Automatic Detection of Conserved RNA Structure Elements in Complete RNA Virus Genomes

Ivo L. Hofacker[a], Martin Fekete[a],
Christoph Flamm[a], Martijn A. Huynen[c,d],
Susanne Rauscher[a], Paul E. Stolorz[e],
and Peter F. Stadler[a,b,*]

[a]Institut f. Theoretische Chemie, Univ. Wien, Austria
[b]The Santa Fe Institute, Santa Fe, New Mexico, U.S.A.
[c]EMBL Heidelberg, Germany
[d]Max Delbrück Center, Berlin, Germany
[e]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California, U.S.A.

[*]Mailing Address:
Institut für Theoretische Chemie, Universität Wien
Währingerstrasse 17, A-1090, Wien, Austria
Phone: **43 1 40480 665 Fax: **43 1 40480 660
Email: `studla@tbi.univie.ac.at` Www: `http://www.tbi.univie.ac.at/~studla`

Abstract

We propose a new method for detecting conserved RNA secondary structures in a family of related RNA sequences. Our method is based on a combination of thermodynamic structure prediction and phylogenetic comparison. In contrast to purely phylogenetic methods, our algorithm can be used for small data sets of about 10 sequences, efficiently exploiting the information contained in the sequence variability. The procedure constructs a prediction only for those parts of sequences that are consistent with a single conserved structure. Our implementation produces reasonable consensus structures without user interference. As an example we have analyzed complete HIV-1 and Hepatitis C virus genomes as well as the small segment of Hanta virus. Our method confirms the known structures in HIV-I and predicts previously unknown conserved RNA secondary structures in HCV.

1. Introduction

One of the major problems facing computational molecular biology is the fact that sequence information is available in far greater quantities than information about the three-dimensional structure of biopolymers. While the prediction of three-dimensional RNA structures from sequence data is infeasible at present (see however [36] for a promising ansatz), the prediction of secondary structure is in principle tractable even for large molecules. Functional secondary structures are conserved in evolution, see for instance [14], and they represent a qualitatively important description of the molecules, as documented by their application to the interpretation of molecular evolution data.

Almost all RNA molecules — and consequently also almost all subsequences of a large RNA molecule — form secondary structures. The presence of secondary structure in itself therefore does not indicate any functional significance. In this contribution we show that potentially functional RNA structures can be identified by a purely computational procedure that combines structure prediction and sequence comparison. RNA viruses are an ideal proving ground for testing such a method:

(1) Distant groups of RNA viruses have very little or no detectable sequence homology and oftentimes very different genomic organization. Thus we can test our approach on essentially independent data sets.

(2) RNA viruses show an extremely high mutation rate, on the order to 10^{-5} to 10^{-3} mutations per nucleotide and replication. Due to this high mutation rate they form *quasispecies*, that is, diffuse “clouds” in sequence space [10] and their sequences evolve at a very high rate. In contrast functional secondary structures are strongly conserved.

Due to the high sequence variation, the application of classical methods of sequence analysis is therefore difficult or outright impossible. Indeed, except for the family *Mononegavirales* (negative-stranded RNA viruses), there is no accepted taxonomy above the *Genus* level.

(3) The high mutation rate of RNA viruses also explains their short genomes of less than some 20,000 nucleotides [10]. A large number of *complete* genomic sequences is available in data bases. The non-coding regions are most likely functionally important since the high selection pressure acting on viral replication rates makes "junk RNA" very unlikely. So far a number of relevant secondary structures have been determined that play a role during the various stages of the viral life cycle in a variety of different classes of viruses, for instance lentiviruses [59, 1, 17], RNA phages [3, 43], flaviviruses [50], pestiviruses [5, 8], picorna viruses [9, 19, 27, 34, 44, 47], hepatitis C viruses [53, 5], or hepatitis D virus [56].

Three unrelated groups of viruses, which contain a variety of human pathogens of global medical importance, will serve as examples, see figure 1 for details:
HIV-1 is a highly complex retrovirus. Its genome is dense with information for the coding of proteins and biologically significant RNA secondary structures. The latter play a role in both the entire genomic HIV-1 sequence and in the separate HIV-1 messenger RNAs which are basically (combined) fragments of the entire genome.
Flaviviridae are small enveloped particles with an unsegmented, plus-stranded RNA genome. This virus family contains the genera flavivirus (which includes the viruses causing Japanese Encephalitis, Dengue, Yellow Fever, and Tick-Borne Encephalitis), pestivirus, hepatitis C, and the recently discovered hepatitis G viruses, see [41] for a recent summary.
Hantaviruses are serologically-related members of the family Bunyaviridae [11]. They are enveloped viruses with a tripartite negative-sense RNA genome. The three genome segments are called L, M and S, encoding the viral transcriptase, envelope glycoproteins, and nucleocapsid protein, respectively. In this contribution we shall be concerned only with the small (S) segment. Hantaviruses have been implicated as etiologic agents for two acute diseases: hemorrhagic fever with renal syndrome (HFRS) and hantavirus pulmonary syndrome (HPS). Both diseases are carried by rodent vectors.

The total length of the genomic sequences of both virus families, on the order of 10000 nucleotides, makes experimental analysis of the secondary structure of full genomes infeasible. For RNAs of this size, structure prediction based on thermodynamic constraints is the only approach that is available at present.

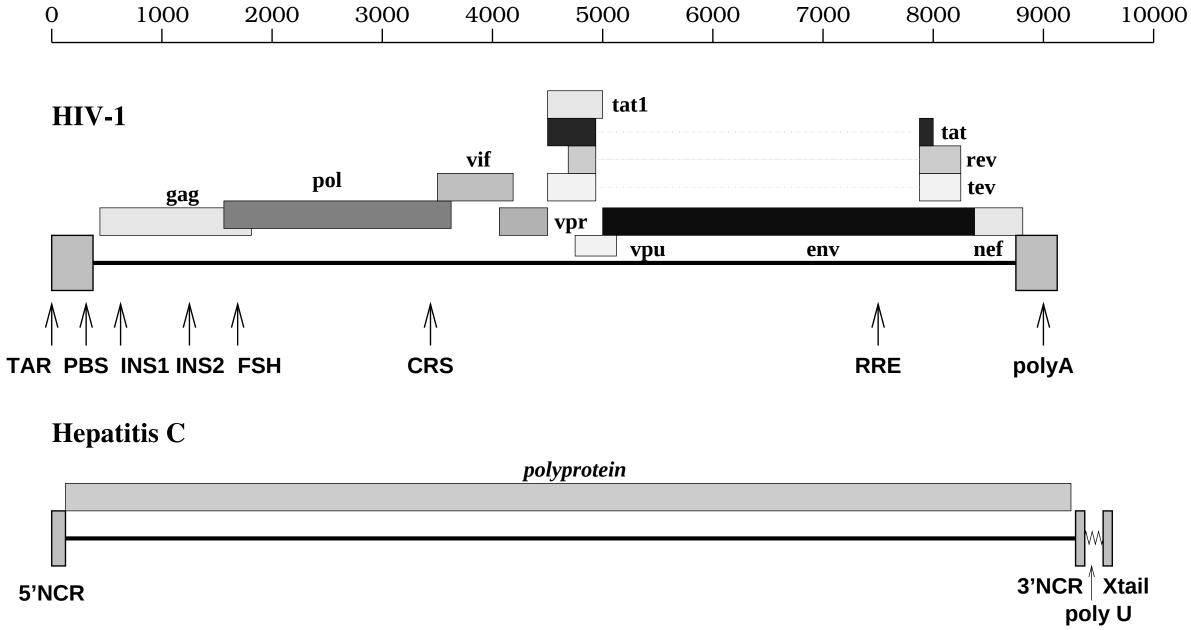

Figure 1: (top) Organization of a retrovirus genome (HIV-1) and a Flaviviridae genome (Hepatitis C). Proteins are shown on top, known features of the RNA are indicated below.

The major genes of HIV-1 are *gag*, *pol*, *env*, *tat* and *rev*. The *gag* gene codes for structural proteins for the viral core. The *pol* gene codes among others for the reverse transcriptase and the protein that integrates the viral DNA (after reverse transcription) into the host DNA. The *env* gene codes for the envelope proteins. The *tat* and *rev* genes code for regulatory proteins Tat and Rev that can bind to TAR and the RRE respectively. INS1, INS2 and CRS are RNA sequences that destabilize the transcript in the absence of the Rev protein. FSH refers to the hairpin that is involved in the ribosomal frameshift from *gag* to *pol* during translation. PolyA refers to the polyadenylation signal. PBS is the primer binding site. For references see [22]

(bottom) About 90% of the approximately 10kb-long genomes of flaviviridae is taken up by a single long open reading frame that encodes a polyprotein which is co- and post-translationally cleaved by viral and cellular proteases into 10 viral proteins (for review, see [46]). The flanking non-coding regions (NCR) are believed to contain cis-acting elements important for replication, translation and packaging. The X-tail, a highly conserved sequence of 98nt beyond a poly-U stretch of variable length, might play an important role in the initiation of genomic replication [53].

