Exploring Secondary Structure Features of RNA Virus Genomes

Diplomarbeit

zur Erlangung des akademischen Grades

Magistra rerum naturalium

an der Formal- und Naturwissenschaftlichen Fakultät

der Universität Wien

eingereicht von

Susanne Rauscher

Institut für Theoretische Chemie

Wien, im Jänner 1997

Abstract

Many RNA viruses exhibit conserved secondary structure motifs whereas their sequences are highly variable. It has been observed that even viruses belonging to the same family show little sequence homology. Many different sequences fold into the same structure. This permits populations of sequences to split and drift apart in sequence space without changing their dominant phenotype. Hence secondary structures may provide additional and more detailed information on the evolution of viruses.

The prediction of the complete matrix of base pairing probabilities was applied to the 3' non-coding region (NCR) of *flavivirus* genomes and an *influenza virus* genome segment. 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.

Influenza viruses are negative-stranded RNA viruses with a segmented genome. Their RNA segments are of inverted complementarity between the termini that could result in a panhandle structure. However, we did not find evidence for conserved secondary structure features in the genomic RNA.

Zusammenfassung

Viele RNA-Viren zeigen konservierte Sekundärstrukturelemente, während sie sich in ihrer Sequenz deutlich unterscheiden. Sogar Viren, die zur selben Familie gehören, weisen nur geringe Sequenzhomologien auf. Viele verschiedene Sequenzen falten in eine Struktur. Das erlaubt Sequenzpopulationen sich zu teilen und im Sequenzraum auseinander zu entwickeln, ohne ihren dominanten Phänotyp zu verändern. Sekundärstrukturen können daher zusätzliche, genauere Informationen zur Erforschung der Evolution von Viren liefern.

Die Vorhersage der kompletten Matrix von Basenpaarungswahrscheinlichkeiten wurde auf die 3'-nicht-kodierenden Regionen von Flavivirusgenomen und auf Influenzavirusgenomsegmente angewandt. Mit dieser Methode können nicht nur wohldefinierte Sekundärstrukturelemente, sondern auch Regionen großer struktureller Flexibilität identifiziert werden.

Flaviviren, von denen viele Vertreter wichtige Humanpathogene sind, haben eine gemeinsame Genomorganisation, während sich ihre RNA-Sequenzen in der funktionell wichtigen 3'-nicht-kodierenden Region signifikant unterscheiden. Wir zeigen, daß Sekundärstrukturen existieren, die in allen *Flaviviren* gleich sind, und auch solche, die für Gruppen von Viren innerhalb des Genus charakteristisch sind. Dieses Ergebnis ist konsistent mit dem etablierten Klassifikationsschema. Die Signifikanz der meisten der vorhergesagten Strukturen wird durch kompensatorische Mutationen untermauert. Da es infektiöse Klone für einige *Flaviviren* gibt, kann man die Bedeutung dieser Strukturen für bestimmte Prozesse des viralen Lebenszykluses, wie z. B. Replikation und Assembly, untersuchen.

Influenzaviren sind negativ-strängige Viren mit einem segmentierten Genom. Die einzelnen Segmente sind an den Enden invers komplementär und können deshalb eine Panhandle-Struktur bilden. Wir haben jedoch keinen Hinweis auf konservierte Sekundärstrukturelemente in ihren RNA-Genomen gefunden.

1. Introduction

Information in biology has a quality that distinguishes it from information in chemistry and physics. It comes in encoded form and it is processed in a way that is closely related to information in technology and computer science. Biological information is essentially stored in genotypes and transfered to future generations through inheritance, and less directly through epigenetic processes. Cellular metabolism is interpreted straightforwardly as information processing. Any comprehensive understanding of biological phenomena requires an interpretation in evolutionary terms. "Nothing in biology makes sense exept in the light of evolution", as Theodosius Dobzhansky pointed out [14]. Understanding the complexity of biological systems is thus always incomplete if nothing is known about the origin [69].

In the case of viruses, on present evidence it seems probable that different groups of present-day viruses have originated in different ways. Some of the very large DNA viruses infecting animals are probably descendend by a degenerative process from very simple cellular parasites. Other, smaller viruses, especially those with RNA genomes, most probably evolved from host genes via transposons or by other means. Some present-day RNA viruses or parts of them might be direct descendants from a prebiotic RNA world [61], a hypothetical, early pre-DNA environment in which RNA replicated by itself and the genetic code and protein synthesis arose [67, 31]. Viruses reproduce at high rate and can adapt to changes by mutation but are completely dependent on the metabolic activity of their host. Their high mutability makes viruses excellent objects for studying evolution.

Even though viral genome sequences are highly variable, their secondary structure seems to be more conserved. Former studies showed that secondary structure elements are conserved if they are of functional importance [52, 47, 36]. Many different sequences fold into the same structure. This permits populations of sequences to split and drift apart in sequence space without changing their dominant phenotype [70]. It has been observed that even viruses belonging to the same family show little sequence homology. Hence studying evolution and origin of viruses should focus not only on sequence but also on secondary structure similarities.

In this thesis we examined the secondary structure motifs in *flaviviruses* and *influenza viruses*. *Flaviviruses* have a common genomic organisation but exhibit a significant degree of RNA sequence diversity in the functionally important 3' non-coding regions (NCRs). 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, for further information refer to Chapter 4. *Influenza viruses* are viruses with a segmented, negative-stranded RNA genome. We found no evidence for conserved secondary structure elements within the *influenza virus* genome. More detailed information on *influenza viruses* is given in Chapter 5.

Many RNA viruses exhibit strongly conserved sequences and/or secondary structure motifs in their terminal NCRs. It is very likely that such highly conserved elements are crucial for the interaction with viral and cellular factors during viral replication. So far a number of significant secondary structures, that is, the pattern of Watson-Crick and GU base pairs, have been determined that play a role during the various stages of the viral life cycle. Elucidation of all the significant secondary structures is necessary for the understanding of the molecular biology of a virus.

Secondary structures account for the major part of the free energy of the spatial structures of nucleic acids. Knots and pseudoknots are usually excluded from the definition of secondary structure. For RNA molecules, if one is willing to accept secondary structures, i.e., base pairing patterns, as a suitable (coarse grained) description of the structures, one can actually compute the structure of minimum free energy (MFE) for, in principle, arbitrary sequences [42, 62, 92, 93]. These algorithms are based on a simple thermodynamic model of RNA (secondary) structures, for which the majority of parameters have been measured directly on small oligonucleotides [29]. The simplicity of the energy model and the relatively small number of contributions in a given sequence allow this approach to be applied successfully. Secondary structures are often well conserved and hence evolve more slowly than the underlying sequences. This fact for instance allows the construction of structural models from the comparison of sequence data [36]. A detailed analysis of RNA structures might therefore contribute to a more comprehensive understanding of viral evolution and the phylogenetic relationships among seemingly unrelated viruses.

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 MFE 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. 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. Almost all secondary structure predictions in the literature have so far only considered the MFE structure and/or a fairly small sample of suboptimal structures, as provided, e.g., by Zuker's `mfold` package [92, 90, 89]. The large size of some genomes implies that there is a huge number of low energy states. In addition, knowledge of the uncertainty of the predicted structure in different regions is most useful for a meaningful interpretation of the data. Hence the computation of a complete set of structures that is suitable for describing the actual structure is too costly. McCaskill's partition function approach [62], which allows for an exact computation of the complete matrix of all base pairing probabilities p_{ij} in thermodynamic equilibrium. This method provides more 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 a HIV-1 virus [48]. 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.

All computations reported in this work were performed using the `Vienna RNA package`, which contains a variety of programs for the computation and comparison of RNA secondary structures [42]. This public domain software can be obtained by anonymous ftp [41]. The `Vienna RNA package` is suitable as a routine tool allowing for a comparative analysis of the complete set of available sequence data for RNA viruses. In this thesis we apply McCaskill's partition function algorithm [62] to explore the secondary structures of RNA virus genomes 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 [59, 81]. 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 mosquito-borne *flaviviruses* by Brinton and coworkers [71].

(iv) The core element of the 3'NCR of *tick-borne encephalitis (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 *Powassan (POW) virus*.

(vi) We did not find conserved secondary structure features in *influenza virus* genome segment 4, and could not confirm its panhandle structure.

2. Viruses

2.1. General Properties

Viruses are defined as a set of one or more nucleic acid template molecules, normally encased in a protective coat or coats of protein or lipoprotein, that is able to organize its own replication only within suitable host cells. Within such cells, virus replication is

(i) dependent on the host's protein synthesizing machinery,

(ii) organized from pools of the required materials rather than by binary fission,

(iii) located at sites that are not separated from the host cell contents by a lipoprotein bilayer membrane, and

(iv) continually giving rise to variants through various kinds of change in the viral nucleic acid.

They are the most efficient of the self-reproducing intracellular parasites. Viruses are unable to generate metabolic energy or to synthesize proteins. They differ from cells in having either DNA or RNA. The complete extracellular form of a virus is called *virion* (or virus particle). In a virion, the viral nucleic acid is covered by a protein capsid, which protects it from enzymatic attack and mechanical breakage and delivers it to a susceptible host. In some of the more complex animal viruses, the capsid itself is surrounded by an envelope containing membrane lipids and glycoproteins [76]. To be identified positively as a virus, an agent must normally be shown to be transmissible.

The structure and replication of viruses have the following features [61]:

1. The nucleic acid may be DNA or RNA and single- or double-stranded. If the nucleic acid is single-stranded it may be of positive or negative sense. (Positive sense has the sequence that would be used in an mRNA for translation to give a viral-coded protein.)

2. The mature virus particle may contain polynucleotides other than the genomic nucleic acid.

3. Where the genetic material consists of more than one nucleic acid molecule, each may be housed in a separate particle or all may be located in one particle.

4. The genomes of viruses vary widely in size, encoding between 1 and about 250 proteins. The viral-coded proteins may have functions in virus replication, in virus movement from cell to cell, in virus structure, and in transmission by vertebrates or fungi.

5. Viruses undergo genetic change. Point mutations occur with high frequency as a result of nucleotide changes brought about by errors in the copying process during genome replication. Other kinds of genetic change may be due to recombination, reassortment of genome pieces, loss of genetic material, or acquisition of nucleotide sequences from unrelated viruses or the host genome.

6. Enzymes specified by the viral genome may be present in the virus particle. Most of these enzymes are concerned with nucleic acid synthesis.

7. Replication of many viruses takes place in distinctive virus-induced regions of the cell, known as viroplasms.

8. With certain nonviral nucleic acid molecules some viruses share the property of integration into host-cell genomes and translocation from one integration site to another.

9. A few viruses require the presence of another virus for their replication.

A virus species might be defined simply as a collection of strains with similar properties. Sometimes it is hard to decide whether two similar virus isolates are identical or not; or whether two isolates are different viruses or strains of the same virus. Two kinds of properties are available for the recognition and delineation of virus strains — structural criteria based on the properties of the virus particle itself and its components, and biological criteria based on various interactions between the virus and its hosts and vectors. Serological properties are based on the structure of the viral protein or proteins.

2.2. Speculations on the Origins of Viruses

Although much relevant information has become available from studies on the structure and replication of viruses and on the molecular biology of cells, the origin and evolution of viruses is only now beginning to emerge from the realm of speculation. Nevertheless, the topic is one of general interest, and one that will be relevant to the problem of classification. There is no compelling reason to suppose that all viruses arose in the same way. Furthermore, it is possible that viruses that originated in one major group of organisms may now exist as agents infecting another group [61]. Three theories are being discussed:

A. Descendants of Primitve Precellular Life-forms

The viruses as we know them today are highly developed parasites. They use the same genetic code as cellular organisms and are dependent for their protein synthesis on ribosomes, tRNAs, and associated enzymes provided by the host cell. Amino acids occur in viral proteins with the same frequency as they do in the globular proteins of other groups of organisms. Thus, although virus particles are relatively simple in structure, there is nothing known about their chemistry or mode of replication to show that they are more primitive in an evolutionary sense than cellular organisms. Nevertheless, current views on the origin of life open up the possibility that some RNA viruses may have descended from prebiotic polymers.

Until fairly recently, precellular life was considered to involve both proteins and RNAs. However, following the discovery of introns in eukaryotic genes, it was found that the splicing out of introns from RNA transcripts and ligation of the exons could take place in absence of protein. Furthermore, the ligation reaction can join exons from different RNA transcripts [87]. Thus self-inserting introns can create transposons to shuttle exons. This provides RNA with a vital evolutionary capacity it would otherwise lack — the ability to produce new combinations of genes [32]. These observations, and reports of other enzymatic activities for intron RNAs, led to propose an evolutionary scheme in which RNAs were the only polymers in the prebiotic stage of evolution [10]. The tRNA-like structures found at the 3' terminus of some RNA plant viruses are proposed to be molecular fossils

from the original RNA world [85], retained because of an essential function. A *Tetrahymena* ribozyme can splice together multiple oligonucleotides aligned on a template strand to give a fully complementary product strand, thus demonstrating the feasibility of RNA-catalyzed RNA replication [17]. On this theory, RNA viruses might represent greatly modified descendants of prebiotic RNAs that later parasitized the earliest cells.

B. Development from Normal Constituents of Cells

It has often been considered that viruses may arise from some cell constituent that has escaped from the normal control mechanisms and become a self-replicating entity. A normal cell component in one organism may develop into a virus when introduced into the cells of another. Because viruses carry genetic information it has been proposed that they arose as host genes that escaped from the control mechanisms of the cell. They are considered then to have developed means of being transfered efficiently to other host cells and of replicating independently of cell division [61].

C. Origin by Degeneration from Cells

The idea that viruses are extremly degenerate parasitic forms that have evolved from cellular organisms has lost favor in recent years, mainly because of the similarities between the behaviour of certain viruses and plasmids. The properties that distinguish all viruses from cells reduce to three:

(i) lack of a complete membrane separating the virus replication site from the host cytoplasm (or nucleoplasm for some eukaryote viruses);

(ii) use of host protein-synthesizing machinery by viruses; and

(iii) binary fission in cell reproduction but not with viruses.

If any viruses did arise by degeneration from cells, these three properties may have been lost at about the same stage. Absence of a bounding membrane during replication would allow the parasite to become dependent on the host protein-synthesizing machinery. It is believed that the large size of the poxviruses, their

complex structure, including many enzymes within the virus particle, and their ability to replicate in the cytoplasm independently of host nuclear functions make it plausible that these viruses arose from cells [21]. They might form the most degenerate member of the series bacterium → rickettsia → chlamydia → poxvirus. If one family of viruses originated in this way, others could have done so too.

The true origin of viruses may never be known for certain unless

(i) they arose in a modular way from host genes;

(ii) amino acid sequences have remained sufficiently conserved in both viruses and hosts; and

(iii) in due course sufficient cellular gene modules are sequenced, especially from lower organisms, to enable convincing identification of the modules from which particular viruses arose.

2.3. Evolution

A. Mechanisms for Virus Evolution

The molecular mechanisms underlying genetic variation in viruses [88] include

(i) mutation due to a base change,

(ii) additions or deletion of bases,

(iii) deletions or duplications of nucleic acid sequences,

(iv) genetic recombination,

(v) reassortment among genome pieces in viruses where the genetic material is in several pieces of nucleic acid, and

(vi) the acquisition of exogenous host or viral genes.

We can envisage viral evolution proceeding in both a micro- and a macro- manner. In microevolution, existing viral genes accumulate small changes by mechanisms

such as nucleotide substitutions, additions, or deletions. In macroevolution of a virus, a sudden major change may take place by recombination with a related virus, by duplication of an existing gene, or by acquisition of a gene from the host or an unrelated virus, processes that have been termed modular evolution [68]. In summary, multiple mechanisms exist for virus evolution, both on a micro and a macro scale with respect to the size of the individual changes involved.

B. Evidence for Virus Evolution

No fossil viruses have yet been discovered. In the meantime, evidence for virus evolution must come from the study of present-day viruses in present-day hosts.

1. General properties of viruses within families and between closely related groups

On present evidence there is a lack of intermediate types between most of the virus families and groups delineated by ICTV [64]. The close similarities in particle morphology, genome strategy, and the three-dimensional structure of proteins leave no doubt that individual viruses within some groups delineated by the ICTV, had a common ancestor at some time in the past. In some instances, the evidence suggests that, from an evolutionary point of view, two or more of the existing ICTV groups of plant viruses would be better placed in a single family [61].

2. Nucleotide and Amino Acid Sequences

Knowledge of nucleotide sequences in viral genomes, and the corresponding sequences of amino acids in the encoded proteins, has provided confirmation of the evolutionary basis for most of the families and groups of viruses delineated by the ICTV. Sequence data can be used to estimate degrees of evolutionary relationship, to develop "trees" indicating possible lines of evolution for viruses within a family, and to discover unexpected relationships [39]. It should be remembered that there are significant difficulties in deriving and interpreting trees constructed from sequence data [83]. Sequence similarity between two genes does not necessarily indicate evolutionary relationship (homology) [75]. Without other evidence it may be impossible to establish whether sequence similarity between two genes is due to a common evolutionary origin or to convergence. Amino acid sequence similarities that are sufficient to lead to a serological cross-reaction may sometimes arise by chance [13].

3. Examples of Evolutionary Change

In various experiments both *in vitro* and *in vivo* with derived or constructed viral RNAs, there has been a trend towards survival of shorter RNAs [30]. An effect of selection pressure is that functional viral genes will be retained only if they are needed for survival of the virus. An example of the effect of selection pressure provides the reversion of viroid point mutations to wild type [61].

4. Rates of Evolution

The rate of point mutation for RNA viruses has been estimated to be approximately 10^{-3}–10^{-4} per nucleotide per round of replication with some variation between different viruses, which is thought to be due to the lack of any error-correcting mechanism in RNA-dependent RNA polymerases (which is three to four orders of magnitude higher than the rate for DNA polymerases) [46]. This fact, coupled to the very high rate of virus replication that is possible, can potentially lead to rapid change in viral genomes [74].

A virus culture is an extremely heterogeneous population of related variants (quasispecies) [19, 46, 16, 15]. All the RNA genomes that have been examined have been found to exist not as a single nucleotide sequence but as a distribution of sequence variants around a consensus sequence. Selection represses non-viable sequences and favors those which are "alive". The concept of quasispecies explains the rapid changes observed in RNA viruses. The high frequency of mutation provides RNA viruses with the variability necessary to adapt to new hosts and environments when transmitted as large populations.

Most of the variants in a culture of a particular virus strain will normally consist of base substitutions at various sites perhaps with some deletions or additions of nucleotides. However, more substantial variation can occur when one or more sequences of a multipartite genome are not under selection pressure [25, 48]. Variants may arise quite rapidly, and these could lead to confusion in strain identification. This potential for rapid change must be kept in mind when considering nucleic acid differences as criteria for recognizing virus strains. It is very difficult to relate mutation rates to the actual rates of change in viruses that might be occuring in the field at present, or over past evolutionary time. The reasons for this include the following:

(i) *Selection Pressure by Hosts*
Viral genomes and gene products must interact in highly specific ways with host macromolecules during virus replication and movement. These host molecules, changing at a rate that is slow compared with the potential for change in a virus, will act as a brake on virus evolution. This led to the proposition that RNA viral proteins evolve with an amino acid substitution rate that is similar to that of their host [54].

(ii) *Uneven Rates of Change in Different Parts of a Viral Genome*
Noncoding regions of viral genomes, particularly at the 5' and 3' termini, which function as recognition sites in viral RNA translation and replication, may be highly conserved in the members of a virus family or group. On the other hand, in viruses with multipartite genomes, one genome segment may be conserved and the other highly variable depending on the functional requirement of the gene product.

(iii) *Variation in Rates of Change over a Time Period*
The environment that dictates the selection pressure on the replication and movement of a virus within a host consists almost entirely of the internal milieu of this host. Other selection pressures on survival involve transmission from one host to another by vectors. It is possible that a switch to a new host species may induce rapid evolution of a virus over quite a short period. Recombination between viral genomes with mutations in different parts of the genome may speed this process [20]. One way in which a virus may gain foothold in a new host species is through coinfection with another virus that normally infects that species.

5. Coevolution of Viruses, Hosts, and Invertebrate Vectors

Fahrenholtz' rule postulates that parasites and their hosts speciate in synchrony [18]. Thus there is a prediction that phylogenetic trees of parasites and their hosts should be topologically identical. In view of the known wide host ranges of many present-day viruses, it is not to be expected that Fahrenholtz' rule will be followed closely for viruses and their hosts. Nevertheless, it is now widely accepted that viruses have had a long evolutionary history and have coevolved with their host organisms.

