Automatic Detection of Conserved Base Pairing Patterns in RNA Virus Genomes

IVO L. HOFACKER[a], 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.

[*]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

Keywords

RNA Secondary Structure Prediction, Structure Alignment, Conserved Substructures, Compensatory Mutations, RNA Virus Genomes

Abstract

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 fact, we may not expect a conserved secondary structure for all parts of a viral genome or a mRNA, even if there is a significant level of sequence conservation.
We present a novel method for detecting conserved RNA secondary structures in a family of related RNA sequences. Our method is based on a combination of predicting the matrices of all equilibrium base pairing probabilities and comparative sequence analysis. In contrast to purely phylogenetic methods, our algorithm can be used for small data sets of about ten sequences, efficiently exploiting the information contained in the sequence variability.
The procedure was tested on artificial data sets showing that it indeed detects very few false positives. Applications to some RNA viruses (HIV1, and Hanta virus) use the complete genomic RNAs. In all cases we have been able to identify most of the known secondary structure features. In addition, we predict a substantial number of conserved structural elements which have not been described so far.

1. Introduction

Many functional RNA molecules have distinctive secondary structures; examples include tRNA, rRNA, RNAse P RNA, and elements of RNA virus genomes like HIV's RRE or the IRES region of picorna viruses. While the homologies of the primary sequences of, say 23S RNAs, is easily recognizable, there is a substantial sequence variation across the "tree of life". In the case of rRNA this sequence variation has been used to infer the secondary structure using sequence covariation, see e.g. [12].

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. Extensive computer simulations [11, 43] showed that a small number of point mutations is very likely to cause large changes in the secondary structures. About 10% difference in the nucleic acid sequence leads almost surely to unrelated structures if the mutated sequence 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 elements then it must of course have some function.

In this contribution we show how this observation can be used to design an algorithm that reliably detects conserved, and therefore most likely functional, RNA secondary structure elements in viral genomes. Of course, our method cannot tell us what the function could be, but it can be used to guide, for instance, deletion studies [29].

Several algorithms exist for the prediction of RNA secondary structures based on thermodynamic rules. The most widely used methods compute a single minimum energy structure through dynamic programming [34, 50, 55]. Because of inaccuracies of the energy model and the measured parameters, the accuracy of these predictions is often insufficient. In cases where the correct structure is known from phylogenetic analysis it has been found that predicted structures contain only 30% to 80% of the correct base pairs [24, 21]. The correct structure can, however, be found within a relatively small energy interval above the ground state.

There are variants of the folding algorithm for computing a sample of suboptimal folds [54], or even *all* structures within a prescribed energy range [53]. A particularly elegant solution is the computation of the complete matrix of base paring probabilities [31], which contains suitably weighted information about all possible secondary structures and therefore reduces the impact of inaccuracies in the structure prediction. The disadvantage of these methods is of course that they leave it up to the user to decide which of the proposed structures to believe.

In contrast to thermodynamic predictions, the phylogenetic approach is based exclusively on the analysis on sequence variation [12], but usually requires a very large number of sequences. With the exception of rRNAs and tRNAs, such data sets are rarely available.

The method presented in this paper aims to utilize the information contained in a multiple alignment of a small set of related sequences to extract conserved features from the pool of plausible structures generated by thermodynamic prediction for each sequence, Figure 1. Related procedures are described in [28, 14]. This ansatz is different from efforts to combine alignment and structure prediction [41, 46, 4]. One disadvantage of these methods in the much higher computational cost which makes them unsuitable for long sequences such as viral genomes. Furthermore they assume implicitly that all sequences have a common structure, not just a few conserved structural features.

