*RNA* (1997), 3:779–791. Cambridge University Press. Printed in the USA.

# Secondary structure of the 3′-noncoding region of flavivirus genomes: Comparative analysis of base pairing probabilities

SUSANNE RAUSCHER,[1] CHRISTOPH FLAMM,[1] CHRISTIAN W. MANDL,[2]
FRANZ X. HEINZ,[2] and PETER F. STADLER[1,3]

[1] Institut für Theoretische Chemie, Universität Wien, Austria
[2] Institut für Virologie, Universität Wien, Austria
[3] The Santa Fe Institute, Santa Fe, New Mexico, USA

## ABSTRACT

**The prediction of the complete matrix of base pairing probabilities was applied to the 3′ noncoding region (NCR) of flavivirus genomes. This approach identifies not only well-defined secondary structure elements, but also regions of high structural flexibility. Flaviviruses, many of which are important human pathogens, have a common genomic organization, but exhibit a significant degree of RNA sequence diversity in the functionally important 3′-NCR. We demonstrate the presence of secondary structures shared by all flaviviruses, as well as structural features that are characteristic for groups of viruses within the genus reflecting the established classification scheme. The significance of most of the predicted structures is corroborated by compensatory mutations. The availability of infectious clones for several flaviviruses will allow the assessment of these structural elements in processes of the viral life cycle, such as replication and assembly.**

**Keywords: RNA secondary structure; structural alignment; base pairing probabilities; flavivirus; 3′-noncoding region**

## INTRODUCTION

Secondary structures, that is, the pattern of Watson-Crick and GU base pairs, account for the major part of the free energy of the spatial structures of nucleic acids. They can be predicted fairly reliably—although not with perfect accuracy—using the standard energy model in which additive energy contributions are assigned to the stacking of base pairs and to loops (Freier et al., 1986). Knots and pseudoknots are usually excluded from the definition of secondary structure for a number of reasons. (1) Very little is known about the thermodynamics of pseudoknots, hence there are no reliable energy parameters. (2) The most efficient folding algorithms, which are based upon dynamic programming, cannot deal with knots or pseudoknots (Zuker & Sankoff, 1984). (3) Pseudoknots can in some cases be understood as an additional feature that is formed on top of the conventional secondary structure.

The comparative method, for the right class of RNA molecules, does very well at predicting the secondary structure and some of the tertiary interactions (Gutell, 1993). Unfortunately, this method requires numerous phylogenetically related and distant sequences for the same RNA molecule. Consequently, the data sets available for most groups of viruses are far too sparse for this technique. On the other hand, thermodynamic methods do not provide completely accurate predictions. In this contribution, we show that a hybrid of thermodynamic predictions and comparative methods for verifying the predictions can be applied successfully to viral sequences.

Of course, the additive energy model is an approximation and the experimentally determined energy parameters suffer from inaccuracies. It is not sufficient, hence, to predict the minimum free energy structure only. In addition, there is no guarantee that the global energy minimum will be found by a folding RNA molecule, in particular with long sequences. Kinetic folding approaches were proposed that are designed to simulate the folding pathway (Gultyaev et al., 1995; van Batenburg et al., 1995). Not being limited to dynamic programming, this approach allows to incorporate pseudoknots (at which point it suffers, of course, from the sparse data sets on their thermodynamic properties). It is, therefore, desirable to include additional structural information, for instance from phylogenetic

Reprint requests to: Peter F. Stadler, Inst. f. Theoretische Chemie, Universität Wien, Währingerstr. 17, A-1090 Wien, Austria; e-mail: studla@tbi.univie.ac.at or stadler@santafe.edu.

comparisons or from chemical probing, in the structure prediction. This is straightforward in the energy minimization.

Predicting a single structure by any approach will, in general, not provide a completely correct answer. In addition, knowledge of the uncertainty of the predicted structure in different regions is most useful for a meaningful interpretation of the data. Two approaches are used routinely to overcome the limitations of single structure predictions: Zuker (1989) devised a version of the folding algorithm that computes a set of suboptimal structures in a certain energy range (see also Jacobson & Zuker, 1993). McCaskill (1990) designed an algorithm that produces the complete matrix of base pairing probabilities $p_{ij}$ in thermodynamic equilibrium based on the computing of the equilibrium partition function. This method provides rather detailed information not only on the structure, but also on the local structural flexibility. It was applied successfully in a recent analysis of the complete genomic RNA of HIV-1. In this work, the entire genome was folded in a single piece on a CRAY Y-MP supercomputer (Huynen et al., 1996a). A large number of known secondary structure elements in different regions of the molecules were present in very good resolution in the data, indicating that secondary structure prediction is indeed a meaningful enterprise with RNAs as large as entire viral genomes.

The genus flaviviruses, family Flaviviridae, comprises almost 70, mostly mosquito- or tick-borne viruses, including a number of human pathogens of global medical importance (for summary, see Monath & Heinz, 1996). Flaviviruses are small enveloped particles with an unsegmented, plus-stranded RNA genome. Based on sequence homologies and serological data, the flaviviruses are subdivided in several serocomplexes, such as (1) the Dengue (DEN) virus types 1–4, (2) Japanese encephalitis (JE), West Nile (WN), Kunjin (KUN) and other viruses, (3) yellow fever (YF) virus, and (4) tick-borne flaviviruses. The main representative of the tick-borne group of flaviviruses is tick-borne encephalitis (TBE) virus, which is endemic in many parts of Europe (European subtype) and Asia (Far Eastern subtype). The most distantly related member of this complex is Powassan (POW) virus, which shares 76% protein sequence homology with TBE virus (Mandl et al., 1993). POW virus is endemic in parts of Canada and Far East Asia and causes sporadic cases of encephalitis in humans.

About 90% of the approximately 11-kb long flavivirus genome is taken up by a single, long, open reading frame that encodes a polyprotein that is co- and post-translationally cleaved by viral and cellular proteases into 10 viral proteins (for review, see Rice, 1996). The flanking noncoding regions (NCR) are believed to contain *cis*-acting elements important for replication, translation, and packaging. In this context, most attention has focused on the 3′-NCR, which is considerably longer than the only approximately 100-nt long 5′-NCR. Short conserved primary sequence motifs were identified in the 3′-NCRs of mosquito-borne flavivirus genomes (Hahn et al., 1987), but these were found to be absent in tick-borne flaviviruses (Mandl et al., 1993). Sequence analysis of a number of TBE virus strains recently revealed a surprising heterogeneity in length of the 3′-NCRs, even among closely related strains (Wallner et al., 1995). The 3′-NCR of TBE virus is subdivided into a variable region and a 3′-terminal core element. The former can range in size from less than 50 nt to more than 400 nt, and includes, in the case of some strains, an internal poly(A) sequence element, whereas the latter is 350-nt long and exhibits a high degree of sequence conservation (Wallner et al., 1995). A secondary structure was proposed for the 3′-terminal 106 nt of this core element, which is also found in the sequence of POW virus (Mandl et al., 1993). Very similar structures were reported for the sequences of mosquito-borne flaviviruses (Grange et al., 1985; Brinton et al., 1986; Wengler & Castle, 1986; Hahn et al., 1987; Shi et al., 1996) in spite of little sequence conservation, suggesting a functional importance of this secondary structure, which may interact with viral or cellular proteins during the initiation of the minus-strand synthesis (Blackwell & Brinton, 1995).

In this contribution, we apply McCaskill's (1990) partition function algorithm to explore the secondary structures of complete genomic 3′-NCRs of flaviviruses and report:

1. The 3′-NCRs of flavivirus genomes form conserved secondary structure motifs, but there are considerable differences in RNA folding between the different flavivirus serocomplexes.
2. Our analysis confirms the existence of the stem-loop structure at the very 3′-end that is described in previous investigations. It is present in almost the same form in all flaviviruses.
3. The 3′-terminal secondary structure was shown to include an ill-defined part consistent with the formation of a pseudoknot as reported for the WN, DEN-3, and YF virus sequences by Brinton and co-workers (Shi et al., 1996).
4. The core element of the 3′-NCR of TBE virus folds into a highly conserved secondary structure independent of the adjacent variable region.
5. A particular structural element distinguishes the 3′-NCRs of European subtype TBE virus strains from Far Eastern strains and POW virus.

## RESULTS

### Tick-borne flaviviruses

We were able to confirm the characteristic secondary structure at the very 3′ end of the genome that was found in all tick-borne flaviviruses (Mandl et al., 1993; Wallner et al., 1995). The dot plot of this region is shown in Figure 1. The base pairing is very well-defined in this region, the plot shows only few alternatives to the minimum free energy structure, with one exception: stem 3 and 4 of the ground state structure can be replaced by an elongated stem 4 at the expense of opening stem 3 completely. This structural detail is involved in formation of the pseudoknot described by Shi et al. (1996) (see also Fig. 4). The minimum free energy structure is shown in bold in Figure 2.

Immediately upstream of the known conserved 3′-terminal stem-loop, we find a well-conserved new motif consisting of six stems. The corresponding sequences are quite conserved in this region. In particular, not all of the six stems are confirmed by compensatory mutations. The numbering of the stems is defined in Figure 2. The positions with compensatory mutations are indicated by circles. I, The sequence is conserved among all sequences. II, The sequence is conserved among all sequences. III, POW virus: 3: AU → GC 6: GC → AU 7: AU → GC. IV, The sequence is conserved except for a deletion in POW virus. V, The sequence is conserved except for POW virus. This virus forms the same secondary structure elements V and VI with quite different sequence motifs. VI, 5: Far Eastern: GC; *Aina*: AU; European: GU.

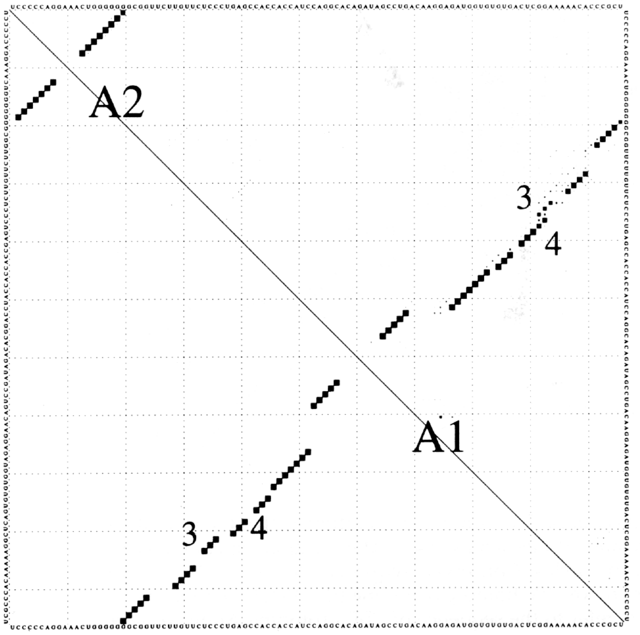

**FIGURE 1.** Dot plot of the conserved secondary structure formed by the 106 nt at the 3′ terminus of the POW virus sequence. This structure is mostly unambiguous, except for the third and fourth stem. In the minimum free energy structure, we have two stems of length 3. Alternatively, the fourth stem may be elongated by two base pairs and stem 3 opens up. This area is involved in the formation of the pseudoknot described by Shi et al. (1996). The area of the squares is proportional to $p_{ij}$ (not to $\log p_{ij}$ as in McCaskill's work).

Compensatory mutations are commonly interpreted as the result of selection pressure, and, thus, they are indicative of a functional secondary structure element. We conjecture hence that the stems III and VI, and, in fact, most likely the entire domain consisting of the stems I–VI, are important for the viral life cycle. Deletion mutants could be used to test this prediction.