2.4. Taxonomic Classification and Phylogeny

The earliest efforts to classify viruses were based upon perceived common pathogenic properties, common organ tropisms, and common ecological and transmission characteristics. For example, viruses that share the pathogenic property of causing hepatitis would have been brought together as *the hepatitis viruses*. The evidence of the structure and composition of virions proposed the grouping of viruses on the basis of shared virion properties. In the 1950s and 1960s, there was an explosion in the discovery of new viruses. The result was confusion over competing, and conflicting schemes of virus classification and nomenclature.

In 1966 the International Committee on Nomenclature of Viruses (ICNV) was established. The ICNV became the International Committee on Taxonomy of Viruses (ICTV) in 1973. In the universal scheme developed by the ICTV, virion characteristics are considered and weighted as criteria for making divisions into one order *Mononegavirales* (contains all viruses with a negative stranded ssRNA genome), families, in some cases subfamilies, and genera. In each case the relative hierarchy and weight assigned to each characteristic used in defining taxa is set arbitrarily and is still influenced by prejudgments of relationships that "we would like to believe (from the evolutionary standpoint), but are unable to prove". The Sixth Report of the ICTV [64], records an universial taxonomy scheme comprising one order, 71 families, 9 subfamilies, and 164 genera, including 24 floating genera, and more than 3,600 virus species. The system still contains hundreds of unassigned viruses, largely because of a lack of data. A listing of RNA virus families and floating genera can be found in Tables 1 to 4.

Table 1. The Positive Stranded ssRNA Viruses

Family	**Floating Genus**	**Nucleic Acid Configuration**	**Genomic Size** [kb]	**Host**[a]
	Arterivirus	1 + linear	13	V
Astroviridae		1 + linear	6.8 - 7.9	V
Barnaviridae		1 + linear	4.4	F
Caliciviridae		1 + linear	7.4 - 7.7	V
	Capillovirus	1 + linear	6.5	P
	Carlavirus	1 + linear	7.4 - 7.7	P
	Closterovirus	1 + linear	15.5	P
Coronaviridae		1 + linear	20 - 30	V
Flaviviridae		1 + linear	9.5 - 12.5	V,I
Leviviridae		1 + linear	3.5 - 4.3	B
	Luteovirus	1 + linear	5.6 - 5.9	P
	Machlomovirus	1 + linear	4.4	P
	Marafivirus	1 + linear	8 - 8.9	P
	Necrovirus	1 + linear	3.8	P
Picornaviridae		1 + linear	7 - 8.5	V,I
	Potexvirus	1 + linear	5.8 - 7	P
Potyviridae		1 + linear	8.5 - 10	P
Sequiviridae		1 + linear	9 - 12	P
	Sobemovirus	1 + linear	4.2	P
	Tobamovirus	1 + linear	6.4	P
Togaviridae		1 + linear	9.7 - 11.8	V,I
Tombusviridae		1 + linear	4 - 4.7	P
	Trichovirus	1 + linear	6.3 - 7.6	P
	Tymovirus	1 + linear	6.3	P
	Umbravirus	1 + linear	4.5	P
Tetraviridae		1,2 + linear	5.5	I
Comoviridae		2 + linear	3.5 - 8.4	P
	Dianthovirus	2 + linear	1.4/3.9	P
	Enamovirus	2 + linear	4.3/5.7	P
	Furovirus	2 + linear	0.6 - 7.1	P
	Idaeovirus	2 + linear	1 - 5.5	P
Nodaviridae		2 + linear	1.4 - 3.3	I
	Tobravirus	2 + linear	1.8 - 6.8	P
Bromoviridae		3 + linear	0.9 - 3.6	P
	Hordeivirus	3 + linear	2.5 - 3.8	P

[a] A: algae; B: bacteria; F: fungi; I: invertebrates; M: mycoplasma; P: plants; Pr: protozoa; V: vertebrates

Table 2. The Negative Stranded ssRNA Viruses

Family	**Nucleic Acid Configuration**	**Genomic Size** [kb]	**Host**[a]
Bunyaviridae	4-5 – linear	11 - 20	V,I,P
Filoviridae	1 – linear	19.1	V
Paramyxoviridae	1 – linear	15.2 - 15.9	V
Rhabdoviridae	1 – linear	11 - 15	V,I,P
Arenaviridae	2 – linear	3.3 - 4.2	V
Orthomyxoviridae	8 – linear	10.0 - 13.6	V

Table 3. The dsRNA Viruses

Family	**Nucleic Acid Configuration**	**Genomic Size** [kbp]	**Host**[a]
Totiviridae	1 + linear	4.6 - 7.0	F, Pr
Hypoviridae	1 linear	10 - 13	F
Birnaviridae	2 linear	2.8 - 3.1	V,I
Partitiviridae	2 linear	1.4 - 3.0	F,P
Cystoviridae	3 linear	2.9 - 6.4	B
Reoviridae	10-12 linear	36 - 60	V,I,P

Table 4. The DNA and RNA Reverse Transcribing Viruses

Family	**Floating Genus**	**Nucleic Acid Type / Configuration**	**Genomic Size** [kb/kbp]	**Host**[a]
Hepadnaviridae		ssDNA / 1 circular	3 - 3.3	V
	Badnavirus	dsDNA / 1 circular	7.5 - 8	P
	Caulimovirus	dsDNA / 1 circular		P
Retroviridae		ssRNA / dimer 1 + linear	7 - 11	V

[a] A: algae; B: bacteria; F: fungi; I: invertebrates; M: mycoplasma; P: plants; Pr: protozoa; V: vertebrates

RNA viruses use different strategies for the replication of their genomes. A special problem arises because uninfected host cells lack enzymes for synthesizing RNA according to the instructions of an RNA template. Consequently, RNA viruses must contain genetic information for the synthesis of an RNA-directed RNA polymerase (also called an RNA replicase or an RNA synthetase) or for an RNA-directed DNA polymerase (also called a reverse transcriptase). It is informative to classify RNA viruses according to the relation between their virion RNA and messenger RNA (mRNA). By convention, mRNA is defined as (+) RNA and its complement as (–) RNA.

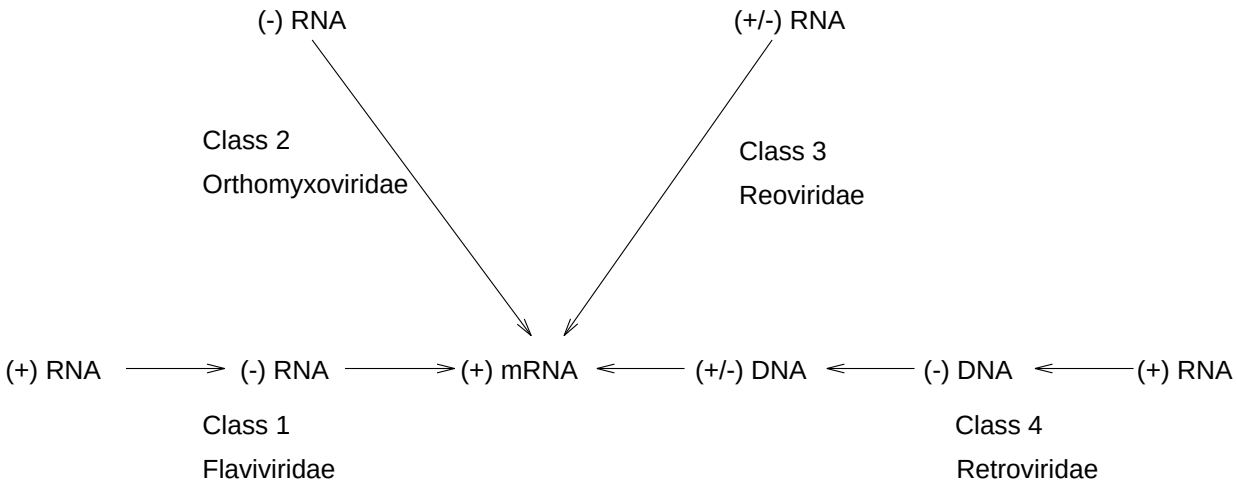


Figure 1: Modes of gene expression of RNA viruses.

Four pathways of replication and transcription of RNA viruses are known, see Figure 2:

Class 1 viruses (e.g., *Flaviviridae*) are positive-strand RNA viruses. They synthesize (–) RNA, which then serves as the template for the formation of (+) mRNA.

Class 2 viruses (e.g., *Orthomyxoviridae*) are negative-strand RNA viruses in which virion (–) RNA is the template for the synthesis of (+) mRNA.

Class 3 viruses (e.g., *Reoviridae*) are double-stranded RNA viruses in which the virion (±) RNA directs the asymmetric synthesis of (+) mRNA.

Class 4 viruses (e.g., *Retroviridae*) express the genetic information in their virion (+) RNA through a DNA intermediate that serves as the template for the synthesis of (+) mRNA. The flow of information is from RNA to DNA and back to RNA. They seem to be related to reverse transcribing DNA viruses, see Table 8. For instance, sequence similarities have been found between *caulimoviruses* and animal retroviruses [80, 60].

RNA virus superfamilies

Several properties of RNA viruses were analyzed and relationships between seemingly unrelated virus families were identified. This finding suggests three superfamilies of viruses which may have a common evolutionary origin [61]:

(i) *picorna*-like viruses

(ii) *Sindbis*-like viruses

(iii) and the *Luteovirus* group.

The idea of superfamilies has been investigated mainly for RNA plant viruses. Given the very similar particle structures of *rhabdoviruses* and their replication in both plants and animals, it is not surprising that some genome and amino acid sequence similarities have been revealed [1]. With this family, there is good reason to believe that the sequence similarities are homologies, that is, have a common evolutionary origin. Amino acid sequence similarities in nonstructural proteins have also been revealed between various RNA plant virus groups having diverse particle morphology and between these viruses and certain viruses infecting vertebrates. This has led to the idea that many plus-strand RNA virus groups may be classified into two major superfamilies and that viruses within these superfamilies may have a common evolutionary origin [33, 88]. A third supergroup or superfamily centered on the *Luteovirus* group has been proposed [37].

Figure 2 summarizes the similarities in genome organization and sequence that have been revealed between *poliovirus* (a member of the *Picornaviridae*), *comoviruses (CPMV)*, *nepoviruses (TBRV)*, and *potyviruses (TVMV)*, see Figure 2.

These viruses have the following features in common:

(i) positive sense ssRNA genomes;

(ii) a VPg at the 5' terminus and a poly(A) tract at the 3' terminus;

(iii) single long open reading frames (ORFs) coding for polyproteins that are processed by viral-coded proteases to give the functional gene products;

(iv) these viruses encode several non-structural proteins that have similar functions and significant amino acid sequence similarity (>20%); and

(v) the genes for these conserved proteins have a similar arrangement for all the genomes.

Several plant RNA viruses have been found to show similarities in amino acid sequence and genome organization with *Sindbis virus*, an *Alphavirus* with a lipoprotein envelope that infects vertebrates. This collection of virus groups is quite variable in genome structure and strategy. They are the *tobamoviruses (TMV)*, *tombusviruses (CuNV)*, *carmoviruses (CarMV)*, *tobraviruses (TRV)*, *hordeiviruses (BSMV)*, and the three closely related groups — *bromoviruses (BMV)*, *cucumoviruses (CMV)*, and *alfalfa mosaic viruses (AMV)*, see Figure 3. *Furoviruses (BNYVV)* [6] and *potexviruses* [28, 73] also have sequence similarities that place them in the superfamily of *Sindbis*-like viruses. All of these viruses have a 5'cap but the 3'termini vary. Most of them specify three proteins with significant sequence similarity to the three nonstructural proteins nsP_1, nsP_2, and nsP_4 of *Sindbis virus*. The strongest sequence similarity is found between *BMV*, *CMV*, and *AMV*, making it virtually certain that these viruses share a common ancestry. The *carmoviruses* and *tombusviruses* have very small genomes encoding only one of the three conserved domains found in the other *Sindbis*-like viruses.

Sequence motifs of nucleic acid helicases and RNA polymerases previously considered to be specific for one of the two superfamilies were found within the new *Luteovirus*-like superfamily. It is suggested that this new superfamily provides an evolutionary link between the other two [37]. Arranging the virus groups into three superfamilies based on the RNA polymerase sequence motif alone gave rise to exactly the same arrangement as that based on the helicase motif alone. This supports the idea that there may be a real basis for the three superfamilies.

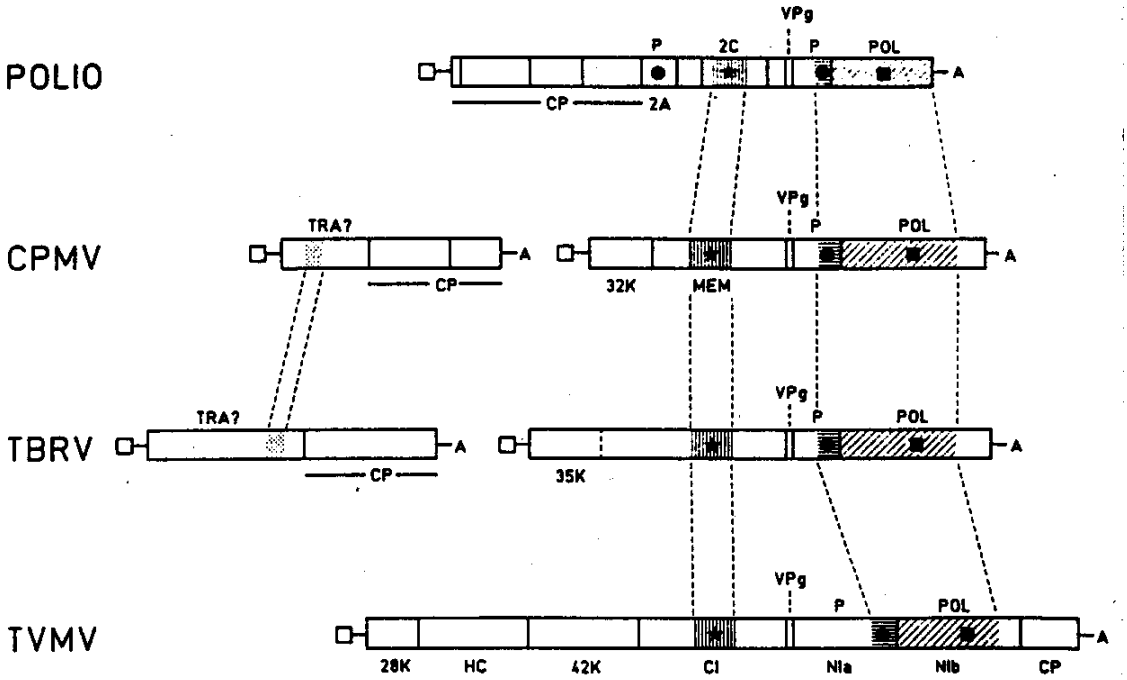


Figure 2: Comparison of the genomic RNAs of *picornaviruses* (POLIO) and the *picorna*-like plant viruses. The "superfamily" of *picorna*-like viruses includes plant viruses, such as *comoviruses (CPMV)*, *nepoviruses (TBRV)*, and *potyviruses (TVMV)*. ORFs are represented as open bars, VPg as open squares, and poly(A) tails as A. Regions of significant (>20%) amino acid sequence similarity in the gene products are indicated by similar shading. TRA = Transport function; CP = capsid protein(s); P = protease; MEM = membrane binding; POL core RNA-dep. RNA polymerase; HC = helper component; ⋆ = nucleotide binding domain; • = cysteine protease domain; ■ = polymerase domain, adapted from [61].

However, the discussed genome and amino acid sequence similarities suggest relationships that cut across many of the accepted criteria for classifying viruses:

(i) host groups — plant and vertebrate viruses;

(ii) morphology — viruses with rod-shaped particles, icosahedral particles, and particles with a lipoprotein envelope;

(iii) kind of nucleic acid — DNA and RNA; and

(iv) numbers of genome segments — viruses with monopartite, bipartite, and tripartite RNA genomes.

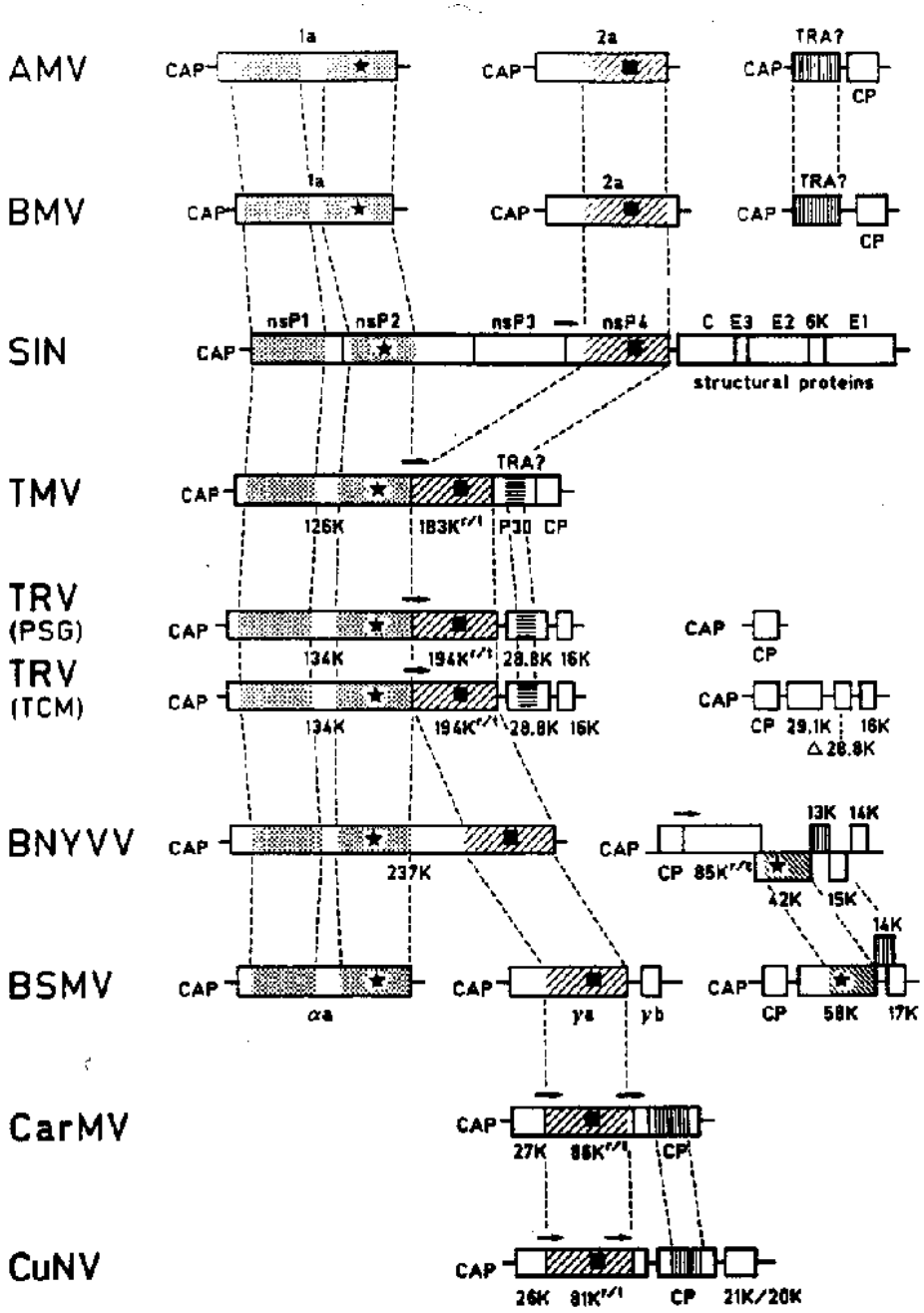


Figure 3: Comparison of the genomic RNAs of *Sindbis virus* (SIN) and the *Sindbis*-like plant viruses. The "superfamily" for *Sindbis*-like viruses includes plant viruses, such as the *alfalfa mosaic virus* group *(AIMV)*, *bromoviruses (BMV)*, *tobamoviruses (TMV)*, *tobraviruses (TRV*, strains *PSG* and *TCM)*, *furoviruses (BNYVV)*, *hordeiviruses (BSMV)*, *carmoviruses (CarMV)*, and *tombusviruses (CuNV)*. ORFs are represented as open bars, and regions of amino acid sequence similarity in the gene products are indicated by similar shading. For reasons of simplicity, closely adjoining or slightly overlapping genes in the *TMV* (p30 and CP) and *CarMV* (CP) genome are drawn contiguously. → = Leaky termination codon; r/t = read-through. For other symbols see the legend in Figure 2, adapted from [61].