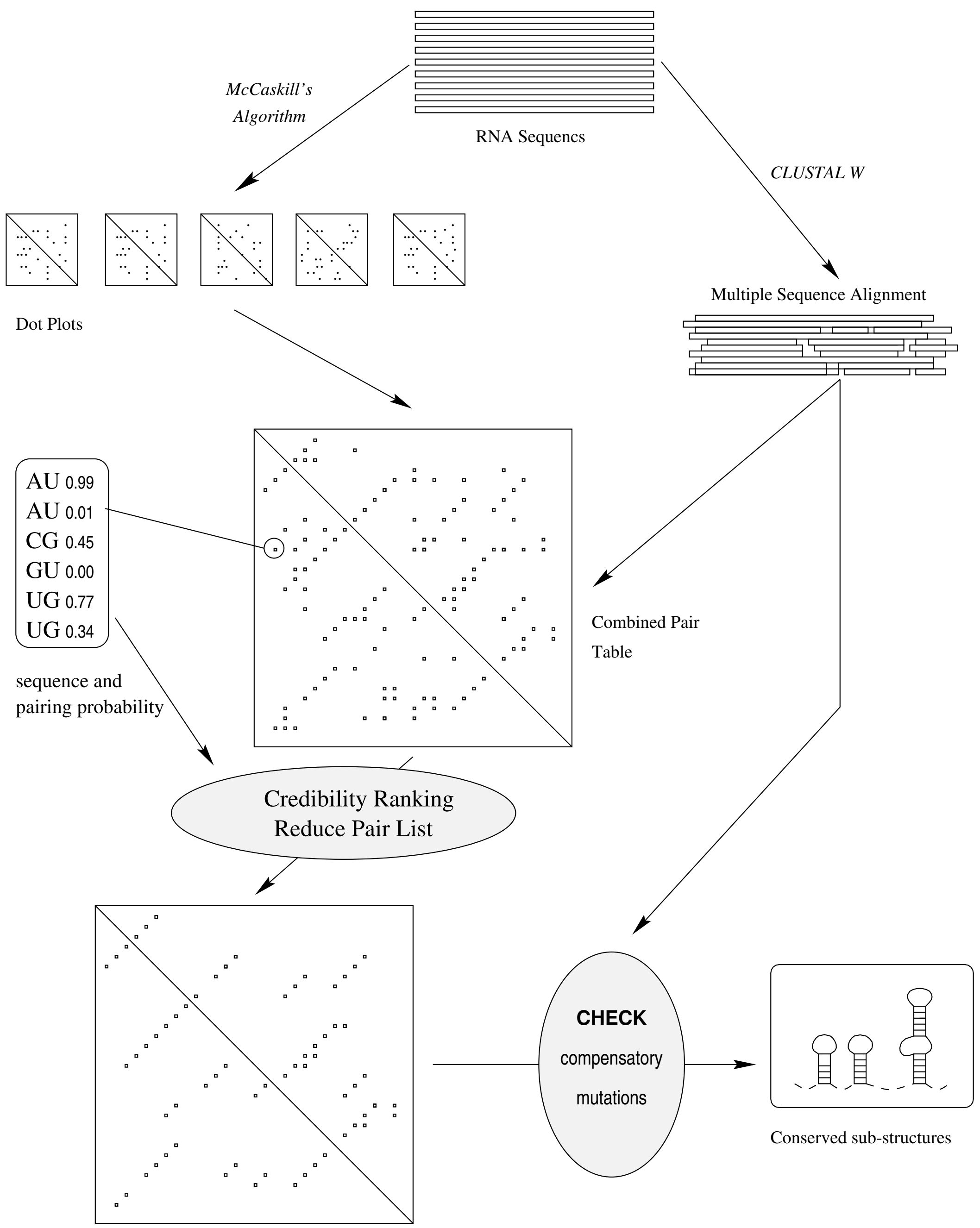

Figure 1: Flow diagram of the algorithm `pfrali`.
A multiple sequence alignment is calculated using `CLUSTAL W` [48]. RNA genomes are folded using McCaskill's partition function algorithm as implemented in the `Vienna RNA Package`. The sequence alignment is then used to align the predicted structures. From this structural alignment we extract putative conserved regions. In the final step the sequence information, in particular compensatory mutations, are used for validating or rejecting predicted structure elements.

2. The Algorithm `pfrali`

The basic two inputs for our algorithm are a multiple sequence alignment and the base pair probabilities from McCaskill's algorithm.

We calculate the multiple sequence alignment using `CLUSTAL W` [48]. No attempt is made to improve the alignment based on predicted secondary structures. While this might increase the number of predicted structural elements, it would also compromise the use of the sequence data for verifying these structures. Furthermore we find that most regions that have functional secondary structure tend to align fairly well, at least locally.

In an earlier paper we restricted ourselves to minimum energy structures [14], i.e., one structure per sequence. Here we extend the method to use base pairing probabilities as obtained from McCaskill's partition function algorithm [31]. Since the base pairing probabilities contain information about a large number of plausible structures, this approach is less likely to miss parts of the correct structures. A similar approach underlies the program `ConStruct` developed by Detlev Riesner's group [28]. The main difference between `ConStruct` and our program `pfrali` is that we make explicit use of the sequence variation to select the credible parts of the predicted structures. Furthermore, we do not assume *a priori* that there is a conserved secondary structure for all (or even most) parts of the sequence.

Base pair probability matrices are conveniently displayed as "dot plots". The `Vienna RNA Package`[1] [15] contains an efficient implementation of McCaskill's algorithm that produces dot plots in `PostScript` format, Figure 2.

Our program reads the pair probabilities from these files as well as a multiple sequence alignment in `CLUSTAL W` format. The gaps in the alignment are inserted into the corresponding probability matrices. We can now superimpose the probability matrices of the individual sequences to produce a *combined dot plot*. To keep the number of base pairs manageable we keep only pairs that occur with a probability of at least $p^* = 10^{-3}$ for at least one sequence. In the combined dot plot the area of a dot at position i, j is proportional to the mean probability $\bar{p}_{i.j}$ (averaged over all sequences). In addition we use a color coding to represent the sequence information.