2. Methods

2.1. RNA Structure Prediction

RNA secondary structures are predicted as minimum energy structures by means of dynamics programming techniques [57, 62]. An efficient implementation of this algorithm is part of the `Vienna RNA Package` [16]. The energy parameters used by the `Vienna RNA Package` are based on [13, 28, 15]. They are identical with the parameters in Zuker's `mfold 2.2` with the exception stacking energies involving **GU** pairs which were taken from [15]. Complete RNA genomes were folded on CalTech's `Delta` using the message passing version of the minimum folding algorithm described in [17, 18]. In this version we do not consider dangling ends.

2.2. Sequence and Structure Comparison

While the computation of the secondary structures is a straightforward (yet computationally demanding) task, their comparison is less obvious.

A variety of combined alignment plus structure prediction procedures have been proposed [48, 52, 6]. The problem with this approach is threefold for our task: (1) The computational efforts become prohibitive for longer sequence: CPU time scales as $\mathcal{O}(n^4)$ in the approximate algorithm [6], and as $\mathcal{O}(n^{3m})$ in the exact version [48], where n is the length of the sequence and m the number of sequences, by far exceeding the available resources. (2) Viral sequences show a large variation in sequence similarity along the chain. Furthermore, we do not expect a conserved secondary structure for all parts of the sequence, even if there is a significant level of sequence conservation. Combined folding and alignment algorithms will therefore produce poor alignments in such cases. (3) The use of a combined algorithm for predicting structure and alignment would not allow independent verification of the predicted structural elements. A possibility to verify the predicted structures, however, is particularly important when dealing with the relatively sparse datasets that are available.

We start the comparison procedure with an alignment of the sequences that is obtained without any reference to the predicted structures. The multiple sequence alignments are calculated using `CLUSTAL W` [54]. We do not attempt to improve the alignment based on visual inspection or based on predicted secondary structures. While this might increase the number of compensatory mutations and possibly also the number of detected structural elements, it would also have compromised the use of the sequence data for verifying the predicted structures.

The sequence alignment is then used to produce an alignment of the secondary structures by introducing the appropriate gaps into the minimum energy foldings. Up to this point our procedure is essentially the same as Riesner's `ConStruct` [35]. The evaluation of the structure alignment, however, is quite different from these approaches: (1) We do not assume *a priori* that there is a conserved secondary structure for all parts of the sequence. Hence we cannot simply search for the secondary structure that maximizes the sum of the predicted base pairing probabilities. (2) We explicitly use the sequence information contained in the multiple alignment to confirm or reject predicted base pairs. A flow diagram of our approach is shown in figure 2.

A quick overview of the data is conveniently obtained from the *mountain representation*. In the mountain representation [20] a single secondary structure is represented in a two dimensional graph, in which the x-coordinate is the position k of a nucleotide in the sequence and the y-coordinate the number $m(k)$ of base pairs that enclose nucleotide k. The mountain representation allows for a straightforward comparison of secondary structures and inspired a convenient algorithm for structure based alignments of secondary structures [20, 32]. Structural features are easily identified in these plots. *Peaks* correspond to hairpins; the symmetric slopes represent the stack enclosing the unpaired bases in the hairpin loop, which appear as a plateau; *plateaus* represent unpaired bases. When interrupting sloped regions they indicate bulges or interior loops, depending on whether they occur alone or paired with another plateau on the other side of the mountain at the same height respectively. *Valleys* indicate the unpaired regions between the branches of a multi-stem loop or, when their height is zero, they indicate unpaired regions separating the components of secondary structures.

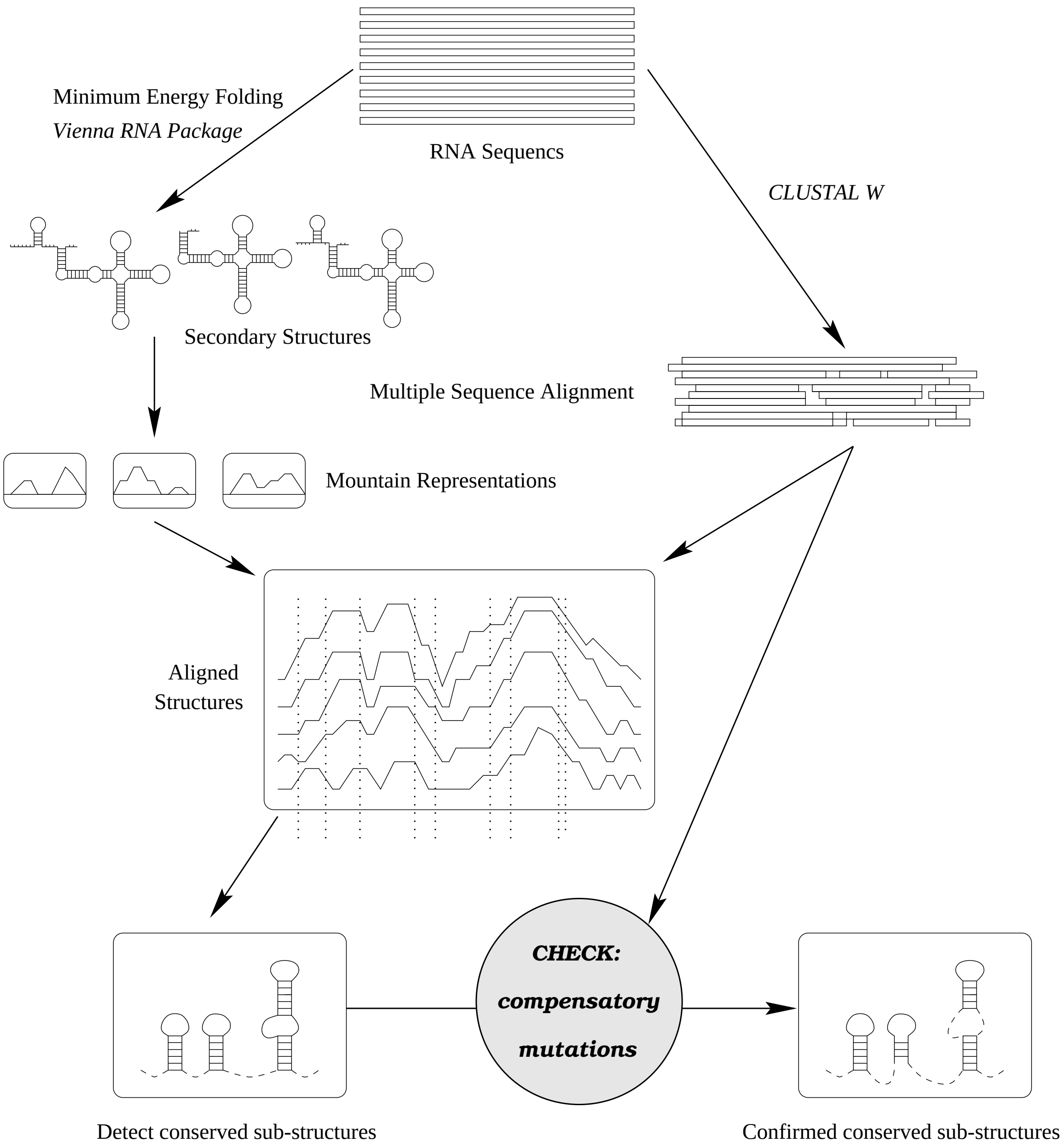

Figure 2: Scheme of the secondary structure analysis of viral genomes.