Many more sequences will need to be determined before the significance of the proposed superfamilies can be established.

There is a low level of similarity between certain sites in proteins of viruses belonging to the *picorna*-like and the *Sindbis*-like superfamilies.

(i) They contain a stretch of amino acids similar to the nucleotide triphosphate binding site of several GTP- or ATP-using enzymes.

(ii) In members of both the *picorna* and *Sindbis* superfamilies a conserved RNA-dependent RNA polymerase domain was identified [50]. A similar conserved sequence was found in retroviral reverse transcriptase, members of viruses like *influenza virus*, *CaMV*, and *hepatitis B virus*, suggesting that the sequence was an active site or a recognition site for polymerases in general.

(iii) A set of conserved amino acid residues has been identified in the chymotrypsinlike serine proteases and in the cysteine proteases of some positive strand RNA viruses [34].

3. Folding and Comparison of RNA

3.1. RNA Secondary Structures

RNA structure can be broken down conceptually into a secondary structure, and a tertiary structure. The secondary structure is a pattern of complementary base pairings. The Watson-Crick base paired regions of the folded structure of an RNA molecule are stabilized both by hydrogen bonding and stacking between the aromatic bases on strands oriented in antiparallel directions. These double-stranded stems and the non-paired regions between them form the structural elements, see Figure 4a. These structural elements include hairpin loops, internal loops, bulge loops, junctions and single-stranded regions, see Figure 5.

The tertiary structure is the three-dimensional configuration of the molecule, see Figure 4b. Tertiary interactions are hydrogen bonding or stacking interactions between structure elements. The pseudoknot is a simple tertiary interaction (not found in tRNA). A pseudoknot is a configuration in which nucleotides that are inside a loop pair with nucleotides outside this loop.

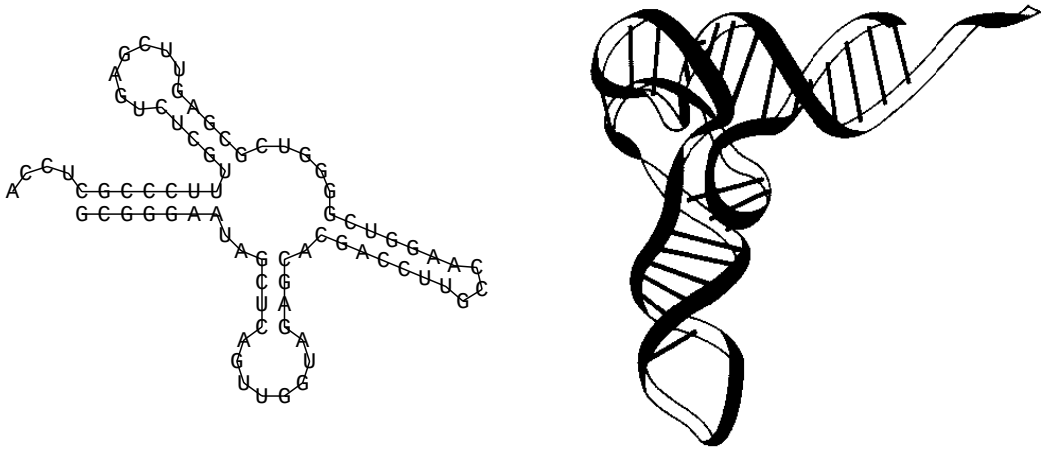


Figure 4: a) The secondary structure extracts the most important information about the structure, namely the pattern of base pairings.
b) The spatial structure of the phenylalanine tRNA from yeast is one of the few known three dimensional RNA structures.

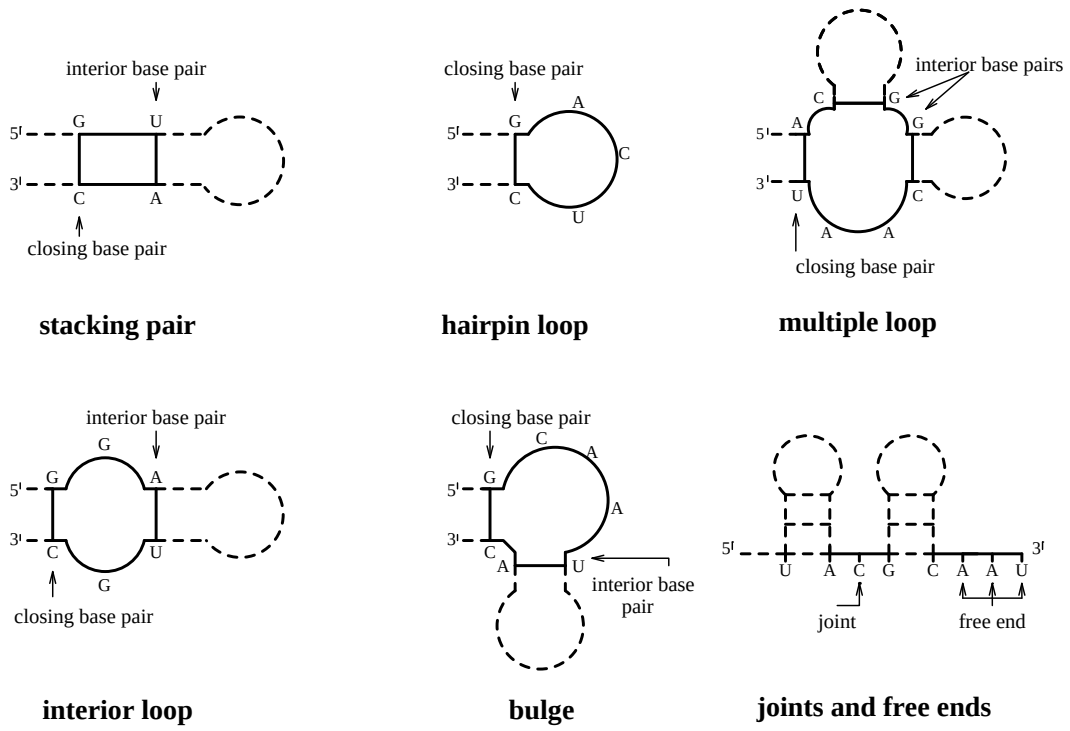


Figure 5: Basic structure elements on nucleic acid secondary structures.
Every structure within the secondary structure model can be decomposed into the basic elements: stems, hairpins, interior loops, bulges, multi-stem loops, joints, and free ends.

As opposed to the protein case, the secondary structure of RNA sequences is well defined; it provides the major set of distance constraints that guide the formation of tertiary structure, and covers the dominant energy contribution to the 3D structure. Secondary structures are conserved in evolutionary phylogeny, and they represent a qualitatively important description of the molecules, as documented by their extensive use for the interpretation of molecular evolution data.

A secondary structure on a sequence is a list of base pairs i, j with $i < j$ such that for any two base pairs i, j and k, l with $i \leq k$ holds:

$$\begin{aligned} i = k &\iff j = l \\ k < j &\implies i < k < l < j \end{aligned} \tag{1}$$

The first condition implies that each nucleotide can take part in not more than one base pair, the second condition forbids knots and pseudoknots. Knots and

pseudoknots are excluded by the great majority of folding algorithms which are based upon dynamic programming concepts.

A base pair k, l is *interior* to the base pair i, j, if $i < k < l < j$. It is *immediately interior* if there is no base pair p, q such that $i < p < k < l < q < j$. For each base pair i, j the corresponding *loop* is defined as consisting of i, j [82] itself, the base pairs immediately interior to i, j and all unpaired regions connecting these base pairs. The energy of the secondary structure is assumed to be the sum of the energy contributions of all loops. (Note that a stacked base pair constitutes a loop of size 4; the smallest hairpin loop has three unpaired bases, i.e., size 5 including the base pair.) The types of structural elements are defined in Figure 5. Experimental energy parameters are available for the contribution of an individual loop as functions of its size, of the type of its delimiting base pairs, and partly of the sequence of the unpaired strains [29, 49]. For the base pair stacking the enthalpic and entropic contributions are known separately. Contributions from all other loop types are assumed to be purely entropic. In the case of multiloops[1] the energy parameters are only an approximation and thus not very reliable, which is important for Chapter 4 and 5. We use a recent version of the parameter set published in [29].

3.2. Representing the Structure

Bracket Notation

The unique decomposition of secondary structures outlined above suggests a simple string representation of structures by identifying a base pair with a pair of matching brackets and denoting an unpaired digit by a circle (upstream is understood in 5'→3' direction in accord with the IUPAC convention; downstream refers to the opposite direction), see Figure 8:

(	upstream paired base
)	downstream paired base
.	single-stranded base.

This bracket notation is coding for a tree [26].

[1] A multiloop is a loop with a degree larger than 2.

Mountain Representation

A convenient way of displaying the size and distribution of secondary structure elements is the *mountain representation* [45]. In this representation a '(' is drawn as a step up, a ')' corresponds to step down, and an unpaired base '.' is shown as horizontal line segment, see Figure 6. The resulting graph looks like a mountain-range where:

- *Peaks* correspond to hairpins. The symmetric slopes represent the stack enclosing the unpaired bases in the hairpin loop, which appear as a plateau.
- *Plateaus* represent unpaired bases. When interrupting sloped regions they indicate bulges or interior loops, depending on whether they occur alone or paired with another plateau on the other side of the mountain at the same height respectively.
- *Valleys* indicate the unpaired regions between the branches of a multi-loop or, when their height is zero, they indicate unpaired regions separating the components of secondary structures.

The height of the mountain at sequence position k is simply the number of base pairs that enclose position k, i.e., the number of all base pairs (i, j) for which $i < k$ and $j > k$.

The mountain representation allows for a straightforward comparison of secondary structures and inspired a convenient algorithm for alignment of secondary structures [53]. A modified version of the mountain representation [45] can be constructed easily from the base pairing probability matrix. The number

$$m(k) \stackrel{\text{def}}{=} \sum_{i<k} \sum_{j>k} p_{ij} \tag{2}$$

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 (MFE) 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

[2] In the terminology of Zuker and Sankoff [92] these are all base pairs to which sequence position k is interior.

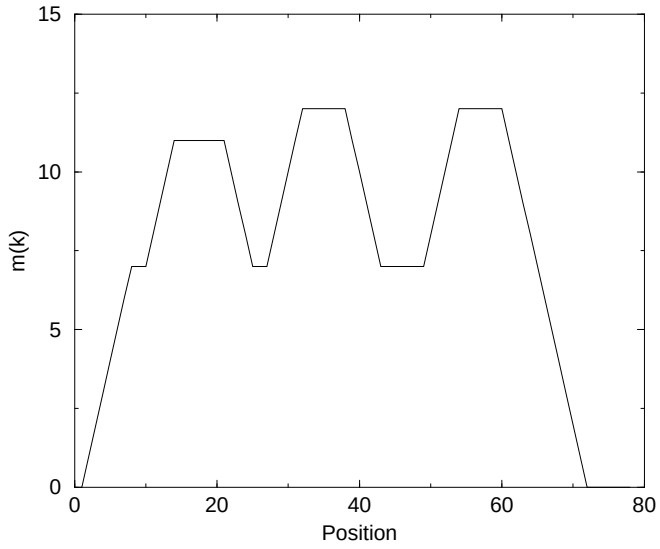


Figure 6: Mountain representation of the tRNA secondary structure shown in Figure 4. The three plateaus correspond to the three hairpin loops of the clover leave structure.

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.

Dot Plots

A dot plot is a two-dimensional graph in which the size of the dot at position i, j within the graph represents the probability of the ij base pair. Thus, in principle, dot plots contain base pairing information. In practice, we suppress the dots corresponding to base pairs that occur with a probability of less than 10^{-5}.

The plot is divided into two triangles. The upper right triangle contains the base pairing probability matrix (p_{ij}); the size of the squares is proportional to the pairing probability. The lower-left triangle displays the MFE structure for comparison. Here only the base pairs that occur in the MFE are indicated, see Figure 7. 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 the diagonal patterns.

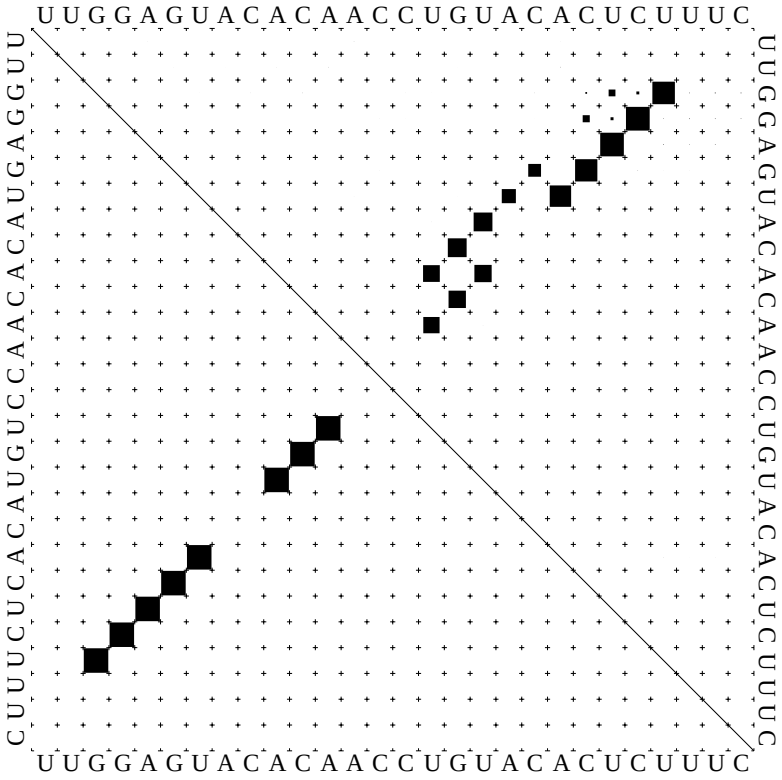


Figure 7: The upper right triangle contains the base pairing probability matrix (p_{ij}); the size of the squares is proportional to the pairing probability. The lower-left triangle displays the MFE structure for comparison.

3.3. RNA Folding Programs

Minimum Free Energy Folding

As a consequence of the additivity of the energy contributions, the MFE can be calculated recursively by dynamic programming [82, 84, 93, 92]. The essential part of the energy minimization algorithm is shown in Table 5. The basic logic of this scheme is derived from sequence alignment: In fact, folding of RNA can be regarded as a form of alignment of the sequence to itself [92].
The algorithm in Table 5 can be parallelized and allows to compute the structure of complete RNA virus genomes such as HIV-1 [43] efficiently.

Because of the simplification in the energy model and the uncertainties in the energy parameters, predictions are not always as accurate as one would like. It is

Table 5. Pseudo Code of the MFE folding algorithm.

```
for(d=1...n)
   for(i=1..n-d)
      j=i+d
      C[i,j] = MIN(
          Hairpin(i,j),
          MIN( i<p<q<j : Interior(i,j;p,q)+C[p,q] ),
          MIN( i<k<j : FM[i+1,k]+FM[k+1,j-1]+cc )       )
      F[i,j] = MIN( C[i,j], MIN(i<k<j : F[i,k]+F[k+1,j]))
      FM[i,j]= MIN( C[i,j]+ci, FM[i+1,j]+cu, FM[i,j-1]+cu,
                    MIN( i<k<j : FM[i,k]+FM[k+1,j] )       )
free_energy = F[1,n]
```

`F[i,j]` denotes the MFE for the subsequence consisting of bases i through j. `C[i,j]` is the energy given that i and j pair. The array `FM` is introduced for handling multi-stem loops. The energy parameters for all loop types except for multi-stem loops are formally subsumed in the function `Interior(i,j;p,q)` denoting the energy contribution of a loop closed by the two base pairs $i-j$ and $p-q$. We have assumed that multi-stem loops have energy contribution `F=cc+ci*I+cu*U`, where `I` is the number of interior base pairs and `U` is the number of unpaired digits of the loop. The time complexity here is $\mathcal{O}(n^4)$.
It is reduced to $\mathcal{O}(n^3)$ by restricting the size of interior loops to some constant $\mathcal{M}$, i.e., $p-i \leq \mathcal{M}$ and $j-q \leq \mathcal{M}$. In general we use $\mathcal{M}=40$. This can be regarded as a minor correction since loops of that size are extremely rare.
The structure (list of base pairs) leading to the minimum energy is usually retrieved later on by *backtracking* through the energy arrays.

therefore, desirable to include additional structural information from phylogenetic or chemical data.

The MFE algorithm allows to include a variety of constraints into the secondary structure prediction by assigning bonus energies to structures honoring the constraints. One may enforce certain base pairs or prevent bases from pairing. Additionally, the algorithm can deal with bases that have to pair with an unknown pairing partner.

Large RNA molecules in general decompose into components, that is, into continous sequence pieces that form base pairs only inside themselves and which are not interior to any other base pairs. Components play a special role in the calculation of the MFE structure: they can be folded independently from each other. The biophysical significance of components is that they form well separated substructures which are only loosely tied to each other. Since $m(k)$ drops to zero at component boundaries this measure could be used to identify them.

Calculation of the Partition Function

The partition function for the ensemble of all possible secondary structures

$$Q = \sum_{\text{all structures } S} e^{-\frac{\Delta G(S)}{RT}} \tag{3}$$

can be calculated analogously [62]. A pseudocode is given in Table 6. The base pairing probabilities can be abtained by "backtracking". For details refer to [62].

Table 6. Pseudocode for the calculation of the partition function.

```
for(d=1...n)
   for(i=1...d)
      j=i+d
      QB[i,j] = EHairpin(i,j) +
         SUM( i<p<q<j : EInterior(i,j;p,q)*QB[p,q] ) +
         SUM( i<k<j : QM[i+1,k-1]*QM1[k,j-1]*Ecc )
      QM[i,j] =
         SUM( i<k<j : (Ecu^(k-i)+QM[i,k-1])*QM1[k,j] )
      QM1[i,j]= SUM( i<k<=j : QB[i,k]*Ecu^(j-k)*Eci )
      Q[i,j]  = 1 + QB[i,j] +
         SUM( i<p<q<j : Q[i,p-1]*QB[p,q] )
partition_function = Q[1,n]
```

Here Ex:=exp$(-x/RT)$ denotes the Boltzmann weights corresponding to the energy contribution x. `Q[i,j]` denotes the partition function Q_{ij} of the subsequence i through j. The array `QM` contains the partition function Q^b_{ij} of the subsequence subject to the fact that i and j form a base pair. `QM` and `QM1` are used for handling the multiloop contributions. `x^y` means x^y. For details see [62].

Both folding algorithms have been integrated into a single interactive program including postscript output of the minimum energy structure and the base pairing probability matrix.

Vienna RNA Package

The `Vienna RNA Package` was developed by the Theoretical Biochemistry group at the "Institut für Theoretische Chemie und Strahlenchemie" [41] and contains computer codes for prediction and comparison of RNA secondary structures. It consists of a library and some standalone programs and allows to

(i) predict MFE secondary structures

(ii) calculate the partition function for the ensemble of structures

(iii) predict melting curves

(iv) search for sequences folding into a given structure

(v) compare secondary structures including pairwise alignment.

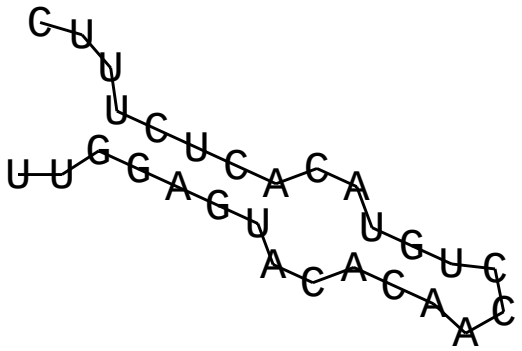
The `Vienna RNA Package` is public domain software and can be obtained by anonymous ftp. `RNAfold` is the core of the package. It reads RNA sequences from *stdin* and calculates their MFE structure, partition function and base pairing probability matrix. It returns the MFE structure in bracket notation, its energy, the free energy of the thermodynamic ensemble and the frequency of the MFE structure in the ensemble to *stdout*. It also produces PostScript files with plots of the resulting secondary structure graph and a *dot plot* of the base pairing matrix, see Figure 8 for an example.

