PERGAMON

Computers & Chemistry 23 (1999) 401–414

Computers & Chemistry

Automatic detection of conserved base pairing patterns in RNA virus genomes

Ivo L. Hofacker[a], Peter F. Stadler[a,b,*]

[a]*Institut für Theoretische Chemie, Universität Wien, Austria*
[b]*The Santa Fe Institute, Santa Fe, New Mexico, USA*

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 cannot 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. The method is based on combining the prediction of base pair probability matrices and comparative sequence analysis. It can be applied to small sets of long sequences and does not require a prior knowledge of conserved sequence or structure motifs. As such it can be used to scan large amounts of sequence data for regions that warrant further experimental investigation. Applications to complete genomic RNAs of some viruses show that in all cases the known secondary structure features are identified. In addition, we predict a substantial number of conserved structural elements which have not been described so far.

Keywords: RNA secondary structure prediction; Structure alignment; Conserved substructures; Compensatory mutations; RNA virus genomes

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, are 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. (Gutell, 1993).

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 (Fontana et al., 1993; Schuster et al., 1994) 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%

* Corresponding author. Tel.: +43-1-40480-665; fax: +43-1-40480-660.

E-mail address: studla@tbi.univie.ac.at (P.F. Stadler)

0097-8485/99/$ - see front matter
PII: S0097-8485(99)00013-3

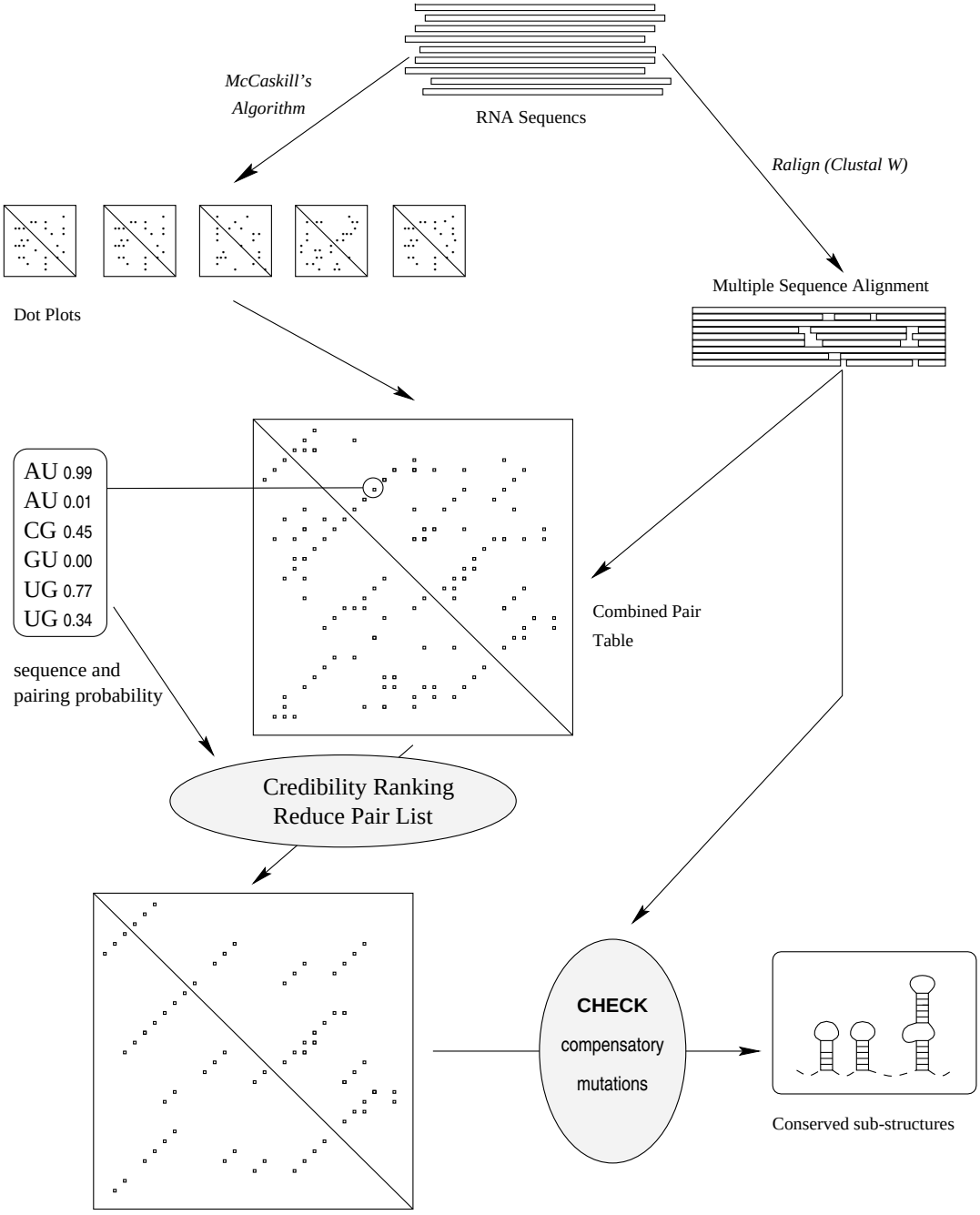

Fig. 1. Flow diagram of the algorithm `pfrali`. A multiple sequence alignment is calculated using `CLUSTAL W` (Thompson et al., 1994). RNA genomes are folded using McCaskill's partition function algorithm as implemented in the `Vienna RNA Package`. The sequence alignment is then used to align the predicted 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.

average pairwise identity are therefore most likely the result of stabilizing selection, not a consequence of the high degree of sequence homology. If selection acts to preserve a structural element then it must, of course, have some function. This observation can be used to design an algorithm that reliably detects conserved

RNA secondary elements in a small sample of related RNA sequences (Hofacker et al., 1998). Of course, we cannot tell what the function of the conserved structure elements might be. Nevertheless, knowledge about their location can be used to guide, for instance, deletion studies (Mandl et al., 1998).