As shown in Figures 3 and 4, this motif seems to appear in two variants that correlate with the serological classification. In order to check the significance of this finding, we calculate the free energy of folding each sequence into the secondary structure of the other subtype. Except for the *Ljubljana* strain of TBE virus, the energy differences are well above the thermal energy (see Table 1). The chance that the assignment of the two secondary structure variants to the serological subtypes is accidental is only $1{:}2^8 \approx 0.4\%$, even if the individual energy differences were not significant.

The minimum free energy structure of POW virus resembles the Far Eastern subtype except for a variation in stems V and VI. Although the overall shape remains the same, we find a shorter stem VI and a longer stem V in POW virus. The sequence of the hairpin loop on top of stem VI contains six conserved nucleotides, AAGGC--A. The calculated energy difference to the European subtype structure (again allowing for the energetically optimal fold in the V/VI region) is +6.66 kcal, more than 10 times the thermal energy.

Even further toward the 5′ end, we find ample evidence for a conserved Y-shaped motif consisting of some 90 nt (see Fig. 3). The stems are labeled VII, VIII, and IX. In addition, there is evidence for an isolated hairpin X in most sequences. This motif is quite well conserved in European subtype sequences, but shows a substantial variation in Far Eastern subtype. In particular, the size of the stems and loops may vary considerably; the overall shape seems to be well-conserved, however.

A more detailed analysis of this region reveals substantial variations in the structural variability as measured by the well-definedness parameter $d(k)$. Figure 4 compares $d(k)$ for the Far Eastern and European subtype sequences, respectively. The region around position 11,070 (approximately 70 nt from the 3′ end) is ill-defined in all sequences. It corresponds to a pseudoknot formed by the nucleotides in the hairpin loop of A2, together with their counterparts in the long stem-loop structure A1. Potential pseudoknots that compete with other secondary structure elements sometimes appear as ill-defined regions in well-definedness plots, despite the fact that the computational model does not make explicit use of pseudoknots. The pseudoknotted structure formed by A2 and A1 was discussed in detail by Brinton and co-workers (Shi et al., 1996; see also the sketch in Fig. 4).

Not surprisingly, the consensus mountains of the two subtypes of TBE viruses differ only by the stem enclos-

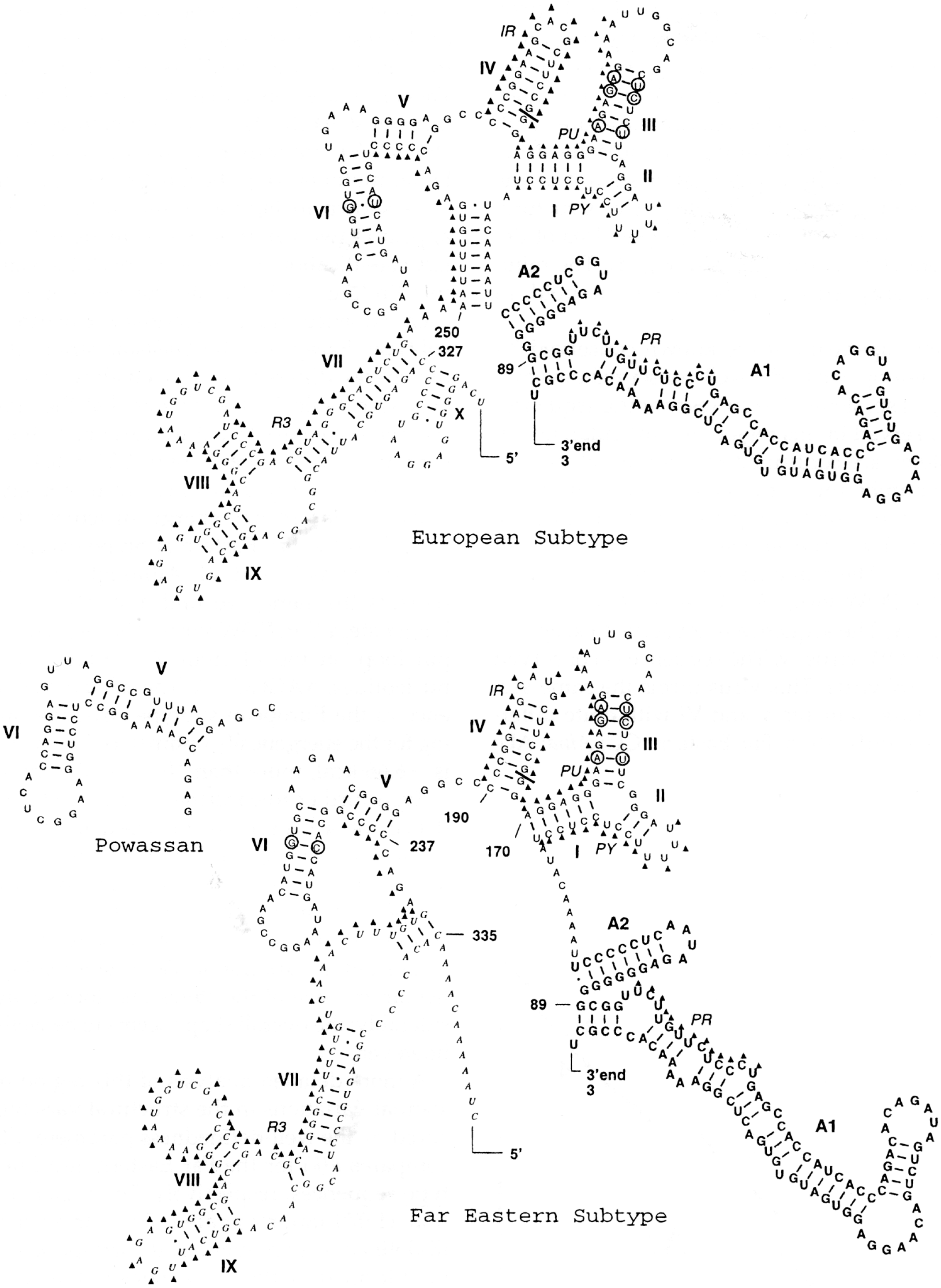

**FIGURE 2.** Secondary structure of the 3′-NCR of tick-borne flaviviruses. The conserved structure at the 3′ end is shown in bold. POW virus differs from the other structures in stems V and VI (shown as inset). The locations of compensatory mutations in at least one of the nine sequences are indicated by circles. Conserved sequence elements adapted from Wallner et al. (1995) are indicated by triangles. PR, pyrimidine-rich box; PY, homo-pyrimidine box; PU, homo-purine box; IR, inverted repeat; R3, imperfect direct repeat. Position numbers refer to the sequence of TBE prototype strain *Neudoerfl.*

ing the stems III–VI. The variances of the slopes are very small almost everywhere else, indicating that we are confronted with a very well-conserved structure. In other words, the classical picture shown in Figure 2 is also the final outcome of the comparative partition function method. The region VII–X, on the other hand, shows both a rather large variance and a small well-definedness. It is interesting to note that European subtype sequences are substantially less well-defined in the region [VII, VIII, IX] than Far Eastern subtype.

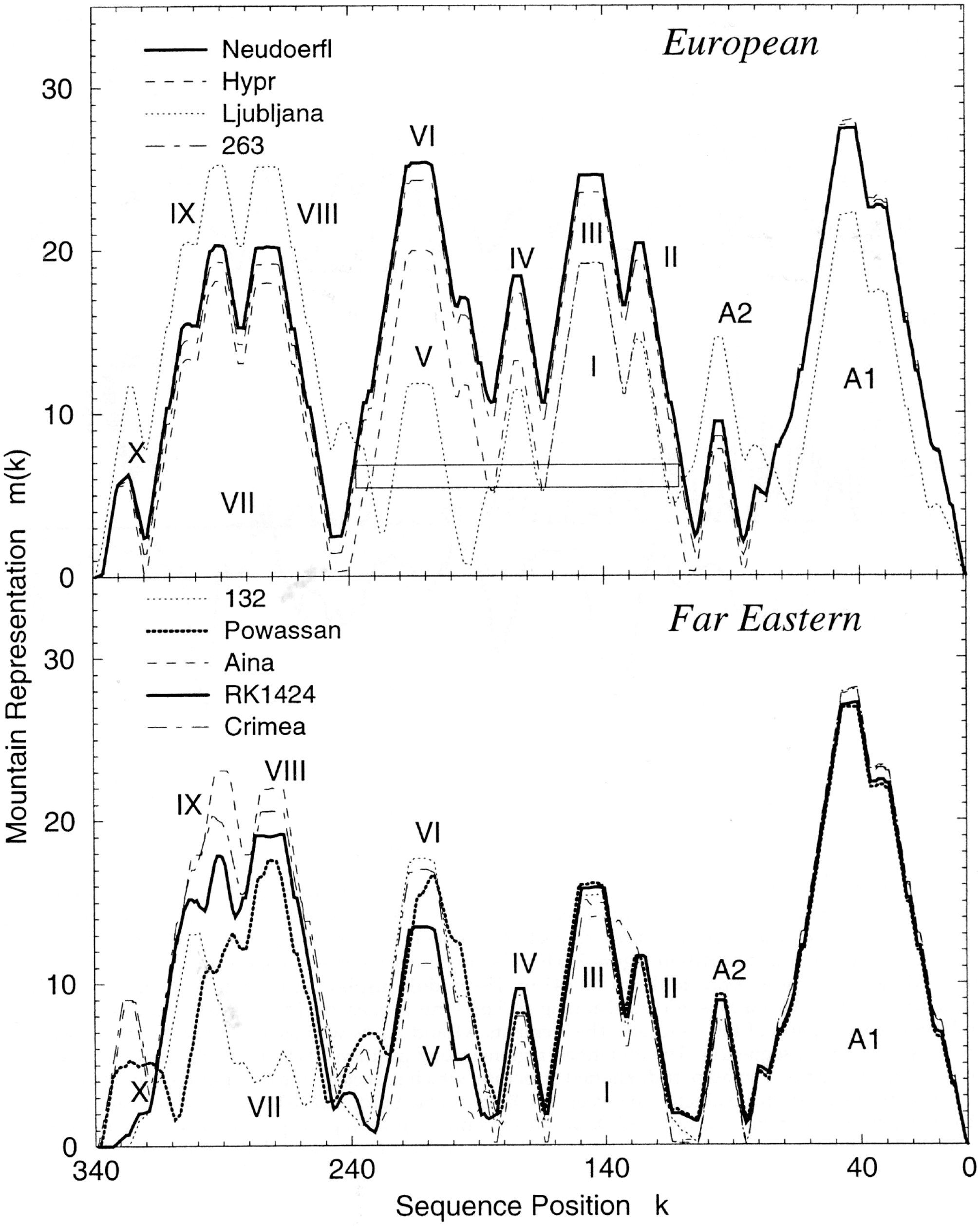

**FIGURE 3.** Mountain representations of the 3′-NCR of TBE viruses. Position numbers are counted from the 3′ end according to a multiple alignment. The stem discriminating Far Eastern from European subtype sequences is indicated in the upper plot.

## Other flavivirus serocomplexes

The analyses of the 18 YF virus sequences are summarized in Figure 5. The structural domain A at the 3′ end, first described by Grange et al. (1985), closely resembles its counterpart A1 in tick-borne flaviviruses. It is a very well-defined stem-loop structure with a large bulge that is almost perfectly conserved. A second domain, B, about 180–280 nt away from the 3′ end, consists of four hairpins. Domain C is located between 350 and 400 nt away from the 3′ end and is composed of two hairpins. Regions A and B are separated by a piece of sequence that is characterized by exceptionally low values of $d(k)$ and that does not exhibit a preferred secondary structure. This region possibly acts as a "spacer" separating region A from the rest of the genome. Beyond some 370 nt from the 3′ end, we do not find unambiguously predicted structures.