The `Vienna RNA package` [41] has already been used successfully for computing the base pairing probability matrix of the complete genome of the coli-phage Qβ on a UNIX workstation and the complete genome of a HIV-1 RNA on a CRAY Y-MP super-computer [48]. It is being used by a number of research groups see e.g. [3, 48, 40].

The length of the complete *flavivirus* and *influenza virus* genome has been prohibitive for computing the equilibrium partition functions and base pairing probabilities of the entire genome. Secondary structure predictions were therefore based on folding subsequences in fairly small windows. This approach has two disadvantages:

```
monet> RNAfold -T 42 -p1
Input string (upper or lower case); @ to quit
.........1.........2.........3.........4.........5.........6.........7 .........
UUGGAGUACACAACCUGUACACUCUUUC
length = 28

UUGGAGUACACAACCUGUACACUCUUUC
..((((((..(((...)))..))))))...
minimum free energy = -3.71
..(((((([[(,,...)))..))))))...
free energy of ensemble = -4.39
frequency of mfe structure in ensemble 0.337231
```



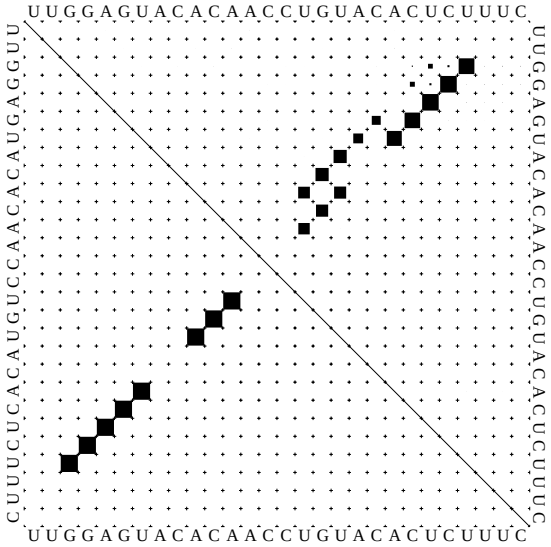


Figure 8: Interactive example run of RNAfold for a random sequence. When the base pairing probability matrix is calculated the symbols `.` `,` `|` `{` `}` `(` `)` are used for bases that are essentially unpaired, weakly paired, strongly paired without preferred direction, weakly upstream (downstream) paired, and strongly upstream (downstream) paired, respectively. Apart from the console output the two postscript files `rna.ps` and `dot.ps` are created. The lower left part of `dot.ps` shows the minimum energy structure, while the upper right shows the pair probabilities. The area of the squares is proportional to the binding probability.

(i) by definition it cannot be used for long-range interactions that span more than the window size, and

(ii) the results depend crucially on the exact location of the boundaries of the sequence window.

Subsequences can be folded independently of the rest of the structure only if they form a component by themselves, i.e., if there are no base pairs to the outside of the sequence window. The only way of identifying the component boundaries is, however, MFE folding the sequence in its entirety. Consequently, the structure can be approximated quite well by independently folding of the identified subsequences.

3.4. Interpreting Computed Structures

Well-Definedness

The well-definedness $d(k)$ of the structure in a certain region is

$$d(k) \doteq \max \left\{ \max_i \{p_{ik}, p_{ki}\},\ 1 - \sum_i p_{ik} \right\} \tag{4}$$

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 high 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, see Figure 14. 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 probablility 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 the well-definedness measure does not depend on the well-definedness of all nucleotides within a secondary structure. As long as most of the nucleotides in a region of the sequence are

well defined, the presence of a few nucleotides with variable base-pairing behavior ("breathing") will not decrease the well-definedness of the whole region.

Several alternative secondary structures may be accessible to a given RNA sequence. To judge the relevance of a particular secondary structure motif, it becomes, therefore, important to assess how well defined it is with respect to alternative structures. The `Vienna RNA Package` [41] allows such assessment by computing the entire partition function (i.e. the equilibrium ensemble) of structural states available to an RNA sequence [62]. From the partition function the complete base pairing probability matrix can be obtained (that is, the probability that base i pairs with base j along the chain). From this matrix measures of *well-definedness* are derived and applied to elucidate novel structural features in flaviviral RNAs.

Huynen and coworkers [47] 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 in [91] that is based on the energy differences between the MFE 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 [27] 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 [5]. 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 [53]. Since the generalized mountain represention $m(k)$

defined in equ.(2) 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

$$q(k) \doteq m(k) - m(k-1) \tag{5}$$

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.

4. Flaviviruses

4.1. Introduction

The family *Flaviviridae* includes three genera, the *flaviviruses*, the *pestiviruses* and the *hepatitis C viruses*. These three genera have diverse biological properties and show no serological cross-reactivity, but appear to be similar in terms of virion morphology, genome organization, and presumed RNA replication strategy.

The genus *flaviviruses* comprises almost 70, mostly arthropod-borne viruses including a number of human pathogens of global medical importance, such as *yellow fever (YF) virus*, *Japanese encephalitis (JE) virus*, the *Dengue (DEN) viruses*, and *tick-borne encephalitis (TBE) virus* [63], see Table 7. Most *flaviviruses* are transmitted to vertebrates by chronically infected tick- or mosquito-vectors. The spectrum of diseases caused by flaviviruses ranges from a mild fever to hepatitis, hemorrhagic disease, and encephalitis.

Flaviviruses are small enveloped particles with an unsegmented, plus-stranded RNA genome. Mature *flavivirus* virions contain three structural proteins: a nucleocapsid or core protein (C; 13kd), a nonglycosylated membrane protein (M; 8kd), and an envelope protein (E; 55kd) which is usually glycosylated. The M and E proteins are both associated with the lipid envelope by means of hydrophobic anchors. The E protein is the major component of the virion surface; it is the main target of immune response. Structural elements of the E protein determinants are assumed to be involved in the binding of virions to cell receptors and in intraendosomal fusion at low pH.

Using serological methods *flaviviruses* can be subdivided into a number of serocomplexes and this classification has generally been confirmed by the genomic sequence data that became available for many *flaviviruses* during the past few years. The construction of evolutionary trees reflects the established classification, shown in Figure 9. The amino acid sequence comparisons of protein E yield a picture that perfectly matches that of the *flavivirus* serocomplexes defined by cross-neutralization using polyclonal immune sera. Phylogenetic trees based on

Table 7. Serological Classification of the Genus *Flavivirus*

Group	**Type member**[b]	**Vector**
Tick-borne encephalitis	TBE flavivirus	Tick
	European subtype	Tick
	Neudörfl*	Tick
	Far Eastern subtype	Tick
	Sofyn*	Tick
	Louping ill virus	Tick
	Langat virus	Tick
	Powassan virus	Tick (Mosquito)
Rio Bravo	Rio Bravo virus	
Japanese encephalitis	Japanese encephalitis virus	Mosquito
	West nile virus	Mosquito (Tick)
	Kunjin virus	Mosquito
Tyuleniy	Tyuleniy virus	Tick
Ntaya	Ntaya virus	Mosquito
Uganda S	Uganda S virus	Mosquito
Dengue	Dengue virus 1	Mosquito
	Dengue virus 2	Mosquito
	Dengue virus 3	Mosquito
	Dengue virus 4	Mosquito
Modoc	Modoc virus	
ungrouped	Yellow fever virus	Mosquito

[b] list incomplete
* indicates prototype strains

sequence comparisons yield information about the time of divergence between different viral types and suggest a subdivision in several serocomplexes, such as (i) the *Dengue (DEN) viruses* types 1 to 4, (ii) *Japanese encephalitis (JE) virus*, *West nile (WN) virus*, *Kunjin (KUN) virus*, *Murray Valley encephalitis (MVE) virus* and others, (iii) *YF virus* and (iv) *tick-borne encephalitis (TBE) virus*. The TBE serocomplex comprises exclusively *flaviviruses* transmitted by ticks in contrast to the majority of other *flaviviruses* which utilize mosquitos as their principal arthropod-vector. The main representative of this serocomplex is *TBE virus*, which is endemic in many parts of Europe (European subtype) and Asia (Far Eastern subtype). *Langat (LGT)*, and *louping ill (LI) virus* have a substantially lower degree of sequence homology. The most distantly related member of the tick-borne serocomplex is *Powassan (POW) virus* which shares 76% protein sequence homology with *TBE virus* [59]. *POW virus* is endemic in parts of Canada and Far East Asia and causes sporadic cases of encephalitis in humans.

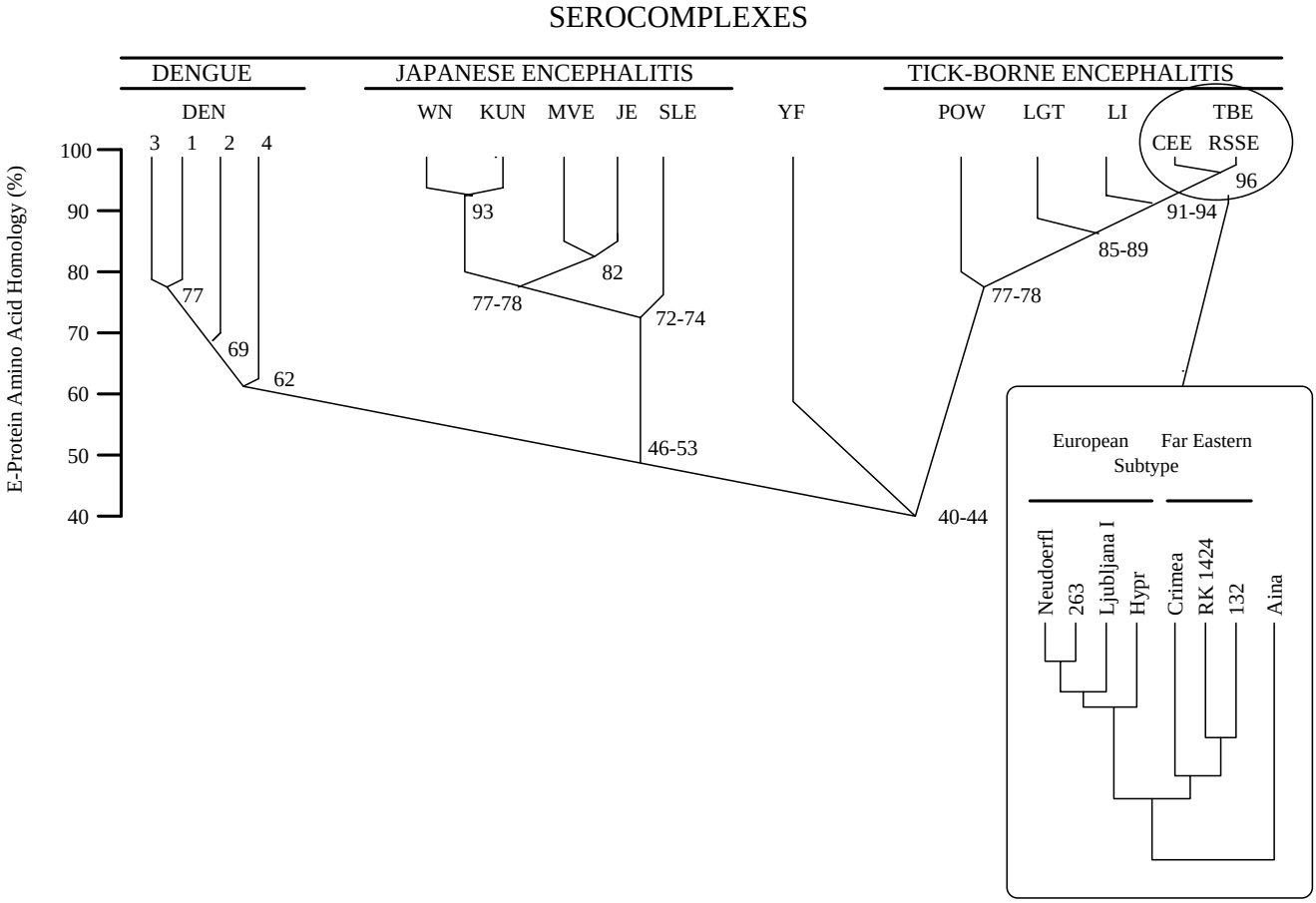


Figure 9: Evolutionary tree of *flaviviruses* drawn on the basis of their *E* protein amino acid homologie. *DEN, WN, KUN, MVE, JE, SLE (St. Louis encephalitis), YF, POW, LGT, LI, TBE.* Evolutionary relationships of *TBE viruses*, adapted from [22, 81].

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 [9]). The flanking noncoding regions (NCRs) are believed to contain cis-acting elements important for replication, translation and packaging. During the flavivirus replication cycle, the plus-strand genomic RNA is first replicated into minus-strand RNA which serves as the template for the synthesis of more genomic RNA. The conserved 3'-terminal structures as well as some short conserved sequences within the 3'NCR of the genomic RNA may function as *cis*-acting replication signals and interact with viral and, possibly, also cellular proteins during the initiation of the minus-strand RNA synthesis [71].

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 [38], but these were found to be absent in tick-borne *flaviviruses*. Sequence analysis of a number of *TBE virus* strains recently revealed a surprising heterogeneity in the length of the 3'NCRs even among closely related strains [81]. Thus, 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 in some cases of includes an internal poly(A) sequence element, whereas the latter is 350 nucleotides long and exhibits a high degree of sequence conservation.

Table 8. List of Sequences for the analysis of the 3'NCR.

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

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 [59]. Very similar structures were reported for the sequences of mosquito-borne *flaviviruses* [35, 86] 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 [4].

For our analysis we used the 44 *flaviviruses* sequences listed in Table 8. From each sequence we extracted the 3'NCR of the genomic RNA. Using a set of longer sequences with up to 1000nts we checked whether 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.

4.2. New Conserved Secondary Structure Motifs at the 3' End

All *flaviviruses* exhibit a well-conserved 3'-terminal secondary structure extending over 106 bases [81]. It consists of a large stem-loop (A1) and a small hairpin (A2) structure, see Figure 10.
Using structural alignment techniques based on the equilibrium ensemble, we were able to confirme the characteristic secondary structure at the very 3'end of the genome. Figure 11 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.
The base pairing is very well defined in this region, the plot shows only few alternatives to the minimum free energy (MFE) 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 in [71].

Immediately upstream of the known conserved 3'-terminal secondary structure we discovered a new well-conserved motif consisting of six stems. The numbering of the stems is defined in Figure 12. Conserved sequence elements adapted from [81] are:

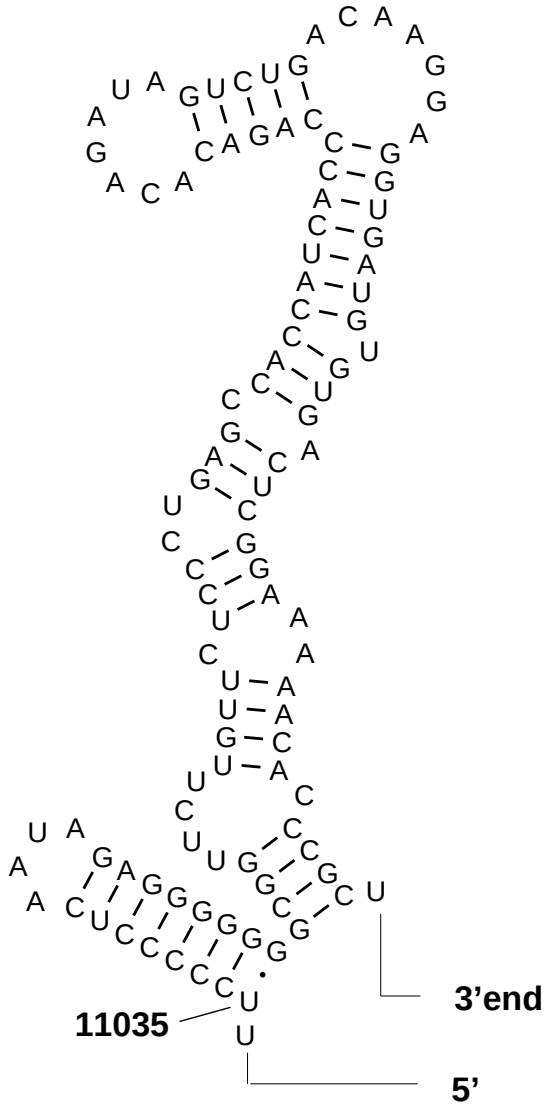


Figure 10: The conserved 3'-terminal secondary structure of *TBE flavivirus Neudörfl* strain consisting of a stem-loop (A1) and a hairpin (A2).

PR pyrimidine rich box,

PY homo-pyrimidine box,

PU homo-purine box,

IR inverted repeat,

R3 imperfect direct repeat.

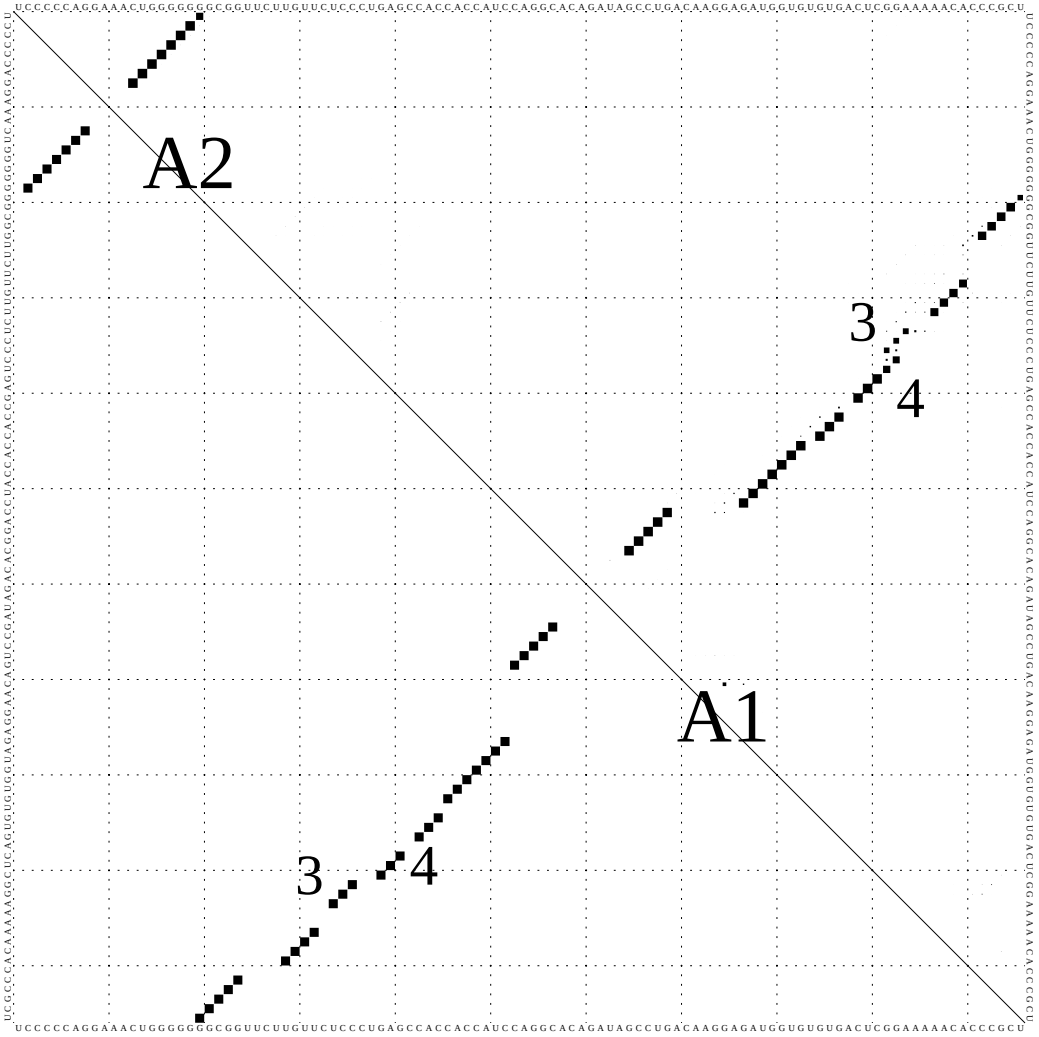


Figure 11: Dot plot of the conserved secondary structure formed by the 106 nucleotides at the 3'terminus of the *POW virus* sequence. This structure is mostly unambiguous, except for the 3rd and 4th stem: In the MFE structure we have two stems of length 3. Alternatively the 4th stem may be elongated by two base pairs and stem 3 opens up. This area is involved in the formation of the pseudoknot described in [71].

