71:8759–8765, 1997.

[48] J. Wang, J. M. Bakkers, J. M. Galama, H. J. Bruins Slot, E. V. Pilipenko, V. I. Agol, and W. J. Melchers. Structural requirements of the higher order RNA kissing element in the enteroviral 3'utr. *Nucl. Acids Res.*, 27:485–490, 1999.

[49] G. Wutz, N. Nowotny, B. Grosse, T. Skern, and E. Kuechler. Equine rhinovirus serotypes 1 and 2: relationship to each other and to aphthoviruses and cardioviruses. *J. Gen. Virol.*, 77:1719–1730, 1996.

[50] R. Zell, M. Dauber, A. Krumbholz, A. Henke, E. Birch-Hirschfeld, A. Stelzner, D. Prager, and R. Wurm. Porcine teschoviruses comprise at least eleven distinct serotypes: molecular and evolutionary aspects. *J. Virol.*, 75:1620–1631, 2001.

[51] R. Zell and A. Stelzner. Application of genome sequence information to the classifation of bovine enteroviruses: the importance of 5'- and 3'-nontranslated regions. *Virus Res.*, 51:213–229, 1997.