[1] `http://www.tbi.univie.ac.at/~ivo/RNA/`

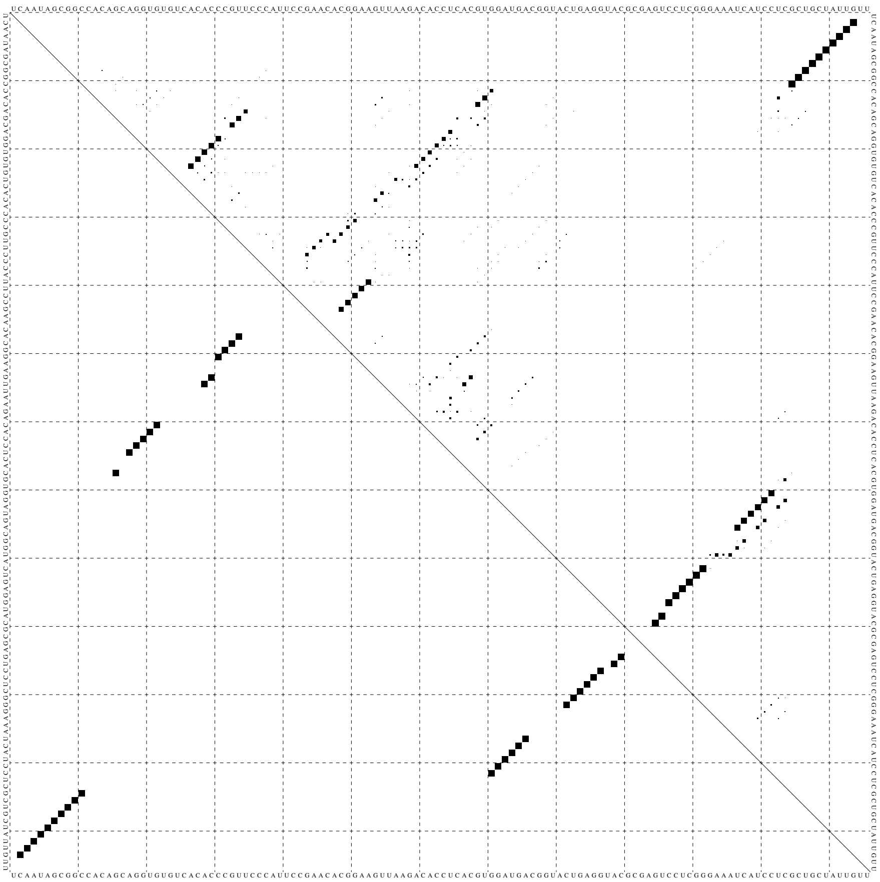

Figure 2: Dot plot of the 5S RNA of *methanospirillum hungatii*. The upper right part shows the predicted base pair probabilities computed with the partition function algorithm of the `Vienna RNA Package`. The area of the squares is proportional to the pairing probability. The lower left part gives the phylogenetic structure from the `Berlin RNA Databank` for comparison. Note that the 5' stem-loop has only low probability in the thermodynamic prediction.

A sequence is *compatible* with base pair $(i.j)$ if the two nucleotides at positions i and j of the multiple alignment can form either a Watson-Crick **GC**, **CG**, **AU**, or **UA**) pair of a wobble (**GU**, **UG**) pair. When different pairing combinations are found for a particular base pair $(i.j)$ we 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.

Phylogenetic methods in general consider only compensatory mutations even though **GU** base pairs are clearly important as evidenced by the fact that **RY**→**YR** conversions are rare [13]. While compensatory mutations of the type **RY→RY**, such as **GC→AU**, can be obtained by two subsequent consistent point mutations, for instance **GC→GU→AU**, a double mutation is required for **RY→YR** mutations. We argue therefore that all consistent mutations, not only compensatory ones, should be seen as support for a proposed structure.

The sequence variation, the number of non-compatible sequences, and the number $c_{i.j}$ of different pairing combinations is incorporated in the combined dot plot as color information. For the details of the encoding scheme see the caption the color plate Figure 3.

The base pairs contained in the combined dot plot will in general not be a valid secondary structure, i.e., they 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, there may not be two base pairs $(i.j)$ and $(k.l)$ such that $i < k < j < l$.

In the reminder of this section we describe how to extract credible secondary structures from the list of base pairs. The procedure is similar in spirit to the algorithm `alidot` which detects conserved structures using a multiple alignment and minimum energy structures as input [14].

In essence, we rank the individual base pairs by their "credibility", using the following criteria:

(1) The more sequences are non-compatible with $(i.j)$, the less credible is the base pair.
(2) If the number of non-compatible sequences is the same, then the pairs are ranked by the product $\bar{p}_{i.j} \times c_{i.j}$ of the mean probability and the number of different pairing combinations.

Then we go through the sorted list and remove all base pairs that conflict with a higher ranked pair by violating conditions (i) or (ii).

The list now represents a valid secondary structure, albeit still containing ill-supported base pairs. We therefore remove all pairs with more than two non-compatible sequences (for large samples this number may have to be increased), as well as pairs with two non-compatible sequences adjacent to a pair that also has non-compatible sequences. Next we omit all isolated base pairs. The remaining pairs are collected into helices and in the final filtering step only helices are retained that satisfy the following conditions:
(i) the highest ranking base pair must not have non-compatible sequences.

(ii) for the highest ranking base pair the product $\bar{p}_{i.j} \times c_{i.j}$ must be greater than 0.3.
(iii) if the helix has length 2, it must not have more non-compatible sequences than consistent mutations.
The resulting structure is the prediction of the `pfrali` program.

The final output of the program consists of a color coded dot plot in PostScript format, as well as a text output containing the sorted list of all base pairs and the final structure. Additional tools are provided to produce annotated secondary structure plots from these data.

3. A Simple Example: 5S RNAs

As a first test of our procedure we applied our method to samples from the `Berlin RNA Databank` [45] which contains 5S RNA sequences and their phylogenetic structures. In order to "simulate" a real life problem, we randomly selected 9 to 20 sequences from different subgroups with different sequences heterogeneities. The base pairs of the phylogenetic structure on average appear with a probability of some 66% in the pair probability matrix. However, different pairs are predicted with very different probabilities and the quality of the predictions differs greatly for different sequences.

For a relatively homogeneous sample such as *Methanomicrobiales* we obtain an almost perfect prediction with only one small helix of two base pairs missing. The mean pairwise sequence identity in this set is about 75%. As sequences become more heterogeneous one finds more compensatory mutations, but the quality of the alignment often suffers, leading to seemingly non-compatible sequences even when a conserved secondary structure exists. Nevertheless, for a heterogeneous sample, such as a random selection of 10 sequences from the kingdom *Archaea* with only 58% mean pairwise identity, we still find most of the correct structure missing only one larger helix (see fig. 4). Still the method does not produce false positives.

We use this example also to explain the output of `pfrali`. A detail from the combined dot plot is shown in Figure 3. This display contains two distinct data sets. In the upper right half the combined dot plot of all predicted bases pairs is shown. The list of these pairs is