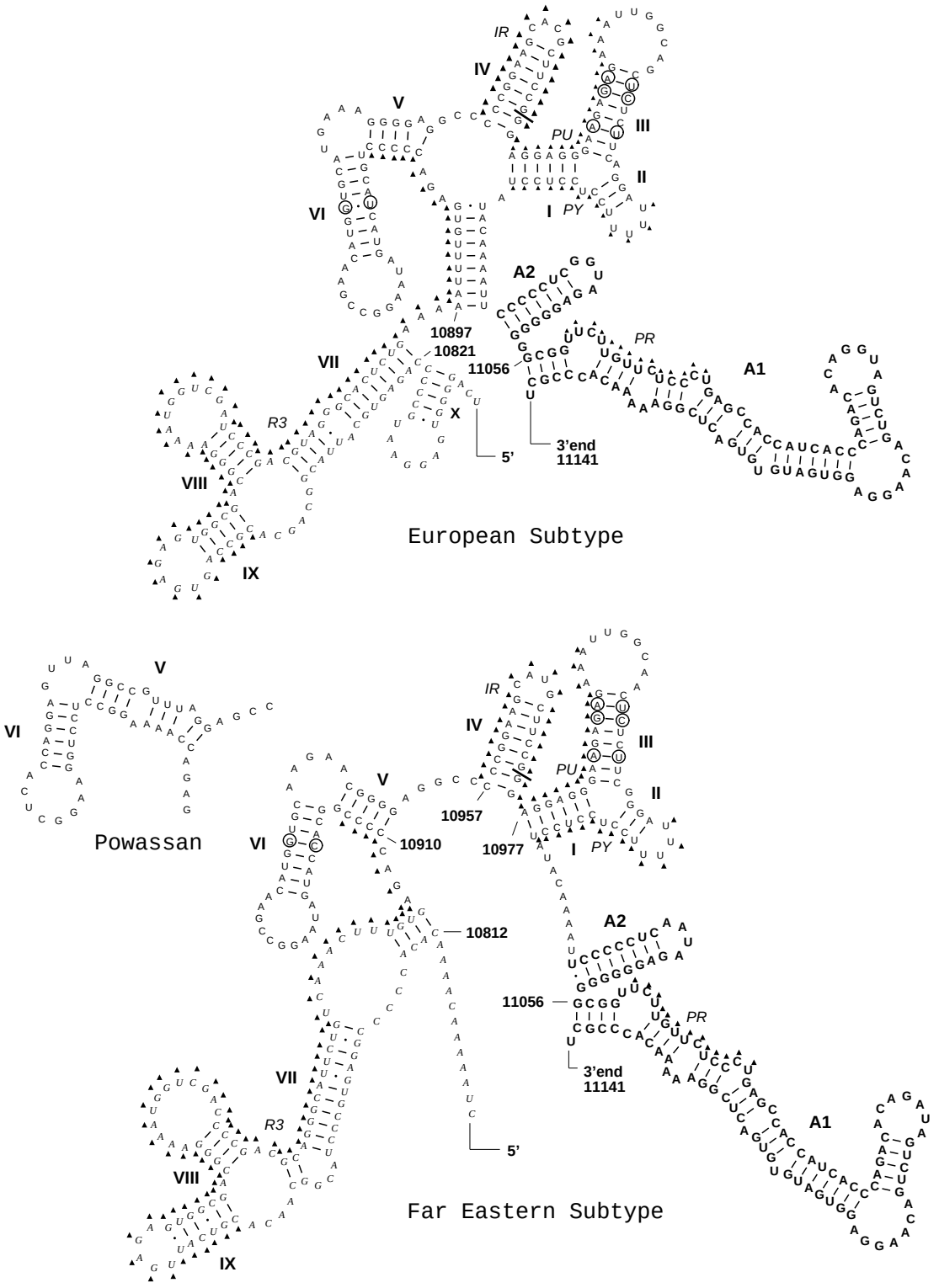


Figure 12: Secondary structure of the 3'UTR of tick-borne *flaviviruses*

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 [81] 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 *Neudörfl*.

stem I The sequence is conserved among all sequences.

stem II The sequence is conserved among all sequences.

stem III *POW virus*: 3: AU→GC 6: GC→AU 7: AU→GC

stem IV The sequence in conserved except for a deletion in *POW virus*.

stem 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.

stem VI 5: *Far Eastern*: GC; *Aina*: AU; *European*: GU.

The corresponding sequences are quite conserved in this region. In particular, not all of the six stems are confirmed by compensatory mutations. Compensatory mutations were identified by calculating a multiple sequence alignment using `CLUSTALW` [79], see Table 11. The corresponding alignment of MFE structures is shown in Table 12. Compensatory mutations are commonly interpreted as the result of selection pressure, and, thus, 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 12 and 13, this new motif decomposes into three components and seems to appear in two variants that correlate with the serological classification. For the *European* subtype (*Neudörfl, 263, Ljubljana* and *Hypr* strains) the elements composing this motifs are located within a multiloop, whereas for the *Far Eastern* subtype (*Crimea, 132, RK1424* and *Aina* strains) the same elements are external. In order to check the significance of this finding we calculated 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 10. The chance that the assignment of the 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 (MFE) 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*.

Table 9. CLUSTAL W(1.60) multiple sequence alignment.
The multiple sequence alignment was used to identify compensatory mutations (at the 3'end of the genome), indicated by numbers and letters, respectively. Conserved positions are indicated by stars.

```
Neudoerfl --TCAGGGGTGAGGAATGCCCCCAGAGTGCATTACGGCAGCACGCCAGTGAGAGTGGCGA
263       --TCAGGGGTGAGGGATGCCCCCAGAGTGCATTACGGCAGCACGCCAGTGAGAGTGGCGA
Ljubljana --TCAGGGGTTGGGGATGCCCCCAGAGTGCATTACGGCAGCACGCCAGTGAGAGTGGCGA
Hypr      --TCAGGGGTGAGGGATGCCCCCAGAGTGCATTACGGCAGCACGCCAGTGAGAGTGGCGA
Crimea    -TGGCCGGGGTAGAAACACCCCCGGAGTGCCCCACGGCAGCACGTCAGTGAGAGTGGCGA
132       -AAAAAAAAACAAACACACCCCCGT-GTATTCCACGGCAACACGTCAGTGAGAGTGGCGA
Aina      GGTCAGGGGTGAG-AACACCCCCAGAGTGCACCACGTCAACACGCCAGTGAGAGTGGCGA
RK1424    --CTAAAAACAAAACACACCCCCGGAGTGCCCTACGGCAACACGTCATTGAGAGTGGCGA
Powassan  --AAGGGCACCAGTCGTGTAGTAAGAAGGCCCTGGCCCAGTGCGGCAGCACACTCAGTGA
                                                **   ** **         * **

Neudoerfl CGGGAAAATGGTCGATCCCGACGTAGG-GCACTCTGAAAAATTTTGTGAGACCCCCTGCA
263       CGGGAAAATGGTCGATCCCGACGTAGG-GCACTCTGAAAAATTTTGTGAGACCCCCTGCA
Ljubljana CGGGAAAATGGTCGATCCCGACGTAGG-GCACTCTGAAAAACTTTGTGAGACCCCCTGCA
Hypr      CGGGAAAATGGTCGATCCCGACGTAGG-GCACTCTGAAAAACTTTGTGAGACCCCCTGCA
Crimea    CGGGAAAATGGTCGATCCCGACGAAGG-GTACTCTGAAAAATTT-GTGAGACCCCCGGCA
132       CGGGAAAATTGTCGACCCCGACGTAGG-GCATTCTGTCCAATTTCGTGAGGCCCCCTGCA
Aina      CGGGAAAATGGTCGATCCCGACGTAGG-GCACTCTGTAAAACTTTGTGAGACCCCCTGCA
RK1424    CGGGAAAATGGTCGACCCCGACGCAGG-GCATTCTGTCAAACTTTGTGAGACCCCCGGCA
Powassan  CGGGAAAGTGGTCGCTCCCGACGTAACTGGGTAAAAACGAACTTTGTGAGACCAAAAGGC
          ******* * ****  ******* *   *          ** ** ***** **    *

Neudoerfl TCATGATAAGGCCGAACATGGTGCATGAAAGGGGAGGCCCCCGGAAGCACGCTTCCGGGA
263       TCATGATAAGGCCGAACATGGTGCATGAAAGGGGAGGCCCCCGGAAGCTCGCTTCCGGGA
Ljubljana TCATGATAAGGCCGAACATGGTGCATGAAAGGGGAGGCCCCCGGAAGCACGCTTCCGGGA
Hypr      TCATGATAAGGCCGAACATGGTGCATGAAAAGGGAGGCCCCCGGAAGTACGCTTCCGGGA
Crimea    CCATGATAAGGCCGAACATGGTGCAAAAACGGGGAGGCCCCCGGAAGCATGCTTCCGGGA
132       CCATGATAAGGCCGAACATGGTGCAAGATAGGGGAGGCCCCCGGAAGCATGCTTCCGGGA
Aina      TCATGACAAGGCCTAACATGATGCACGAA-GGGGAGGCCCCCGGAAGCATGCTTCCGGGA
RK1424    CCATGATAAGGCCGAACATGGTGCAAGAACGGGGAGGCCCCCGGAAGCATGCTTCCGGGA
Powassan  CTCCTGGAAGGCTCACCA-GGAGTTAGGCCGTTTAGGAGCCCCCGAGCATAACTC-GGGA
          5      *****  * ** *5 *           ***  ***   **      ** ****

Neudoerfl GGAGGGAAGAGAGAAATTGGCAGCTCTCTTCAGGATTTTTCCTCCTCCTATACAAAATTC
263       GGAGGGAAGAGAGAAATTGGCAGCTCTCTTCAGGATTTTTCCTCCTCCTATACAAAATTC
Ljubljana GGAGGGAAGAGAGAAATTGGCAGCTCTCTTCAGGATTTTTCCTCCTCCTATACAAAATTC
Hypr      GGAGGGAAGAGAGAAATTGGCAGCTCTCTTCAGGATTTTTCCTCCTCCTATACAAAATTC
Crimea    GGAGGGAAGAGAGAAATTGGCAACTCTCTTCGGGATTTTTCCTCCTCCTATACCAAATTC
132       GGAGGGAAGAGAGAAATTGGCAACTCTCTTCGGGATTTTTCCTCCTCCTATACCAAATTC
Aina      GGAGGGAAGAGAGAAATTGGCAGCTCCCTTCAGGATTTTTCCTCCTCCTATACTAAATTC
RK1424    GGAGGGAAGAGAGAAATTGGCAACTCTCTTCGGGATTTTTCCTCCTCCTATACAAAATTC
Powassan  GGAGGGAGGAAGAAAATTGGCAATCTTCCTCGGGATTTTTCCGCCTCCTATACTAAATTT
          *******3**67 *********  76**3** ********** ********** *****

Neudoerfl CCCCTCGGTAGA-GGGGGGGCGGTTCTTGTTCTCCCTGAGCCACCATCACCCAGACACAG
263       CCCCTCGGTAGA-GGGGGGGCGGTTCTTGTTCTCCCTGAGCCACCATCACCCAGACACAG
Ljubljana CCCCTCGGCAGA-GGGGGGGCGGTTCTTGTTCTCCCTGAGCCACCATCACCCAGACACAG
Hypr      CCCCTCGGTAGA-GGGGGGGCGGTTCTTGTTCTCCCTGAGCCACCATCACCCAGACACAG
Crimea    CCCCTCAATAGA-GGGGGGGCGGTTCTTGTTCTCCCTGAGCCACCATCACCCAGACACAG
132       CCCCTCAATAGA-GGGGGGGCGGTTCTTGTTCTCCCTGAGCCACCATCACCCAGACACAG
Aina      CCCCTCAACAGA-GGGGGGGCGGTTCTTGTTCTCCCTGAGCCACCATCACCCAGACACAG
RK1424    CCCCTCAATAGA-GGGGGGGCGGTTCTTGTTCTCCCTGAGCCACCATCACCCAGACACAG
Powassan  CCCCCAGGAAACTGGGGGGGCGGTTCTTGTTCTCCCTGAGCCACCACCATCCAGGCACAG
          ****     *   **********************************a**b****c*****

Neudoerfl GTAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT
263       ATAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT
Ljubljana ATAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT
Hypr      ATAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT
Crimea    ATAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT
132       ATAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT
Aina      ATAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT
RK1424    ATAGTCTGACAAGGAGGTGATGTGTGACTCGGAAAAACACCCGCT
Powassan  ATAGCCTGACAAGGAGATGGTGTGTGACTCGGAAAAACACCCGCT
           ***c***********b**a************************
```

Table 10. Alignment of the MFE structures in bracket notation.
Compensatory mutations are indicated by numbers and letters, respectively.

```
Neudoerfl --...(((((.......)))))(((((((((.(((((.......(((((.......))))).
263       --...(((((.......)))))(((((((((.(((((.......(((((.......))))).
Ljubljana --...(((((.......)))))(((((((((.(((((.......(((((.......))))).
Hypr      --...(((((.......)))))(((((((((.(((((.......(((((.......))))).
Crimea    -...((((((.........))))))..((((....))))...(((((.......))))).
132       -.....................(((((-.((((((((((.....))).))).))))))...)
Aina      ((((.((((((..-..))))))(((((((((.(((((((....(((((.......))))).
RK1424    --...........(((((....((((((((((...((.....(((((.....)))))).
Powassan  --.(((((.....................)))))..((((((((((....(((.....))).

Neudoerfl (((((..........)))))..)))))..)-)))))))...(((((((((...(((((((((
263       (((((..........)))))..)))))..)-)))))))...(((((((((...(((((((((
Ljubljana (((((..........)))))..)))))..)-)))))))......((((((...(((((((((
Hypr      (((((..........)))))..)))))..)-)))))))......((((((.(.(((((((((
Crimea    (((((..........)))))......(((-(.(((((((...))))-.)))))))...(((
132       ))))).....(((((....)))))..(((-((...)))))..........((((((((((
Aina      .((((..........)))))))))..)))-)))))))....((...))..))))...((((
RK1424    ((((............))))..))))))-)))))))........)))).....(((((((((
Powassan  (((((..........)))))..))))..)))))))....((((...))))....(((((.(((

Neudoerfl (((((..........)))))))))).....))))..)...(((((((((....)))))))))(
263       (((((..........)))))))))).....))))..)...(((((((((....)))))))))(
Ljubljana (((((..........)))))))))).....))))..)...(((((((((....)))))))))(
Hypr      (((((..........)))))))))......))).)))..(((((((((....)))))))))(
Crimea    (((((..........)))))))).......(((((((((((((((....)))))))).
132       (((((..........)))))))))......))).))))(((((((((....)))))))))(
Aina      (((((..........)))))))))....-.((((((..(((((((((....)))))))))).
RK1424    (((((..........)))))))).....)))))......(((((((((....)))))))))(
Powassan  (((((((........)))-)))....))))..)))..))....((((((.....)))-))))(
          5                 5

Neudoerfl ((((((((((((.........))))))))).(((....))).))))))..)))))))))(
263       ((((((((((((.........))))))))).(((....))).))))))..)))))))))(
Ljubljana ((((((((((((.........))))))))).(((....))).))))))..))))))..((
Hypr      ((((((((((((.........))))))))).(((....))).))))))..))))))..((
Crimea    ))...((((((((.........))))))))).(((....))).))))))..........((
132       ((((((((((((.........))))))))).(((....))).))))))..........((
Aina      ((((((.................)))))))).(((....))).))))))..........((
RK1424    ((((((((((((.........))))))))).(((....))).))))))..........((
Powassan  ((((((((((((.........))))))))).(((....))).))))))..........(
            3  67           76  3

Neudoerfl ((((((....))-)))))..((((....((((.(((..(((.(((((((((((((((....
263       ((((((....))-)))))..((((....((((.(((..(((.(((((((((((((((....
Ljubljana ((((((....))-))))))((((....((((.(((..(((.(((((((((((((((....
Hypr      ((((((....))-))))))((((....((((.(((..(((.(((((((((((((((....
Crimea    ((((((....))-))))))((((....((((.(((..(((.(((((((((((((((....
132       ((((((....))-))))))((((....((((.(((..(((.(((((((((((((((....
Aina      ((((((....))-))))))((((....((((.(((..(((.(((((((((((((((....
RK1424    ((((((....))-))))))((((....((((.(((..(((.(((((((((((((((....
Powassan  (((((((....)))))))))((((....((((.(((..(((.(((((((((((((((....
                                                        a  b    c

Neudoerfl ...)))))).......)))))))..)))..))))))..))))..))))).
263       ...)))))).......)))))))..)))..))))))..))))..))))).
Ljubljana ...)))))).......)))))))..)))..))))))..))))..))))).
Hypr      ...)))))).......)))))))..)))..))))))..))))..))))).
Crimea    ...)))))).......)))))))..)))..))))))..))))..))))).
132       ...)))))).......)))))))..)))..))))))..))))..))))).
Aina      ...)))))).......)))))))..)))..))))))..))))..))))).
RK1424    ...)))))).......)))))))..)))..))))))..))))..))))).
Powassan  ...)))))).......)))))))..)))..))))))..))))..))))).
              c           b  a
```

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 +6.66kcal/mol, more than ten times the thermal energy.

A statistical study of the distribution of secondary structures over the space of possible sequences of fixed length [70] shows that

(i) many different sequences do fold into the same MFE structure, and

(ii) sequences with identical structure form a network of paths along which structure-neutral sequences are separated by one or at most two point mutations.

This permits populations of sequences to split and drift apart from one another in sequence space without changing their dominant phenotype [48]. 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.

Even further towards the 5'end we find ample evidence for a conserved Y-shaped motif consisting of some 90 nucleotides, see Figure 11. 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 14 compares $d(k)$ for the *Eastern* and *Western* 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 coworkers [71].

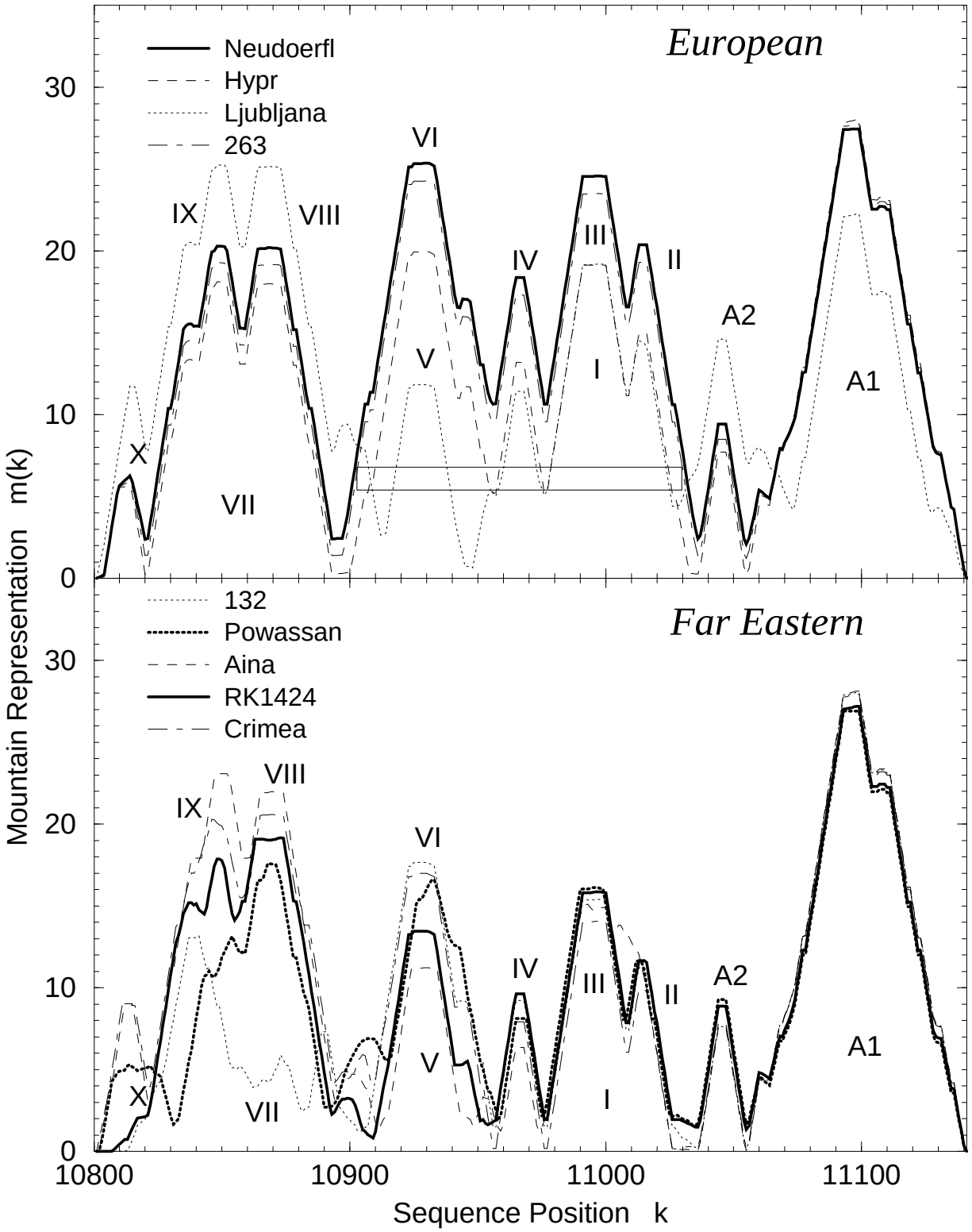


Figure 13: Mountain representations of the 3'NCR of *TBE viruses*.
Nucleotide positions are indicated for TBE virus strain *Neudörfl*. The stem discriminating *Far Eastern* from *European* subtype sequences is indicated in the upper plot.

