Secondary Structure of the 3'-Non-Coding Region of Flavivirus Genomes

Comparative Analysis of Base Pairing Probabilities

SUSANNE RAUSCHER[a], CHRISTOPH FLAMM[a],
CHRISTIAN W. MANDL[b], FRANZ X. HEINZ[b], AND PETER F. STADLER[a,c,*]

[a]Institut für Theoretische Chemie, Universität Wien, Austria
[b]Institut für Virologie, Universität Wien, Austria
[c]The Santa Fe Institute, Santa Fe, New Mexico, U.S.A.

[*]Mailing Address: Peter F. Stadler
Inst. f. Theoretische Chemie, Universität Wien
Währingerstr. 17, A-1090 Wien, Austria
Phone: **43 1 40480 665 Fax: **43 1 40480 660
E-Mail: studla@tbi.univie.ac.at or stadler@santafe.edu

Abstract

The prediction of the complete matrix of base pairing probabilities was applied to the 3' non-coding 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 organisation 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 to assess the involvement of these structures in specific processes of the viral life cycle, such as replication and assembly.

1. 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 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: (i) Very little is known about the thermodynamics of pseudoknots, hence there are no reliable energy parameters. (ii) The most efficient folding algorithms, which are based upon dynamic programming, cannot deal with knots or pseudoknots (Zuker & Sankoff, 1984). (iii) Pseudoknots can in many cases be understood as an additional feature that is formed on top of the conventional secondary structure.

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 (van Batenburg *et al.*, 1995a; Gultyaev *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 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 reliable 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 routinely used 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 successfully applied in a recent analysis of the complete genomic RNA of HIV-1. In this work the entire genome was folded as whole on a `CRAY Y-MP` super-computer (Huynen *et al.*, 1996). 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 and 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 (i) the Dengue (DEN) virus types 1 to 4, (ii) Japanese encephalitis (JE), West nile (WN), Kunjin (KUN) and other viruses, (iii) yellow fever (YF) virus, and (iv) tick-borne flaviviruses. The main representativ 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 11kb-long flavivirus genome is taken up by a single long open reading frame that encodes a polyprotein which is co- and post-translationally cleaved by viral and cellular proteases into 10 viral proteins (for review, see Rice (1996)). The flanking non-coding regions (NCR) are believed to contain cis-acting elements important for replication, translation and packaging. In this context, most attention has focussed on the 3'-NCR, which is considerably longer than the only approximately 100nts-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 nucleotides to more than 400 nucleotides and includes in the cases of some strains an internal poly(A) sequence element, whereas the latter is 350 nucleotides long and exhibits a high degree of sequence conservation (Wallner *et al.*, 1995). A secondary structure was proposed for the 3'-terminal 106 nucleotides 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 (Shi *et al.*, 1996; Hahn *et al.*, 1987; Brinton *et al.*, 1986; Wengler & Castle, 1986; Grange *et al.*, 1985) 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 partition function algorithm (McCaskill, 1990) to explore the secondary structures of complete genomic 3'-NCRs of flaviviruses and report:

(i) The 3'-NCRs of flavivirus genomes form conserved secondary structure motifs, but there are considerable differences of RNA folding between the different flavivirus serocomplexes.

(ii) 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.

(iii) 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).

(iv) The core element of the 3'-NCR of TBE virus folds into a highly conserved secondary structure independent of the adjacent variable region.

(v) A particular structural element distinguishes the 3'-NCRs of European subtype TBE virus strains from Far Eastern strains and POW virus.

2. Methods

2.1. Software and Sequences

For our analysis we used the 44 flaviviruses sequences listed in Table 1. From each sequence we extracted the 3'-NCR of the genomic RNA. Using a set of longer sequences with up to 1000nt 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 towards the 5'end. Very long range interaction, spanning more than 1kb, cannot be excluded by this method.

Table 1. 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

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 secondary structures (Hofacker *et al.*, 1994). This public domain software can be obtained by anonymous ftp[1].

A multiple sequence alignment was calculated using `CLUSTALW` (Thompson *et al.*, 1994). It was used to identify compensatory mutations.