Viruses of the JE and DEN serocomplexes yield qualitatively similar results. Our predictions of conserved structures in their 3′-NCRs are summarized in Figure 6. Consistent with previous findings, we find a well-defined and well-conserved stem-loop structure

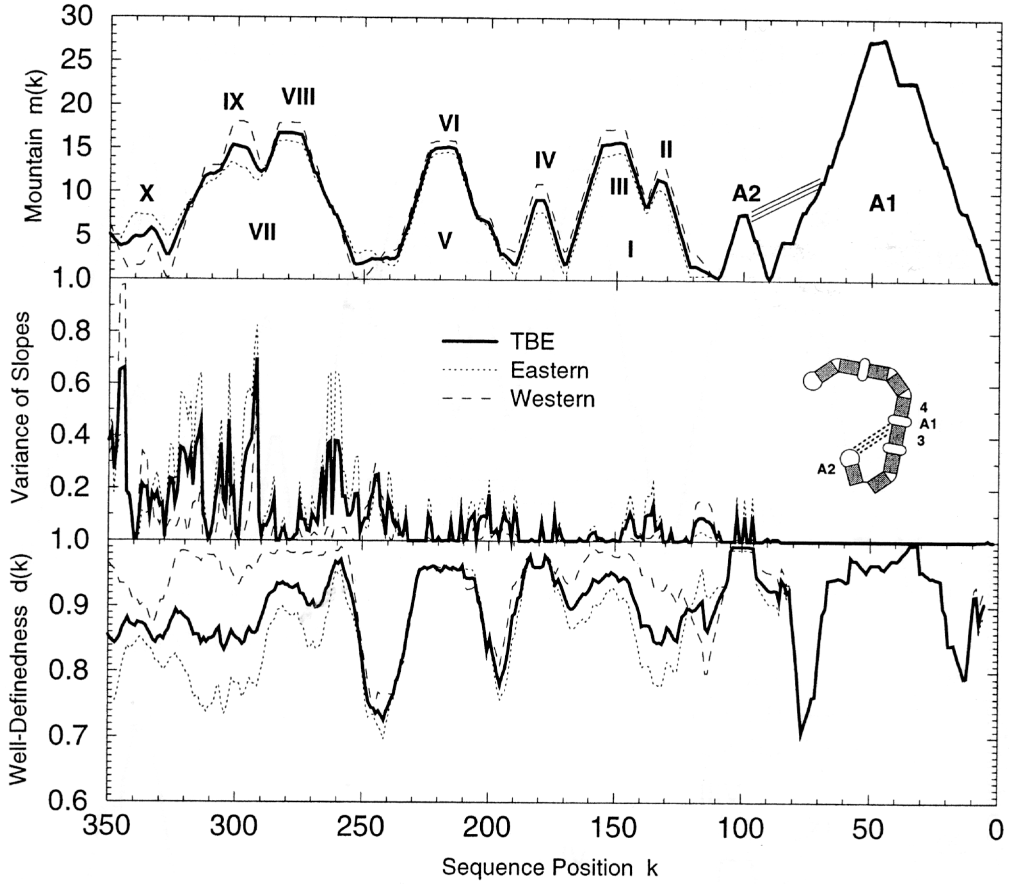

**FIGURE 4.** Well-definedness and conservedness of the 3′-NCR secondary structure of TBE viruses. The upper part of the figure compares the mountain representation of the consensus of all TBE virus sequences (Table 2) with the consensus of the Far Eastern and European subtype structures, respectively. Sequence positions are numbers from the 3′ end according to a multiple sequence alignment shown in Figure 7. This alignment inserts about 10 gaps into the original sequences, the range displayed here thus corresponds to the distal 341 nt. The middle display contains the variances of the slopes. Note that the structural element A1 is almost identical among all TBE virus sequences. The well-definedness, $d(k)$, shown at the bottom indicates flexible parts of the molecule. The elements A1 and A2 as well as a number of other hairpins are very well-defined, $d(k) \approx 1$, in all sequences. The ill-defined region around position 70 (11,070 in the strain *Neudoerfl* sequence) corresponds to the location of a small pseudoknotted structure, which is sketched in the middle panel (see Shi et al., 1996 for details). The most significant distinction between Far Eastern and European subtype is the region between positions 350 and 250, comprising the stems VII, VIII, and IX, where the European subtype sequences seem to be much more flexible than the Far Eastern subtype. Note that not much correlation exists between well-definedness and the variance of the slopes.

at their 3′ end. In the cases of DEN viruses, however, the stem of this structure is shorter than for other flaviviruses.

A particularly interesting feature is the T-shaped element B, which is shared by JE and DEN, but appearently is not conserved in TBE or YF viruses. A large number of compensatory mutations confirms this structural element (see Fig. 6). It occurs at different genomic positions in DEN and JE sequences, and includes, respectively, the highly conserved, mosquito-borne flavivirus-specific sequence elements CS2 and RCS2 (Hahn et al., 1987). The sequence of the adjacent hairpin loop is highly conserved in DEN viruses, but shows variations among the members of the JE serocomplex. We were not able to confirm a similar structure for the CS2 sequence of YF viruses; RCS2 is absent in this group of viral sequences.

## DISCUSSION

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 hence does not imply any functional significance. It is important, therefore, to develop methods for identifying potentially functional parts of a secondary structure prior to experiments.

Elucidation of all the significant secondary structures is a necessary prerequisite for the understanding of the molecular biology of a virus. So far, a number of relevant secondary structures have been determined that play a role during the various stages of the viral life cycle in a variety of different classes of viruses, for instance lentiviruses (Baudin et al., 1993; Hofacker et al., 1996), RNA phages (Biebricher, 1994; Olsthoorn

**TABLE 1**. Folding energies (in kcal/mol) for nine flavivirus sequences.[a]

| Sequence | GenBank | $E$(mfe) (kcal/mol) | $\Delta E$(far east) (kcal/mol) | $\Delta E$(europ) (kcal/mol) |
|---|---|---|---|---|
| Neudoerfl | U27495 | −84.39 | +2.04 | 0 |
| 263 | U27491 | −84.39 | +2.04 | 0 |
| Ljubljana | U27494 | −82.61 | +0.07 | 0 |
| Hypr | U39292 | −84.48 | +1.26 | 0 |
| Crimea | U27493 | −85.12 | 0 | +2.80 |
| 132 | U27490 | −88.94 | 0 | +7.54 |
| RK1424 | U27496 | −86.07 | 0 | +0.88 |
| Aina | U27492 | −86.24 | 0 | +7.23 |
| Powassan | L06436 | −86.24 | 0 | +6.66 |

[a] $\Delta E$(far east) and $\Delta E$(europ) are the energy differences to the minimum free energy structure when the sequences are forced to fold with or without the stem enclosing the motifs I–VI (see Fig. 2). A re-evaluation of the energies using the parameters in Walter et al. (1994), which include co-axial stacking, yields the same qualitative result.

et al., 1995), flaviviruses (Shi et al., 1996), pestiviruses (Brown et al., 1992; Deng & Brock, 1993), and hepatitis C viruses (Brown et al., 1992; Tanaka et al., 1996).

A large number of secondary structure predictions in the literature are based on phylogenetic comparison without considering energetics. Most predictions based on thermodynamics so far only consider the minimum energy structure and/or a fairly small sample of suboptimal structures, as provided, e.g., by Zuker's *mfold* package (Zuker & Sankoff, 1984; Zuker, 1989). McCaskill's partition function approach (McCaskill, 1990), which allows for an exact computation of the complete matrix of all base pairing probabilities, provides more complete and reliable structural information. By calculating the probability distribution of all base pair interactions, we have a tool that allows us to predict the structure and estimate the reliability of the prediction at the same time.

The data sets available for most groups of viruses are too small for the phylogenetic method. Hence, we have combined thermodynamic predictions with the comparative approach in the contribution; energy-driven folding is used for structure prediction, compensatory mutations allow for a verification of the predicted base pairs in many cases.

A particularly important point is the fact that the well-definedness and the variance of the slopes of the mountain representation are not strongly correlated. Ill-defined regions with small variance could thus indicate flexible parts of the molecule that are possibly of functional importance or regions that are involved in pseudoknots, rather than being an artefact of inaccurate predictions. Again, we interpret conservedness of predicted structure as a sign of functional importance. Ill-defined regions with a high variance between different sequences, on the other hand, suggest that structural features are not significant for RNA function. In contrast to earlier approaches (Zuker & Jacobson, 1995; Huynen et al., 1996a, 1997), functional importance is thus not tantamount to thermodynamic stability in our scheme. It is also an advantage of our method that it does not necessarily predict secondary structures for all parts of the molecules. The averaging of the mountain representations for the individual sequences amplifies conserved elements only, whereas variable regions disappear in the background.

This technique was applied to the 3′-NCR of flavivirus genomes. A previously described secondary structure motif formed by the 3′-terminal approximately 100 nt (Grange et al., 1985; Brinton et al., 1986; Wengler & Castle, 1986; Hahn et al., 1987; Mandl et al., 1993; Shi et al., 1996) was confirmed for all flavivirus sequences. However, in the cases of DEN viruses, our analysis predicted a somewhat different structure with a significantly shorter stem than present in other flavivirus sequences. In addition, we found well-defined secondary structures in the 3′-NCRs of mosquito-borne and tick-borne flaviviruses. One structural element (termed B in Fig. 6) was found to be present in both the DEN viruses and the members of the JE serocomplex, but surprisingly, located at different genomic positions. The 3′-NCRs of these viruses contain two copies of a sequence motif, termed CS2 and RCS2 in Hahn et al. (1987), that are highly conserved among mosquito-borne flaviviruses. Structure B includes CS2 in the cases of DEN viruses, but RCS2 in the JE group. A potential functional importance of this structure is suggested by its high degree of conservation and the presence of a large number of compensatory mutations.

The core element of the TBE virus 3′-NCR (Wallner et al., 1995), i.e., the 3′-terminal 341 nt, was found to fold into a conserved structural pattern irrespective of the presence of various sequence elements in the adjacent variable region. This observation is compatible with the idea that the core element represents a minimal, but functionally sufficient 3′-NCR of tick-borne flaviviruses. Interestingly, the European strains of TBE virus are distinct from the other tick-borne flavivirus sequences by a particular structural element, which, however, is shared by the Far Eastern TBE virus strains and POW virus, suggesting a somewhat closer evolutionary link between POW virus (which is also endemic in Far East Asia) and the Far Eastern subtype of TBE virus than between POW and the European TBE subtype. A statistical study of the distribution of secondary structures over the space of possible sequences of fixed length (Schuster et al., 1994) shows the following. (1) Many different sequences do fold into the same (minimum energy) structure. (2) Neutral paths percolate sequence space along which all sequences fold into the same secondary structure. In fact, there are extended *neutral networks* of sequences folding into the same structure (Grüner et al., 1996; Reidys et al., 1997).