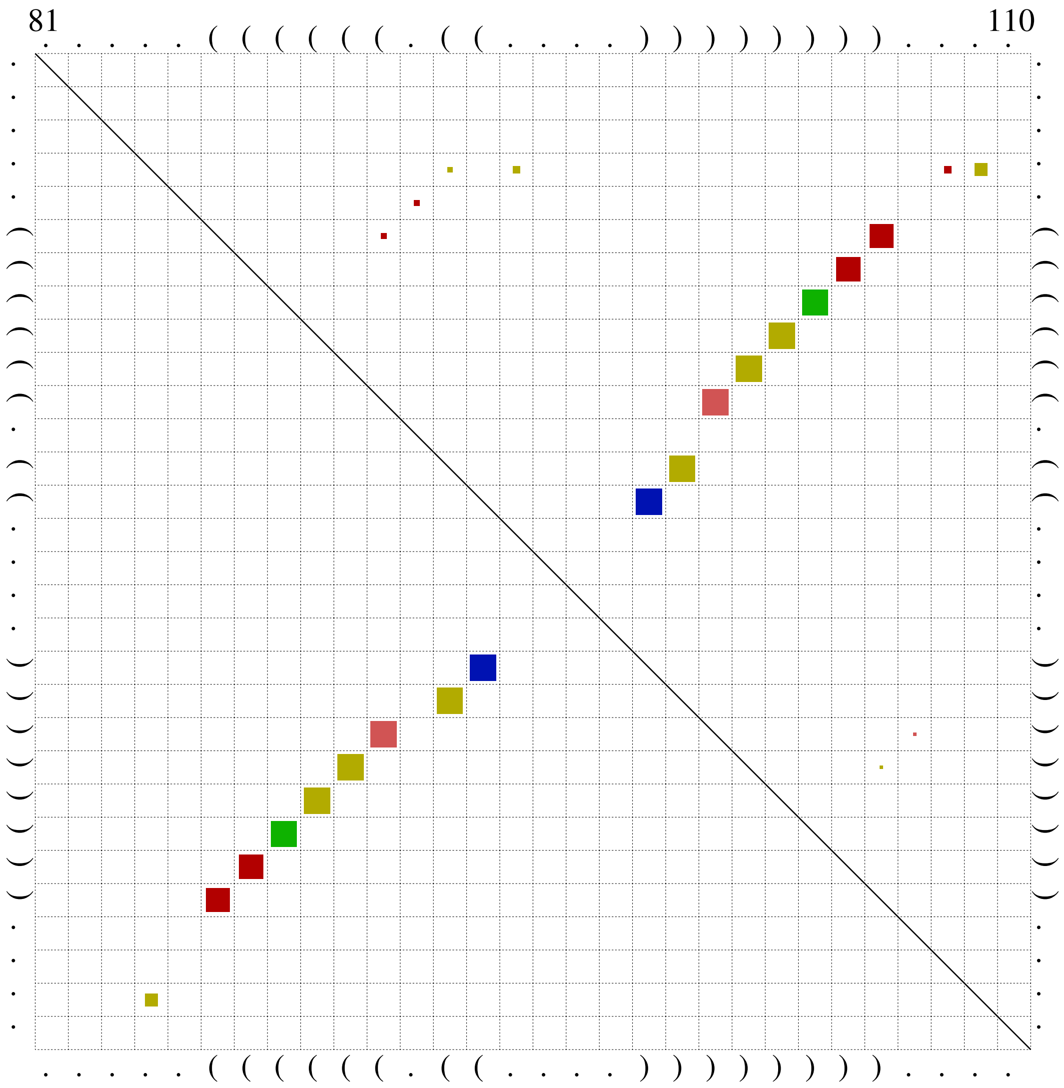

Figure 3: Detail from the combined dot plot computed for a sample of 21 5S rRNA sequences, 12 from halobacteriales and 9 from methanobacteriales.
A square in row i and column j of the dot plot indicates a predicted pair $(i.j)$. Its size and color indicates the frequency and "credibility" of the base 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, 4, 5, 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: combined dot plot of the entire sample.
Lower left triangle: The secondary structure that contains the most credible base pairs that are consistent with a single secondary structure.

sorted as described in the previous section and then we prune all base pairs that conflict with a higher ranked pairs in the list. The result, a valid secondary structure, is given in the lower left part of the color dot plot.

Dot plots become difficult to read for longer sequences. Hence we have adapted the so called mountain representation [19] to incorporate the information of sequence variation, pairing probabilities, and non-compatible sequences. The mountain representation allows for a straightforward comparison of secondary structures and inspired a convenient algorithm for structure based alignments of secondary structures [19, 25].

In the original mountain representation a single secondary structure is represented in a two dimensional graph in which the x-coordinates are the positions in the sequence, whereas the y-coordinates are proportional to the number of base-pairs by which every nucleotide is enclosed. A generalization to arbitrary pairing probabilities was introduced in [22] as

$$m_k = \sum_{i<k} \sum_{j>k} p_{ij}.$$

By definition, m_k counts all base pairs which contain k, weighted with their respective pairing probabilities. In our color mountain plots we represent a single structure such that each base pair (i, j) appears as a slab from i to j, Figure 4 (top), with a thickness proportional to p_{ij}. We use the same color code as for the color dot plots, Figure 3, to indicate the sequence information. Color mountain plots, which are produced directly from the color dot plots using a `perl` script, can be used to display the predicted structure for sequences of the size of entire virus genomes.

Finally, an annotated drawing of the filtered secondary structure that is the final output of the `pfrali` algorithm is generated, Figure 4 (bottom). Positions that support the predicted secondary structure through consistent mutations are marked by circles. Base pairs with non-compatible sequences are shown in gray.