2.2. 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 (Figure 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 & 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

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

[1] `ftp://ftp.itc.univie.ac.at/pub/RNA`

counts all base pairs containing[2] 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 which 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* $d(k)$ 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 should 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 (Huynen *et al.*, 1996).

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.

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

2.3. 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 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). Since the generalized mountain represention $m(k)$ defined in equ.(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 consenus 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, while $q(k) = 0$ 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 shall see in the following 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.

3. Results

3.1. 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 the formation of the pseudoknot described by Shi *et al.* (1996). 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 in 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.

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 through 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

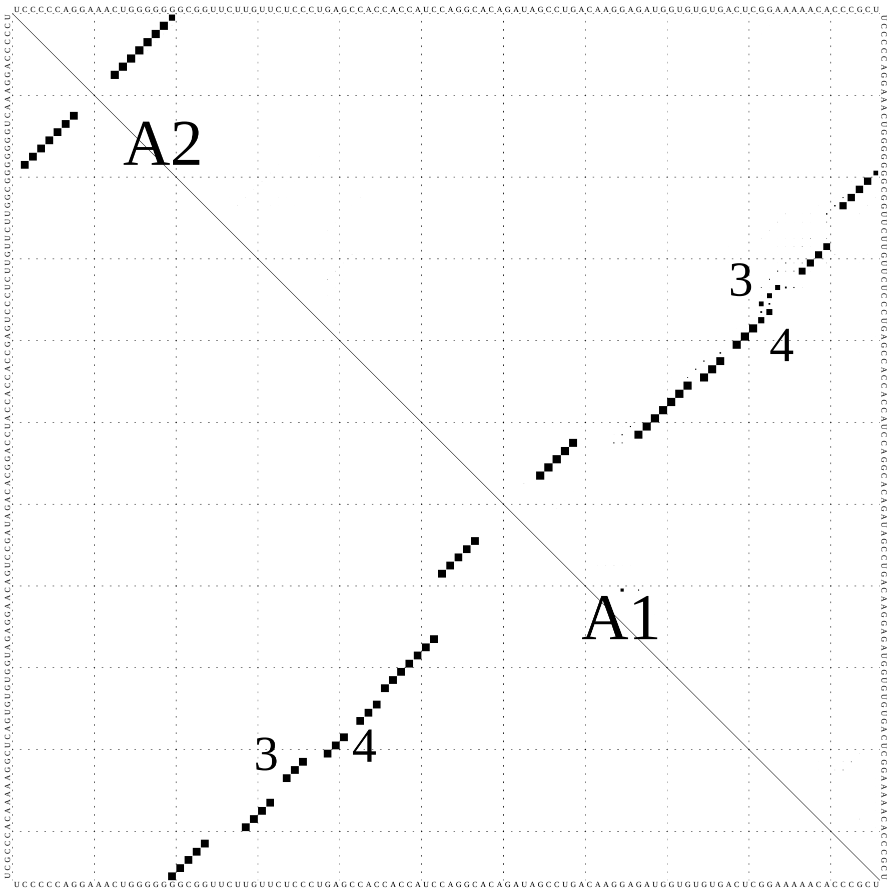

Figure 1: Dot plot of the conserved secondary structure formed by the 106 nucleotides at the 3'terminus of the POW virus sequence. This structure is mostly unambiguous, except for the 3rd and 4th stem: In the minimum free energy structure we have two stems of length 3. Alternatively the 4th stem may 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).

structure of the other subtype. Except for the *Ljubljana* strain of TBE virus the energy differences are well above the thermal energy, see Table 2. 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. While 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,