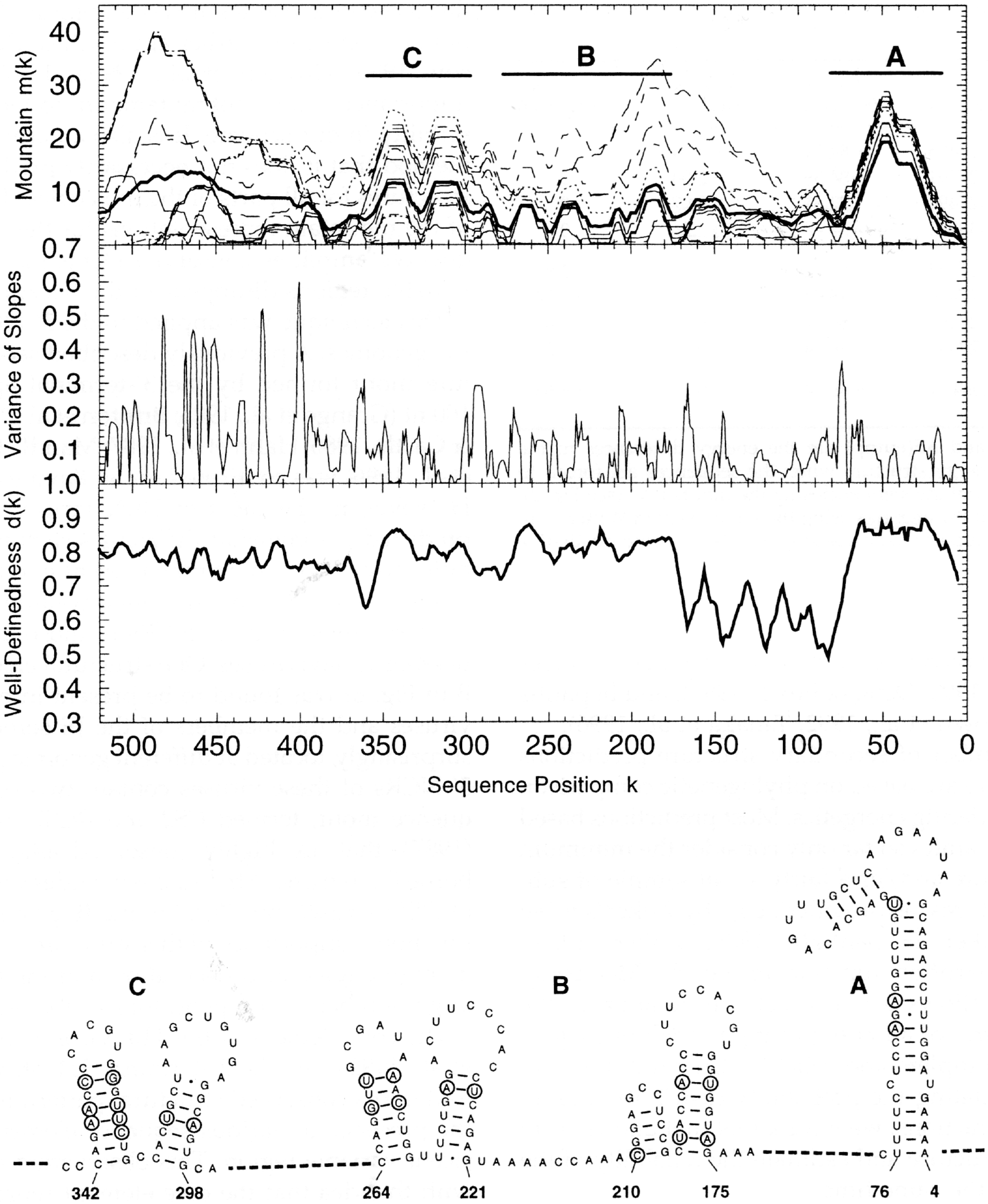

**FIGURE 5.** Secondary structure of the YF virus 3′-NCR. The upper part of the figure shows the generalized mountain representations of all 18 YF virus sequences and their consensus (bold). Below the conservedness (variance of the slopes) and the well-definedness are shown. Position numbers are counted from the 3′ end. For details see the text. The lower part of the figure is a conventional display of the consensus secondary structure as determined from the upper part of the figure. Our method does not predict a defined structural model for all of the sequence; dashed lines indicate undetermined pieces. Compensatory mutations are indicated by circles. Circles on only one side of a stem indicate GC-GU or AU-GU mutations.

This permits populations of sequences to split and drift apart from one another in sequence space without changing their dominant phenotype (Huynen et al., 1996b). Drift in sequence space therefore does not necessarily imply drift in phenotype space. This finding seems to explain the discrepancy between sequence-based and structure-based phylogeny of flaviviruses. The *Powassan* case may serve as an example for this scenario.

The functional importance of the secondary structures described in this communication will have to be verified by direct biological testing. Infectious cDNA clones that became available recently for several flaviviruses (Rice et al., 1989; Lai et al., 1991; Sumiyoshi et al., 1992; Khromykh & Westaway, 1994; Mandl et al., 1997) allow us to assess the effects of specific mutations on the biology of these viruses. The structural predictions presented here

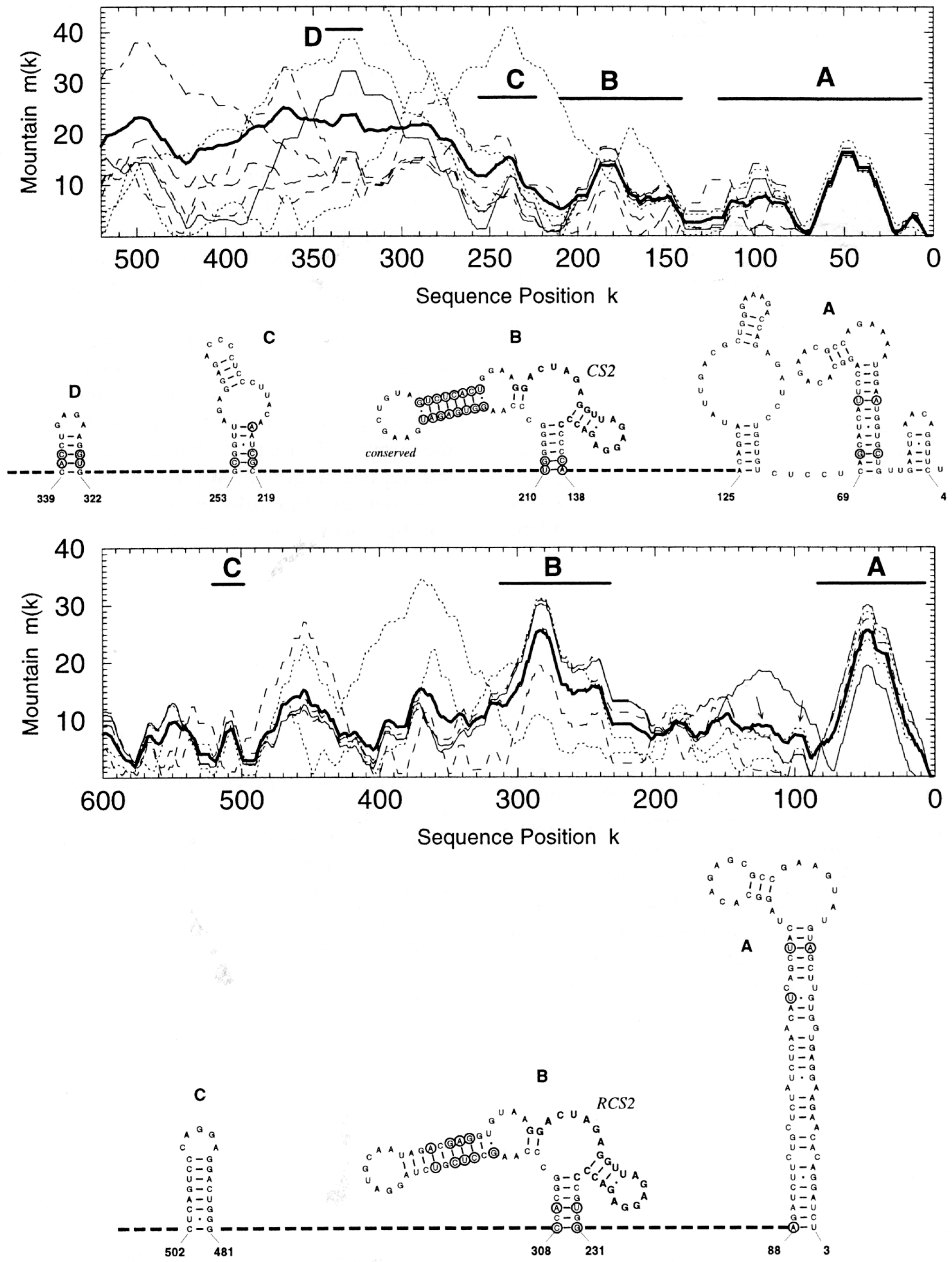

**FIGURE 6.** Conserved secondary structures in the DEN (above) and JE (below) serocomplexes. We show the superposition of the mountain representations and the conventional diagrams of conserved consensus structures. Sequences that are unpaired consistently but occur in variable structural contexts are indicated by arrows. A large number of compensatory mutations (indicated by circles) confirms the proposed structures. Note that element B has the same fold in both DEN and JE serocomplexes, although it occurs at different genomic positions. The left hairpin loop of this element has a highly conserved sequence in DEN viruses, hinting at its functional importance.

can serve as a rational basis for future mutagenesis experiments.

## MATERIALS AND METHODS

### Software and sequences

For our analysis, we used the 44 flaviviruses sequences listed in Table 2. From each sequence, we extracted the 3′-NCR of the genomic RNA. Using a set of longer sequences with up to 1,000 nt, we checked that this portion of the 3′-NCR folds as a distinct unit, i.e., that the terminal nucleotides form the same structure irrespective of additional fragments further toward the 5′ end. Very long-range interaction, spanning more than 1 kb, cannot be excluded by this method.

All computations reported in this paper were performed using the Vienna RNA package, which contains a variety of programs for the computation and comparison of RNA sec-

**TABLE 2**. List of sequences used in this contribution.

| Tick-borne encephalitis | Japanese encephalitis | Dengue | Yellow fever |
|---|---|---|---|
| Far Eastern | JE | Type 1 | U17066 |
| U27490 | U14163 | M87512 | U17067 |
| U27492 | U15763 | | U21055 |
| U27493 | M18370 | Type 2 | U52393 |
| U27496 | M55506 | M29095 | U52396 |
| | D90194 | M84728 | U52399 |
| European | D90195 | M84727 | U52401 |
| U27491 | L48961 | M20558 | U52405 |
| U27494 | | M19197 | U52407 |
| U27495 | West Nile | | U52411 |
| U39292 | M12294 | Type 3 | U52414 |
| | | M93130 | U52417 |
| Powassan | Kunjin | | U52420 |
| L06436 | L24512 | Type 4 | U52423 |
| | | M14931 | U54798 |
| | | | K02749 |
| | | | X02807 |
| | | | X03700 |

ondary structures (Hofacker et al., 1994). This public domain software can be obtained by anonymous ftp (ftp://ftp.itc.univie.ac.at/pub/RNA).

The energy parameters used by the Vienna RNA package are based on Freier et al. (1986), Jaeger et al. (1989), and He et al. (1991). They are identical with the parameters in Zuker's *mfold 2.2*, with the exception of stacking energies involving GU pairs, which were taken from He et al. (1991).

Walter et al. (1994) show that the inclusion of co-axial stacking can improve the quality of the prediction. Currently, there is no algorithm available that includes this effect. However, the partition function algorithm approximates co-axial stacking by always taking the dangling end contributions into account even if the base preceeding or following a helix is paired.

A multiple sequence alignment was calculated using CLUSTAL W (Thompson et al., 1994). The result is shown as Figure 7. We have not attempted to improve the alignment based on visual inspection or based on predicted secondary structures. Although this might have increased the number of compensatory mutations and possibly also the number of detected structural elements, it would also have compromised the use of the sequence data for verifying the predicted structures. We have therefore kept the sequence data and the structure data strictly separated. The unimproved CLUSTAL W alignment is used to align the structure data shown in Figures 3–6. All subsequent analysis, such as consensus mountains and conservedness measures, are based on this alignment.