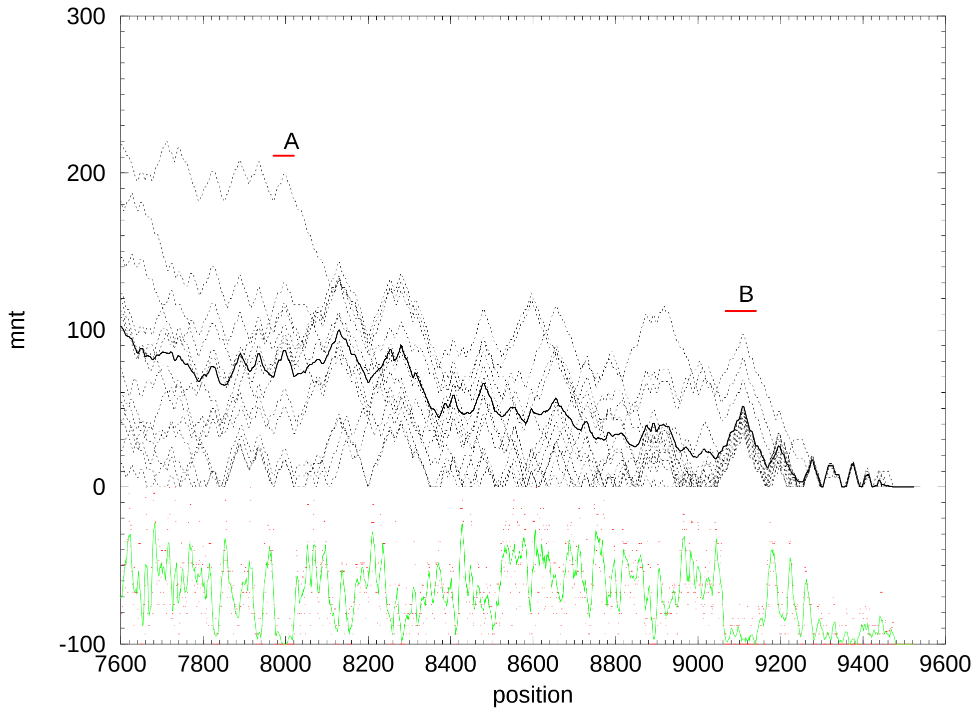

Figure 3: Aligned mountain representations $m(k)$ of the RNA secondary structure of 13 complete Hepatitis C genomes.
The folds were computed with CalTech's `Intel Delta`, a distributed memory parallel computer with 512 nodes and roughly 12MBytes of memory per node. The thick full line is the average mountain representation.
In the lower part of the sequence we plot the variance of the slopes (scattered dots) and a running average (full green line). Deep minima of the green curve correspond to consistently predicted parts of the structure, such as the two regions labelled A and B.

Mountain representation plots, such as the one in figure 3, can be used to identify conserved substructures. The consensus mountain of a set of N sequences can be defined as

$$m(k) = \frac{1}{N} \sum_{s=1}^{N} m_s(k) \tag{1}$$

The quality of a consensus mountain can be assessed at each position by comparing the slopes $q_s(k) := m_s(k) - m_s(k-1)$ of the different sequences [45]. These one-dimensional representations such as $m(k)$ provide a global overview of the structure and can be used to guide a manual reconstruction of consensus secondary structure elements. This approach turned out to be rather tedious already for the 3'NCRs of flaviviruses with a chain length of only 200-300nt and is certainly not feasible for the analysis of entire genomes. On the other hand, the data contained in these simplified one-dimensional representations are not detailed enough to allow for the *automatic* reconstruction of conserved patterns.

2.3. Automatic Detection of Conserved Structural Elements

The starting point of a more detailed analysis is a list of all predicted base pairs. This list will in general not be a valid secondary structure, i.e., it will violate one or both of the following two conditions:
(i) No nucleotide takes part in more than one base pair.
(ii) Base pairs never cross, that is, for two base pairs $(i.j)$ and $(k.l)$ we have either $i < j < k < l$, $k < l < i < j$, $i < k < l < j$, or $k < i < j < l$, but never $i < k < j < l$ or $k < i < l < j$.
Therefore we rank the individual base pairs by their "credibility" (see below). Then we go through the sorted list and weed out all base pairs that violate conditions (i) or (ii). Finally, base pairs that are deemed unlikely are removed.

Clearly the sorting procedure is of crucial importance. For each predicted base pair $(i.j)$ we store the nucleotides occurring in the corresponding positions in the sequence alignment. We shall call a sequence *non-compatible* with a base pair $(i.j)$ if the two nucleotides at positions i and j would form a non-standard base pair such as **CA** or **UU**. A sequence is *compatible* with base pair $(i.j)$ if the two nucleotides form either one of the following six combinations: **GC**, **CG**, **AU**, **UA**, **GU**, **UG**.

When different standard combinations are found for a particular base pair $(i.j)$ we may speak of *consistent* mutations. If we find combinations such as **GC** and **CG** or **GU** and **UA**, where both positions are mutated at once we have compensatory mutations. The occurrence of consistent and, in particular, compensatory mutations strongly supports a predicted base pair, at least in the absence of non-consistent mutations.

From the frequencies f_{ij} with which $(i.j)$ is predicted in the sample of sequences we derive the *pseudo-entropy*

$$S_{ij} = -\sum_k f_{ik} \ln f_{ik} - \sum_k f_{kj} \ln f_{kj} + f_{ij} \ln f_{ij} \, , \tag{2}$$

where $(i.k)$ and $(k.j)$ are the alternative predicted base pairs involving i and j, respectively. The pseudo-entropy is a measure for the reliability with which $(i.j)$ is predicted.

We call a base pair $(i.j)$ *symmetric* if j is the most frequently predicted pairing partner of i and if i is the most frequently predicted pairing partner of j. Note that for each sequence position i there is at most one symmetric base pair involving i. A symmetric base pair $(i.j)$ has necessarily a rather large value f_{ij}; in particular, it does not allow a large number of structural alternatives.

In a first preprocessing step we remove for each i all but the most frequent pair $(i.j)$ from the list of predicted base pairs. The list is then sorted according to the following hierarchical criteria:

(1) The more sequences are non-compatible with $(i.j)$, the less credible is the base pair.
(2) Symmetric base pairs are more credible than other base pairs.
(3) A base pair with more consistent mutations is more credible.
(4) Base pairs with smaller values of the pseudo-entropy S_{ij} are more credible.

Scanning the sorted list from the top, we remove a base pairs if it conflicts with a higher-ranking one that has already been accepted. Finally, base pairs with f_{ij} below some threshold are removed. This ensures that structures are predicted only for regions in which we have a strong signal. The threshold value is a conservative estimate for the reliability of the secondary structure prediction [23]. In this study a value of 0.3 gave good results. The final output can be displayed as a color-coded dot plot (as shown in figures 4 and 5) or as a color-coded mountain plot (figure 6).

The virtue of this approach can be tested quite easily. In figure 4 we compare the predicted consensus structure for two sets of 2-error mutants of the same sequence. If all sequences fold into the same structure we obtain of course a perfect prediction that is supported by a small number of compensatory mutations. A set of randomly generated 2-error mutants does not lead to an acceptable prediction. Even this small amount of sequence heterogeneity leads to a quite diverse set of structures, and hence no unambiguous secondary structure is predicted. Our approach is therefore capable of distinguishing conserved secondary structure elements from pieces of sequence with high degrees homology but without conserved structural features.

Sometimes well-predicted stacked regions are interrupted by individual “holes” or show a single base pair with a few non-compatible sequences. While in many cases these features