Our method combines structure prediction and *motif search* (Dandekar and Hentze, 1995). Programs such as `RNAmot` (Gautheret et al., 1990) or S. Eddy's `RNAbob` scan a database of sequences for RNA motifs that are specified in terms of sequence as well as secondary structure constraints. In contrast, our approach does not require any prior knowledge about the structural motifs: their structures are predicted during the search.

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 (Nussinov et al., 1978; Waterman, 1978; Zuker and Stiegler, 1981). Approaches to kinetic folding (Mironov et al., 1985; Gultyaev et al., 1995) are also based on the thermodynamic rules. 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 (Konings and Gutell, 1995; Huynen et al., 1997). 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 (Zuker, 1989), or even *all* structures within a prescribed energy range (Wuchty et al., 1999). Non-deterministic kinetic folding algorithms (Gultyaev et al., 1995) can produce ensembles of structures by repeatedly running them with different random numbers. A much more elegant and efficient solution is the computation of the complete matrix of base paring probabilities (McCaskill, 1990) which contains suitably weighted information about all possible secondary structures and therefore reduces the impact of inaccuracies in the structure predicition. 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 of sequence variation (Winker et al., 1990), but usually requires a very large number of sequences. Such data sets are thus rarely available outside a very small number of very well-studied classes of molecules such as tRNA, rRNAs (Gutell, 1993), RNAseP, or self-splicing introns. In the latter case the dataset was sufficient for the construction of a three-dimensional model (Michel and Westhof, 1990).

Our method aims at utilizing 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, (see Fig. 1). Related procedures are described in (Lück et al., 1996; Hofacker et al., 1998). Our approach is different from efforts to simultaneously compute alignment and secondary structures (Sankoff, 1985; Tabaska and Stormo, 1997; Corodkin et al., 1997). One disadvantage of these methods is 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.

This contribution is organized as follows: After describing the algorithm (Section 2) we use 5S RNA as a reference case, where the correct structures are well known, to demonstrate the reliability of the structure prediction (Section 3). Applications to large genomic RNAs from different classes of viruses (Flavivirus, Lentivirus, and Hantavirus) are presented in Section 4.

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` (Thompson et al., 1994). 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 (Hofacker et al., 1998), i.e., one structure per sequence. Here we extend the method to use base pairing probabilities as obtained from McCaskill's partition function algorithm (McCaskill, 1990). 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 (Lück et al., 1996). 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

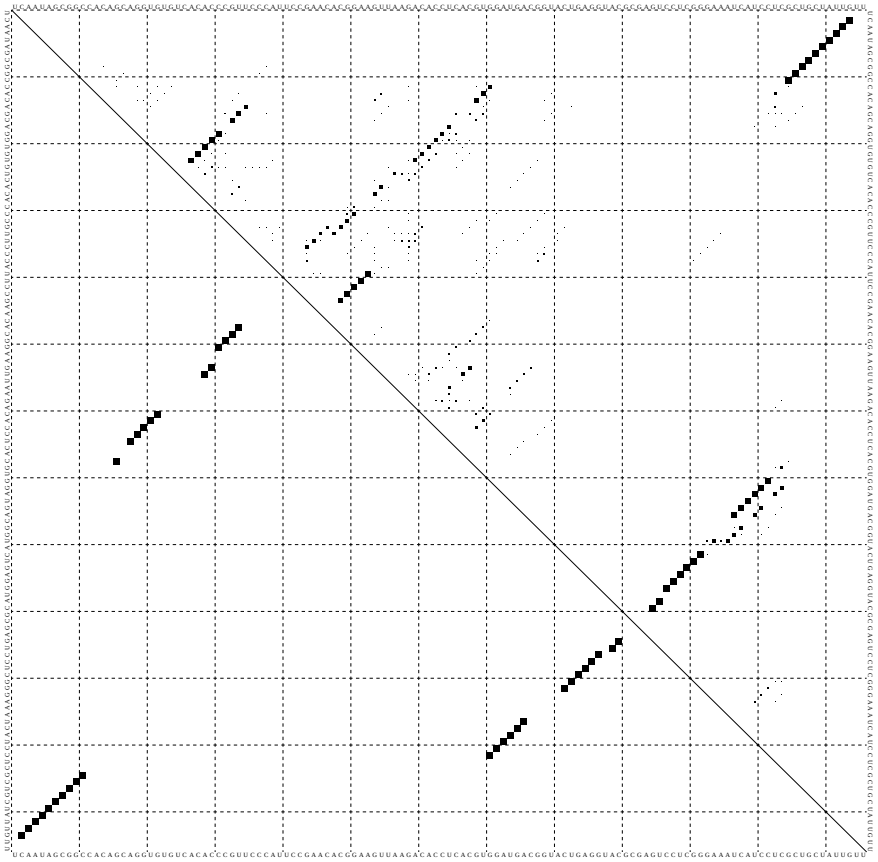

Fig. 2. Dot plot of the 5S RNA of *methanospirillum hungatii*. The upper right part shows the predicted base pair probabilities computed with the partition function algorithm of the `Vienna RNA Package`. The area of the squares is proportional to the pairing probability. The lower left part gives the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

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] (Hofacker et al., 1994) contains an efficient implementation of McCaskill's algorithm that produces dot plots in `PostScript` format, see Fig. 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. Base pairs with even lower probabilities are very unlikely to be part of an important structure. 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.