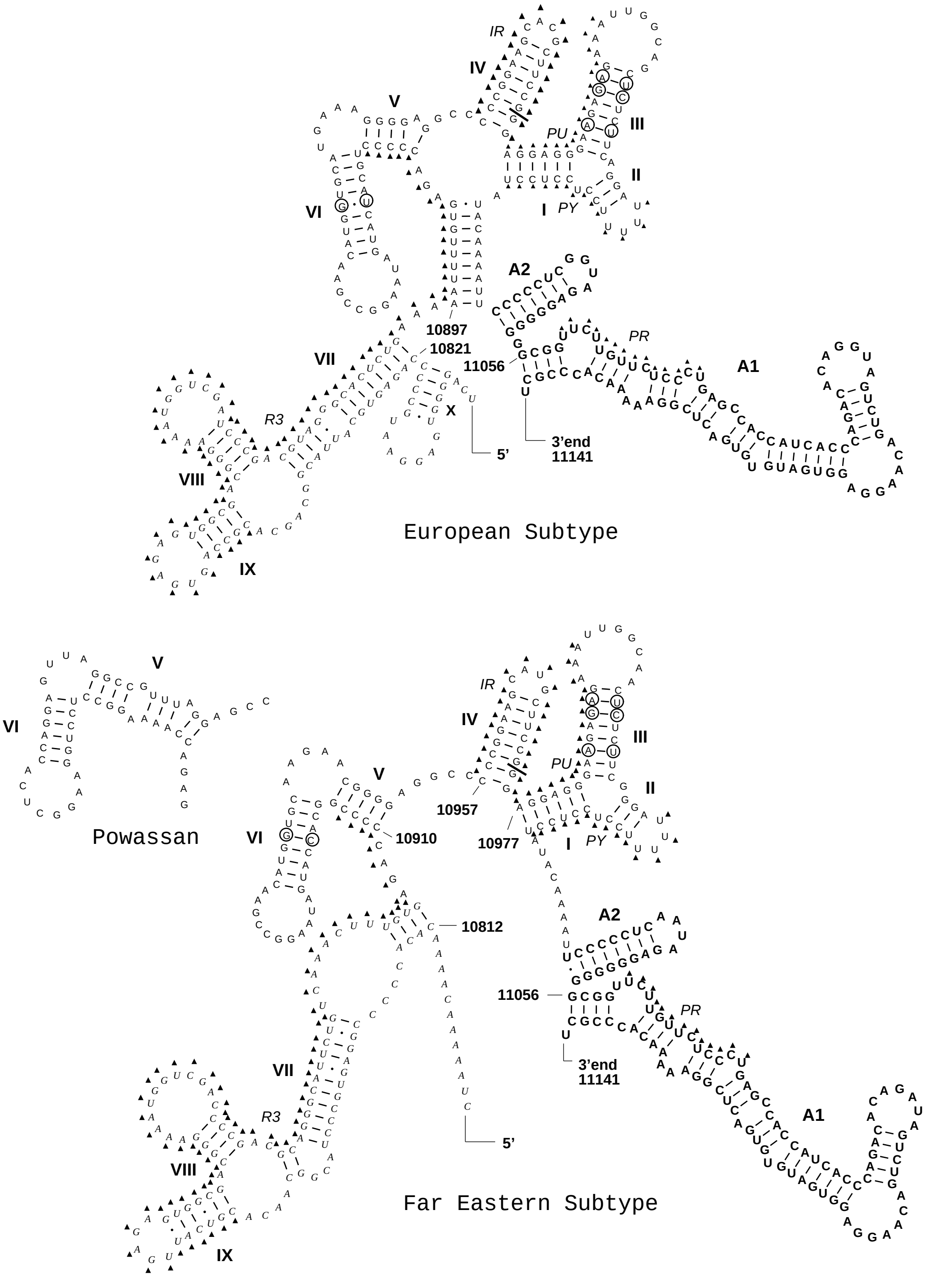

Figure 2: Secondary structure of the 3'-NCR of tick-borne flaviviruses.
The conserved structure at the 3'end is shown in bold. POW virus differs from the other structures in stems V and VI (shown as inset). The locations of compensatory mutations in at least one of the nine sequences are indicated by 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*.