## Dot plots, mountain representation, and well-definedness

A dot plot is a two-dimensional graph in which the area of the dot at position $(i,j)$ represents the probability $p_{ij}$ that the base pair $(i,j)$ occurs in thermodynamic equilibrium. (In practice, base pairs that occur with a probability of less than $10^{-5}$ are suppressed in our program.) The plot is divided into two triangles (Fig. 1). The upper-right triangle contains the base pairing probability matrix $(p_{ij})$. In the lower-left triangle, we display the minimum free energy structure for comparison.

Figure 1 shows the dot plot of the 3′-terminal structure. Hairpin loops appear as diagonal patterns close to the separating line between the two triangles, with the distance from this line indicating the loop size. Internal loops and bulges appear as shift and gaps in these diagonal patterns.

A very convenient way of displaying the size and distribution of secondary structure elements is a modified version of the mountain representation introduced in Hogeweg and Hesper (1984). In the original mountain representation, a single secondary structure is represented in a two-dimensional graph in which the $x$-coordinate is the position $k$ of a nucleotide in the sequence and the $y$-coordinate is proportional to the number of base pairs that enclose nucleotide $k$. The mountain representation allows for a straightforward comparison of secondary structures and inspired a convenient algorithm for alignment of secondary structures (Konings & Hogeweg, 1989).

A modified version of the mountain representation can be constructed easily from the base pairing probability matrix. The number

$$m(k) \overset{\text{def}}{=} \sum_{i<k} \sum_{j>k} p_{ij} \tag{1}$$

counts all base pairs[4] containing nucleotide $k$, weighted with their respective pairing probabilities. In order to see that $m(k)$ is, in fact, a close relative of the mountain representation, we assume for a moment that $(p_{ij})$ is the pairing matrix of a minimum free energy structure. In this case, $m(k)$ is the number of base pairs that contain $k$, i.e., it is constant for any position in a loop, increases by one at each paired position at the 5′ side of a stack, and decreases by one at each paired position at the 3′ side of a stack. $m(k) = 0$ if $k$ is either an external base or the outermost base pair of a component. The well-definedness of the structure in a certain region is

$$d(k) \overset{\text{def}}{=} \max\left\{\max_i\{p_{ik}, p_{ki}\}, 1 - \sum_i p_{ik}\right\}, \tag{2}$$

i.e., $d(k)$ is the probability of the most probable base pair involving $k$, or the probability that $k$ is unpaired, whichever is larger. Thus, $d(k)$ is large when a base either has a high probability of pairing with one specific other base, or it has a high probability of not interacting at all. A plot of $d(k)$ versus nucleotide position reveals information on the stability of small scale patterns. The idea behind measuring $d(k)$ is that the well-definedness of a region provides information about its functional significance. A secondary structure that is important for the function of a molecule will oftentimes have a high probability of occurring in the thermodynamic ensemble of alternative secondary structures *and* should not just be one of the many alternative structures that have a near equal probability of occurring. Note that this does not exclude that a (small) part of a well-defined structural ele-