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 or 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 occurence 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

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

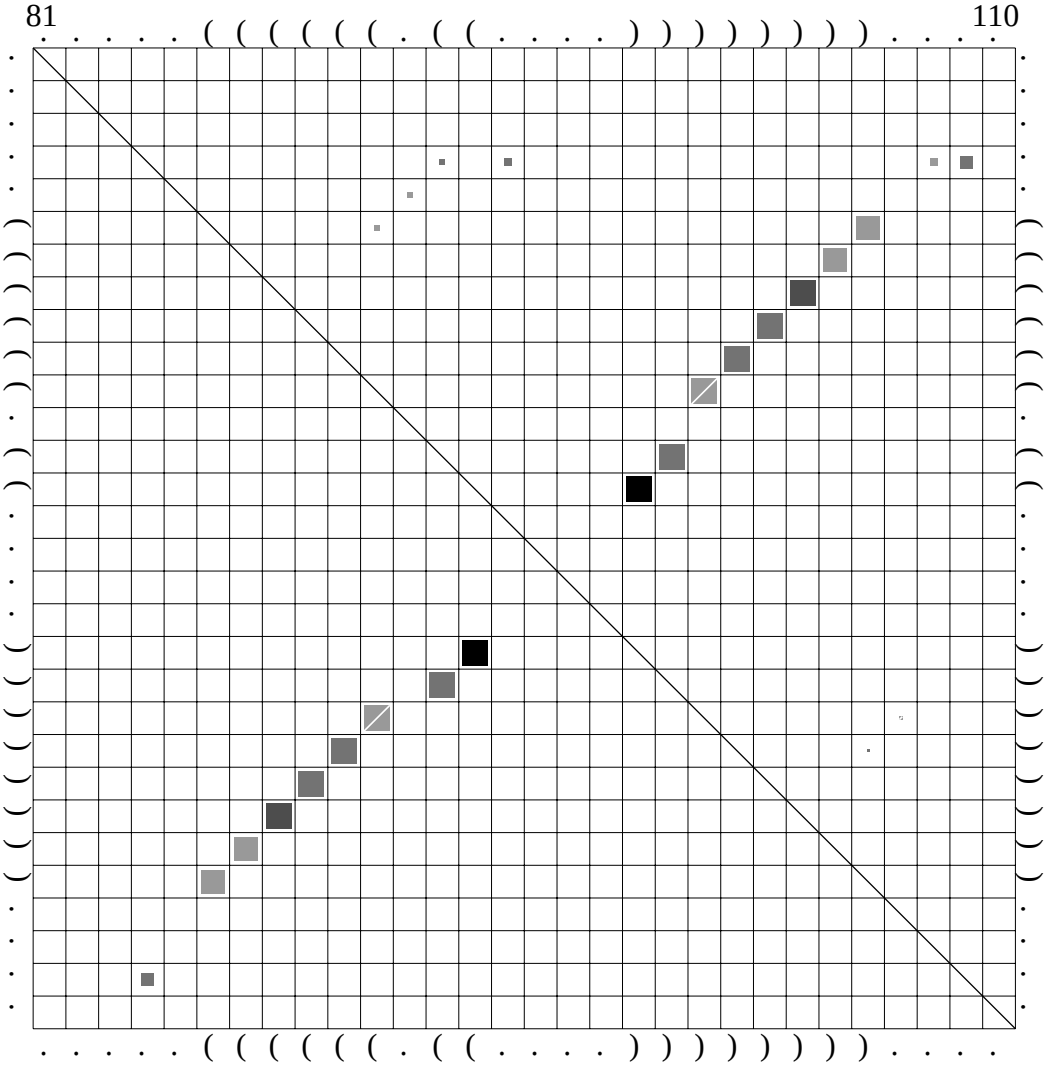

Fig. 3. Detail from the combined dot plot computed for a sample of 21 5S rRNA sequences, 12 from halobacteriales and 9 from methanobacteriales. The original colors of the `pfrali`-output have been converted into grayscales and fill patterns in order to avoid color plates.A square in row i and column 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. Grayscales, from light to dark, indicate an increasing number of consistent mutations. The base pair marked by a white stripe has one non-compatible sequence. If there are more than two 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.

compensatory mutations, even though **GU** base pairs are clearly important as evidenced by the fact that **RY** $\rightarrow$ **YR** conversions are rare (Higgs, 1998). While compensatory mutations of the type **RY** $\rightarrow$ **RY**, such as **GC** $\rightarrow$ **AU**, can be obtained by two subsequent consistent point mutations, for instance **GC** $\rightarrow$ **GU** $\rightarrow$ **AU**, a double mutation is required for **RY** $\rightarrow$ **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. In this paper the color information is translated to grayscales and fill patterns (see Fig. 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:

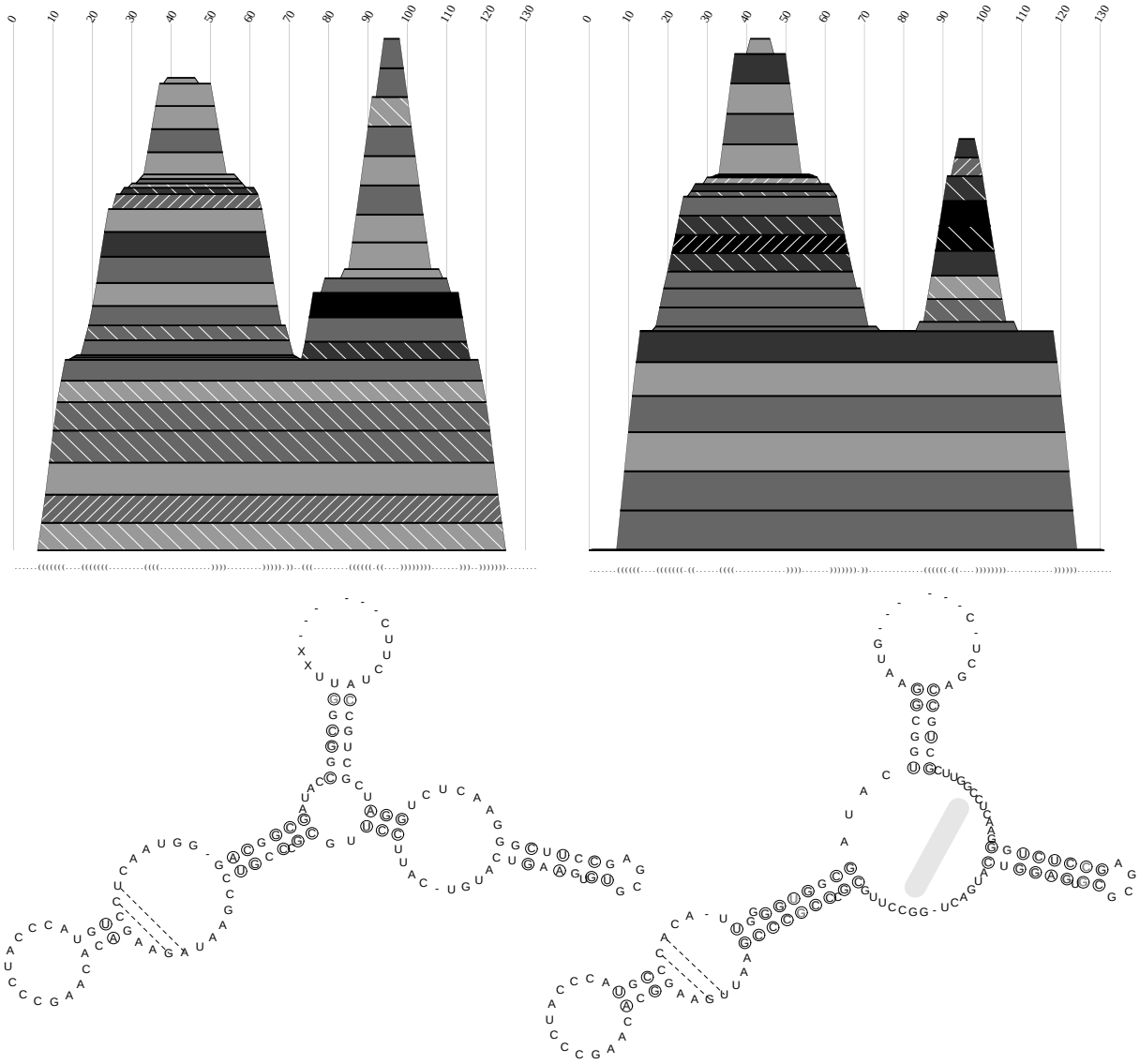

Fig. 4. Predicted structures for two sets of 5S RNA. Mountain plots of the most credible pairs (above) and filtered secondary structures (below) are shown. The height of each slab is proportional to the frequency of the corresponding base pair. Grayscales, from light to dark, indicate an increasing number of consistent mutations. The base pairs marked by sparse and dense white stripes have one or two non-compatible sequences, respectively. Again, we do not display base pairs with more than two non-compatible sequences. LHS: 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. RHS: prediction for 10 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).

(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 $(k \cdot l)$ such that $i<k<j<l$.

In the remainder 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 (Hofacker et al., 1998).

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. Our goal is to produce a list of well-supported secondary structure features that contains as few false positive as possible. We therefore use a series of additional 'filtering' steps. First, we remove all pairs with more than two non-compatible sequences, as well as pairs with two non-compatible sequences adjacent to a pair that also has noncompatible sequences. Helices with so many non-compatible sequences can hardly be called 'conserved' (For large samples these rules might have to be modified to tolerate somewhat larger numbers of 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:

1. The highest ranking base pair must not have non-compatible sequences.
2. For the highest ranking base pair the product $\bar{p}_{i.j} \times c_{i.j}$ must be greater than 0.3.
3. If the helix has length 2, it must not have more non-compatible sequences than consistent mutations.

In general, these filtering steps only remove insignificant structural motifs that one would have disregarded upon visual inspection anyways. The remaining list of base pairs is the conserved structure predicted by the `pfrali` program.