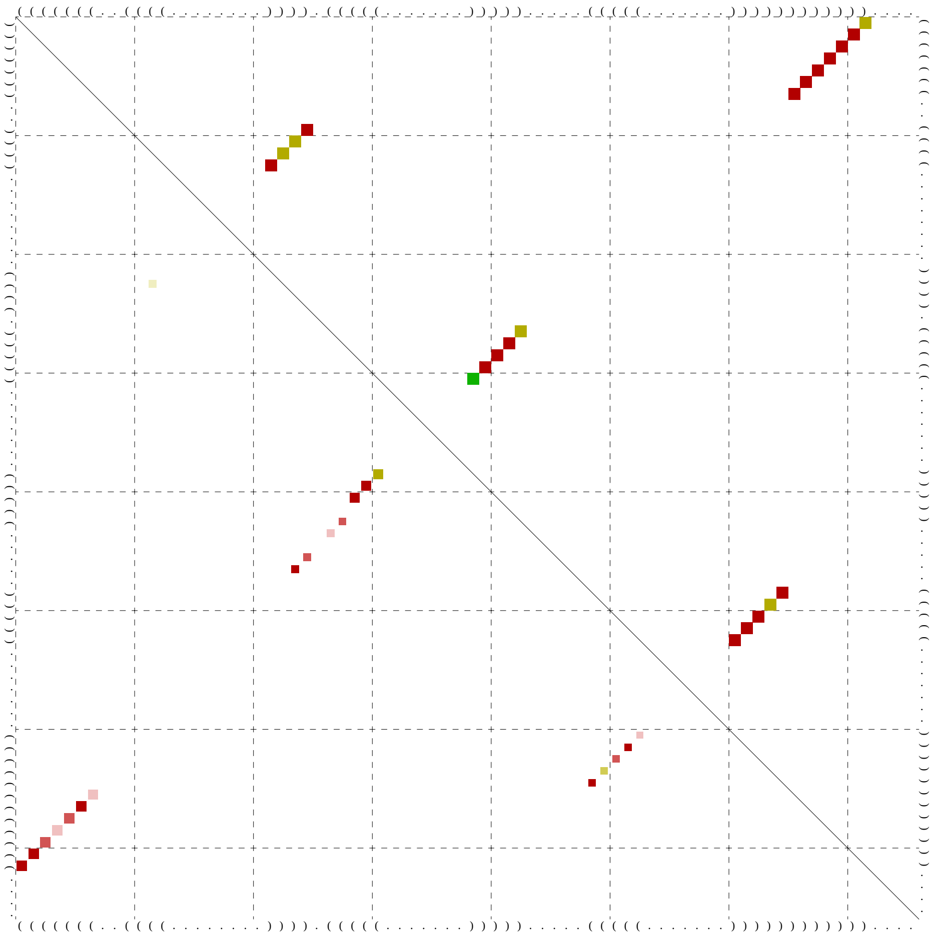

Figure 4: An artificial example: Two samples of 2-error mutants of the tRNA-phe sequence were subjected to our procedure.
A square in row i and column j of the dot plot indicates a predicted pair $(i.j)$. Its size and color indicates the frequency and "credibility" of the base pair. The area of the square is proportional to the frequency f_{ij} with which $(i.j)$ is predicted. Colors indicate the number of consistent mutations: 1, 2, 3 different types of base pairs. Saturated colors, , indicate that there are only compatible sequences. Decreasing saturation of the colors indicates an increasing number of non-compatible sequences: 1, 2 sequences that cannot form a (i, j). If there are more than 2 non-compatible sequences the entry is not displayed.
Upper right triangle: The 29 sequences that fold into the familiar clover-leaf structure of tRNAs lead to a perfect reconstruction of the secondary structure, supported by a (small) number of compensatory mutations. 47 of the 76 sequence positions are conserved.
Lower left triangle: A sample of 20 randomly generated 2-error mutants of the tRNA-phe sequence of yeast does not produce a reasonable prediction of the clover leaf structure: One stack of the clover-leaf is not predicted at all, and another stack does not confirm the wild-type structure. In this sample we have 43 conserved positions. Note that the predictions are weak in this case, there are inconsistent mutants, and there is only a single compensatory mutation.
The comparison of the two data sets shows that our algorithm is very sensitive in detecting regions with *non*-conserved structure even in highly conserved sequences.

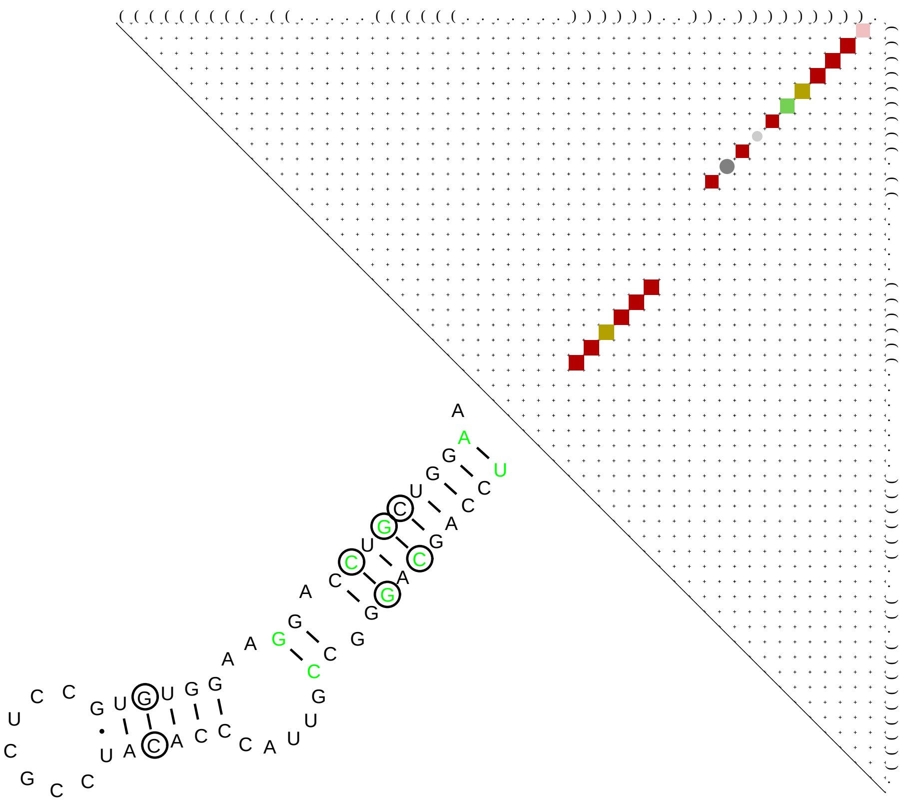

Figure 5: Comparison of predicted minimum energy structures in *region A* (around position 8000) of the HCV genome.

The color coding of the dot plot is explained in the caption of figure 4. The lower left part of the plot shows a conventional picture of the predicted structure. Base pairs marked in green have non-consistent mutations, circles indicate compensatory mutations.

The extended outer stem contains a number of compensatory mutations supporting its existence. Nevertheless, there are two "holes" and one bleached square that at first glance would tempt one to reject the prediction. A close examination shows, however, that the bleached green square belongs to a base pair that is almost always predicted, exhibits 3 different types of standard base pairs and is **UU** in a single sequence. As **UU** mismatches are perfectly stable within extended helices this observation does not contradict a predicted base pair. Similarly, the first hole (shown as a light grey circle in the figure) in this stem is **AC** in 4 sequences and forms three different types of base pairs, namely **GC**, **GU**, and **AU**. The second hole (large grey circle) is a conserved **GA** mismatch that might well be present in the secondary structure.

reflect structural variability or the existence of an internal loop, they can be attributed to non-standard base pairs in other cases. A rather convincing example is shown in figure 5.

3. Results

As first applications of our method we have investigated the minimum energy folding of 13 complete Hepatitis C (HPC) sequences, a sample of 13 complete HIV1 sequences and the S segment of 19 strains of Hantavirus (access codes are listed in the appendix). Minimum free energy structures of the complete HIV and HPC genomes were obtained on CalTech's `Intel Delta`. For details of the parallel computer implementation of the folding algorithm see [17]. The Hanta sequences were folded using the serial version of the folding program, which uses a slightly more recent parameter set [55]. Multiple sequence alignments were obtained using `CLUSTAL W`. Both the folding outputs and the multiple alignments were processed without further modifications.

Hepatitis C sequences have chain lengths around 9500nt. The main differences in length stem from the 3'-end of the genome, where a poly-U region separates a 98nt sequence from the rest of the genome [30, 53]. This so-called X-tail is not present in the published "complete" genomes with a single (very recent) exception (Genbank accession number D85516). In this sequence we found no long range interactions involving the X-tail, and hence no evidence for the panhandle structure postulated in Figure 7B of [30]. We find that the poly-U region acts as a spacer causing the X-tail to fold as a separate domain. It is justified, therefore, to consider the main part of the genome (before the poly-U region) and the X-tail separately.

The length of the `CLUSTAL W` alignment of the main part of the genome, up to the poly-U, is 9538. Insertions or deletions appear in 267 positions before the poly-U. 4919 positions are conserved, the mean pairwise homology of sequences is 80%.

The 5'NCR has recently been studied using a combination of thermodynamic prediction and biochemical methods [5, 51, 21]. Unfortunately, the sequences at the 5'-end are highly

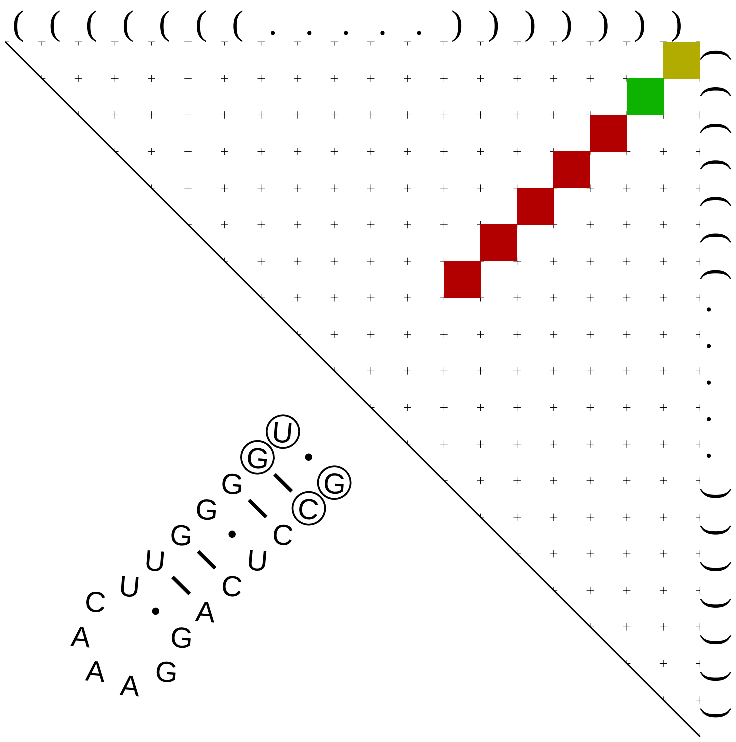

Figure 6: Predicted conserved minimum energy structure of *region B* (see figure 3) of the HCV genome. The color coding of the dot plot is explained in the caption of figure 4. The lower left part of the plot shows a conventional picture of the predicted structure. Circles indicate a compensatory mutation.

conserved (97% pairwise homology, 89% of the 342 positions are conserved). As a consequence of the very small sequence variation our approach is not much better than thermodynamic predictions on a single sequence in this case. We find most of hairpins appearing in the model of Brown [5], but predict no longer range base pairs.