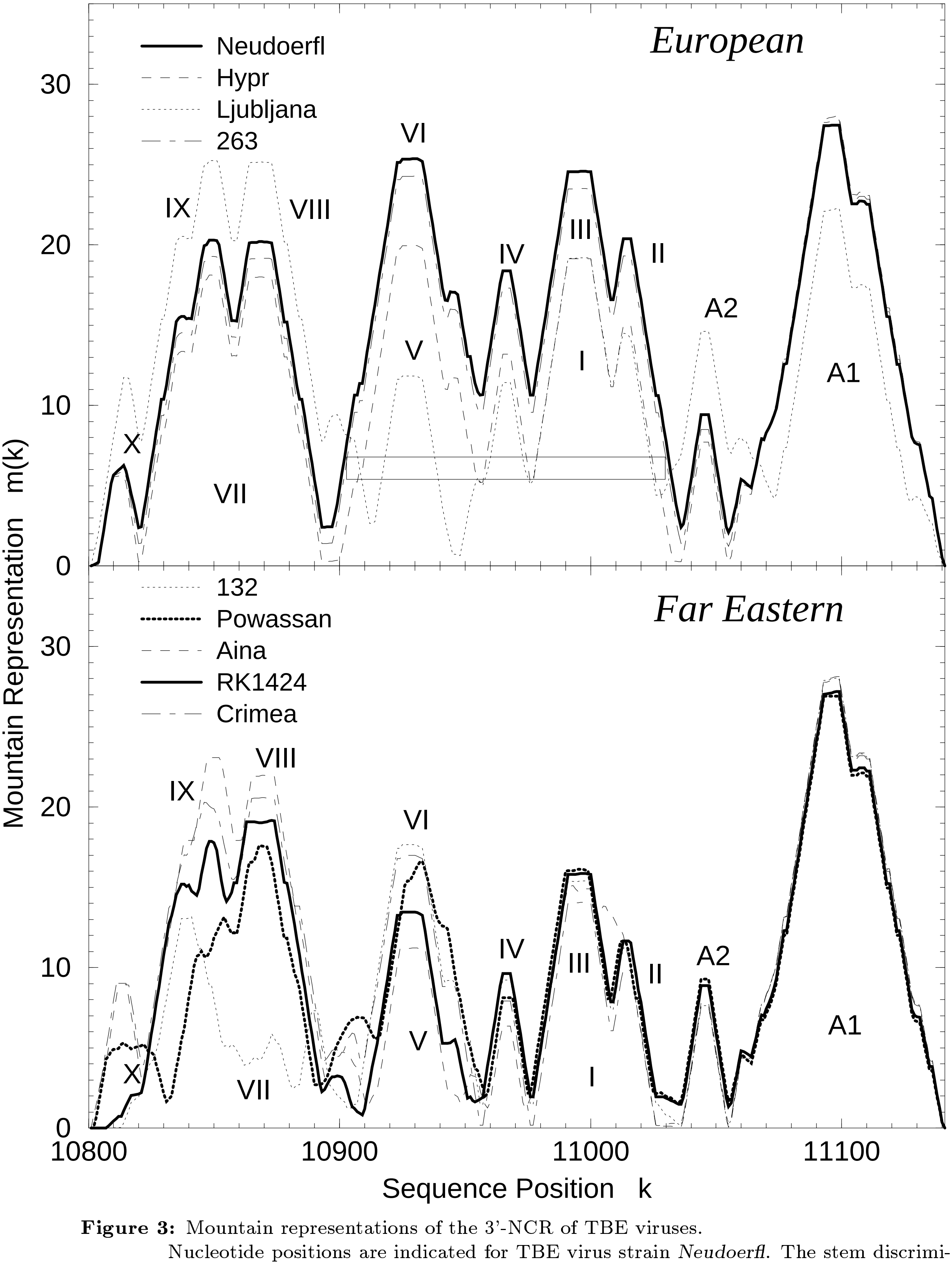

Figure 3: Mountain representations of the 3'-NCR of TBE viruses.
Nucleotide positions are indicated for TBE virus strain *Neudoerfl*. The stem discriminating Far Eastern from European subtype sequences is indicated in the upper plot.