The final output of the program consists of a color coded dot plot in `PostScript` format (reproduced here in a black and white version), 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. 5S RNAs as a test case

Before applying our procedure to large viral RNA sequences, let us first demonstrate the feasibility of the approach on a small test case, where the correct structures are well known. The `Berlin RNA Databank` (Specht et al., 1991) 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 `pfrali` 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 Fig. 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 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 (Hogeweg and Hesper, 1984) 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

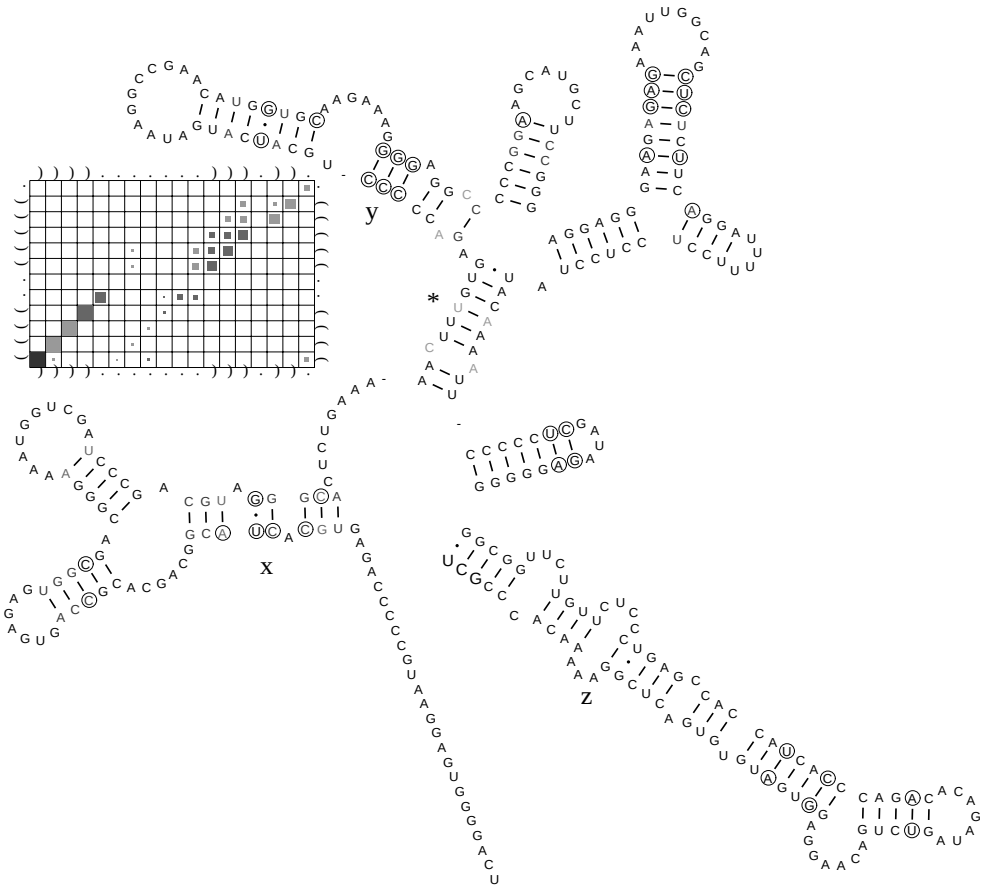

Fig. 5. Consensus structure of the 3′UTR of flaviviruses. The structure is virtually identical with previous analyses (Rauscher et al., 1997; Proutski et al., 1997b), except for minor local deviations in the stems denoted by *x, y* and *z*. In these cases the structural alternatives are clearly visible in the dot plots. The inset shows the situation for the region marked *y*, where the two strands of the helix (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 (Rauscher et al., 1997).

of secondary structures (Hogeweg and Hesper, 1984; Konings and Hogeweg, 1989).

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 (Huynen et al., 1996) 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, Fig. 4 (top), with a thickness proportional to p_{ij}. We use the same color code as for the color dot plots (see Fig. 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, Fig. 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.