A similar situation is encountered for the X-tail: From the 98 nucleotides only 5 show any sequence variability. There are only 3 different sequences on these 98nt among the 8 database entries currently available. Our data agree with three stem-loops predicted in [4, 26] based on chemical probing. It is also supported by two consistent mutations in the long helix at the very 3'-end.

We do, however, find convincing structural motifs within the coding region of the viral genome. Two examples of which are shown in figures 5 and 6. Although both regions have not been investigated before, the large number of compensatory mutations clearly indicates that these structural motifs are conserved.

The minimum free energy structures of the 13 sequences contain a total of 23186 base pairs. Preprocessing leaves only 2805 list entries. Of these, 432 entries are inconsistent with higher-ranking entries in the sorted list, 572 entries are removed because there are more than 2 inconsistent sequences, and the frequency f_{ij} of 298 of the remaining base pairs is the below the threshold value 0.3. This leaves us with 1503 pairs in the dot plots. Of these, 985 have only compatible sequences, 179 have a single incompatible sequence, 339 have two.

Table 1. Predicted Secondary Structure Elements.

	HPC	HIV 1
Number of sequences N	13	13
min. sequence length	9400	9074
max. sequence length	9502	9292
Alignment length	9538	9535
Conserved positions	4919	4779
Average sequence homology	80%	83%
Different base pairs	23186	20667
Credible base pairs	1503	1121
1 consistent mutations	460	300
2 consistent mutations	80	44
3 consistent mutations	8	2
4 consistent mutations	2	0

As a second example we have re-analyzed a sample of HIV1 sequences from an earlier study [17]. The number of predicted conserved base pairs is similar to the Hepatitis C case, see Table 1 for details. In the following we discuss the automatically generated predictions for two well-understood secondary structure motifs, namely TAR and RRE, in some detail.

At the 5'-end of the viral HIV-1 RNA molecule resides the *trans*-Activating Responsive (TAR) element [2], which interacts with the regulatory Tat protein. The binding of the Tat protein to TAR is responsible for the activation and/or elongation of transcription of the provirus [12, 29]. On the basis of biochemical analysis [1] and computer prediction of the 5'-end of the genome it is known that the TAR region in HIV-1 forms a single, isolated stem loop structure of about 60 nucleotides with about 20 base pairs interrupted by two bulges.

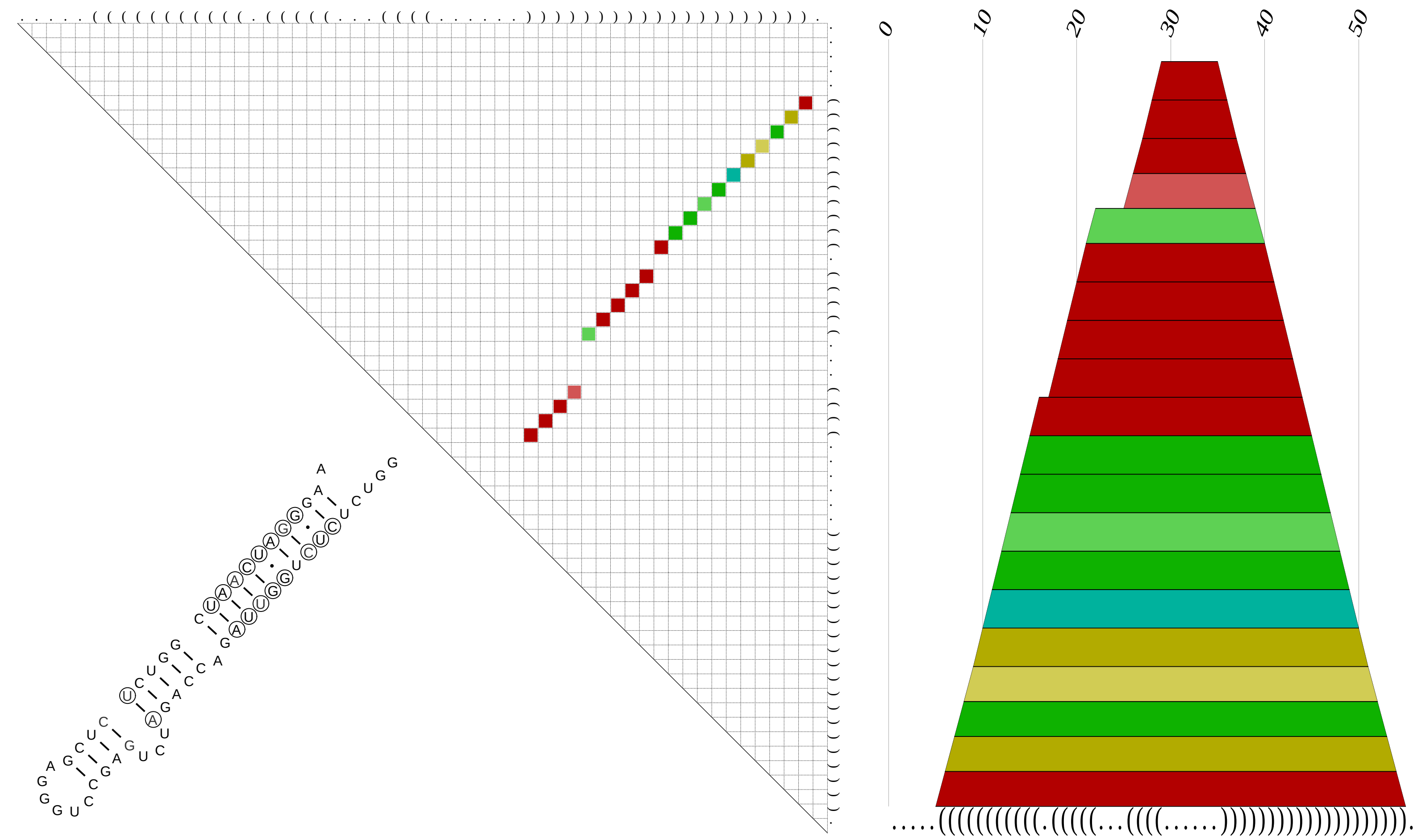

Figure 7: The TAR structure of HIV1. Almost all predicted base pairs are consistent with all 13 sequences, most of them are predicted in at least 11 sequences. A large number of compensatory mutations supports the thermodynamic predictions. Our computed consensus structure (lower left) matches the structure determined by probing and phylogenetic reconstruction [1]. We display here the consensus dot plot, the classical secondary structure, and a mountain representation. The latter is a convenient alternative to dot plots for larger structural motifs. Base pairs are represented by slabs connecting the two sequence positions. The width and color of a slab corresponds to size and color of the corresponding dot plot entry.