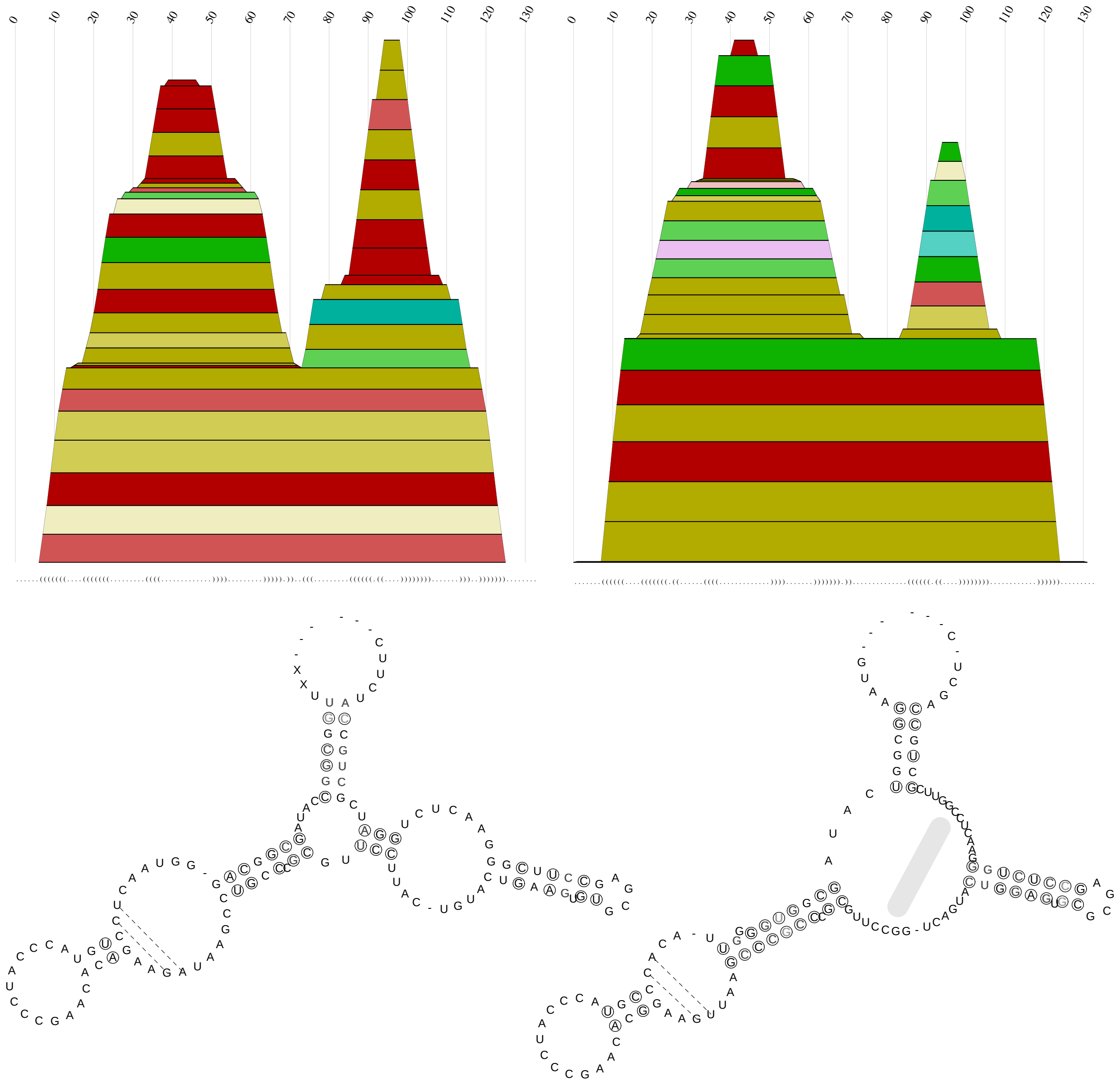

Figure 4: Predicted Structures for two sets of 5S RNA. Color mountain plots of the most credible pairs (above) and filtered secondary structures (below) are shown.
L.h.s.: prediction for a set of nine 5S RNA sequences from Methanomicrobiales. The structure is identical to the phylogenetic model except for a short helix of two pairs which is missing (indicated by dashed lines). All predicted helices are supported by at least one consistent mutation as indicated by the circles; circles on only one side of the pair indicate consistent mutations involving **GU** pairs, while two circles refer to compensatory mutations.
R.h.s.: prediction for ten randomly chosen 5S RNAs from the kingdom Archaea. There are even more compensatory mutations, but also more non-compatible sequences (gray letters) as a result of the lower quality of the alignment. As a consequence one helix in the right arm is missing entirely (indicated by the gray bar).

4. Applications

Secondary structures of single stranded RNA viruses are known to play an important role in the regulation of the viral life cycle. So far a number of functional secondary structure elements have been determined in a variety of different classes of viruses, such as lentiviruses [52, 1, 16], RNA phages [2, 35], flaviviruses [44], pestiviruses [3, 6], picorna viruses [9, 18, 23, 27, 36, 40], hepatitis C viruses [47, 3], or hepatitis D virus [49].

The method described here was designed to scan the moderate size data sets of long RNA sequences that are available for many groups of viruses. These sequences are far too long to allow reliable structure predictions based on a single sequence, yet there are too few sequences available for a purely phylogenetic approach. RNA viruses exhibit a high mutation rate on the order to 10^{-5} to 10^{-3} mutations per nucleotide and replication [8, 7]. As a consequence, different strains of the same virus species show sufficient sequence variation to justify our approach.

In the following we will present examples from three unrelated groups of viruses which contain a variety of human pathogens of global medical importance: The 3' untranslated regions of Tick Borne Encephalitis virus, the RRE region of HIV-1, and the panhandle structure of Hanta virus.

Tick Borne Encephalitis (TBE) virus belongs to the genus flavivirus (which also includes the viruses causing Japanese Encephalitis, Dengue, and Yellow Fever). Flaviviruses are small enveloped particles with an unsegmented, plus-stranded RNA genome of approximate 10kb that encodes a single polyprotein. The 5' and 3' untranslated regions (UTR) are known to contain highly conserved secondary structures, see [32] for a recent summary.

We have reanalyzed the 3' UTR of 9 TBE virus sequences used in [39, 29]. Because sequences for the full length genome are not available for all strains, we have folded only the 3' UTRs (some 340nt). For some strains the full genomes were folded and it was found that the 3' UTR does not interact strongly with the rest of the sequence. The structure shown in figure 5 is virtually identical to the ones produced manually in [39] and to the one independently obtained by Holmes and coworkers [38] using a different folding algorithm. The only discrepancies are local shifts involving a few bases.