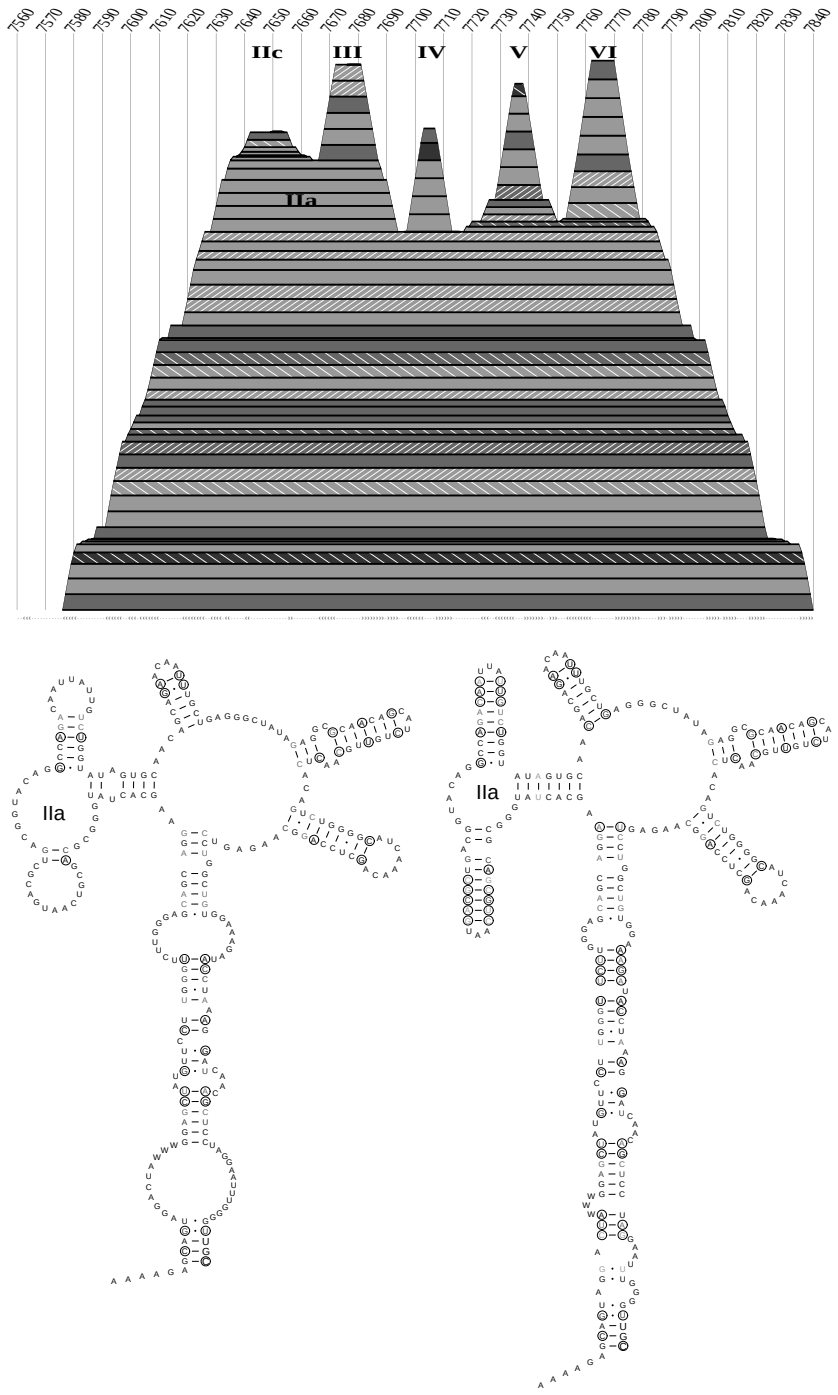

Fig. 6. Prediction of the conserved secondary structure of the RRE region of HIV-1. Top: mountain plot of the RRE region of HIV-1. This figure is a detail from an analysis of the complete HIV-1 genome. Below: the structure on the LHS is the output generated by `pfrali`. It only contains the highly conserved base pairs. The more realistic structure on the RHS 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 (Dayton et al., 1992).

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 (Wills and Hughes, 1990; Baudin et al., 1993; Hofacker et al., 1996a), RNA phages (Biebricher, 1994; Olsthoorn et al., 1995), flaviviruses (Shi et al., 1996), pestiviruses (Brown et al., 1992; Deng and Brock, 1993), picorna viruses (Duke et al., 1992; Hoffman and Palmenberg, 1995; Jackson and Kaminski, 1995; Le et al., 1993; Pilipenko et al., 1989; Rivera et al., 1988), hepatitis C viruses (Tanaka et al., 1996; Brown et al., 1992), or hepatitis D virus (Wang et al., 1986).

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 ranging from 10^{-5} to 10^{-3} mutations per nucleotide and replication (Drake, 1993; Domingo et al., 1995). 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 (TBE) virus, the RRE region of HIV-1, and the panhandle structure of Hanta virus.

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 10 kb that encodes a single polyprotein. The 5′ and 3′ untranslated regions (UTR) are known to contain highly conserved secondary structures, see (Monath and Heinz, 1996) for a recent summary.

We have reanalyzed the 3′ UTR of 9 TBE virus sequences used in (Rauscher et al., 1997; Mandl et al., 1998). Because sequences for the full length genome are not available for all strains, we have folded only the 3′ UTRs (some 340 nt). 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 Fig. 5 is virtually identical to the ones produced manually in (Rauscher et al., 1997) and to the one independently obtained by Holmes and coworkers (Proutski et al., 1997b), using a different folding algorithm. The only discrepancies are local shifts involving a few bases.

Manual reconstruction of a consensus structure proved to be a time-consuming and error-prone task. In contrast, the structure in Fig. 5 was produced without human intervention except for the layout of the structure drawing using the program `XRNA` (Weiser and Noller, 1995).

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 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 Huynen and Konings, 1998).

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 benchmark for structure prediction programs. It is located within the env gene, more than 7000 nt from the 5′ end and more 1500 nt 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, Fig. 6, and seem to allow substantial variations (Hofacker et al., 1996a; Huynen et al., 1996).

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 outside of the sequence window. Even though this problem can be reduced by employing larger windows, the only reliable way of identifying the component boundaries is to fold the sequence in its entirety.