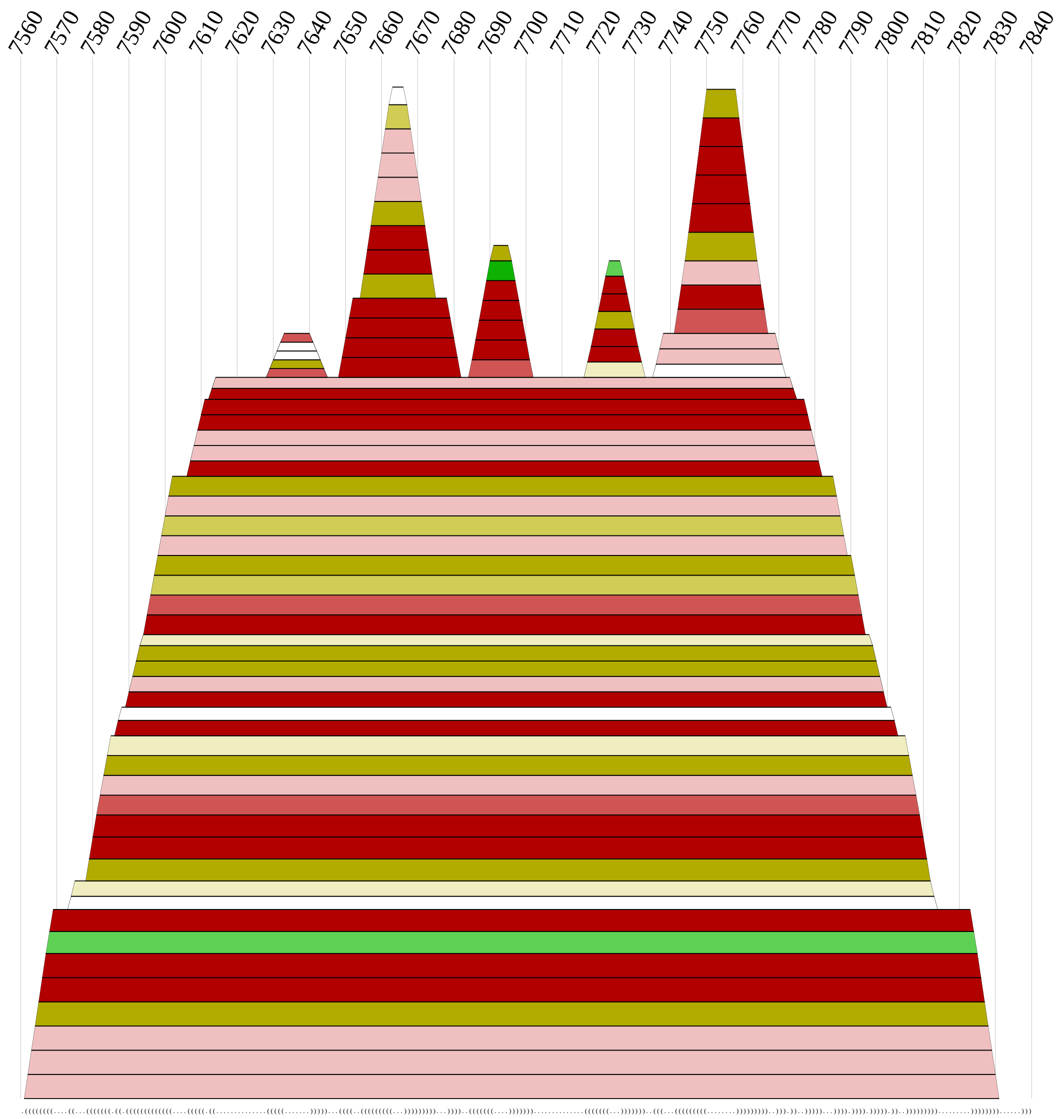

Figure 8: Color-coded mountain plot of the RRE region. The five-fingered structure is clearly visible. The peaks are, from left to right: IIc, III, IV, V, and VI, in the notation of [7]. For structure motifs longer than some 100nt dot plots become hard to interpret.

This structure is indeed predicted in the minimum free energy structures of 11 of the 13 sequences analyzed here. The consensus prediction, figure 7, is identical to the structure reported in the literature.

The Rev response element (RRE), is an important conserved RNA structure that is located within the *env* gene. The interaction of RRE with the Rev protein reduces splicing and increases the transport of unspliced and single-spliced transcripts to the cytoplasm, which

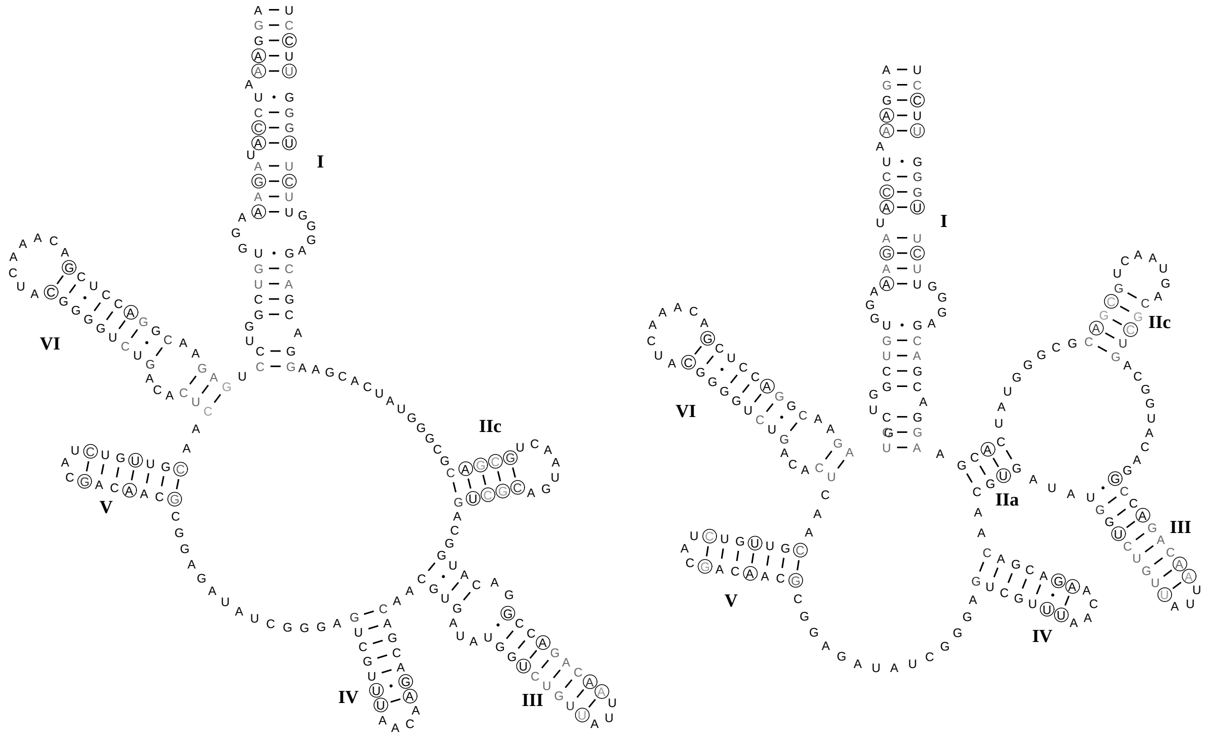

Figure 9: Consensus structures of the HIV1 RRE region from a set 13 sequences and from the 21 sequences reported in [17]. The main hairpins are present in both predictions, the only difference is hairpin **IIa** which is supported by a single compensatory base pair in the larger data set. The predictions agree very well with an experimentally supported structure [7] that also contains **IIa**. The sequence in the **IIa** region is conserved in the smaller data set, and purely thermodynamic considerations favor the short stack extending stack **III** in the right hand structure (this stem is called IIc in [39, 60]). Interestingly, earlier studies [24, 17] indicate a substantial structural versatility in this region which may explain minor disagreements between different published structures, e.g. [7, 39, 60].

is necessary for the formation of new virion particles [37].

A long stem-root structure (I) separates the binding region very well from the rest of the RNA. The long stem-loop structure furthermore indicates that the structure is easily accessible. The consensus secondary structure of the RRE in HIV-1 is a multi-stem loop structure consisting of five hairpins supported by a large stem structure [7], see figures 8 and 9. An alternative structure of only 4 hairpins, in which the hairpins III and IV of the consensus model merge to form one hairpin has also been proposed [39, 60]; it matches the minimum energy structure for some sequences, e.g. HIVLAI [24]. A comparison of minimum energy structures [17, 18] shows that there appears to be a third structure in which hairpin III is relatively large and a few of the other hairpins have disappeared from the minimum free energy structure. A comprehensive analysis of the base pairing probabilities in the RRE shows that the hairpins II, IV, and V, as well as the basis of hairpin III are meta-stable in the sense that they allow for different structures with comparable probabilities [25].

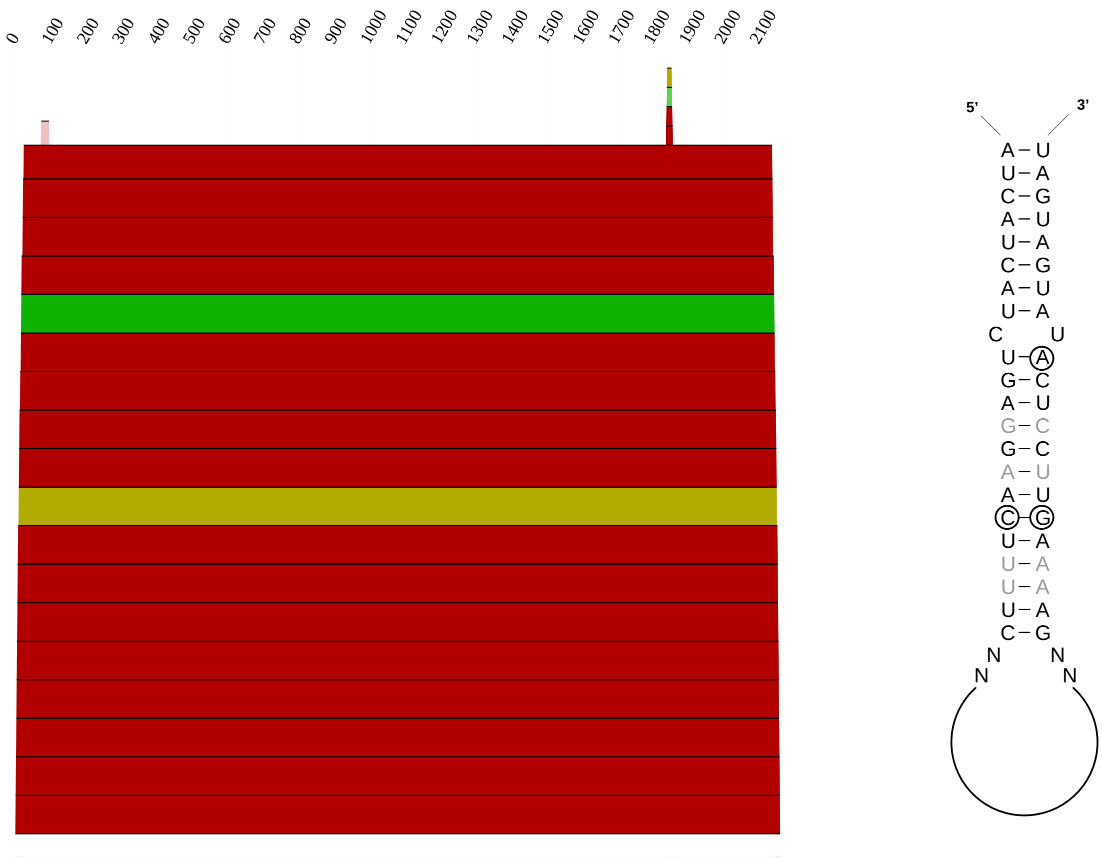

Figure 10: Consensus structures for the minus strand of the S segment of hantavirus. The only consistently predicted structure is a *panhandle* formed by 5' and 3'-ends.