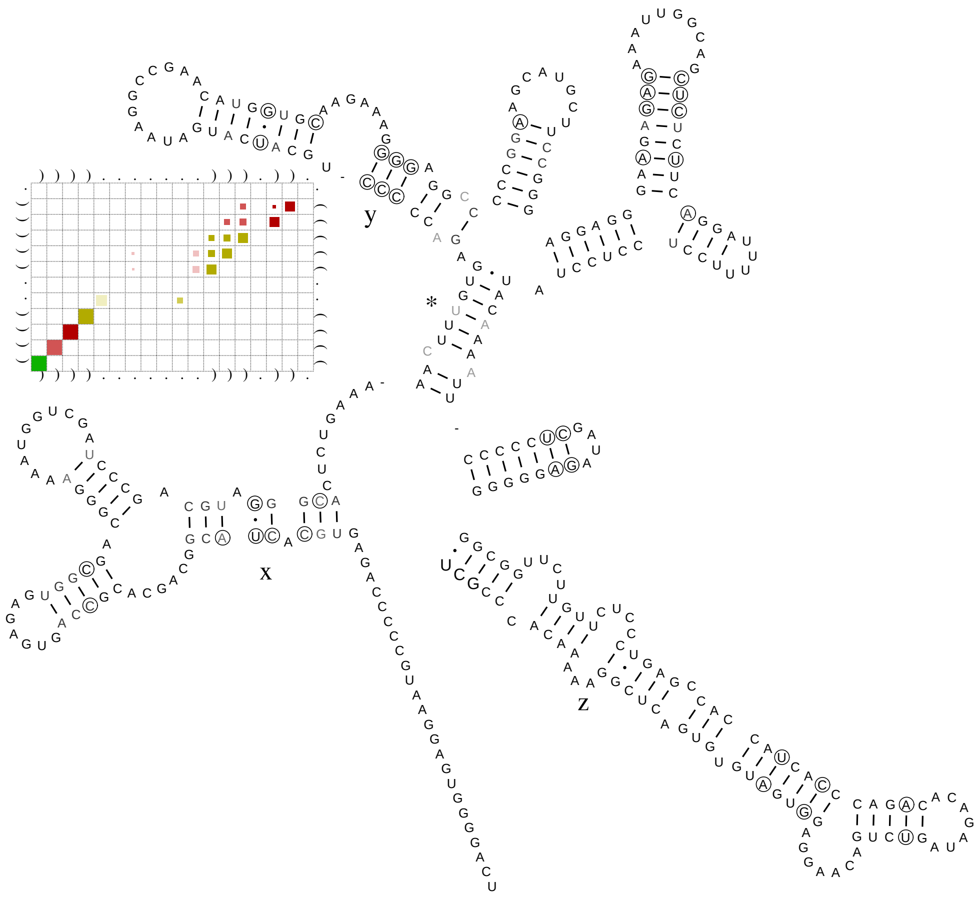

Figure 5: Consensus structure of the 3'UTR of flaviviruses.
The structure is virtually identical with previous analyses [39, 38], except for minor local deviations in the stems denoted by x, y and z. In these cases the structural alternatives are clearly visible in the dot plots. The inset shows the situation for the region marked y, where the two strands of the helix (ochre boxes in right half of inset) can shift by one or two bases.
The stem $*$ is consistently predicted in sequences of the European subtype, while it seems to be absent from Far Eastern subtype and Powassan virus [39].

Manual reconstruction of a consensus structure proved to be a time-consuming and error-prone task. In contrast, the structure in figure 5 was produced without human intervention except for the layout of the structure drawing using the program `XRNA` [51].

The 9200 nucleotides of the HIV-1 genome are dense with information for the coding of proteins and RNA secondary structures that are known to have regulatory functions. 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. The