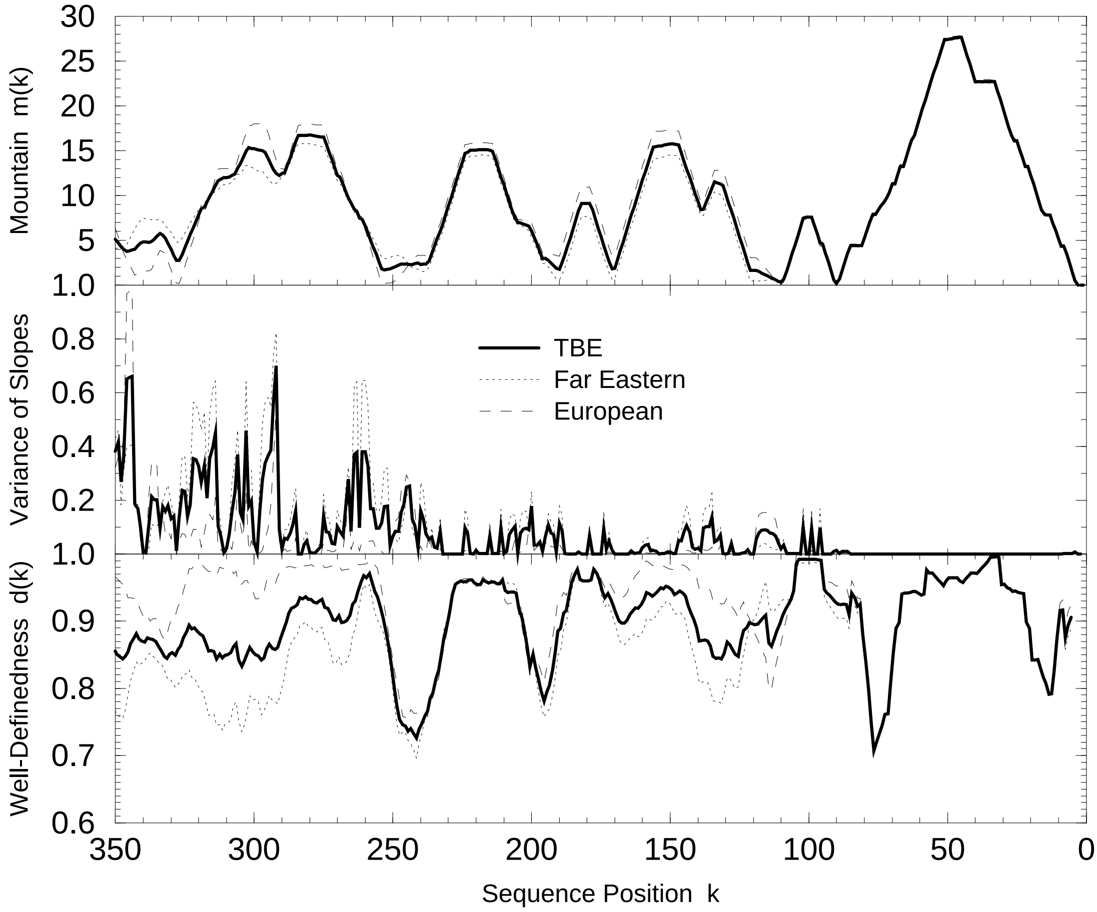

Figure 4: Well-definedness and conservedness of the 3'-NCR secondary structure of TBE viruses. The upper part of the figure compares the mountain representation of the consensus of all TBE virus sequences (Table 1) with the consensus of the Far Eastern and European subtype structures, respectively. Sequence positions are numbers from the 3'end according to a multiple sequence alignment. This alignment inserts about 10 gaps into the original sequences, the range displayed here thus corresponds to the distal 341 nucleotides.

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 (11070 in the strain *Neudoerfl* sequence) corresponds to the location of a small pseudoknotted structure, see text for more 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 Far Eastern subtype. Note that there is not much correlation between well-definedness and the variance of the slopes.

Table 2. Folding Energies (in kcal/mol) for 9 flavivirus sequences. ΔE(far east) and ΔE(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 through VI.

Sequence	GenBank	E(mfe) (kcal/mol)	ΔE(far east) (kcal/mol)	ΔE(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

`AAGGC--A`. The calculated energy difference to the European subtype structure (again allowing for the energetically optimal fold in the V/VI region) is +6.66kcal, more than ten times the thermal energy.

Even further towards the 5'end we find ample evidence for a conserved Y-shaped motif consisting of some 90 nucleotides, see Figure 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 11070 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 oftentimes 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).

Not surprisingly, the consensus mountains of the two subtypes of TBE viruses differ only by the stem enclosing the stems III through 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 through 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.