As a final example we consider the S segment of Hantavirus. The approximately 1700nt long S segment contains a single ORF encoding a nucleocapsid (N) protein. In contrast to other members of the family Bunyaviridae there is no evidence for a second non-structural (NS_s) protein coded by the S segment. We used the 19 sequences listed in the appendix, which have a mean pairwise homology of 63.9%. The only detected structural feature in this case is a 19 base-pair stem-loop structure formed by the 5' and 3' ends.

This *panhandle structure* is highly significant: All sequences are compatible with the structure and it is part of the minimum energy prediction in 16 of the 19 minus strands and 14 plus strands. There are two positions which show compensatory mutations, see figure 10. The panhandle structure was postulated already in the eighties for all bunyaviridae [42, 49].

4. Discussion

We have presented a combination of secondary structure prediction based on thermodynamic criteria and sequence comparison that is capable of reliably identifying conserved structural features in a set of related RNA molecules. The method has been designed for routine investigations of large RNA molecules such as complete viral genomes. Indeed, the procedure does not require any interventions: `Clustal W` alignments and minimum energy structures (as obtained from the `Vienna RNA Package` or Zucker's `mfold` [61]) can be used "as is".

Conserved secondary structures are likely to be functional, thus our method can be used to find functional secondary structures. Since our method emphasizes sequence variation, it can complement other methods for finding functional RNA secondary structures based on thermodynamic prediction such as [33, 24].

We have applied this technique to complete genomes of three quite different species of RNA viruses: HIV-1, Hepatitis C virus (HPC), and the small segment of hantavirus. In all cases we have been able to identify most of the known secondary structure features. In addition, we predict a large number of conserved structural elements which have not been described so far.

We have designed our approach in such a way that it does not predict a structure for all parts of a molecule, the filtering procedure outlined in Methods is designed in such a way that only base pairs that may occur in almost all sequences and that are predicted in a sizeable fraction of the sequences will be accepted. It is not surprising, therefore, that the predicted RRE structures in figure 9 do not contain every single base pair of the published, experimentally supported structures. Rather, we obtain a subset of base pairs that is consistent with known features. This suggest that we are not producing a large number of false positives. On the other hand, we recover most of the structures described for both HIV-1 and HPC, as well as the panhandle structure of hantavirus.

Extensive computer analysis of the RRE region of HIV-1 and HIV-2 has shown that the sequence alignment does not completely coincide with the alignment at the level of the secondary structure [31]. This has two important implications: 1) methods that predict secondary structure of RNA on the basis of co-variation of positions within the sequence [14] cannot provide an unambiguous answer here, and 2) the RRE has structural versatility. As a consequence we obtain slightly different predictions for the conserved structure depending on the set of sequences used for the analysis. This structural versatility could also play a role in a single HIV clone. Using McCaskill's algorithm [40] for predicting the matrix of base pairing probabilities we have indeed identified a spectrum of alternative structures for the RRE of HIV-LAI in a previous communication [24]. Similar features have been detected in the 3'NCR of flaviviruses [45, 38]. These facts make it worthwhile to generalize the present approach to using base pairing probability matrices instead of minimum energy structures despite the substantial increase in required computer resources. Preliminary data indicate a promising increase in the accuracy of predicted structures.

Acknowledgments

This research was performed in part using the CACR parallel computer system operated by Caltech on behalf of the Center for Advanced Computing Research. Access to this facility was provided by the California Institute of Technology. Partial financial support by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Proj. No. P 12591-INF, is gratefully acknowledged. We'd like to thank Andreas Wagner for his comments.

Appendix

In this study we have used the following viral RNA sequences. Genbank accession numbers are given in parenthesis.

HIV-1:

HIVANT70 (M31171, L20587), HIVBCSG3C (L02317), HIVCAM1 (D10112, D00917), HIVD31 (X61240, X16109 U23487), HIVELI (K03454, X04414), HIVLAI (K02013), HIVMAL (K03456), HIVMVP5180 (L20571), HIVNDK (M27323), HIVOYI (M26727), HIVRF (M17451, M12508), HIVU455 (M62320), and HIVZ2Z6 (M22639).

HPC: complete genomes (except for the X-tail):

HCU16362 (U16362), HCU45476 (U45476), HPCCGAA (M67463), HPCCGENOM (L02836), HPCCGS (D14853), HPCEGS (D17763), HPCHCJ1 (D10749), HPCJ483 (D13558, D01217), HPCJRNA (D14484, D01173), HPCJTA (D11168, D01171), HPCK3A (D28917), HPCPP (D30613), and HPCRNA (D10934).

X-tail sequences (accession numbers only):

D63922, D67091, D67092, D67093, D67094, D67095, D67096, D85516.

The last sequence is a complete genome including the X-tail.

Hantavirus sequences:

AF004660, HNVNPSS, HVU37768, HMU32591, HSU29210, KHU35255, AF005727, PHU47136, PHVSSEG, PSU47135, PUUSNP, PUVSVIN83, PUVSVIRRT, PVSZ84204, PVU22423, VRANICAS, RMU52136, HPSNUPR, TUVS5302.

The comparison algorithm described in section 3.3 is implemented as an `ANSI C` program `alidot`. It generates a text file with information on all predicted base pairs and a postscript file of the dot plot of the predicted conserved base pairs. Alternative representations, such as the aligned mountain plots, input files for XRNA [58] and post-processing of XRNA output is handled by a collection of `perl` scripts. This software is available upon request from the authors.

References

[1] F. Baudin, R. Marquet, C. Isel, J. L. Darlix, B. Ehresmann, and C. Ehresmann. Functional sites in the 5' region of human immunodeficiency virus type 1 RNA form defined structural domains. *J. Mol. Biol.*, 229:382–397, 1993.

[2] B. Berkhout. Structural features in TAR RNA of human and simian immunodeficiency viruses: a phylogenetic analysis. *Nucl. Acids Res.*, 20:27–31, 1992.

[3] C. Biebricher. The role of RNA structure in RNA replication. *Ber. Bunsenges. Phys.Chem.*, 98:1122–1126, 1994.

[4] K. J. Blight and C. M. Rice. Secondary structure determination of the conserved 98-base sequence at the 3' terminus of hepatitis C virus genome RNA. *J. Virol.*, 71:7345–7352, 1997.

[5] E. A. Brown, H. Zhang, L.-H. Ping, and S. M. Lemon. Secondary structure of the 5' nontranslated regions of hepatitis C virus and pestivirus genomic RNAs. *Nucl. Acids Res.*, 20:5041–5045, 1992.

[6] 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.

[7] E. T. Dayton, D. A. Konings, D. M. Powell, B. A. Shapiro, L. Butini, J. V. Maizel, and A. I. Dayton. Extensive sequence-specific information throughout the CAR/RRE, the target sequence of the human immunodeficiency virus type 1 Rev protein. *J. Virol.*, 66:1139–1151, 1992.

[8] R. Deng and K. V. Brock. 5' and 3' untranslated regions of pestivirus genome: primary and secondary structure analyses. *Nucl. Acids Res.*, 21:1949–1957, 1993.

[9] 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.

[10] M. Eigen, J. McCaskill, and P. Schuster. The molecular quasispecies. *Adv. Chem. Phys.*, 75:149–263, 1989.

[11] R. M. Elliott, C. S. Schmaljohn, and M. S. Collett. Bunyavirus genome structure and gene expression. *CurrentTopics in Microbiology and Immunology*, 169:91–141, 1991.

[12] S. Feng and E. Holland. HIV-1 tat trans-activation requires the loop sequence within tar. *Nature*, 334:165–167, 1988.

[13] S. M. Freier, R. Kierzek, J. A. Jaeger, N. Sugimoto, M. H. Caruthers, T. Neilson, and D. H. Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci., USA*, 83:9373–9377, 1986.