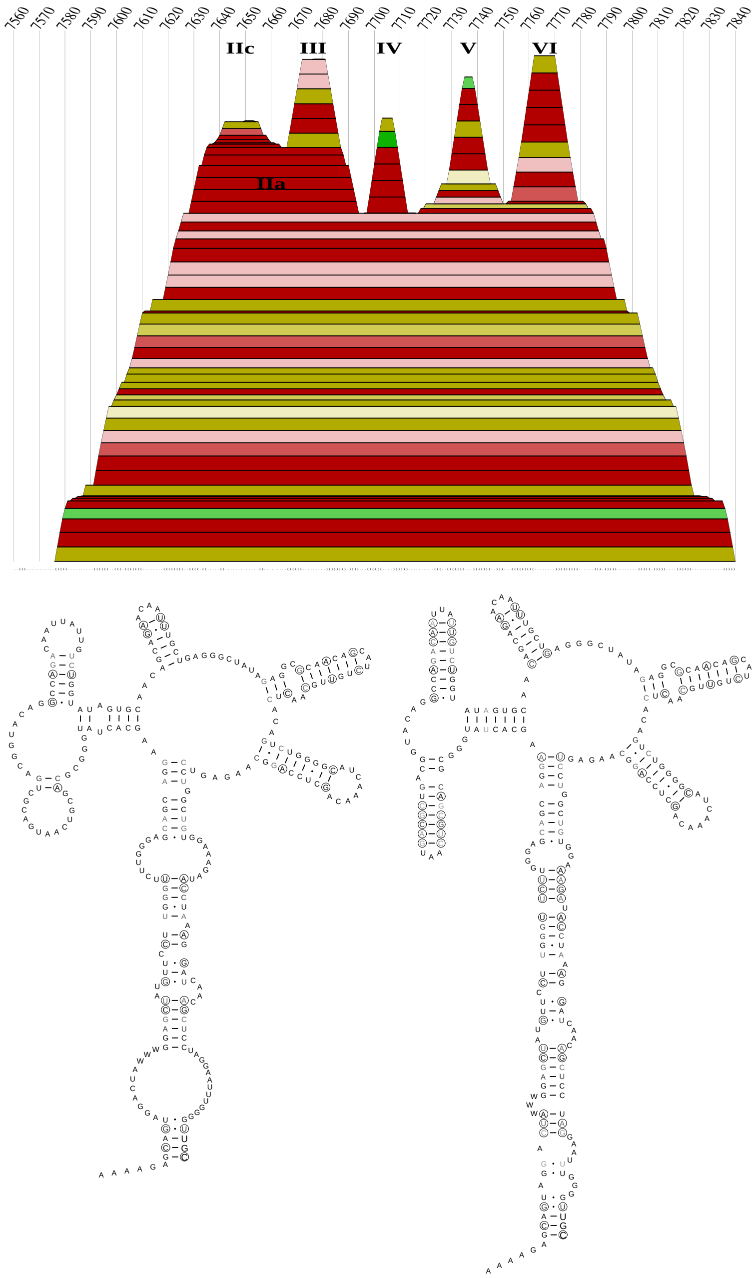

Figure 6: Prediction of the conserved secondary structure of the RRE region of HIV-1.
Top: Color Mountain Plot of the RRE region of HIV-1. This figure is a detail from an analysis of the complete HIV-1 genome.
Below: The structure on the l.h.s. is the output generated by `pfrali`. It only contains the highly conserved base pairs. The more realistic structure on the r.h.s. is obtained by using the `pfrali`-structure as a constraint for computing the minimum energy structure of the consensus sequence. This procedure inserts additional base pairs that do not conflict with the `pfrali` structure. The Roman numerals correspond to the numbering of the hairpins in [5].

best known secondary structure motifs are the TAR and the RRE regions which serve as binding sites to the Tat and Rev proteins. For details and references see [20].

The TAR hairpin structure, being located within the first 60 nucleotides at the 5' end of the HIV genome, is easily predicted correctly (data not shown). The RRE structure, on the other hand, may serve as a bench-mark for structure prediction programs. It is located within the env gene, more than 7000nt from the 5' end and more 1500nt from the 3' end of the genome, involving more than 250 nucleotides. It features a "five-fingered" motif with a large multi-loop. Parts of the structure are most likely flexible, in particular around the IIa region, figure 6, and seem to allow substantial variations [16, 22].

Secondary structure predictions of large RNA molecules with a few thousand nucleotides are usually performed by folding fairly small subsequences. Such an approach has two disadvantages, however, (i) by definition it cannot be used for long-range interactions that span more than the size of the sequence window, and (ii) the results depend crucially on the exact location of the boundaries of the subsequence, since subsequence fold independently of the rest of the structure only if they form a component by themselves, i.e., if there are no base pairs to the out-side of the sequence window. The only way, however, of identifying the component boundaries is to fold the sequence in its entirety.

Our analysis uses the 13 complete HIV-1 sequences listed in the appendix, which have a mean pairwise sequence identity of 83%. Each of the foldings took about 26h on a `Dec Alpha server 5/375` and required approximately 1GByte of memory. These requirements have until recently been prohibitive for folding the entire genome.

The combined dot plot contain about 59000 entries. Of these we retain about 2300 that are displayed in the color mountain plots. The final filtering step again reduces this number by a factor 2, leaving 1116 base pairs in the final predicted structure. Many of the remaining structural elements are fairly small helices involving only short range interactions. We suspect that most of them will not be significant, but it seems inappropriate to leave this decision to the computer program.

A small number of predicted features detected by `pfrali`, however, involve more than a dozen base pairs. Among them, the RRE and TAR regions are the most prominent ones. The RRE of HIV-1 is a rare case where relatively long range base pairs, spanning more than 300nt, are predicted consistently. The mean pairwise sequence identity of the RRE region is 86% in our sample.

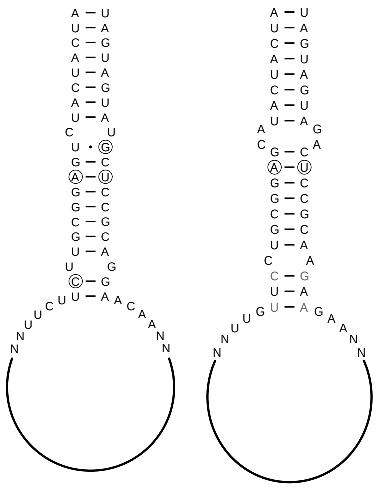

Figure 7: Panhandle structure of the medium segment of Hantavirus. No other significant secondary structures are predicted for this example.
Apparently, both the plus (r.h.s.) and the minus strand (l.h.s.) of Hantaviruses form a panhandle structures. This requirement strongly limits the sequence variability since a **GU** pair in one strand becomes a **C A** mismatch in the other strand. Consequently the sequence is highly conserved here. The structure is confirmed by one compensatory mutation.

The mountain representation reveals a five-fingered motif, Figure 6. A previous computational study [16, 17] has revealed that three distinct folding patterns are possible with comparable energies for the top of the RRE element. Our prediction is almost identical with the five-fingered motif presented in [5], which was based predominantly on phylogenetic reconstructions. An alternative fold, in which IIa interferes with the stem V was proposed in [30].

Hantaviruses are serologically-related members of the family Bunyaviridae [10]. They are enveloped viruses with a tripartite negative-sense RNA genome. Hantaviruses have been implicated as etiologic agents for two acute diseases: hemorrhagic fever with renal syndrome (HFRS) and hantavirus pulmonary syndrome (HPS). The three genome segments are called L, M and S, encoding the viral transcriptase, envelope glycoproteins, and nucleocapsid protein, respectively.