3.2. 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 (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 to 280 nucleotides away from the 3'end consists of four hairpins. Domain C is located between 350 and 400 nucleotides 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 "spacer" separating region A from the rest of the genome. Beyond some 370 nucleotides 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 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 Figure 6. It occurs at different genomic positions in DEN and JE sequences, and includes, respectively,

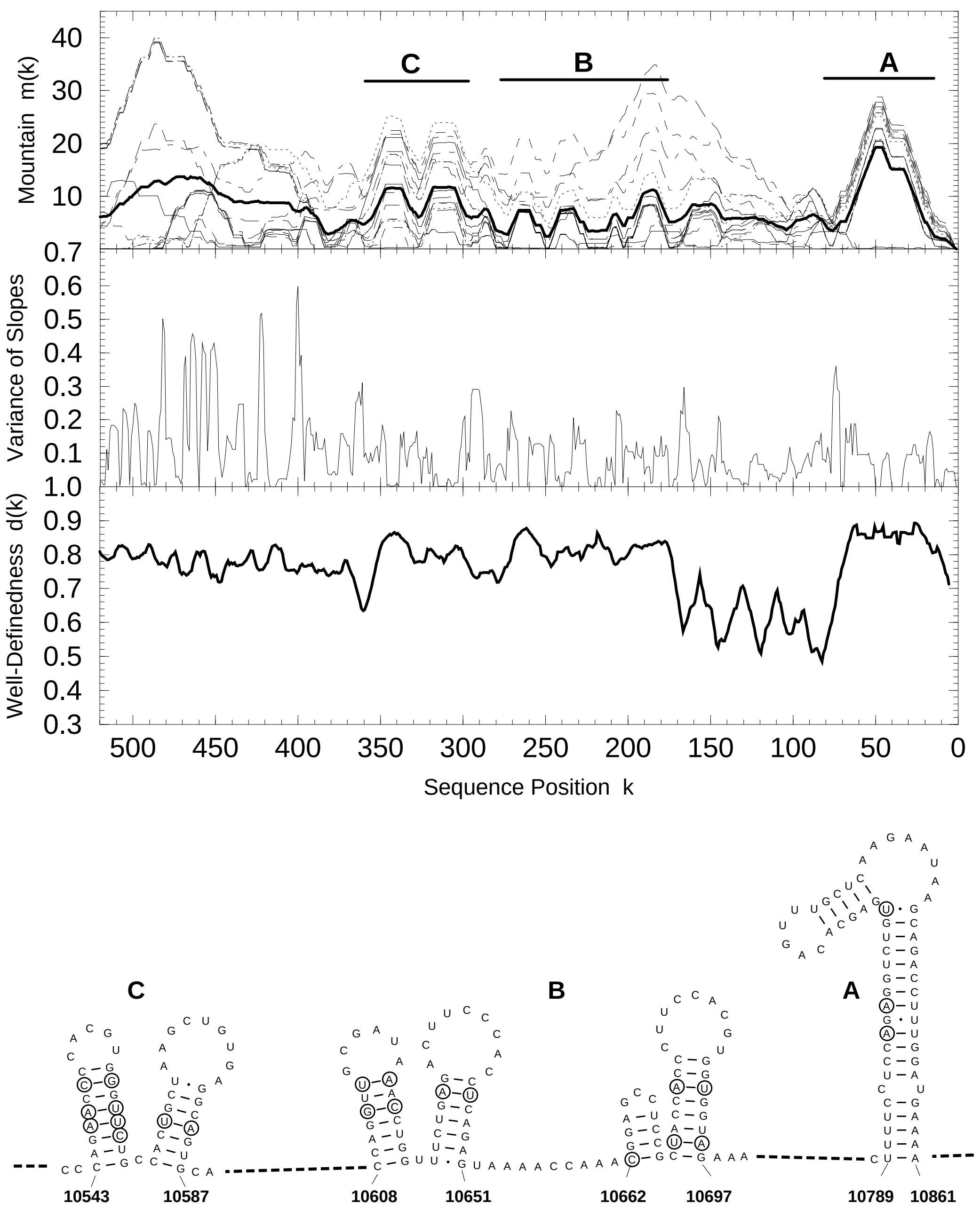

Figure 5: Secondary structure of the YF virus 3'-NCR.