[14] R. R. Gutell. Evolutionary characteristics of RNA: Inferring higher-order structure from patterns of sequence variation. *Curr. Opin. Struct. Biol*, 3:313–322, 1993.

[15] L. He, R. Kierzek, J. SantaLucia, A. E. Walter, and D. H. Turner. Nearest-neighbor parameters for GU mismatches. *Biochemistry*, 30, 1991.

[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, M. A. Huynen, P. F. Stadler, and P. E. Stolorz. Knowledge discovery in RNA sequence families of HIV using scalable computers. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR*, pages 20–25, Portland, OR, 1996. AAAI Press.

[18] I. L. Hofacker, M. A. Huynen, P. F. Stadler, and P. E. Stolorz. RNA folding and parallel computers: The minimum free energy structures of complete HIV genomes. Technical report, SFI, Santa Fe, New Mexico, 1996. # 95-10-089.

[19] M. A. Hoffman and A. C.Palmenberg. Mutational analysis of the J-K stem-loop region of the encephalomyocarditisvirus IRES. *J. Virol.*, 69:4399–406, 1995.

[20] P. Hogeweg and B. Hesper. Energy directed folding of RNA sequences. *Nucl. Acids Res.*, 12:67–74, 1984.

[21] M. Honda, E. A. Brown, and S. M. Lemon. Stability of a stem-loop involving the initiator AUG controls the efficiency of internal initiation of translation on hepatitis c virus RNA. *RNA*, 2:955–968, 1996.

[22] M. Huynen and D. Konings. Questions about RNA structures in HIV and HPV. In G. L. Myers, editor, *Viral Regulatory Structures and Their Degeneracy*, volume XXVIII of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 69–82, Reading, MA, 1998. Addison Wesley Longman.

[23] M. A. Huynen, R. Gutell, and D. A. M. Konings. Assessing the reliability of RNA folding using statistical mechanics. *J. Mol. Biol.*, 265:1104–1112, 1997.

[24] M. A. Huynen, A. S. Perelson, W. A. Viera, and P. F. Stadler. Base pairing probabilities in a complete HIV-1 RNA. *J. Comp. Biol.*, 3:253–274, 1996.

[25] M. A. Huynen, P. F. Stadler, and W. Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proc. Natl. Acad. Sci., USA*, 93:397–401, 1996.

[26] T. Ito and M. M. C. Lai. Determination of secondary structure and cellular protein binding to the 3'-untranslated region of the hepatitis c virus RNA genome. *J. Virol.*, 71:8698–8706, 1997.

[27] R. J. Jackson and A. Kaminski. Internal initiation of translation in eukaryotes: the picornavirus paradigm and beyond. *RNA*, 1:985–1000, 1995.

[28] J. A. Jaeger, D. H. Turner, and M. Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci., USA*, 86:7706–7710, 1989.

[29] B. Klaver and B. Berkhout. Evolution of a disrupted TAR RNA hairpin structure in the HIV-1 virus. *EMBO J.*, 13:2650–2659, 1994.

[30] A. Kolykhalov, S. Feinstone, and C. M. Rice. Identification of a highly conserved sequence element at the 3'terminus of hepatitis C virus genome RNA. *J. Virology*, 70:3363–3371, 1996.

[31] D. A. M. Konings. Coexistence of multiple codes in messenger RNA molecules. *Comp. & Chem.*, 16:153–163, 1992.

[32] D. A. M. Konings and P. Hogeweg. Pattern analysis of RNA secondary structure, similarity and consensus of minimal-energy folding. *J. Mol. Biol.*, 207:597–614, 1989.

[33] S.-Y. Le, J.-H. Chen, K. Currey, and J. Maizel. A program for predicting significant RNA secondary structures. *CABIOS*, 4:153–159, 1988.

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

[35] 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.

[36] F. Major, M. Turcotte, D. Gautheret, G. Lapalme, E. Fillion, and R. Cedergren. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science*, 253:1255–1260, 1991.

[37] M. H. Malim, J. Hauber, S. Y. Le, J. V. Maizel, and B. Cullen. The HIV-1 Rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature*, 338:254–257, 1989.

[38] 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.

[39] D. Mann, I. Mikaelian, R. Zemmel, S. Green, A. Lowe, T. Kimura, M. Singh, P. Butler, M. Gait, and J. Karn. A molecular rheostat. Co-operative rev binding to stem I of the rev-response element modulates human immunodeficiency virus type-1 late gene expression. *J. Mol. Biol.*, 241:193–207, 1994.

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

[41] T. P. Monath and F. X. Heinz. Flaviviruses. In B. N. Fields, D. M. Knipe, P. M. Howley, R. M. Chanock, J. L. Melnick, T. P. Monath, B. Roizmann, and S. E. Straus, editors, *Fields Virology*, pages 961–1034. Lippincott-Raven, Philadelphia, 3rd edition, 1996.

[42] P. V. N. Paradigon, M. Girard, and M. Bouloy. Panhandles and hairpin structures at the termini of germiston virus RNAs (bunyavirus). *Virology*, 122:191–197, 1982.

[43] R. C. L. Olsthoorn, G. Garde, T. Dayhuff, J. F. Atkins, and J. van Duin. Nucleotide sequence of a single-stranded RNA phage from *pseudomonas aeruginosa*: Kinship to

coliphages and conservation of regulatory RNA structures. *Virology*, 206:611–625, 1995.

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

[45] S. Rauscher, C. Flamm, C. Mandl, F. X. Heinz, and P. F. Stadler. Secondary structure of the 3'-non-coding region of flavivirus genomes: Comparative analysis of base pairing probabilities. *RNA*, 3:779–791, 1997.

[46] C. M. Rice. Flaviviridae: the viruses and their replication. In B. N. Fields, D. M. Knipe, P. M. Howley, R. M. Chanock, J. L. Melnick, T. P. Monath, B. Roizmann, and S. E. Straus, editors, *Fields Virology*, pages 931–959. Lippincott-Raven, Philadelphia, 3rd edition, 1996.

[47] V. M. Rivera, J. D. Welsh, and J. Maizel, J.V. Comparative sequence analysis of the 5' noncoding region of the enteroviruses and rhinoviruses. *Virology*, 165:42–50, 1988.

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

[49] C. S. Schmaljohn, G. B. Jennings, J. Hay, and J. M. Dalrymple. Coding strtegy of the S genome segment of hantaan virus. *Virology*, 155:633–643, 1986.

[50] P.-Y. Shi, M. A. Brinton, J. M. Veal, Y. Y. Zhong, and W. D. Wilson. Evidence for the existence of a pseudoknot structure at the 3' terminus of the flavivirus genomic RNA. *Biochemistry*, 35:4222–4230, 1996.

[51] D. B. Smith, J. Mellor, L. M. Jarvis, F. Davidson, J. Kolberg, M. Urdea, P. Yap, P. Simmonds, and The International HCV Collaborative Study Group. Variation of the hepatitis C virus 5' non-coding region: implications for secondary structure, virus detection and typing. *J. Gen. Virol.*, 76:1749–1761, 1995.

[52] J. E. Tabaska and G. D. Stormo. Automated alignment of RNA sequences to pseudoknotted structures. In T. Gaasterland, P. Karp, K. Karplus, C. Ouzounis, C. Sander, and A. Valencia, editors, *Proceedings of the ISMB-97*, pages 311–318, Menlo Park, CA, 1997. AAAI Press.

[53] T. Tanaka, N. Kato, M.-J. Cho, K. Sugiyama, and K. Shimotohno. Structure of the 3' terminus of the hepatitis C virus genome. *J. Virol.*, 70:3307–3312, 1996.

[54] 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.

[55] A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Müller, D. H. Mathews, and M. Zuker. Co-axial stacking of helixes enhances binding of oligoribonucleotides and improves predicions of RNA folding. *Proc. Natl. Acad. Sci. USA*, 91:9218–9222, 1994.

[56] K. Wang, Q. Choo, A. Weiner, J. Ou, R. Najarian, R. Thayer, G. Mullenbach, K. Denniston, J. Gerin, and M. Houghton. Structure, sequence and expression of the hepatitis delta (delta) viral genome. *Nature*, 323(6088):508–514, 1986.

[57] M. S. Waterman. Secondary structure of single - stranded nucleic acids. *Adv. math. suppl. studies*, 1:167–212, 1978.

[58] B. Weiser and H. Noller. `XRNA`. `ftp://fangio.ucsc.edu/pub/XRNA/`. (Public Domain Software).

[59] P. R. Wills and A. J. Hughes. Stem loops in HIV and prion protein mRNAs. *J. AIDS*, 3:95–97, 1990.

[60] R. W. Zemmel, A. C. Kelley, J. Karn, and P. J. G. Butler. Flexible regions of RNA structure facilitate cooperative rev assembly of the rev-response element. *J. Mol. Biol.*, 258:763–777, 1996.

[61] M. Zuker. `mfold-2.3`. `ftp://snark.wustl.edu/`. (Free Software).

[62] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46:591–621, 1984.