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

The combined dot plot contain about 59,000 entries. Of these we retain about 2300 that are displayed in the

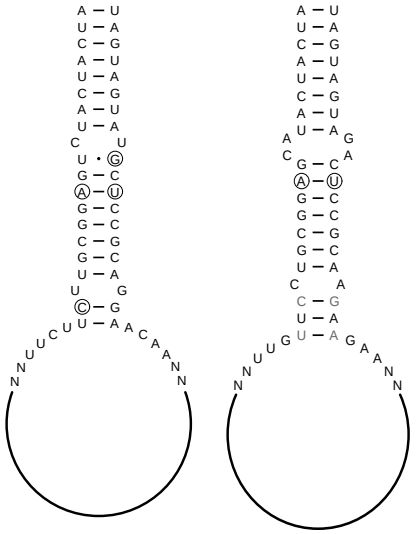

Fig. 7. Panhandle structure of the medium segment of Hantavirus. No other significant secondary structures are predicted for this example.Apparently, both the plus (RHS) and the minus strand (LHS) of Hantaviruses form panhandle structures. This requirement strongly limits the sequence variability since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

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 300 nt, are predicted consistently. The mean pairwise sequence identity of the RRE region is 86% in our sample.

The mountain representation reveals a five-fingered motif (Fig. 6). A previous computational study (Hofacker et al., 1996a, b) 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 (Dayton et al., 1992), which was based predominantly on phylogenetic reconstructions. An alternative fold, in which IIa interferes with the stem V was proposed in (Mann et al., 1994).

Hantaviruses are serologically-related members of the family Bunyaviridae (Elliott et al., 1991). 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 3700 nt. 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 40,000 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, Fig. 7. This so-called panhandle structure was postulated already in the eighties for all bunyaviridae (Paradigon et al., 1982; Schmaljohn et al., 1986). In 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. 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. The algorithm is designed to extract promising structural features without user intervention, and is therefore suitable for scanning long sequences such as viral genomes. 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 (Mandl et al., 1998; Proutski et al., 1997a).

The algorithm is implemented as an `ANSI C` program `pfrali` that reads a `CLUSTAL W` multiple alignment 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 is available from the authors.

We have analyzed the 3′UTR of TBE virus, 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. The strongest signals in HIV-1 correspond to known structural elements such as TAR and RRE. However, several more regions warrant further experimental investigations. For Hanta virus no conserved structures are predicted apart from the well-known panhandle linking the 5′ and 3′ end of each segment.

Our approach emphasizes sequence variation and can therefore complement and extend other methods for finding functional RNA secondary structures that are based solely on thermodynamic prediction (Le et al., 1988; Huynen et al., 1996). On the other hand, it does not require prior knowledge of structure motifs or sequence patterns, as in the case of motif search programs (Gautheret et al., 1990). A further 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` (Lück et al., 1996). The use of base pairing probability matrices instead of single structures (as in `alidot` (Hofacker et al., 1998)) significantly reduces the chance to overlook a structural feature.

In its current implementation, `pfrali` will not predict structures containing pseudo-knots. While the modifications of the `pfrali` program are trivial, including pseudo-knots makes sense only in conjunction with a folding algorithm that explicitly predicts pseudo-knots.

The performance of the `pfrali` program depends on the quality of the multiple sequence alignment that is used as an input. Although the results are surprisingly robust with respect to minor alignment problems, we expect that alignment inaccuracies will become a substantial problem when dealing with more diverse data sets. It is quite possible that the use of a local alignment procedure such `DIALIGN` (Morgenstern et al., 1998) could extend applicability of our method.

Acknowledgements

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 A

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

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

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

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

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

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

Dayton, E.T., Konings, D.A.M., Powell, D.M., Shapiro, B.A., Butini, L., Maizel, J.V., Dayton, A.I., 1992. 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.

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

Domingo, E., Holland, J.J., Biebricher, C.K., Eigen, M., 1995. Quasispecies: The concept and the word. In: Gibbs,

A., Calisher, C.H., García-Arenal, G. (Eds.), Molecular Basis of Virus Evolution. Cambridge University Press, Cambridge, UK, pp. 171–180.

Drake, J.W., 1993. Rates of spontaneous mutations among RNA viruses. Proc. Natl. Acad. Sci., USA 90, 4171–4175.

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

Elliott, R.M., Schmaljohn, C.S., Collett, M.S., 1991. Bunyavirus genome structure and gene expression. Current Topics in Microbiology and Immunology 169, 91–141.

Fontana, W., Konings, D.A.M., Stadler, P.F., Schuster, P., 1993. Statistics of RNA secondary structures. Biopolymers 33, 1389–1404.

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

Gultyaev, A.P., vanbatenburg, F.H.D., Pleij, C.W.A., 1995. The computer simulation of RNA folding pathways using a genetic algorithm. J. Mol. Biol 250, 37–51.

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

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

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

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

Hofacker, I.L., Huynen, M.A., Stadler, P.F., Stolorz, P.E., 1996a. 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. AAAI Press, Portland, OR, pp. 20–25.

Hofacker, I.L., Huynen, M.A., Stadler, P.F., Stolorz, P.E., 1996b. RNA folding and parallel computers: the minimum free energy structures of complete HIV genomes. Technical Report # 95–10–089, SFI Santa Fe, New Mexico.

Hoffman, M.A., Palmenberg, A.C., 1995. Mutational analysis of the J-K stem-loop region of the encephalomyocarditis-virus IRES. J. Virol 69, 4399–4406.

Hogeweg, P., Hesper, B., 1984. Energy directed folding of RNA sequences. Nucl. Acids Res 12, 67–74.