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.

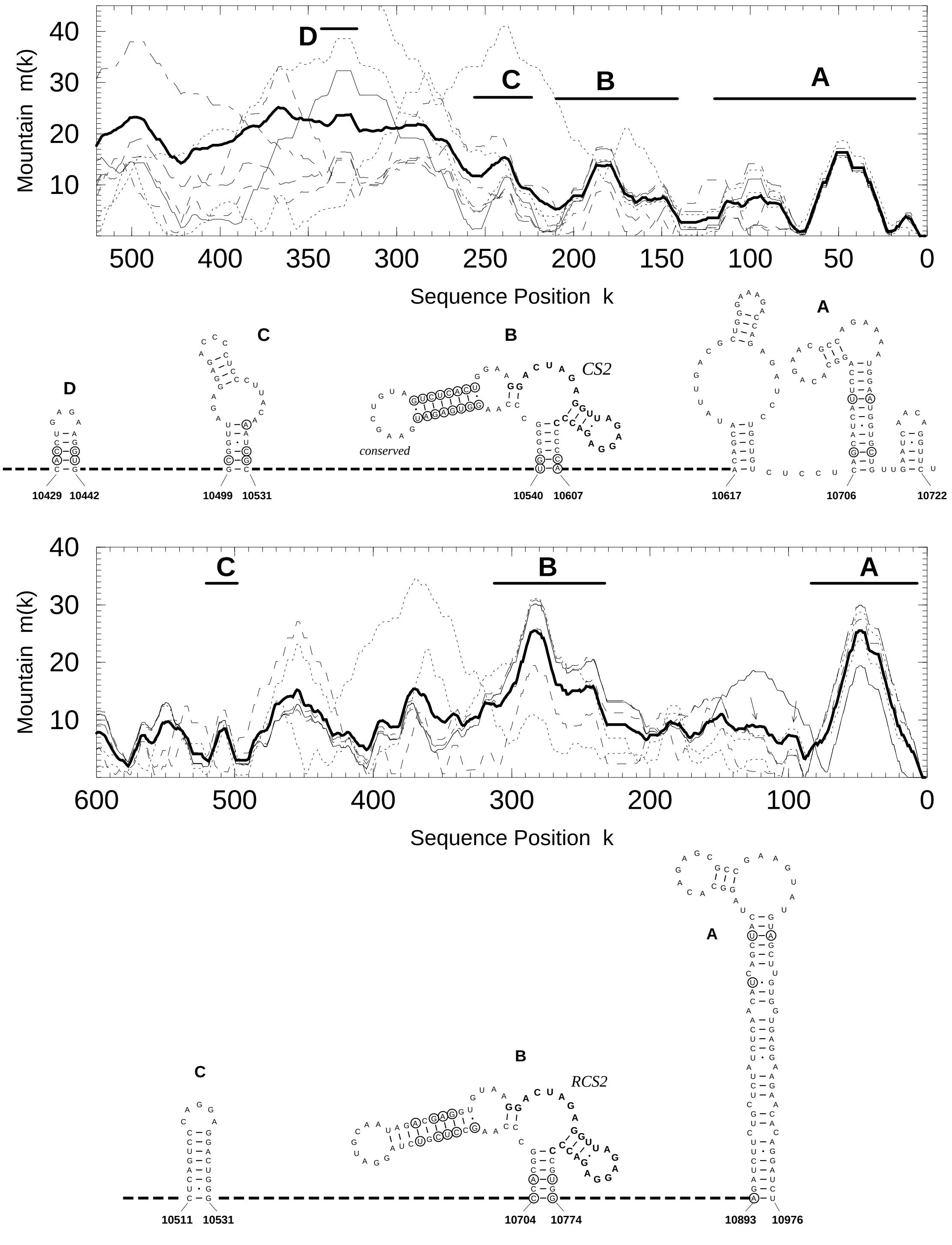

Figure 6: Conserved secondary structures in the DEN (above) and JE (below) serocomplexes. We show the superposition of the mountain representations and the conventional diagrams of conserved consensus structures. Sequences that are consistently unpaired but occur in variable structural contexts are indicated 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.