[4] In the terminology of Zuker and Sankoff (1984), these are all base pairs to which sequence position $k$ is interior.

```
TEU27490    --AAAAAACAAACACA-----------CCCCCGTGTATTCCACGGCAACACGTCAGTGAGAGTGGCGACGGGAAAATTGTCGACCCCGACGTAGG-GCATTCTGTCCAATTTCGTGAGG
TEU27492    --AGCTGGTCAGGGGTG--AG-AACACCCCCAGAGTGCACCACGTCAACACGCCAGTGAGAGTGGCGACGGGAAAATGGTCGATCCCGACGTAGG-GCACTCTGTAAAACTTTGTGAGA
TEU27493    --AGCTGGCCGGGGT----AGAAACACCCCCGGAGTGCCCCACGGCAGCACGTCAGTGAGAGTGGCGACGGGAAAATGGTCGATCCCGACGAAGG-GTACTCTGAAAAA-TTTGTGAGA
TEU27496    ---------------------------CCCCGGAGTGCCCTACGGCAACACGTCATTGAGAGTGGCGACGGGAAAATGGTCGACCCCGACGCAGG-GCATTCTGTCAAACTTTGTGAGA
PWARPT      ACGGCTGGCAAAGGGCACCAGTCGTGTAGTAAGAAGGCCCTGGCCCAGTGCGGCAGCACACTCAGTGACGGGAAAGTGGTCGCTCCCGACGTAACTGGGTAAAAACGAACTTTGTGAGA
TBE (fe)    ~~~~~~~~~~~~~~~~~~~~~~((((.....((((((((((...((.....((((((......)))))).((((............)))).)))))..))))))))........))))..
TEU27491    --AACCGGTCAGGGGTG--AGGGATGCCCCCAGAGTGCATTACGGCAGCACGCCAGTGAGAGTGGCGACGGGAAAATGGTCGATCCCGACGTAGG-GCACTCTGAAAAATTTTGTGAGA
TEU27494    --AACCGGTCAGGGGTT--GGGGATGCCCCCAGAGTGCATTACGGCAGCACGCCAGTGAGAGTGGCGACGGGAAAATGGTCGATCCCGACGTAGG-GCACTCTGAAAAACTTTGTGAGA
TEU27495    --GACCGGTCAGGGGTG--AGGAATGCCCCCAGAGTGCATTACGGCAGCACGCCAGTGAGAGTGGCGACGGGAAAATGGTCGATCCCGACGTAGG-GCACTCTGAAAAATTTTGTGAGA
TEU27497    --GATCGGTCAGGGGTG--AGGGATGCCCCCAGAGTGCATTACGGCAGCACGCCAGTGAGAGTGGCGACGGGAAAATGGTCGATCCCGACGTAGG-GCACTCTGAAAAACTTTGTGAGA
TBE (e)     ~~~~~~~~~~~(((((.......))))).((((((((.(((((......((((((.......)))))).(((((..........))))).))))).))))))))...((((((((...
JEU14163    CGCA---GCCCCAGGAGGACTGGGT-TACCAAAGCCGT----TGAGCCCCCACGGCCCAAGCCT-CGTCTAGGATGCAA-TAGACGAGGTGTAA--GGACT---AGAGGTTAGAGGAGA
JEU15763    CGCA---GCCCCAGGAGGACTGGGT-TACCAAAGCCGT----TGAGCCCCCACGGCCCAAGCCT-CGTCTAGGATGCAA-TAGACGAGGTGTAA--GGACT---AGAGGTTAGAGGAGA
JEVBEICG    CGCA---GCCCCAGGAGGACTGGGT-TACCAAAGCCGT----TGAGCCCCCACGGCCCAAGCCT-CGTCTAGGATGCAA-TAGACGAGGTGTAA--GGACT---AGAGGTTAGAGGAGA
JEVCG       CGCA---GCCCCAGGAGGACTGGGT-TACCAAAGCCGT----TGAGCCCCCACGGCCCAAGCCT-CGTCTAGGATGCAA-TAGACGAGGTGTAA--GGACT---AGAGGTTAGAGGAGA
JEVCMNAA    CGCA---GCCCCAGGAGGACTGGGT-TACCAAAGCCGT----TGAGCCCCCACGGCCCAAGCCT-CGTCTAGGATGCAA-TAGACGAGGTGTAA--GGACT---AGAGGTTAGAGGAGA
JEVSAA      CGCA---GCCCCAGGAGGACTGGGT-TACCAAAGCCGT----TGAGCCCCCACGGCCCAAGCCT-CGTCTAGGATGCAA-TAGACGAGGTGTAA--GGACT---AGAGGTTAGAGGAGA
JEVSAV      CGCA---GCCCCAGGAGGACTGGGT-TACCAAAGCCGT----TGAGCCCCCACGGCCCAAGCCT-CGTCTAGGATGCAA-TAGACGAGGTGTAA--GGACT---AGAGGTTAGAGGAGA
WNFCG       CGATAGTGCCCCAGGTGGACTGGGT-TAACAAAGGCAAAACATCGCCCCACGCGGCCATAACCC-TGGCTATGGTGT---TAACCAGGGAGAAG--GGACT---AGAGGTTAGAGGAGA
KUNNONCODB  CGACAGTGCCCCAGGAGGACTGGGT-GAACAAAGGCGAATCAACGTCCCACGCGGCCCTAGCTC-TGGCAATGGTGT---TAACCAGAGTGAAA--GGACT---AGAGGTTAGAGGAGA
JE          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~[[[(((.((..(([[ [[[(((.........)))]]]]]])))...........))...  ......((((......))
DEN2CGA     TGCA---GCCTGTAGCTCCACCT-----GAGAAGGTGTAAAAAA--TCTGGGAGGCCACAAAC-----CATGGAAGCTG-TACGCATGGCGTAGT-GGACT---AGCGGTTAGAGGAGA
DENCMEMSA   TGCA---GCCTGTAGCTCCACCT-----GAGAAGGTGTAAAAAA--TCCGGGAGGCCACAAAC-----CATGGAAGCTG-TACGCATGGCGTAGT-GGACT---AGCGGTTAGAGGAGA
DENCMEMSB   TGCA---GCCTGTAGCTCCACCT-----GAGAAGGTGTAAAAAA--TCCGGGAGGCCACAAAC-----CATGGAAGCTG-TACGCATGGCGTAGT-GGACT---AGCGGTTAGAGGAGA
DENJAMCG    TGCA---GCCTGTAGCTCCACCT-----GAGAAGGTGTAAAAAA--TCTGGGAGGCCACAAAC-----CATGGAAGCTG-TACGCATGGCGTAGT-GGACT---AGCGGTTAGAGGAGA
DENRCG      TGCA---GCCTGTAGCTC-ACCT-----GAGAAGGTGTAAAAAA--TCCGGGAGGCCACAAAC-----CATGGAAGCTG-TACGCATGGCGTAGT-GGACT---AGCGGTTAGAGGAGA
DEN3CME     TGCT---GCCTGTAGCTCCGTC------GTGGGGACGTAAAG----CCTGGGAGGCTGCAAAC-----CGTGGAAGCTG-TACGCACGGTGTAGC-AGACT---AGTGGTTAGAGGAGA
DEN4STRA    TGCT---GCCTGTAGCTCCGCCAAT-AATGGGAGGCGTAATAATCCCCAGGGAGGCCATGCGC-----CACGGAAGCTG-TACGCGTGGCATATT-GGACT---AGCGGTTAGAGGAGA
DENT1SEQ    TGCT---GCCTGTAGCTTCATC------GTGGGGATGTAAAAA---CCTGGGAGGCTGCAACC-----CATGGAAGCTG-TACGCATGGGGTAGC-AGACT---AGTGGTTAGAGGAGA
DEN         ~~~~~~~~~~~~~~~~~~([[[(.......)))]]]~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~([[(((....((((.
FLYF17D1    CCCAGAACCCCA------CACGAGT-T-TTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCCG-ACCTCCAGGTTGCGA-AAAAC---CTGGTTTCTGGGACC
FLYF17DG    CCCAGAACCCCA------CACGAGT-T-TTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCCG-ACCTCCAGGTTGCGA-AAAAC---CTGGTTTCTGGGACC
YFU17066    CCCAGAACTCCA------CACGAGT-T-TTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCCG-ACCTCCAGGTTGCGA-AAAAC---CTGGTTTCTGGGACC
YFU17067    CCCAGAACCCCA------CACGAGT-T-TTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCCG-ACCTCCAGGTTGCGA-AAAAC---CTGGTTTCTGGGACC
YFU21055    CCCAGAACTCCA------CACGAGT-T-TTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCCG-ACCTCCAGGTTGCGA-AAAAC---CTGGTTTCTGGGACC
YFU52393    CCCAGGACTCTAGAATTCTAGGAGT-C-TTGCCACCGCTAAG-----CTGTGAGGCGGTGCGG---GCTGGGACAGCCG-TTTTCCAGATCACGA-CAGAT---CTGGTTTCTGGGATT
YFU52396    CCCAGGACTCTAGAATTCTAGGAGT-C-TTGCCACCGCTAAG-----CTGTGAGGCGGTGCGG---GCTGGGACAGCCG-TTTTCCAGATCACGA-CAGAT---CTGGTTTCTGGGATT
YFU52399    CCCAGAGACCCA------CGTGGGC-T-CTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCGC-ACTTCCAGGTCGCGA-GAAAC---CTGGTTTCTGGGACT
YFU52401    CCCAGAGCCCCA------CAGGGAC-T-CTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCCG-ACCTCCAGGTTGCGA-AAAAT---CTGGTTTCTGGGACT
YFU52405    CCCAGAACCCCA------CGTGGGT-T-CTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCCG-CCTTCCAGGTTGCGA-TAAAC---CTGGTTTCTGAGACT
YFU52407    CCCAGAACCCCA------CGTGGGT-T-CTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCCG-CCTTCCAGGTTGCGA-TAAAC---CTGGTTTCTGAGACT
YFU52411    CCCAGAACCCCA------TGTGGGT-T-CTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCCG-CCTTCCAGGTTGCGA-TAAAC---CTGGTTTCTGAGACT
YFU52414    CCCAGAACTCCA------CACGAGT-T-TTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCCG-ACCTCCAGGTTGCGA-AAAAC---CTGGTTTCTGGGACC
YFU52417    CCCAGAACCCCA------TGTGGGT-T-CTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCCG-CCTTCCAGGTTGCGA-TAAAC---CTGGTTTCTGAGACT
YFU52420    CCCAGAACCCCA------TGTGGGT-T-CTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCCG-CCTTCCAGGTTGCGA-TAAAC---CTGGTTTCTGAGATC
YFU52423    CCCAGGACTCTAGAATTTAGGGAGT-C-TTGCCACCGCTAAG-----CTGTGAGGCGGTGCGG---GCTGGGACAGCCG-TTTTCCAGATCACGA-CAAAT---CTGGTTTCTGGGATT
YFU54798    CCCAGAGCCCCA------CATGGGC-T-CTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCCG-ACCTCCAGGTTGCGA-AAAAT---CTGGTTTCTGGGACC
YFVCG       CCCAGAACCCCA------CACGAGT-T-TTGCCACTGCTAAG-----CTGTGAGGCAGTGCAG---GCTGGGACAGCCG-ACCTCCAGGTTGCGA-AAAAC---CTGGTTTCTGGGACC
YF          ..((([[([(..      ...)]])] ] ])).(((([(((...       ......))))])))...   ~~~~~~~~~~~~~~~~~~((((([([....  ..])]   )))).((((([(...  338

TEU27490    CCCCCTGCACCATGATAAGGCCGAACATGGTGCAAGATAGGGGAGGCCCCCGGAAGCATGCTTCCGGGAGGAGGGAAGAGAGAAATTGGCAACTCTCTTCGGGATTTTTCCTCCTCCT-
TEU27492    CCCCCTGCATCATGACAAGGCCTAACATGATGCACGAA-GGGGAGGCCCCCGGAAGCATGCTTCCGGGAGGAGGGAAGAGAGAAATTGGCAGCTCCCTTCAGGATTTTTCCTCCTCCT-
TEU27493    CCCCCGGCACCATGATAAGGCCGAACATGGTGCAAAAACGGGGAGGCCCCCGGAAGCATGCTTCCGGGAGGAGGGAAGAGAGAAATTGGCAACTCTCTTCGGGATTTTTCCTCCTCCT-
TEU27496    CCCCCGGCACCATGATAAGGCCGAACATGGTGCAAGAACGGGGAGGCCCCCGGAAGCATGCTTCCGGGAGGAGGGAAGAGAGAAATTGGCAACTCTCTTCGGGATTTTTCCTCCTCCT-
PWARPT      CCAAAAGGCCTCCTGGAAGGCTCACCA-GGAGTTAGGCCGTTTAGGAGCCCCCGAGCATAACTC-GGGAGGAGGGAGGAAGAAAATTGGCAATCTTCCTCGGGATTTTTCCGCCTCCT-
TBE (fe)    .((((((([[((((.............))))))))).....)))))......((((((((....))))))))((((((((((((((.............)))))))).)((......))).))))))
TEU27491    CCCCCTGCATCATGATAAGGCCGAACATGGTGCATGAAAGGGGAGGCCCCCGGAAGCTCGCTTCCGGGAGGAGGGAAGAGAGAAATTGGCAGCTCTCTTCAGGATTTTTCCTCCTCCT-
TEU27494    CCCCCTGCATCATGATAAGGCCGAACATGGTGCATGAAAGGGGAGGCCCCCGGAAGCACGCTTCCGGGAGGAGGGAAGAGAGAAATTGGCAGCTCTCTTCAGGATTTTTCCTCCTCCT-
TEU27495    CCCCCTGCATCATGATAAGGCCGAACATGGTGCATGAAAGGGGAGGCCCCCGGAAGCACGCTTCCGGGAGGAGGGAAGAGAGAAATTGGCAGCTCTCTTCAGGATTTTTCCTCCTCCT-
TEU27497    CCCCCTGCATCATGATAAGGCCGAACATGGTGCATGAAAAGGGAGGCCCCCGGAAGTACGCTTCCGGGAGGAGGGAAGAGAGAAATTGGCAGCTCTCTTCAGGATTTTTCCTCCTCCT-
TBE (e)     ((((((((([((((.............)))))))))......)))).)...(((((((.....)))))))((((((((((((([[[..........]]]))))])).(((......))).))))))
JEU14163    CCCCGTGG-----------AAACAACAACATGCGGCCC---AAGCCCCCTCGAAGCTGTAGAGGAGGTGGAAGGACTAGAGGTTAGAGGAGACCCCGCATTT----GCATCAAA-----
JEU15763    CCCCGTGG-----------AAACAACAACATGCGGCCC---AAGCCCCCTCGAAGCTGTAGAGGAGGTGGAAGGACTAGAGGTTAGAGGAGACCCCGCATTT----GCATCAAA-----
JEVBEICG    CCCCGTGG-----------AAACAACAACATGCGGCCC---AAGCCTCCTCGAAGCTGTAGAGGAGGTGGAAGGACTAGAGGTTAGAGGAGACCCCGCATTT----GCATCAAA-----
JEVCG       CCCCGTGG-----------AAACAACAATATGCGGCCC---AAGCCCCCTCGAAGCTGTAGAGGAGGTGGAAGGACTAGAGGTTAGAGGAGACCCCGCATTT----GCATCAAA-----
JEVCMNAA    CCCCGTGG-----------AAACAACAACATGCGGCCC---AAGCCCCCTCGAAGCTGTAGAGGAGGTGGAAGGACTAGAGGTTAGAGGAGACCCCGCATTT----GCATCAAA-----
JEVSAA      CCCCGTGG-----------AAACAACAATATGCGGCCC---AAGCCCCCTCGAAGCTGTAGAGGAGGTGGAAGGACTAGAGGTTAGAGGAGACCCCGCATTT----GCATCAAA-----
JEVSAV      CCCCGTGG-----------AAACAACAACATGCGGCCC---AAGCCCCCTCGAAGCTGTAGAGGAGGTGGAAGGACTAGAGGTTAGAGGAGACCCCGCATTT----GCATCAAA-----
WNFCG       CCCCGCGT-----------AAAAA--AGTGCACGGCCC---AA-CTTGGCTGAAGCTGTAAGCCAAG-GGAAGGACTAGAGGTTAGAGGAGACCCCGTGCCA-AAAACACCAAAAGAA-
KUNNONCODB  CCCCGCGT-----------TCTGA--AGTGCACGGCCC---AG-CCTGGCTGAAGCTGTAGGTCAGG-GGAAGGACTAGAGGTTAGTGGAGACCCCGTGCCGCAAAACACCACAACAAC
JE          )))))]]] ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~
DEN2CGA     CCCCTCCCTT-ACAAATCGCAGCAACAAT-GGGGGCCC---AAGGTGAGATGAAGCTGTAGTCTCACTGGAAGGACTAGAGGTTAGAGGAGACCCCCCCAA----AACAAAAAA-----
DENCMEMSA   CCCCTCCCTT-ACAAATCGCAGCAACAAT-GGGGGCCC---AAGGCGAGATGAAGCTGTAGTCTCGCTGGAAGGACTAGAGGTTAGAGGAGACCCCCCCGA----AACAAAAAA-----
DENCMEMSB   CCCCTCCCTT-ACAAATCGCAGCAACAAT-GGGGGCCC---AAGGCGAGATGAAGCTGTAGTCTCGCTGGAAGGACTAGAGGTTAGAGGAGACCCCCCCGA----AACAAAAAA-----
DENJAMCG    CCCCTCCCTT-ACAAATCGCAGCAACAAC-GGGGGCCC---AAGGTGAGATGAAGCTGTAGTCTCACTGGAAGGACTAGAGGTTAGAGGAGACCCCCCCAA----AACAAAAAA-----
DENRCG      CCCCTCCCTT-ACAGATCGCAGCAACAAT-GGGGGCCC---AAGGTGAGATGAAGCTGTAGTCTCACTGGAAGGACTAGAGGTTAGAGGAGACCCCCCCAA----AACAAAAAA-----