In this contribution we consider only with the medium (M) segment, which has a sequence length of about 3700nt. Hanta viruses are more heterogeneous than HIV-1, the mean pairwise sequence identity of our sample is only 67%. We have folded both the (+) and the (-) strand. The combined dot plots contain some 40000 entries, only 581 (+) and 549

(-) pairs are in the color mountains (not shown). Of these, only 54 pairs belong to the final predicted structure.

The only prominent feature consists of 18 base pairs joining the 5' and 3' ends of the sequence, figure 7. This so-called panhandle structure was postulated already in the eighties for all bunyaviridae [33, 42]. In the both strands there are no other significant signals: the remaining base pairs form a variety of very short helices. It is of course possible that subgroups of Hanta viruses contain additional conserved structures that are not common to the entire group.

5. Discussion

We have presented a method for elucidating conserved secondary structures in moderate size samples of related RNA sequences. Since only functional secondary structures are likely to be conserved, it can be used to guide mutagenesis or deletion studies, in particular in the context of vaccine development [29, 37].

The algorithm is designed to extract promising structural features without user intervention, and is therefore suitable for scanning long sequences such as viral genomes. It combines thermodynamic structure prediction in the form of Boltzmann-weighted base pairing probabilities with the analysis of sequence covariations in a multiple sequence alignment. Since our approach emphasizes sequence variation, it can complement other methods for finding functional RNA secondary structures that are based solely on thermodynamic prediction [26, 22].

As example applications we have analyzed the 3'UTR of TBE virus, as well as the complete genomes of HIV-1 and the complete M segment of Hanta virus. The amount of detected structural features varies strongly between different virus families. An advantage of our approach is, that it generates predictions only for the parts of the sequence, where there is reasonable support for a conserved structure. This distinguishes our approach from efforts to solve the alignment and folding problem simultaneously, as well as Riesner's `ConStruct`. Thus the prediction for Hanta virus consists essentially only of the well known panhandle linking the 5' and 3' end of each segment. On the other hand the genome of HIV-1 contains significant structural motifs in all parts of the sequence. While the strongest signals correspond to known structural elements such as TAR and RRE, several more regions warrant further experimental investigations.

The algorithm is implemented as an `ANSI C` program `pfrali` that reads a `Clustal W` multiple aligment file and base pairing probabilities as produced by the `Vienna RNA Package` and writes combined dot plots and the sorted list of base pairs. Alternative representations such as color mountain plots, annotated secondary structure drawings, or zooming into sections of dot plots is handled by a collection of `perl` scripts. The software will be available from the authors.

Acknowledgments

Stimulating discussions with Martijn A. Huynen and Andreas Wagner, and partial financial support by the Austrian *Fonds zur Förderung der Wissenschaftlichen Forschung*, Proj. No. P 12591-INF, is gratefully acknowledged.

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).

Flavivirus sequences:

Neudoerfl (U27495), 263 (U27491), Ljubljana (U27494), Hypr (U39292), Crimea (U27493), 132 (U27490), RK1424 (U27496), Aina (U27492), Powassan (L06436).

Hantavirus sequences:

PVMZ84205 (Z84205), PUUMGP (X61034), PUVMVIN83 (Z49214), PVU22418 (U22418), TUVM5302 (Z69993), AF028023 AF028024 AF028022 AF005728 AF030551 AF030552 HPSMSEG (L25783), HANM (M14627), HANG1G2A (L08753).

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

[3] 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.

[4] 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.

[5] E. T. Dayton, D. A. M. 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.

[6] 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.

[7] E. Domingo, J. J. Holland, C. K. Biebricher, and M. Eigen. Quasispecies: The concept and the word. In A. Gibbs, C. H. Calisher, and G. García-Arenal, editors, *Molecular Basis of Virus Evolution*, pages 171–180. Cambridge University Press, Cambridge UK, 1995.

[8] J. W. Drake. Rates of spontaneous mutations among RNA viruses. *Proc. Natl. Acad. Sci., USA*, 90:4171–4175, 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] 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.

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

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

[13] P. Higgs. The influence of RNA secondary structure on the rates of substitution in RNA-encoding genes. Preprint, Univ. Manchester, 1998.

[14] 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.*, 1998. submitted, Santa Fe Institute Preprint 98-03-020.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[21] 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.

[22] 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.

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

[24] D. Konings and R. Gutell. A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNAs. *RNA*, 1:559–574, 1995.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] 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.

[33] 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.

[34] R. Nussinov, G. Piecznik, J. R. Griggs, and D. J. Kleitman. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35(1):68–82, 1978.

[35] 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.

[36] 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.

[37] V. Proutski, M. W. Gaunt, E. A. Gould, and E. C. Holmes. Secondary structure of the 3' untranslated region of yellow fever virus: Implications for virulence, attenuation and vaccine developement. *J. Gen. Virol.*, 78:1543–1549, 1997.

[38] V. Proutski, E. A. Gould, and E. C. Holmes. Secondary structure of the 3' untranslated region of flaviviruses: Similarities and differences. *Nucl. Acids Res.*, 25:1195–1202, 1997.

[39] 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.

[40] 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.

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

[42] 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.

[43] 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.

[44] 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.

[45] T. Specht, J. Wolters, and V. A. Erdmann. Compilation of 5S rRNA and 5S rRNA gene sequences. *Nucl. Acids Res. (Supplem.)*, 19:2189–2191, 1991. `http://userpage.chemie.fu-berlin.de/fb_chemie/ibc/agerdmann/5S_rRNA.html`.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

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

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

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

[53] S. Wuchty, I. L. H. Walter Fontana, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 1998. submitted, Santa Fe Institute Preprint 98-05-040.

[54] M. Zuker. The use of dynamic programming algorithms in RNA secondary structure prediction. In M. S. Waterman, editor, *Mathematical Methods for DNA Sequences*, pages 159–184. CRC Press, 1989.

[55] M. Zuker and P. Stiegler. Optimal computer folding of larger RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.*, 9:133–148, 1981.