Huynen, M., Konings, D., 1998. Questions about RNA structures in HIV and HPV. In: Myers, G.L. (Ed.), Viral Regulatory Structures and Their Degeneracy, Volume XXVIII of Santa Fe Institute Studies in the Sciences of Complexity. Addison Wesley, Longman, Reading, MA, pp. 69–82.

Huynen, M.A., Gutell, R., Konings, D.A.M., 1997. Assessing the reliability of RNA folding using statistical mechanics. J. Mol. Biol 267, 1104–1112.

Huynen, M.A., Perelson, A.S., Vieira, W.A., Stadler, P.F., 1996. Base pairing probabilities in a complete HIV-1 RNA. J. Comp. Biol 3, 253–274.

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

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

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

Le, S.-Y., Chen, J.-H., Currey, K., Maizel, J., 1988. A program for predicting significant RNA secondary structures. Comp. Appl. Biosci 4, 153–159.

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

Lück, R., Steger, G., Riesner, D., 1996. 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.

Mandl, C.W., Holzmann, H., Meixner, T., Rauscher, S., Stadler, P.F., Allison, S.L., Heinz, F.X., 1998. 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.

Mann, D., Mikaelian, I., Zemmel, R., Green, S., Lowe, A., Kimura, T., Singh, M., Butler, P., Gait, M., Karn, J., 1994. 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.

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

Michel, F., Westhof, E., 1990. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J. Mol. Biol 216, 585–610.

Mironov, A.A., Dyakonova, L.P., Kister, A.E., 1985. A kinetic approach to the prediction of RNA secondary structures. Journal of Biomolecular Structure and Dynamics 2, 953.

Monath, T.P., Heinz, F.X., 1996. Flaviviruses. In: Fields, B.N., Knipe, D.M., Howley, P.M., Chanock, R.M., Melnick, J.L., Monath, T.P., Roizmann, B., Straus, S.E. (Eds.), Fields Virology, 3rd ed. Lippincott-Raven, Philadelphia, pp. 961–1034.

Morgenstern, B., Frech, K., Dress, A.W.M., Werner, T., 1998. DIALIGN: Finding local similarities by multiple sequence alignment. Bioinformatics 14, 290–294.

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

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

Olsthoorn, R.C.L., Garde, G., Dayhuff, T., Atkins, J.F., van Duin, J., 1995. Nucleotide sequence of a single-stranded

RNA phage from pseudomonas aeruginosa: Kinship to coliphages and conservation of regulatory RNA structures. Virology 206, 611–625.

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

Proutski, V., Gaunt, M.W., Gould, E.A., Holmes, E.C., 1997a. Secondary structure of the 3′ untranslated region of yellow fever virus: implications for virulence, attenuation and vaccine development. J. Gen. Virol 78, 1543–1549.

Proutski, V., Gould, E.A., Holmes, E.C., 1997b. Secondary structure of the 3′ untranslated region of flaviviruses: similarities and differences. Nucl. Acids Res 25, 1195–1202.

Rauscher, S., Flamm, C., Mandl, C., Heinz, F.X., Stadler, P.F., 1997. Secondary structure of the 3′-non-coding region of flavivirus genomes: comparative analysis of base pairing probabilities. RNA 3, 779–791.

Rivera, V.M., Welsh, J.D., Maizel Jr, J.V., 1988. Comparative sequence analysis of the 5′ noncoding region of the enteroviruses and rhinoviruses. Virology 165, 42–50.

Sankoff, D., 1985. Simultaneous solution of the RNA folding, alignment, and protosequence problems. SIAM J. Appl. Math 45, 810–825.

Schmaljohn, C.S., Jennings, G.B., Hay, J., Dalrymple, J.M., 1986. Coding strategy of the S genome segment of hantaan virus. Virology 155, 633–643.

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

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

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

Tabaska, J.E., Stormo, G.D., 1997. Automated alignment of RNA sequences to pseudoknotted structures. In: Gaasterland, T., Karp, P., Karplus, K., Ouzounis, C., Sander, C., Valencia, A. (Eds.), Proceedings of the ISMB-97. AAAI Press, Menlo Park, CA, pp. 311–318.

Tanaka, T, Kato, N., Cho, M.-J., Sugiyama, K., Shimotohno, K., 1996. Structure of the 3′ terminus of the hepatitis C virus genome. J. Virol 70, 3307–3312.

Thompson, J.D., Higgs, D.G., Gibson, T.J., 1994. CLUSTAL W: 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.

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

Waterman, M.S., 1978. Secondary structure of single-stranded nucleic acids. Adv. Math. Suppl. Studies 1, 167–212.

Weiser, B., Noller, H., 1995. XRNA. ftp://fangio.ucsc.edu/pub/XRNA/. (Public Domain Software).

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

Winker, S., Overbeek, R., Woese, C.R., Olsen, G.J., Pfluger, N., 1990. Structure detection through automated covariance search. Comput. Appl. Biosci 6, 365–371.

Wuchty, S., Fontana, W., Hofacker, I.L., Schuster, P., 1999. Complete suboptimal folding of RNA and the stability of secondary structures. Biopolymers 49, 145–165.

Zuker, M., 1989. The use of dynamic programming algorithms in RNA secondary structure prediction. In: Waterman, M.S. (Ed.), Mathematical Methods for DNA Sequences. CRC Press, Boca Roton, pp. 159–184.

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