DEN3CME     CCCCTCCC---ATGACACAACGCAGCAGC-GGGGCCCG---AGCACTGAGGGAAGCTGTACCTCCTTGCAAAGGACTAGAGGTTATAGGAGACCCCCCGCA----AACAA-AAA-----
DEN4STRA    CCCCTCCCATCACTGAAAAACGCAGCAAAAGGGGGCCC---AAGCCAGGAGGAAGCTGTACTCCTGGTGGAAGGACTAGAGGTTAGAGGAGACCCCCCCAA----CACAA-AAA-----
DENT1SEQ    CCCCTCCC---AAAACATAACGCAGCAGC-GGGGCCC----AACACCAGGGGAAGCTGTATCCTGGTGGTAAGGACTAGAGGTTAGAGGAGACCCCCGGCAT---AACAATAAA-----
DEN         ...)))).    ...))]]])~~~~~~~~[ (((((.((       ..[[[[[[[(.........]]]]]]]]...)).......((((.....))))))))))]]~    ~~~~~~~(
FLYF17D1    TCCCACCCC--AGAGTAAAAA-GAACGGAGCCTCCGCT---ACCACCCTCCCACGTGGTGGTAGAAAGACGGGGTCTAGAGGTTAGAGAAGACCCTCCAGGG---AACAAATAGTGG--
FLYF17DG    TCCCACCCC--AGAGTAAAAA-GAACGGAGCCTCCGCT---ACCACCCTCCCACGTGGTGGTAGAAAGACGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---AACAAATAGTGG--
YFU17066    TCCCACCCC--AGAGTAAAAA-GAACGGAGCCTCCGCT---ACCACCCTCCCACGTGGTGGTAGAAAGACGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---AACAAATAGTGG--
YFU17067    TCCCACCCC--AGAGTAAAAA-GAACGGAGCCTCCGCT---ACCACCCTCCCACGTGGTGGTAGAAAGACGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---AACAAATAGTGG--
YFU21055    TCCCACCCC--AGAGTAAAAA-GAACGGAGCCTCCGCT---ACCACCCTCCCACGTGGTGGTAGAAAGACGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---AACAAATAGTGG--
YFU52393    TCCCAACCC--AGAGTGAAAT-GATTGGAGCCTCCGCC---ACCGCCCACCCACGGGGCAGTGGAAAGATGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---A--AATTAGTGG--
YFU52396    TCCCAACCC--AGAGTGAAAT-GATTGGAGCCTCCGCC---ACCGCCCACCCACGGGGCAGTGGAAAGATGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---A--AATTAGTGG--
YFU52399    TCCCACCTC--AGAGTAAAACCAAATGGAGCCTCCGCT---ACCACCTTCCCACGTGGTGGTAGAAAGACGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---AACAAATAGTGG--
YFU52401    TCCCACCCC--AGAGTAAAAA-GAAAGGAGCCTCCGCT---ACCACCTCCCCACGTGGTGGTAGAAAGACGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---AACAATTAGTGG--
YFU52405    TCCCACCTC--AGAGTAAAACCAAACGGAGCCTCCGCT---ACCACCCTTCCACGTGGTGGTAGAAAGATGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---AACAGATAGTGG--
YFU52407    TCCCACCTC--AGAGTAAAACCAAACGGAGCCTCCGCT---ACCACCCTTCCACGTGGTGGTAGAAAGATGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---AACAGATAGTGG--
YFU52411    TCCCACCTC--AGAGTAAAACCAAACGGAGCCTCCGCT---ACCACCTTCCCACGTGGTGGTAGAAAGACGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---AACAGATAGTGG--
YFU52414    TCCCACCCC--AGAGTAAAAA-GAACGGAGCCTCCGCT---ACCACCCTCCCACGTGGTGGTAGAAAGACGGGGTCTAGAGGTCAGAGGAGACCCTCCAGGG---AACAAATAGTGG--
YFU52417    TCCCACCTC--AGAGTAAAACCAAACGGAGCCTCCGCT---ACCACCTTCCCACGTGGTGGTAGAAAGACGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---AACAGATAGTGG--
YFU52420    TCCCACCTC--AGAGTAAAACCAAACGGAGCCTCCGCT---ACCACCTTCCCACGTGGTGGTAGAAAGACGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---AACAGATAGTGG--
YFU52423    TCCCAACCC--AGAGTAAAAA-TGATGGAGCCTCCGCC---ACCACCCACCCGCGGGGTGGTAGAAAGATGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---A--AATTAGTGG--
YFU54798    TCCCACCCC--AGAGTAAAAAAGAAAGGAGCCTCCGCT---ACCACCTCCCCACGTGGTGGTAGAAAGGTGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---AACAATTAGTGG--
YFVCG       TCCCACCCC--AGAGTAAAAA-GAACGGAGCCTCCGCT---ACCACCCTCCCACGTGGTGGTAGAAAGACGGGGTCTAGAGGTTAGAGGAGACCCTCCAGGG---AACAAATAGTGG--
YF          ......)]) )))).......  ...[(((...))))([    (((([((.........))]))))])...~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~  119

TEU27490    ATACCAAATTCCCCCTCAATAGA-GGGGGGGCGGTTCTTGTTCTCCCTG-AGCCACCATCACCCAGACACAGATAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT--
TEU27492    ATACTAAATTCCCCCTCAACAGA-GGGGGGGCGGTTCTTGTTCTCCCTG-AGCCACCATCACCCAGACACAGATAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT--
TEU27493    ATACCAAATTCCCCCTCAATAGA-GGGGGGGCGGTTCTTGTTCTCCCTG-AGCCACCATCACCCAGACACAGATAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT--
TEU27496    ATACAAAATTCCCCCTCAATAGA-GGGGGGGCGGTTCTTGTTCTCCCTG-AGCCACCATCACCCAGACACAGATAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT--
PWARPT      ATACTAAATTTCCCCCAGGAAACTGGGGGGGCGGTTCTTGTTCTCCCTG-AGCCACCACCATCCAGGCACAGATAGCCTGACAAGGAGATGGTGTGTGACTCGGAAAAACACCCGCT--
TBE (fe)    .........((((((((....)) ))))))((((.....((((.(((..( ((.((((([[[[(((((((........)))))).........]]]]]]).))).))))))..)))).)))).
TEU27491    ATACAAAATTCCCCCTCGGTAGA-GGGGGGGCGGTTCTTGTTCTCCCTG-AGCCACCATCACCCAGACACAGATAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT--
TEU27494    ATACAAAATTCCCCCTCGGCAGA-GGGGGGGCGGTTCTTGTTCTCCCTG-AGCCACCATCACCCAGACACAGATAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT--
TEU27495    ATACAAAATTCCCCCTCGGTAGA-GGGGGGGCGGTTCTTGTTCTCCCTG-AGCCACCATCACCCAGACACAGGTAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT--
TEU27497    ATACAAAATTCCCCCTCGGTAGA-GGGGGGGCGGTTCTTGTTCTCCCTG-AGCCACCATCACCCAGACACAGATAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT--
TBE (e)     .)))))))))((((((((....)) ))))))((((.....((((.(((..( ((.((((([[[[(((((((........)))))).........]]]]]]).))).))))))..)))).)))).
JEU14163    -CAGCATATTGACACCTGGGAATAGACTGGGA-GATCTTCTGCTCTATCTCAACATCAGCTACTAGGCACAGAGCGCCGAAGTATGTAGCTGGTAGTG-AGGAAGAACACAGGATCT--
JEU15763    -CAGCATATTGACACCTGGGAATAGACTGGGA-GATCTTCTGCTCTATCTCAACATCAGCTACTAGGCACAGAGCGCCGAAGTATGTAGCTGGTGGTG-AGGAAGAACAC---------
JEVBEICG    -CAGCATATTGACACCTGGGAATAGACTGGGA-GATCTTCTGCTCTATCTCAACATCAGCTACTAGGCACAGAGCGCCGAAGTATGTAGCTTGTGGTG-AGGAAGAACACAGGATCT--
JEVCG       -CAGCATATTGACACCTGGGAATAGACTGGGA-GATCTTCTGCTCTATCTCAACATCAGCTACTAGGCACAGAGCGCCGAAGTATGTAGCTGGTGGTG-AGGAAGAACACAGGATCT--
JEVCMNAA    -CAGCATATTGACACCTGGGAATAGACTGGGA-GATCTTCTGCTCTATCTCAACATCAGCTACTAGGCACAGAGCGCCGAAGTATGTAGCTGGTGGTG-AGGAAGAACAC---------
JEVSAA      -CAGCATATTGACACCTGGGAATAGACTGGGA-GATCTTCTGCTCTATCTCAACATCAGCTACTAGGCACAGAGCGCCGAAGTATGTAGCTGGTGGTG-AGGAAGAACACAGGATCT--
JEVSAV      -CAGCATATTGACACCTGGGAATAGACTGGGA-GATCTTCTGCTCTATCTCAACATCAGCTACTAGGCACAGAGCGCCGAAGTATGTAGCTGGTGGTG-AGGAAGAACACAGGATCT--
WNFCG       ACAGCATATTGACACCTGGGA-TAGACTAGGG-GATCTTCTGCTCTG---CACAACCAGCCACACGGCACAGTGCGCCGACATAGGTGGCTGGTGGTG-CT--AGAACACAGGATCT--
KUNNONCODB  ACAGCATATTGACACCTGGGA-TAGACTAGGA-GATCTTCTGCTCTG---CACAACCAGCCACACGGCACAGTGCGCCGACAATGGTGGCTGGTGGTG-CG--AGAACACAGGATCT--
JE          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~( ((((((.((.(((.(((((.(([.((([((.. (((.........)))........))]]])).))).)) ).))))).)).)))))))
DEN2CGA     -CAGCATATTGACGCTGGGAA--AGACCAGAG-ATCCTGCTGTCTCCT--CAGCATCATT--CCAGGCACAGAACGCCAGAAAATG-GAATGGTGCTG-TTG-AATCAACAGGTTCT--
DENCMEMSA   -CAGCATATTGACGCTGGGAA--AGACCAGAA-ATACTACTATCTACT--CAGCATCATT--CCAGGCACAGAACGCCAGAAAATG-GAATGGTGCTG-TTG-AATCAACAGGTTCT--
DENCMEMSB   -CAGCATATTGACGCTGGGAC--AGACCAGAA-ATCCAACTATCTACT--CAGCATCATT--CTAGGCACAGAACGCCAGAAAATG-GAATGGTGCTG-TTG-AATCAACAGGTTCT--
DENJAMCG    -CAGCATATTGACGCTGGGAA--AGACCAGAG-ATCCTGCTGTCTCCT--CAGCATCATT--CCAGGCACAGAACGCCAGAAAATG-GAATGGTGCTG-TTG-AATCAACAGGTTCT--
DENRCG      -CAGCATATTGACGCTGGGAA--AGACCAGAG-ATCCTGCTGTCTCCT--CAGCATCATT--CCAGGCACAGGACGCCAGAAAATG-GAATGGTGCTG-TTG-AATCAACAGGTTCT--
DEN3CME     -CAGCATATTGACGCTGGGAG--AGACCAGAG-ATCCTGCTGTCTCCT--CAGCATCATT--CCAGGCACAGAACGCCAGAAAATG-GAATGGTGCTG-TTG-AATCAACAGGTTCT--
DEN4STRA    -CAGCATATTGACGCTGGGAA--AGACCAGAG-ATCCTGCTGTCTCTG--CAACATCAAT--CCAGGCACAGAGCGCCGCAAGATG-GATTGGTGTTG-TTG-ATCCAACAGGTTCT--
DENT1SEQ    -CAGCATATTGACGCTGGGAG--AGACCAGAG-ATCCTGCTGTCTCTA--CAGCATCATT--CCAGGCACAGAACGCCAGAAAATG-GAATGGTGCTG-TTG-AATCAACAGGTTCT--
DEN          (((((........((((...     ...)))).......))))))....    (([[(((((([(   ((((((.........))).......)) ))]]]))))]]))   ..(  ((((.....)))).
FLYF17D1    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGAGGTCTGTGAGCACAGTTTGCTCAAGAATAAGCAGACCTTTGGATGACAAACACAAAACCACT
FLYF17DG    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGAGGTCTGTGAGCACAGTTTGCTCAAGAATAAGCAGACCTTTGGATGACAAACACAAAACCACT
YFU17066    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGAGGTCTGTGAGCACAGTTTGCTCAAGAATAAGCAGACCTTTGGATGACAAACACAAAACCACT
YFU17067    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGAGGTCTGTGAGCACAGTTTGCTCAAGAATAAGCAGACCTTTGGATGACAAACACAAAACCACT
YFU21055    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCGGAGGTCTGTGAGCACAGTTTGCTCAAGAATAAGCAGACCTTTGGATGAAAAACACAAAACCACT
YFU52393    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGGGGTCTGTGAGCACAGTTTGCTCTAGAAGAAGCAGACCTTTGGATGAAAAACACAAAACCACT
YFU52396    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGGGGTCTGTGAGCACAGTTTGCTCTAGAAGAAGCAGACCTTTGGATGAAAAACACAAAACCACT
YFU52399    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCGGAGGTCTGTGAGCACAGTTTGCTCAAGAATAAGCAGACCTTTGGATGAAAAACACAAAACCACT
YFU52401    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGAGGTCTGTGAGCACAGTTTGCTCAAGAATAAGCAGACCTTTGGATGAAAAACACAAAACCACT
YFU52405    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGAGGTCTGTGAGCACAGTTTGCTCAAGAATAAGCAGACCTTTGGATGAAAAACACAAAACCAC-
YFU52407    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGAGGTCTGTGAGCACAGTTTGCTCAAGAATAAGCAGACCTTTGGATGAAAAACACAAAACCAC-
YFU52411    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGAGGTCTGTGAGCACAGTCTGCTCAAGAATAAGCAGACCTTTGGATGAAAAACACAAAACCAC-
YFU52414    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGAGGTCTGTGAGCACAGTTTGCTCAAGAATAAGCAGACCTTTGGATGAAAAACACAAAACCACT
YFU52417    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGAGGTCTGTGAGCACAGTCTGCTCAAGAATAAGCAGACCTTTGGATGAAAAACACAAAACCACT
YFU52420    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGAGGTCTGTGAGCACAGTCTGCTCAAGAATAAGCAGACCTTTGGATGAAAAACACAAAACCACT
YFU52423    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGGGGTCTGCGAGCACAGTTTGCTCTAGAAGAAGCAGACCTTTGGATGAAAAACACAAAACCACT
YFU54798    -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGAGGTCTGTGAGCACAGTTTGCTCAAGAATAAGCAGACCTTTGGTTGACAAACACAAAACCACT
YFVCG       -GACCATATTGACGCCAGGGAA-AGACCGGAGTGGTTCTCTGCTTTTC--CTCCAGAGGTCTGTGAGCACAGTTTGCTCAAGAATAAGCAGACCTTTGGATGACAAACACAAAACCACT
YF          ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~(((((   .(((([([((((((([(((((.....))))).........))))))))))))).)))))~~~~~~~~~~~~~   0
```