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.

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

Most secondary structure predictions in the literature have so far only considered the minimum free 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 the access to an excellent tool that allows us to predict the structure and estimate the reliability of the prediction at the same time.

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 may thus be interpreted as flexible parts of the molecule that are possibly of functional importance instead of being an artefact of inaccurate predictions. Ill-defined regions with a high variance between different sequences, on the other hand, suggest structural features are not significant for RNA function. In contrast to earlier approaches (Huynen *et al.*, 1996; Huynen *et al.*, 1997; Zuker & Jacobson, 1995), 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 molcules. The averaging of the mountain representations for the individual sequences amplifies conserved elements only, while 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 nucleotides (Mandl *et al.*, 1993; Shi *et al.*, 1996; Brinton *et al.*, 1986; Grange *et al.*, 1985; Hahn *et al.*, 1987; Wengler & Castle, 1986) 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 Figure 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 strucure 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 nucleotides, 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.

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

Acknowledgements

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

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.

Blackwell, J. L. & Brinton, M. A. (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, J. S., Stadler, P. F., & Schuster, P. (1993). RNA multi-structure landscapes. A study based on temperature dependent partition functions. *Eur.Biophys.J.* **22**, 13–24.

Brinton, M. A., Fernandez, A. V., & Dispoto, J. H. (1986). The 3'-nucleotides of flavivirus genomic RNA form a conserved secondary structure. *Virology,* **153**, 113–121.

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. *Nucleic Acids Res.* **20**, 5041–5045.

Deng, R. & Brock, K. V. (1993). 5' and 3' untranslated regions of pestivirus genome: primary and secondary structure analyses. *Nucleic Acids Res.* **21**, 1949–1957.

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

Freier, S. M., Kierzek, R., Jaeger, J. A., Sugimoto, N., Caruthers, M. H., Neilson, T., & Turner, D. H. (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 Lett.* **188**, 159–163.

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.

Hahn, C. S., Hahn, Y. S., Rice, C. M., Lee, E., Dalgarno, L., Strauss, E. G., & Strauss, J. H. (1987). Conserved elements in the 3' untranslated region of flavivirus RNAs and potential cyclization sequences. *J. Mol. Biol.* **198**, 33–41.

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. (1996). 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* pp. 20–25, Portland, OR: AAAI Press.

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

Huynen, M. A., Gutell, R., & Konings, D. A. M. (1997). Assessing the reliability of RNA folding using statistical mechanics. *J. Mol. Biol.* . in press.

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

Jacobson, A. B. & Zuker, M. (1993). Structural analysis by energy dot plot of large mRNA. *J. Mol. Biol.* **233**, 261–269.

Khromykh, A. A. & Westaway, E. G. (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, 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.

Lai, C.-J., 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, C. W., Ecker, M., Holzmann, H., Kunz, C., & Heinz, F. X. (1997). Infectious cDNA clones of tick-borne encephalitis virus European subtype prototypic strain Neudoerfl and high virulence strain Hypr (in press). Preprint.

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

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

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

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.

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

Rice, C. M., Grakoui, A., Galler, R., & Chambers, T. J. (1989). Transcription of infectious yellow fever RNA from full-length cDNA templates produced by in vitro ligation. *New. Biol.* **1**, 285–296.

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.

Sumiyoshi, H., Hoke, C. H., & Trent, D. W. (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, 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). 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, F. H. D., Gultyaev, A. P., & Pleij, C. W. A. (1995a). An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theor. Biol.* **174**, 269–280.

Wallner, G., Mandl, C. W., Kunz, C., & Heinz, F. X. (1995). The flavivirus 3'-noncoding region: Extensive size heterogeneity independent of evolutionary relationships among strains of tick-borne encephalitis virus. *Virology,* **213**, 169–178.

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: *Mathematical Methods for DNA Sequences*, (Waterman, M. S., ed) pp. 159–184. CRC Press.

Zuker, M. & Jacobson, A. B. (1995). "Well-determined" regions in RNA secondary structure prediction: analysis of small subunit ribosomal RNA. *Nucl. Acids Res.* **23**, 2791–2798.

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