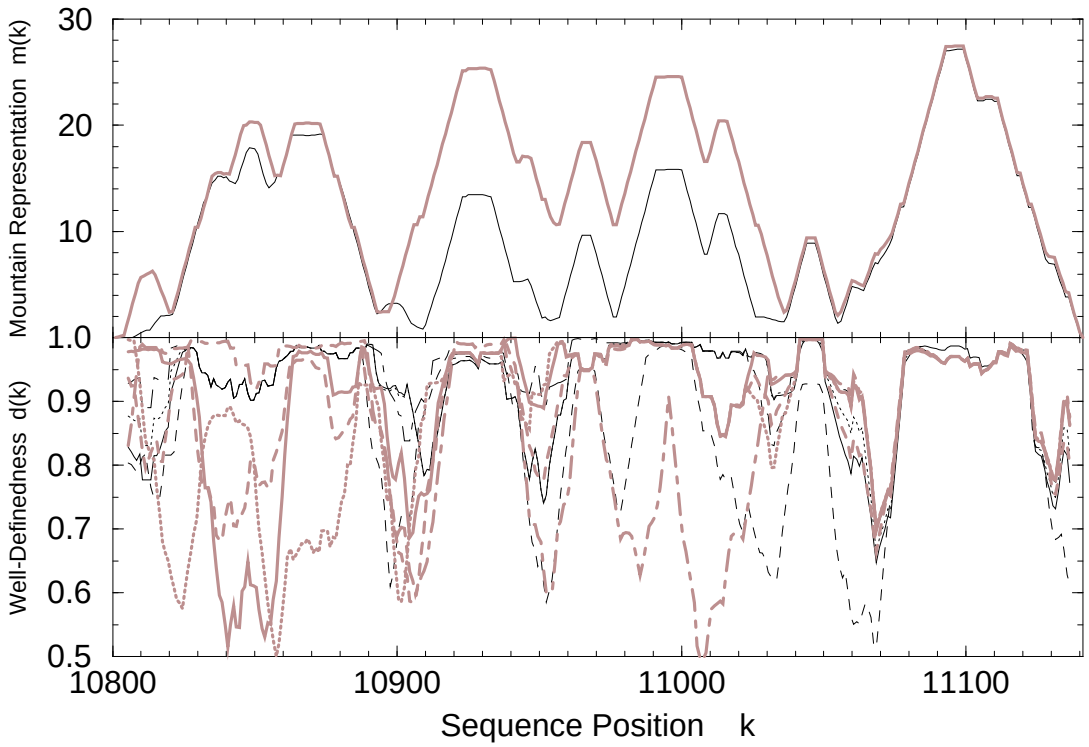


Figure 14: Well-definedness of the 3'UTR.
The plot shows running averages of $d(k)$ taken over 10 nucleotides. *European* subtype sequences are shown in gray: solid: *Neudörfl*; dotted: *263*; dashed: *Ljubljana*; dot-dashed: *Hypr*.
Far Eastern subtype sequences are shown in black: solid: *Crimea*; dotted: *132*; dashed: *RK1424*; dot-dashed: *Aina*.
In the upper part of the figure we give the mountain representation of *Neudörfl* and *RK1424* strain sequences for comparison. 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 11070 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 10800 and 10900 comprising the stems VII, VIII, and IX, where the *European* subtype sequences seem to be much more flexible than *Far Eastern* subtype.

It is interesting to note that *European* subtype sequences are substantially less well-defined in the region [VII,VIII,IX] between sequence position 10800 and 10900 than *Far Eastern* subtype. This is the most significant distinction between *Far*

Table 11. Folding Energies (in kcal/mol) for 9 *flavivirus* sequences. ΔE(far east) and ΔE(europ) are the energy differences to the MFE 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)
Neudörfl	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

Eastern and *European* subtype, where the *European* subtype secondary structures seem to be much more flexible than *Far Eastern* subtype.

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 15 is also the final outcome of the comparative partition function method. The regions VII through X, on the other hand show both a rather large variance and a small well-definedness.

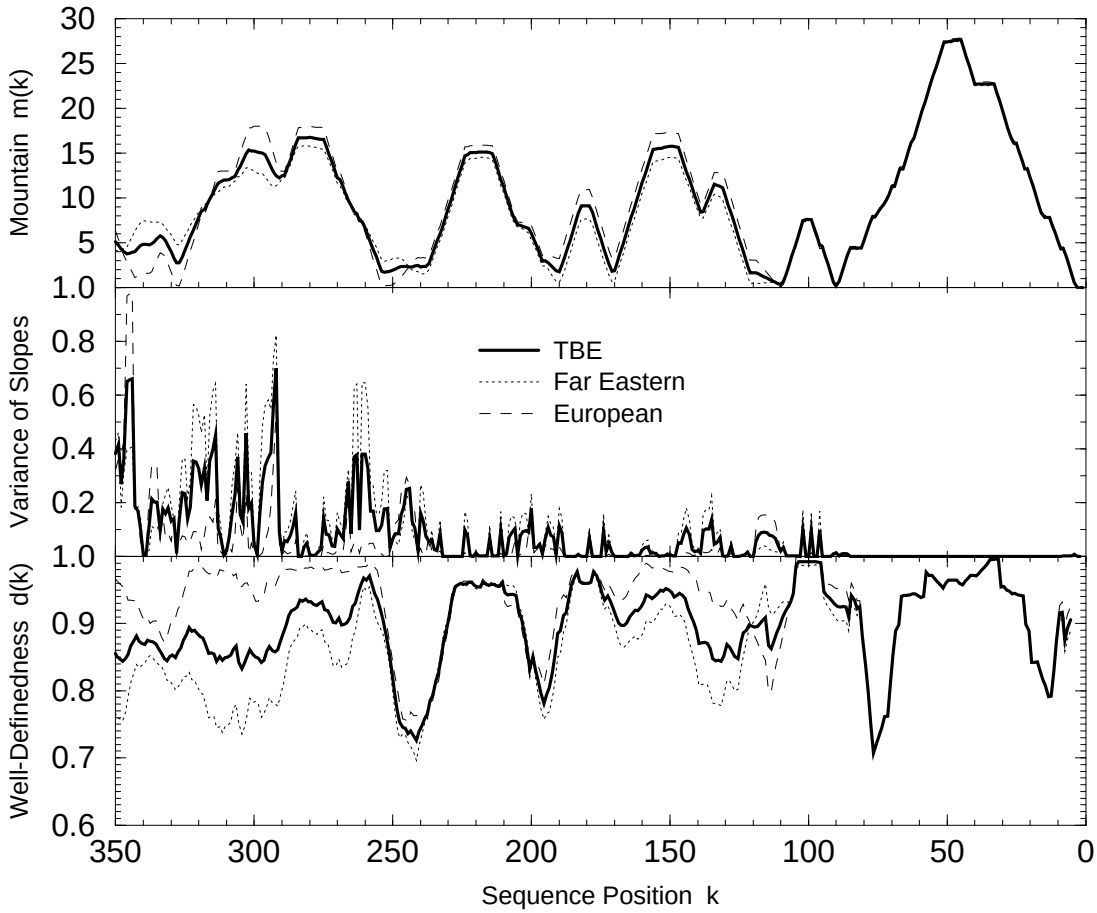


Figure 15: Well-definedness and conservedness of the 3'NCR secondary structure of *TBE virus*.
The upper part of the figure compares the mountain representation of the consensus of all *TBE* sequences (Table 8) with the consensus of the *Far Eastern* and *European* subtype structures, respectively. Sequence 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* 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 *Neudörfl* 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.

4.3. Other Flavivirus Serocomplexes

In this section we shall be concerned with the structural features that are shared among the members of the other *flavivirus* serocomplexes. An attempt to refine the classification based on predicted structures, however, goes beyond the focus of this work.

The analysis of the 18 *YF virus* sequences are summarized in Figure 16. The structural domain A at the 3'end, first described in [35], 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 downstream of 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 a "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 Figures 17 and 18. Consistent with previous findings [7, 38] we characterize a well defined and well conserved stem-loop structure at their 3'end. In the cases of DEN viruses, however, the stem is shorter than for other *flaviviruses*.

A particularly interesting feature is the T-shaped element B which is shared by *JE* and *DEN viruses* but apparently is not conserved in *TBE* or *YF viruses*. A large number of compensatory mutations confirms this structural element, see Figure 17 and 18. It occurs at different genomic positions in *DEN* and *JE* sequences, and includes the highly conserved, mosquito-borne *flavivirus*-specific sequence elements *CS2* and *RCS2* [38]. 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 virus*; *RCS2* is absent in this group of viral sequences.

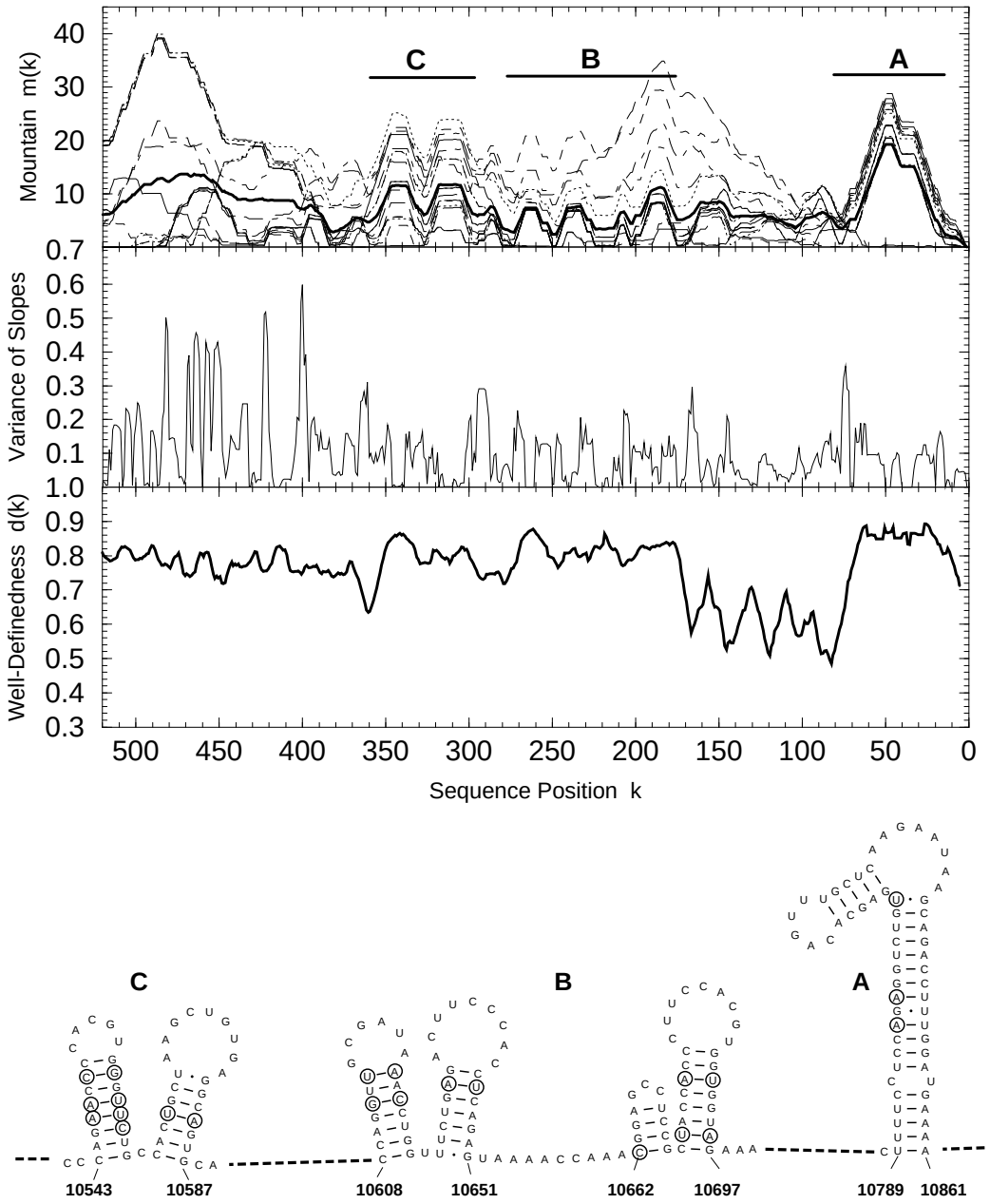


Figure 16: Secondary structure of the *YF virus* 3'NCR.

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.

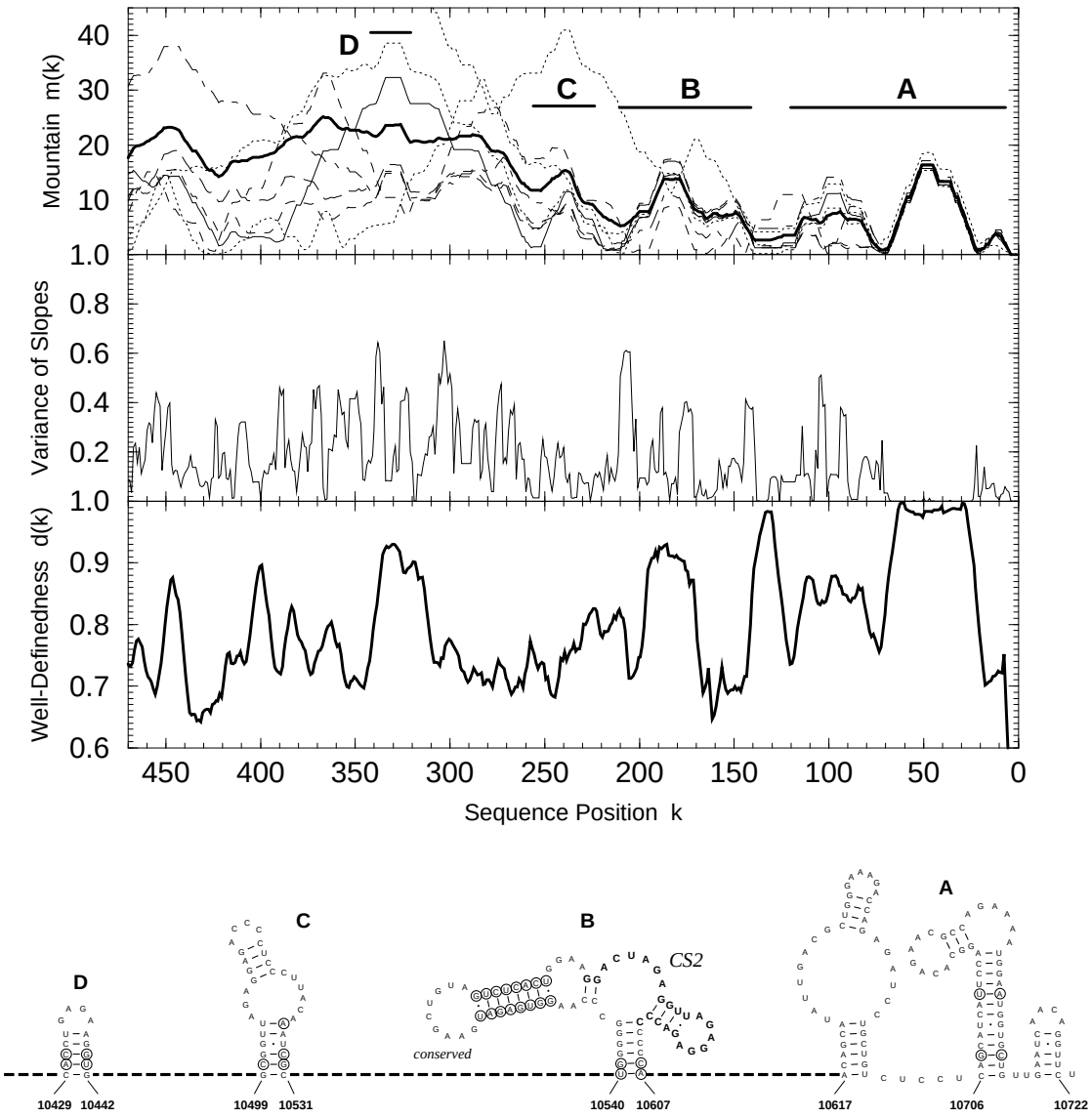


Figure 17: Conserved secondary structures in the *DEN* serocomplex. We show the superposition of the mountain representations and the conventional diagrams of conserved consensus structures. A large number of compensatory mutations (indicated by circles) confirm the proposed structures. Note that element 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 hinting at its functional importance. The conserved box *CS2* is indicated in bold.

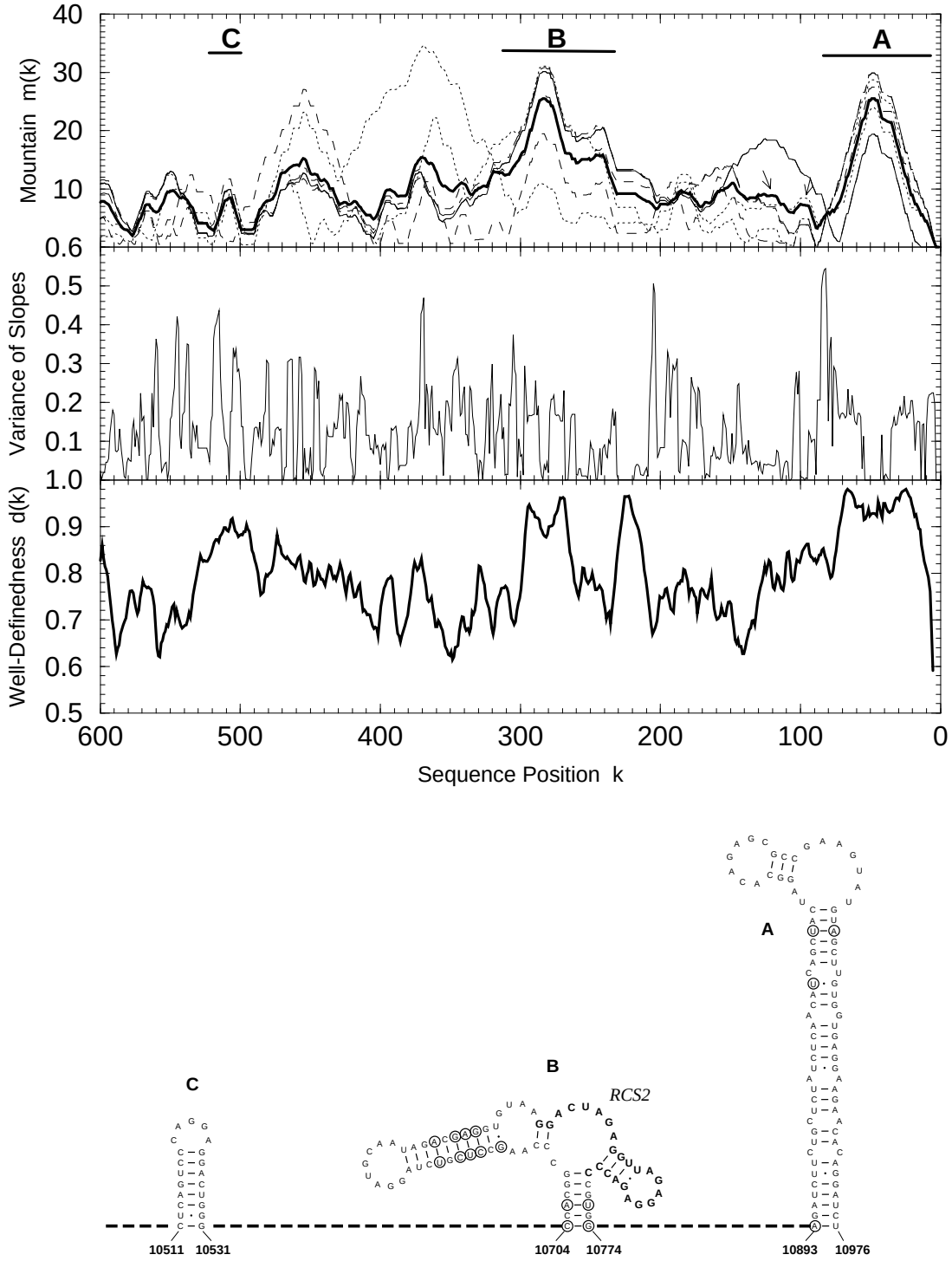


Figure 18: Conserved secondary structures in *JE* serocomplex. We show the superposition of the mountain representations and the conventional diagrams of conserved consensus structures. Sequences that are consistently unpaired but occur in variable structural contexts are indicated by arrows. A large number of compensatory mutations (indicated by circles) confirm 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 conserved box *RCS2* is shown in bold.