**FIGURE 7.** CLUSTAL W alignment of all Flavivirus sequences used in this contribution. Predicted secondary structures are displayed in dot-bracket notation: ~ denotes nonpredicted parts; . denotes predicted unpaired positions; ( ) matching parentheses denote base pairs; [ ] matching brackets are used to indicate base pairs that are corroborated by compensatory mutations.

ment is flexible; a good example is the RRE region of HIV, which forms a characteristic five-fingered motif, whereas the primary binding site itself forms a very ill-defined small substructure (see Huynen et al., 1996a).

Huynen and co-workers (1997) recently proposed an entropy measure with similar properties that is also based on the base pairing probabilities and has a somewhat higher sensitivity. A related notion is the "well-determinedness" introduced by Zuker and Jacobson (1995) that is based on the energy differences between the minimum free energy structure and suboptimal folds. We prefer to use a measure explicitly based on the individual base pairing probabilities, such as $d(k)$, because it allows for a much more detailed quantitative interpretation.

### Structural alignment and consensus mountains

Even a high sequence homology of more than 90% does not necessarily imply structural similarity. A statistical survey (Fontana et al., 1993) shows that a small number of mutations is sufficient to completely alter the secondary structure and at 10% sequence difference, the overwhelming majority of sequences will fold into structures that have at most vague similarities. A similar study using the partition function algorithm leads to the same qualitative results (Bonhoeffer et al., 1993). Conservation of (secondary) structure among related sequences should therefore be seen as a consequence of the functional importance of the structure rather than as a consequence of sequence homology.

Comparing structures by comparing their mountain representations was shown to be a very useful technique (Konings & Hogeweg, 1989). Because the generalized mountain represention $m(k)$ defined in Equation 1 is no longer a one-to-one display of one particular structure, it makes sense to compute the average mountain of all structures in a particular group of sequences. This average mountain is always based on a multiple sequence alignment in order to accomodate insertions and deletions. In order to identify flexible parts of a consensus mountain, we shall use the average well-definedness.

The quality of a consensus mountain can be assessed at each position by comparing the slopes $q(k) = m(k) - m(k-1)$ of the different sequences. The slope of the mountain at the position $k$ describes the preferred behavior of nucleotide $k$. If $q(k) = +1$ or $q(k) = -1$, then position $k$ is paired upstream or downstream in all structures, respectively, whereas $q(k) = 0$ indicates a base that is either unpaired or paired in both directions with equal probability. The variance of $q(k)$ then determines the conservedness of a structural element across a sample of sequences.

Conservedness and well-definedness are independent concepts. We have seen that there are indeed well-conserved structural features that are not well-defined at all. The comparative approach presented in this contribution can therefore be used to identify regions that are potentially important in functional terms without being exceptionally stable or well-defined. Our approach thus goes beyond previous attempts to computationally find functional regions in RNA molecules.

## ACKNOWLEDGMENTS

Partial financial support was provided by the Austrian "Fonds zur Förderung der Wissenschaftlichen Forschung," project no. P11065-CHE.

*Received December 27, 1996; returned for revision February 13, 1997; revised manuscript received April 7, 1997*

## REFERENCES

Baudin F, Marquet R, Isel C, Darlix JL, 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.

Blackwell JL, Brinton MA. 1995. BHK cell proteins that bind to the 3′ stem-loop structure of the West Nile virus genome RNA. *J Virol 69*:5650–5658.

Bonhoeffer S, McCaskill JS, Stadler PF, Schuster P. 1993. RNA multistructure landscapes. A study based on temperature dependent partition functions. *Eur Biophys J 22*:13–24.

Brinton MA, Fernandez AV, Dispoto JH. 1986. The 3′-nucleotides of flavivirus genomic RNA form a conserved secondary structure. *Virology 153*:113–121.

Brown EA, Zhang H, Ping LH, Lemon SM. 1992. Secondary structure of the 5′ nontranslated regions of hepatitis C virus and pestivirus genomic RNAs. *Nucleic Acids Res 20*:5041–5045.

Deng R, Brock KV. 1993. 5′ and 3′ untranslated regions of pestivirus genome: Primary and secondary structure analyses. *Nucleic Acids Res 21*:1949–1957.

Fontana W, Konings DAM, Stadler PF, Schuster P. 1993. Statistics of RNA secondary structures. *Biopolymers 33*:1389–1404.

Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, Neilson T, Turner DH. 1986. Improved free-energy parameters for predictions of RNA duplex stability. *Proc Natl Acad Sci USA 83*:9373–9377.

Grange T, Bouloy M, Girard M. 1985. Stable secondary structures at the 3′-end of the genome of yellow fever virus (17D vaccine strain). *FEBS Letts 188*:159–163.

Grüner W, Giegerich R, Strothmann D, Reidys C, Weber J, Hofacker IL, Stadler PF, Schuster P. 1996. Anaylsis of RNA sequence structure maps by exhaustive enumeration. II. Structures of neural networks and shape space covering. *Monath Chem 127*:375–389.

Gultyaev AP, van Batenburg FHD, Pleij CWA. 1995. The computer simulation of RNA folding pathways using a genetic algorithm. *J Mol Biol 250*:37–51.

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

Hahn CS, Hahn YS, Rice CM, Lee E, Dalgarno L, Strauss EG, Strauss JH. 1987. Conserved elements in the 3′ untranslated region of flavivirus RNAs and potential cyclization sequences. *J Mol Biol 198*:33–41.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh Chem 125*:167–188.

Hofacker IL, Huynen MA, Stadler PF, Stolorz PE. 1996. Know discovery in RNA sequence families of HIV using scalable computers. In: *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon*. Portland, Oregon: AAAI Press. pp 20–25.

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

Huynen MA, Gutell R, Konings DAM. 1997. Assessing the reliability of RNA folding using statistical mechanics. *J Mol Biol*. Forthcoming.

Huynen MA, Perelson AS, Viera WA, Stadler PF. 1996a. Base pairing probabilities in a complete HIV-1 RNA. *J Comp Biol 3*:253–274.

Huynen MA, Stadler PF, Fontana W. 1996b. Smoothness within ruggedness: The role of neutrality in adaptation. *Proc Natl Acad Sci USA 93*:397–401.

Jacobson AB, Zuker M. 1993. Structural analysis by energy dot plot of large mRNA. *J Mol Biol 233*:261–269.

Jeager JA, Turner DH, Zuker M. 1989. Improved predictions of secondary structures for RNA. *Proc Natl Acad Sci USA 86*:7706–7710.

Khromykh AA, Westaway EG. 1994. Completion of kunjin virus RNA sequence and recovery of an infectious RNA transcribed from stably cloned full-length cDNA. *J Virol 68*:4580–4588.

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

Lai CJ, Zhao B, Hori H, Bray M. 1991. Infectious RNA transcribed from stably cloned full-length cDNA of dengue type 4 virus. *Proc Natl Acad Sci USA 88*:5139–5143.

Mandl CW, Ecker M, Holzmann H, Kunz C, Heinz FX. 1997. Infectious cDNA clones of tick-borne encephalitis virus European subtype prototypic strain Neudoerfl and high virulence strain Hypr. *J Gen Virol 78*:1049–1057.

Mandl CW, Holzmann H, Kunz C, Heinz FX. 1993. Complete genomic sequence of powassan virus: Evaluation of genetic elements in tick-borne versus mosquito-borne flaviviruses. *Virology 194*:173–184.

McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers 29*:1105–1119.

Monath TP, Heinz FX. 1996. Flaviviruses. In: Fields BN, Knipe DM, Howley PM, Chanock RM, Melnick JL, Monath TP, Roizmann B, Straus SE, eds. *Fields virology, 3rd ed*. Philadelphia: Lippincott-Raven. pp 961–1034.

Olsthoorn RCL, Garde G, Dayhuff T, Atkins JF, 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.

Reidys C, Stadler PF, Schuster P. 1997. Generic properties of combinatory maps: Neural networks of RNA secondary structure. *Bull Math Biol 59*:339–397.

Rice CM. 1996. Flaviviridae: The viruses and their replication. In: Fields BN, Knipe DM, Howley PM, Chanock RM, Melnick JL, Monath TP, Roizmann B, Straus SE, eds. *Fields virology, 3rd ed*. Philadelphia: Lippincott-Raven. pp 931–959.

Rice CM, Grakoui A, Galler R, Chambers TJ. 1989. Transcription of infectious yellow fever RNA from full-length cDNA templates produced by in vitro ligation. *New Biol 1*:285–296.

Schuster P, Fontana W, Stadler PF, Hofacker IL. 1994. From sequences to shapes and back: A case study in RNA secondary structure. *Proc R Soc Lond B 255*:279–284.

Shi PY, Brinton MA, Veal JM, Zhong YY, Wilson WD. 1996. Evidence for the existence of a pseudoknot structure at the 3′ terminus of the flavivirus genomic RNA. *Biochemistry 35*:4222–4230.

Sumiyoshi H, Hoke CH, Trent DW. 1992. Infectious Japanese encephalitis virus RNA can be synthesized from in vitro-ligated cDNA templates. *J Virol 66*:5425–5431.

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

Thompson JD, Higgs DG, Gibson TJ. 1994. CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucleic Acids Res 22*:4673–4680.

van Batenburg FHD, Gultyaev AP, Pleij CWA. 1995. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J Theor Biol 174*:269–280.

Wallner G, Mandl CW, Kunz C, Heinz FX. 1995. The flavivirus 3′-noncoding region: Extensive size heterogeneity independent of evolutionary relationships among strains of tick-borne encephalitis virus. *Virology 213*:169–178.

Walter AE, Turner DH, Kim J, Lyttle MH, Müller P, Mathews DH, Zuker M. 1994. Co-axial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc Natl Acad Sci USA 91*:9218–9222.

Wengler G, Castle E. 1986. Analysis of structural properties which possibly are characteristic for the 3′-terminal sequence of genomic RNA of flaviviruses. *J Gen Virol 67*:1183–1188.

Zuker M. 1989. The use of dynamic programming algorithms in RNA secondary structure prediction. In: Waterman MS, ed. *Mathematical methods for DNA sequences*. Boca Raton, Florida: CRC Press. pp 159–184.

Zuker M, Jacobson AB. 1995. "Well-determined" regions in RNA secondary structure prediction: Analysis of small subunit ribosomal RNA. *Nucleic Acids Res 23*:2791–2798.

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