4.4. Pseudoknot Structure at the 3'Terminus of the Genomic RNA

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 [92].

(iii) Pseudoknots can in many cases be understood as an additional feature that is formed on top of the conventional secondary structure.

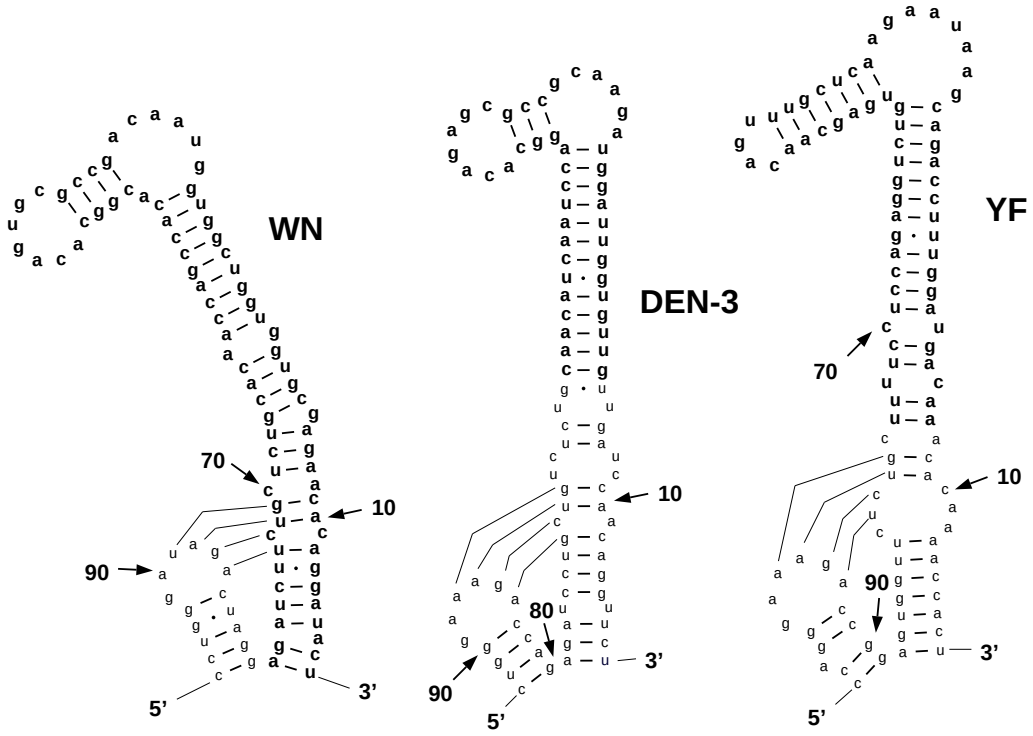


Figure 19: Computer-predicted secondary structures for the 3'-terminal genomic RNAs of mosquito-borne *flaviviruses* by Brinton and coworkers [71]. In spite of the sequence divergence among *flaviviruses*, similar secondary structures were predicted for the 3'-terminal sequences of each of the genomic RNAs, a large 3' hairpin (A1) and a small 5' hairpin (A2). The sequence within the loop of A2 is highly conserved. Conserved features that are consistent with our findings in the previous section are shown in bold.

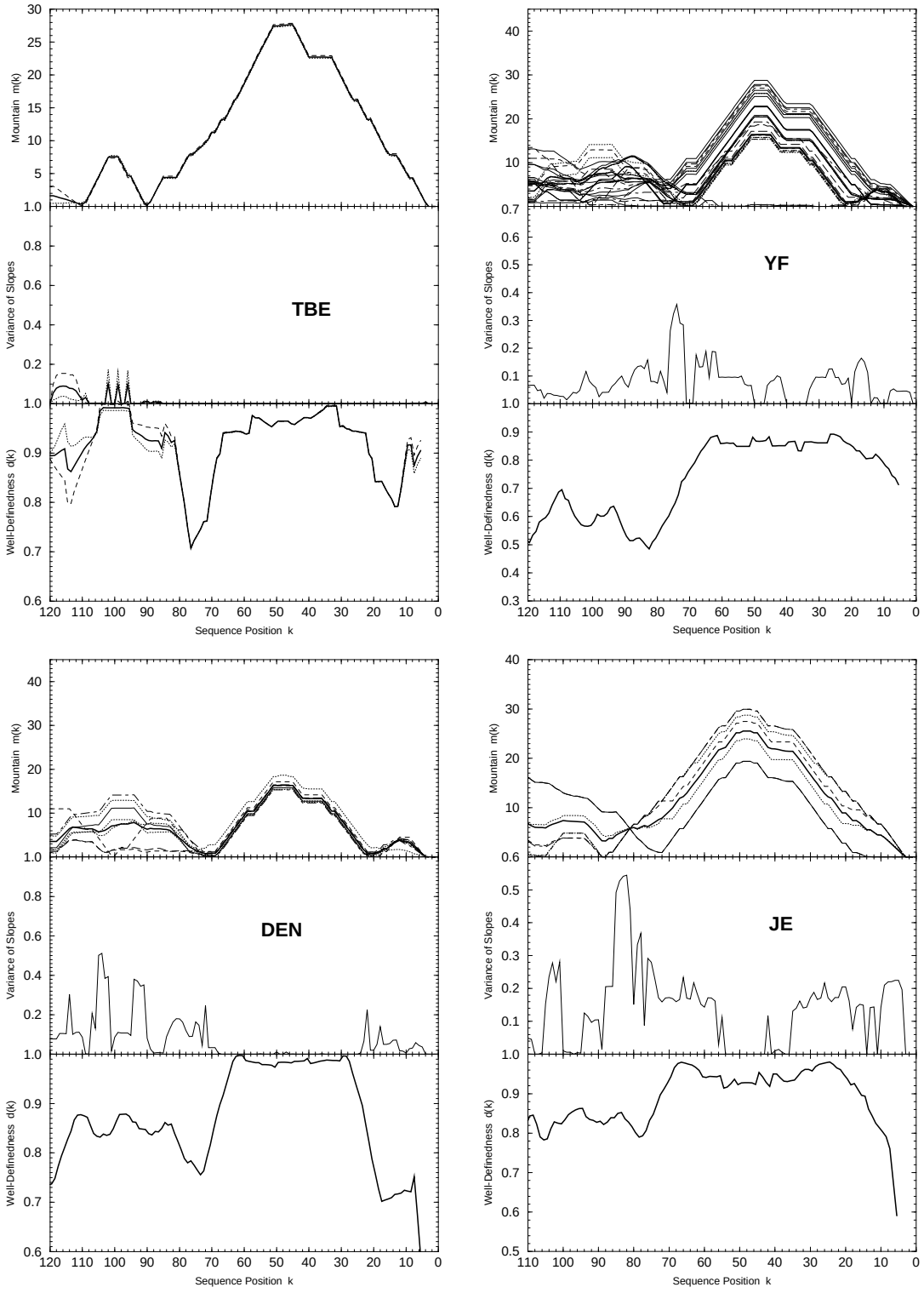


Figure 20: Well-definedness and conservedness of the secondary structure motifs A1 and A2 of the *flaviviruses TBE, YF, DEN,* and *JE.*
The plots show the ill-defined regions corresponding to the pseudoknotted structure between A1 and A2.

We found that the secondary structure on the 5' side of A1 (around nucleotide 11070 in *Neudörfl* strain) is ill-defined in all sequences, see Figure 20. Circular dichroism spectra and ribonuclease probing suggested that base pairing occurs between the two 3'-terminal secondary structures A1 and A2 of *flaviviruses* revealing that tertiary structures could be formed by interactions between them [71]. Figure 19 shows the pseudoknotted structure formed by A2 and A1 of *WN*, *DEN-3*, and *YF virus* as discussed in detail by Brinton and coworkers.

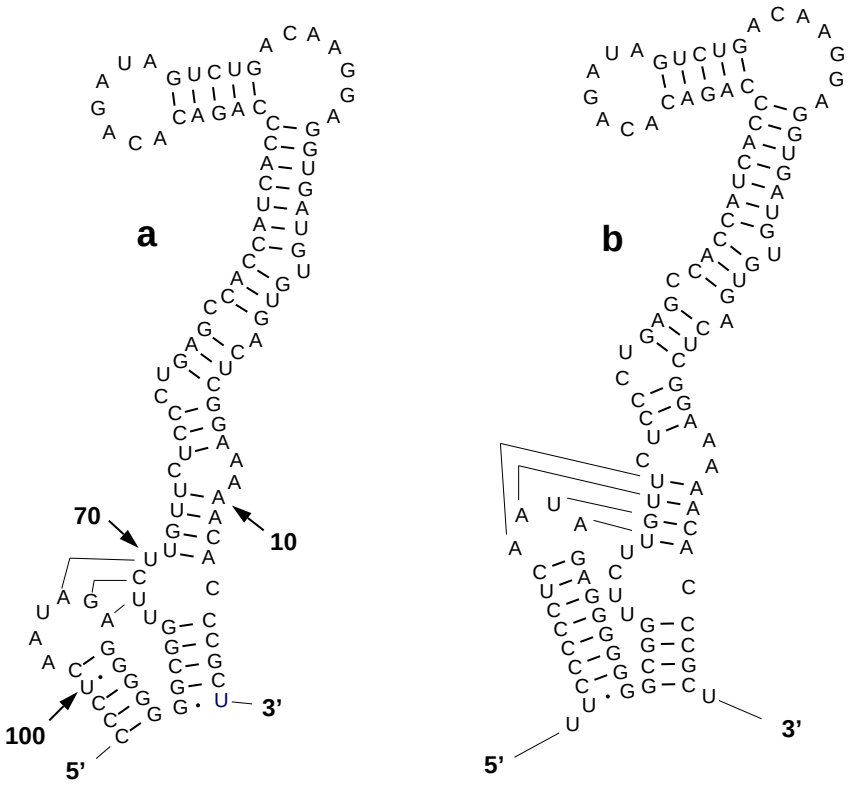


Figure 21: Possible pseudoknot tertiary interaction between 5' side of A1 and the loop of A2 of the *Neudörfl* strain. The structure according to the findings of Brinton and coworkers is given in **a**, whereas the MFE structure of the same element is shown in **b**. The energy difference between these two structures is 6.90kcal/mol.

The ill-defined region around nucleotide 11070 corresponds to a possible pseudoknot structure formed between an unstable region on the 5'side of A1 and some nucleotides in the conserved loop of A2 of *TBE viruses*, shown in Figure 21. `RNAfold` excludes pseudoknots as all folding algorithms based upon dynamic programming. The secondary structure can be represented as a weighted average of the base pair

probabilities. Potential pseudoknots that compete with other secondary structure elements appear as ill-defined regions in well-definedness plots, despite the fact that the computational model does not make explicit use of pseudoknots. Nevertheless, this result supports the existence of a pseudoknot structure.

Possible pseudoknot tertiary interaction formed between an unstable region on the 5' side of A1 and some nucleotides in the conserved loop of A2 of *TBE virus* are shown in Figure 21.

Folding of the MFE structure of complete viral genome

Predicting a single structure by any approach will in general not provide a reliable answer. However, MFE folding of the entire genomic RNA of *TBE viruses* confirms the overall shape and sequence position of the conserved secondary structure features found in our analysis, see Figure 22.

Large RNA molecules decompose into components, that is, into continuous sequence pieces that form base pairs only inside themselves and are not interior to any other base pair. At the components boundaries $m(k)$ drops to zero. By calculation of the MFE structure, they can be folded indepently from each other. This is important for the secondary structure prediction by folding part of the complete molecule. To make sure that the selected portion folds into a distinct unit it is necessary to calculate the folding of the entire sequence.

The MFE algorithm allows to fold structures including constraints, that is, to force base pairs or to prevent certain bases from pairing. This will be executed by honoring constraints with bonus energies. The MFE structure of strain *Neudörfl* genome base pairs across the entire sequence ([nts 4-26] to [nts 11055-11077]) and forms a huge multiloop. The formation of this multiloop can be prevented by introducing constraints. We chose a base that is unpaired in the consensus secondary structure of the 3'NCR and require that it remains unpaired in the complete genomic structure. Introducing such constraints into the MFE folding is a built-in feature of the `Vienna RNA Package`.

The calculated energy difference between the panhandle and the "open" conformation is 6.06kcal/mol (0.2% of the MFE). As the multiloop energy parameter are

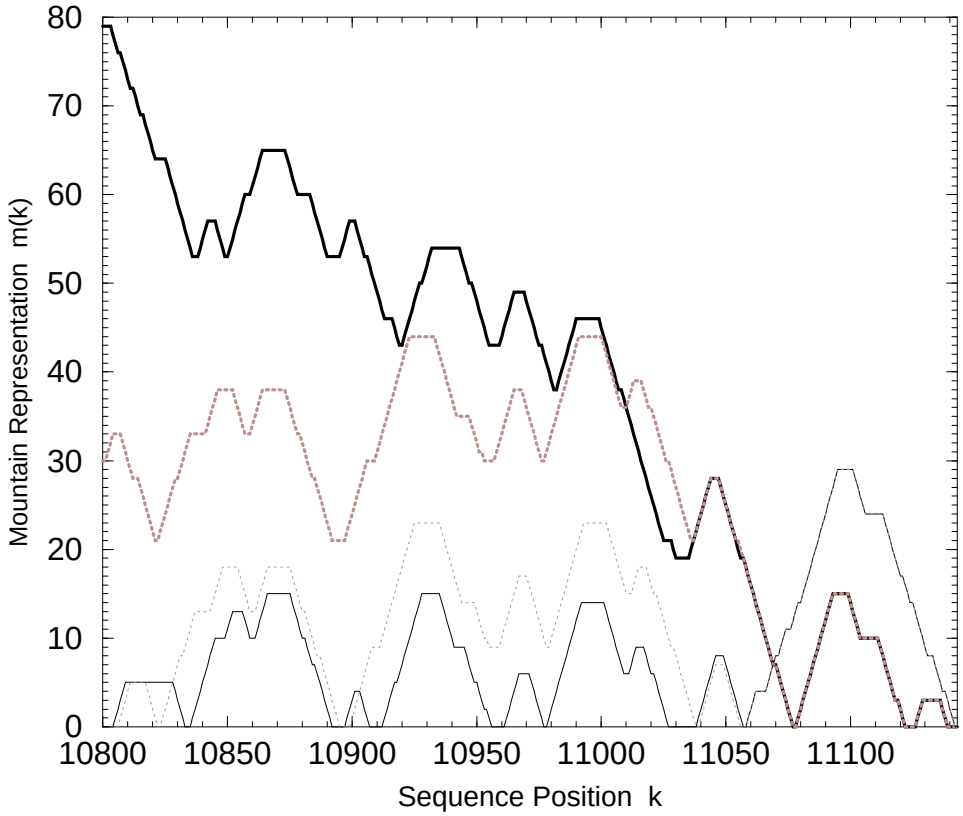


Figure 22: MFE structure of A1 and A2.
POW structures are shown in black; *Neudoerfl* structures are shown in gray. Bold: MFE structure when calculating the entire genomic RNA; thin line: MFE structure when folding the 3' terminus only.

rather crude approximations (see the discussion in the next chapter) the difference is most likely not significant. Hahn discussed the functional importance of a conserved box *CS1* for the circularization of mosquito-borne *flaviviruses* genomes [38]. At the 3'end of tick-borne *flaviviruses* genomes a conserved box PR which has an inverted counterpart at the 5'end was found by Mandl [59]. Nevertheless, the predictability of the long-range base pairing (>200nts) decreases with increasing distance. Computation of the partition function, unfortunately is beyond the capability of the available computer equipment.

5. Influenza Viruses

5.1. Orthomyxoviridae

The family *Orthomyxoviridae* includes three genera: *influenza A* and *B viruses*, *influenza C virus*, and *Thogoto*-like viruses. The *influenza A, B* and *C viruses* can be distinguished on the basis of antigenic differences between their nucleocapsid (NP) and matrix (M) proteins. *Influenza A* and *B viruses* each contain eight distinct RNA segments, see Figure 23, whereas *influenza C viruses* contain seven RNA segments.

The family *Orthomyxoviridae* comprises enveloped viruses with a segmented, single-stranded RNA (ssRNA) genome that has been termed negative-stranded because the viral mRNAs are transcribed from the viral RNA segments; by convention mRNA is plus-stranded. The genomic RNA of negative-strand RNA viruses has to serve as a template for synthesis of the antigenome plus-strand. Negative-strand RNA viruses encode and package their own RNA-dependent RNA transcriptase, but mRNAs are synthesized only after the virus has been uncoated in the infected cell. Viral replication occurs after synthesis of the mRNAs and requires synthesis of viral proteins. The newly synthesized antigenome plus-strand (for *influenza virus* often termed template RNA or cRNA) serves as the template for further copies of the minus-strand genomic RNA. Among the RNA viruses, *influenza virus* is very special in that all of its RNA synthesis - transcription and replication - takes place in the nucleus of the infected cell. The nucleus provides the environment for the synthesis of *influenza virus* mRNAs in an unusual process, as initiation requires mGpppXm-containing capped primers that are generated from a subset of host cell RNAs by an *influenza virus*-encoded cap-dependent endonuclease. In addition to stealing caps, *influenza virus* mRNAs make use of another aspect of host cell nuclear function, namely, the splicing machinery. *Influenza virus* mRNA transcripts provide the only known example of splicing of RNA that is not transcribed from DNA by RNA polymerase II. The *influenza viruses* provide some remarkable examples of genome diversity: spliced mRNAs and overlapping reading frames, bicistronic mRNAs and coupled translation of tandem cistrons. The variety of mechanisms used for the synthesis of proteins by *influenza virus* provides a paradigm of successful exploitation of a genome [22].

Within the envelope of *influenza A* and *B viruses* there are eight segments of single-stranded RNA (ranging from 2,341 to 890 nucleotides) contained in the form of an ribonucleoprotein (RNP). Associated with the RNPs are small amounts of the transcriptase complex, consisting of the proteins PP_1, PB_2 and PA.

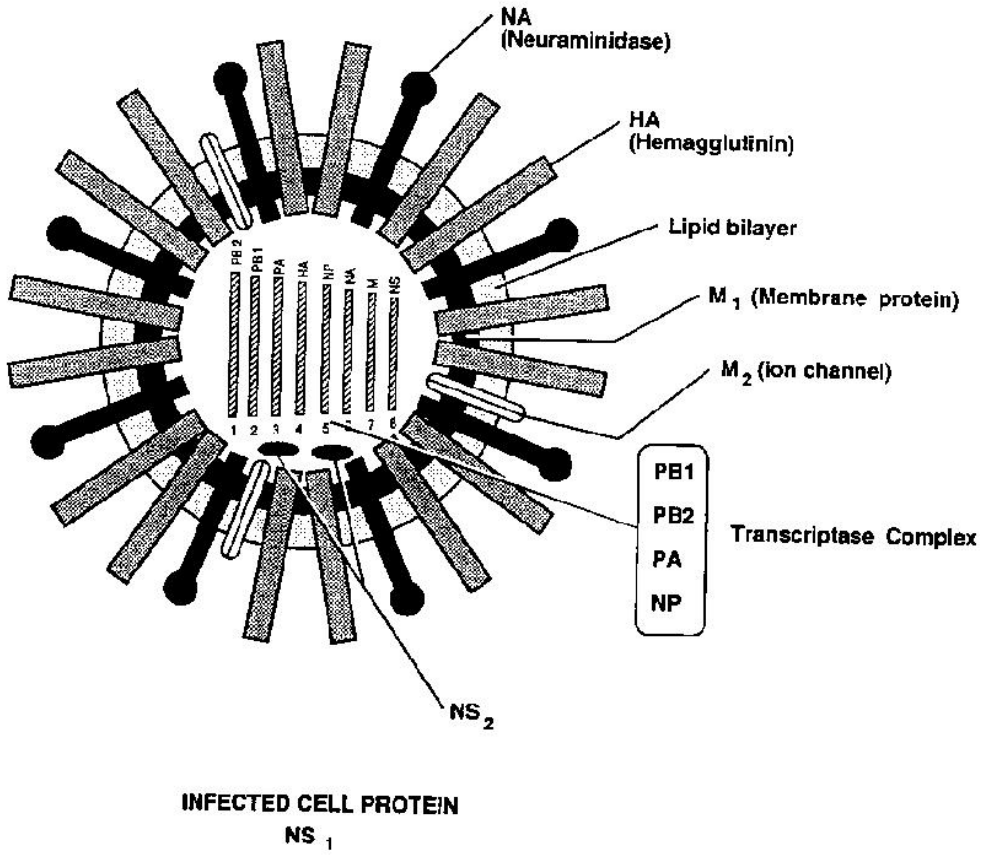


Figure 23: Structure of the *influenza A virus* paricle. Three types of integral membrane protein - hemagglutinin (HA), neuramidase (NA) and small amounts of the M_2 ion channel protein - are inserted through the lipid bilayer of the viral membrane. Within the envelope there are eight segments of single-stranded genome RNA, adapted from [64].

The gene assignment for *influenza A virus* is as follows: RNA segment 1 codes for PB_2, 2 for PB_1, 3 for PA, 4 for HA, 5 for NP, 6 for NA, 7 for M_1 and M_2, 8 for NS_1 and NS_2. *Influenza B virus* does not encode an M_2 integral membrane protein, but NB glycoprotein encoded by RNA segment 6 (which also encodes NA) is of similar structure to M_2.

5.2. The Panhandle Structure

The first nucleotides at the 3' and 5'end of each virion RNA (vRNA) segment are conserved and inverted complementary in all eight RNA segments in isolated virions and infected cells. The number of nucleotides conserved at the 3' and 5'ends are 12 and 13nts, respectively, for *influenza A virus*. In *influenza B virus*, however, only 9 and 10nts are absolutely conserved at the 3' and the 5' ends, respectively. Both influenza A and B virion RNAs (vRNAs) can form a panhandle structure by partial base pairing between the 3'- and 5'-terminal sequences [12], see Figure 24.

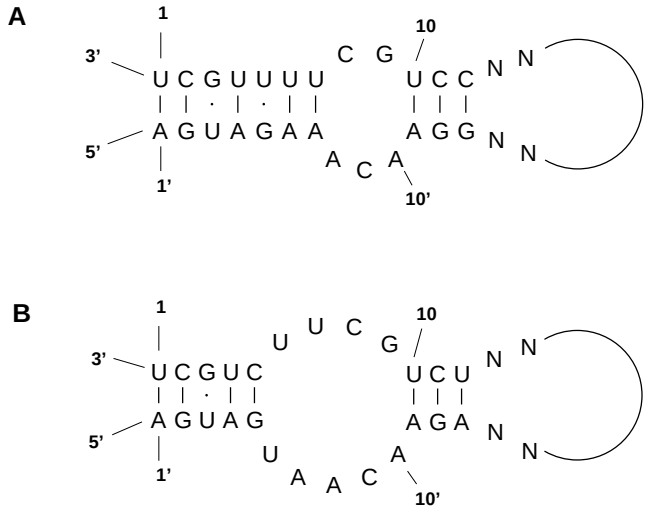


Figure 24: Sequences and potential panhandle structures of the *influenza A* and *B* vRNAs. Both *influenza A* and *B* vRNAs can form a panhandle structure by partial base pairing between the 3'- and 5'-terminal sequences.

The panhandle structure can be divided into two stems, which are joined together by a loop structure. It has been suggested, that the panhandle structure can serve as a regulatory signal for transcription and replication, as well as for packaging of RNA into virus particles [24, 56, 23]. There is experimental evidence that the panhandle is a *cis*-acting signal for polyadenylation, and a recent study indicates that it may be needed for the endonuclease activity of the RNA polymerase complex [57]. Forming the panhandle structure implies that the RNA forms a large multiloop in most cases. Four sequences formed a rod-like structure that does

not contain a multiloop. Multiloops are not very accurately parametized in the standard energy model [52]. They are approximated by the linear function:

$$E_{ml} = a + b(deg - 1) + c * N = 4.60 + 0.10(deg - 1) + 0.40 * N \qquad (6)$$

The energy differences (ΔE) to the MFE structure when the segment 4 is forced to form the panhandle structure is given in Table 12 for *influenza A* and in Table 13 for *influenza B viruses*.

Table 12. Folding Energies (in kcal/mol) for *influenza A virus* genomes
Sequences that form a rod-like structure without a multiloop are indicated by - in the last column.

	GenBank	E_{mfe}	E_{ph}	ΔE	E_{ml}
EIVH3A	L27597	-265.21	-244.78	20.43	9.50
FLAHA3055J	L20406	-265.42	-251.44	13.98	7.60
FLAHA575RI	L20409	-265.13	-251.15	13.98	7.60
FLAHA7N1	M24457	-260.75	-248.14	12.61	7.30
FLAHAEM	M29257	-252.75	-236.05	16.70	-
FLAHAH3A	L39913	-268.15	-249.50	18.65	-
FLAHAJ3055	L20407	-264.37	-250.39	13.98	7.60
FLAHARI557	L20408	-262.48	-248.50	13.98	7.60
FLAHAS157A	L20410	-264.10	-250.12	13.98	7.60
FLAHATS1	M24458	-264.87	-251.61	13.26	7.30
FLAN2HA	M73771	-274.65	-259.59	15.06	6.10
FLAN2HAC	M73774	-297.23	-279.49	17.74	6.40
FLAN2HAD	M73775	-302.62	-279.98	22.64	6.10
FLAN2NAE	M73776	-271.50	-260.49	11.01	5.20
FLAN8HAA	M73772	-301.35	-285.74	15.61	9.10
FLAN8HAB	M73773	-266.95	-250.31	16.64	8.50
IVU02085	U02085	-248.30	-231.21	17.09	-
IVU02464	U02464	-250.24	-232.42	17.82	-
ORA77HA	X05907	-306.00	-285.66	20.34	9.80
ORINF1	M55060	-305.34	-287.12	18.22	6.80
S67220	S67220	-260.74	-250.62	10.12	7.00

Table 13. Folding Energies (in kcal/mol) for *influenza B virus* genomes

	GenBank	E_{mfe}	E_{ph}	ΔE	E_{ml}
FLBHAOC	K02713	-278.65	-270.29	8.36	6.00
FLBHAZO	K00423	-279.52	-260.51	19.01	6.10

The calculated energy differences between the panhandle structure and the thermodynamic optimal fold are significant, and about double the penalty energies for multiloops (E_{ml}). The degree of a multiloop (deg) is the number of stacks and N is the number of unpaired bases within the multiloop (a=4.60kcal/mol is the parameter for the closing base pair). The formation of the panhandle must be stabilized by proteins or kinetically controlled. The higher free energy is not due to the formation of a multiloop, some sequences (FLAHAEM, FLAHAH3A, IVU02085, IVU02464) do not even form a multiloop. RNA folding is not only influenced by thermodynamic, but also by kinetic and other factors.

Calculating the MFE structure and the base pairing probability matrix of *influenza virus* sequment 4 (which codes for hemagglutinin), we could not find conserved secondary structure elements which seem to be of functional significance, see Figure 25. One could speculate that this is due to the fact that they are negative-strand viruses which steal the caps from their host and make use of other aspect of host cell nuclear functions. A possible explanation could be that they do not need secondary structure features which must be recognized by proteins.

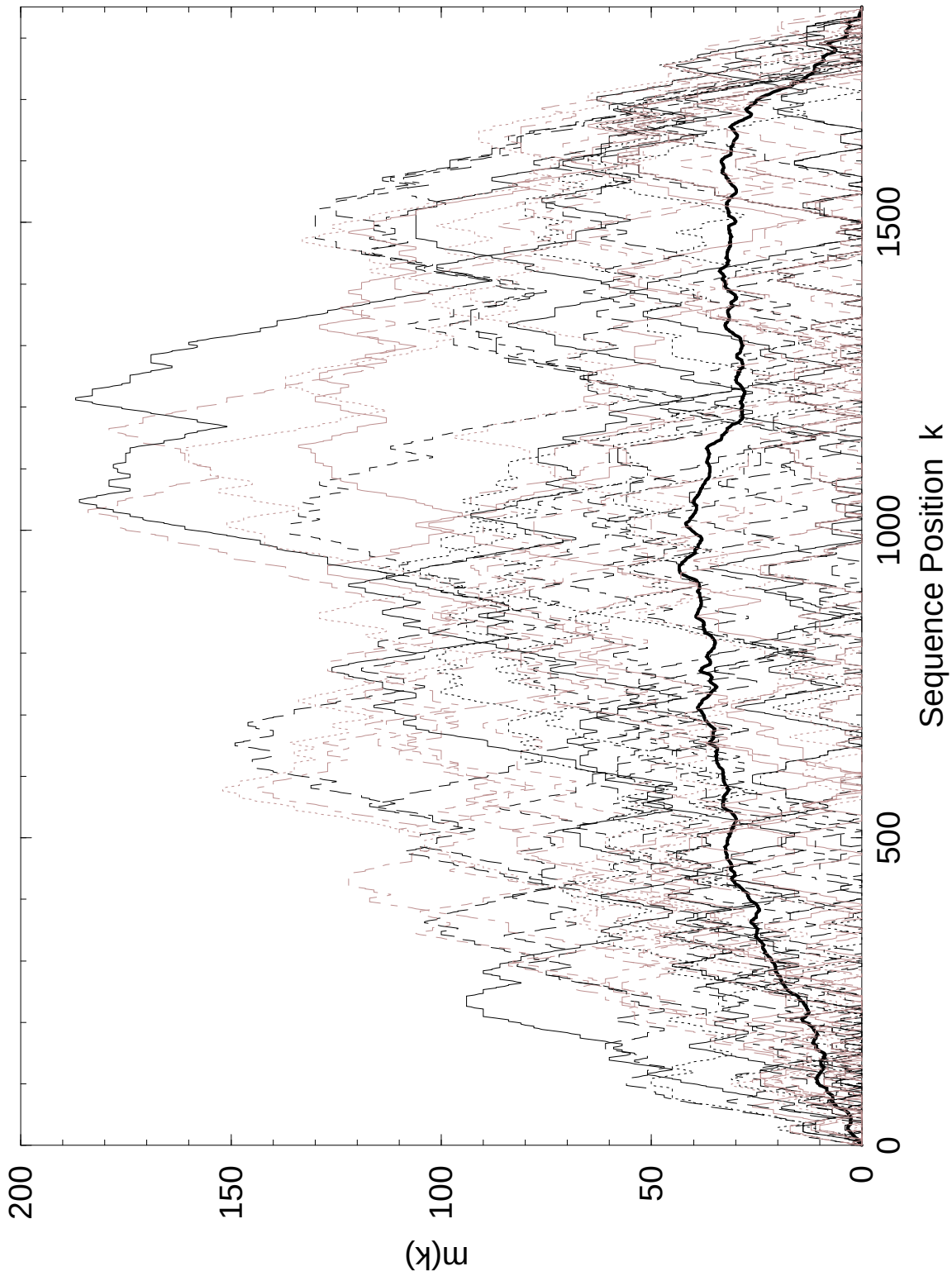


Figure 25: Mountain representation of *influenza A virus* segment 4 sequences. The consensus mountain is shown in bold.

6. Conclusion and Outlook

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* [2, 44], *RNA phages* [3, 65], *flaviviruses* [72], *pestiviruses* [8, 11] and *hepatitis C viruses* [78, 8].

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 [92, 90]. McCaskill's partition function approach [62], 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 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 that are not significant for RNA function. In contrast to earlier approaches [48, 47, 91], 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 [59, 7, 35, 38, 71, 86] was confirmed for all *flavivirus* sequences. However, in the cases of *Dengue (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 25) 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 [38], that are highly conserved among mosquito-borne *flaviviruses*. Structure B includes *CS2* in the cases of *DEN viruses*, but *RCS2* in the *JE* serocmoplex. 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 [81], 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 biolgical testing. Infectious cDNA clones that recently became available for several *flaviviruses* [51, 55, 66, 77, 58] make it possible 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. In *influenza virus* genome segments, however, we could not find conserved secondary structure elements.

The MFE folding of the entire genome of viruses predicted only the thermodynamically most stable secondary structure. Under physiological conditions, however the RNA molecules do not take on only the most stable structure, they seem to

rapidly change their conformation between structures with similar free energies. A realistic investigation of RNA structures has to account for this fact which is of utmost biological importance. The simplest way to do this is to compute not only the optimal structure but all structures within a certain range of free energies by calculating the full matrix of base pairing probabilities p_{ij}. A comparative analysis of base pairing probabilities for a complete viral genome requires a parallelized version of the partition function algorithm and is beyond our computational possibilities at the moment.

7. References

[1] P. Argos, G. Kramer, M. J. H. Nicklin, and E. Wimmer. Similarity in gene organization and homology between proteins of animal picornaviruses and a plant comovirus suggest common ancestry of these virus families. *Nucleic Acids Res.*, 12:7251–7267, 1984.

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

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

[4] J. L. Blackwell and M. A. Brinton. BHK cell proteins that bind to the 3' stem-loop structure of the West nile virus genome RNA. *J. Virol.*, 69:5650–5658, 1995.

[5] S. Bonhoeffer, J. S. McCaskill, P. F. Stadler, and P. Schuster. RNA multi-structure landscapes. A study based on temperature dependent partition functions. *Eur.Biophys.J.*, 22:13–24, 1993.

[6] S. Bouzoubaa, L. Quillet, H. Guilley, G. Jonard, and K. Richards. Nucleotide sequence of beet necrotic yellow vein virus RNA1. *J. Gen. Virol.*, 68:615–626, 1987.

[7] M. A. Brinton, A. V. Fernandez, and J. H. Dispoto. The 3'-nucleotides of flavivirus genomic RNA form a conserved secondary structure. *Virology*, 153:113–121, 1986.

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

[9] T. J. Chambers, C. S. Hahn, R. Galler, and C. M. Rice. Flavivirus genome organization, expression, and replication. *Annu. Rev. Microbiol.*, 44:649–688, 1990.

[10] J. E. Darnell and W. F. Doolittle. Speculations on the early course of evolution. *Proc. Natl. Acad. Sci., USA*, 83:1271–1275, 1986.

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

[12] U. Desselberger, V. R. Racaniello, J. J. Zazra, and P. Palese. The 3'- and 5'-terminal sequences of influenza virus RNA segments are highly conserved and show partial inverted complementarity. *Gene*, 8:315–328, 1980.

[13] R. G. Dietzgen and M. Zaitlin. Tobacco mosaic virus coat protein and the large subunit of the host protein ribulose-1,5-bisphosphate carboxylase share a common antigenic determinant. *Virology*, 155:262–266, 1986.

[14] T. Dobzhansky, F. J. Ayala, G. L. Stebbins, and J. W. Valentine, editors. *Evolution.* W. H. Freemann & Co., San Francisco, CA, 1977.

[15] E. Domingo and J.-J. Holland. High error rates, population equilibrium, and evolution of RNA replication systems. *RNA Genetics*, 3:3–36, 1988.

[16] E. Domingo, E. Martinez-Salas, F. Sobrino, J. C. de la Torre, A. Portela, J. Ortin, C. Lopez-Galindez, P. Perez-Brena, N. Villanueva, R. Najera, S. VandePol, D. Steinhauer, N. DePolo, and J. Holland. The quasispecies (extremly heterogenous) nature of viral RNA genome populations: Biological relevance — a review. *Gene*, 40:1–8, 1985.

[17] J. A. Doudna and J. W. Szostak. RNA catalysed synthesis of complementary-strand RNA. *Nature*, pages 519–522, 1989.

[18] W. Eichler. Some rules on ectoparasitism. *Ann. Mag. Nat. Hist. Ser.*, 1:588–598, 1948.

[19] M. Eigen, J. McCaskill, and P. Schuster. Molecular Quasi-Species. *J.Phys.Chem.*, 92:6881–6891, 1988.

[20] N. Eldredge, editor. *Time Frames. The Rethinking of Darwinian Evolution and the Theory of Punctuated Equilibria.* Simon & Schuster, 1985.

[21] F. Fenner. Portraits of viruses: The poxviruses. *Intervirology*, 11:137–157, 1979.

[22] B. N. Fields, D. M. Knipe, P. M. Howley, R. M. Chanock, J. L. Melnick, T. P. Monath, B. Roizmann, and S. E. Straus, editors. *Fields Virology.* Lippincott, 3rd edition, 1996.

[23] R. Flick, G. Neumann, E. Hoffmann, E. Neumeier, and G. Hobom. Promoter elements in the influenza vRNA terminal structure. *RNA*, 2:1046–1057, 1996.

[24] E. Fodor, D. C. Pritlove, and G. C. Brownlee. The influenza virus panhandle is involved in the initiation of transcrition. *J. Virol.*, 68(6):4092–4096, 1994.

[25] W. Fontana and L. W. Buss. "The arrival of the fittest":towards a theory of biological organisation. *Bull. Math. Biol.*, 56(1):1–64, 1994.

[26] W. Fontana, T. Griesmacher, W. Schnabl, P. Stadler, and P. Schuster. Statistics of landscapes based on free energies, replication and degredation rate constants of RNA secondary structures. *Monatsh. Chem.*, 122:795–819, 1991.

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

[28] R. L. S. Forster, M. W. Bevan, S.-A. Harbison, and R. C. Gardner. The complete nucleotide sequences of the potexvirus white clover mosaic virus. *Nucleic Acids Res.*, 16:291–303, 1988.

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

[30] R. French and P. Ahlquist. Characterisation and engineering of sequences controlling in *vivo* synthesis of brome mosaic virus subgenomic RNA. *J. Virol.*, 62:2411–2420, 1988.

[31] R. F. Gesteland and J. F. Atkins, editors. *The RNA World.* Cold Spring Harbor Laboratory Press, 1993.

[32] W. Gilbert. The RNA world. *Nature*, 319:618, 1986.

[33] R. Goldbach. Genome similarities between plant and animal RNA viruses. *Microbiol. Sci.*, 4:197–202, 1987.

[34] A. E. Gorbalenya, A. P. Donchenko, V. M. Blinov, and E. V. Koonin. Cysteine proteases of positive strand RNA viruses and chymotrypsin-like serine proteases. a distinct protein superfamily with a common structural fold. *FEBS Lett.*, 243, 1989.

[35] T. Grange, M. Bouloy, and M. Girard. Stable secondary structures at the 3'-end of the genome of yellow fever virus (17D vaccine strain). *FEBS Lett.*, 188:159–163, 1985.

[36] R. R. Gutell, M. Larsen, and C. R. Woese. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol. Rev.*, 58:10–26, 1994.

[37] N. Habili and R. H. Symons. Evolutionary relationship between luteoviruses and other RNA plant viruses based on the sequence motifs in their putative RNA polymerases and nucleic acid helicases. *Nucleic Acids Res.*, 17:9543–9555, 1989.

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

[39] L. A. Heaton, B. I. Hillman, B. G. Hunter, D. Zuidema, and A. O. Jackson. A physical map of the genome of sonchus yellow net virus, a plant rhabdovirus with six genes and conserved genejunction sequences. *Proc. Natl. Acad. Sci.*, 86:8665–8668, 1989.

[40] P. Higgs. Thermodynamic properties of transfer RNA. A computational study. *J. Chem. Soc. Faraday. Trans.*, 91:2531–2540, 1995.

[41] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. `Vienna RNA Package`.
`ftp://ftp.itc.univie.ac.at/pub/RNA/`
`http://www.tbi.univie.ac.at/~ivo/RNA/`,
1994. (Public Domain Software).

[42] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125(2):167–188, 1994.

[43] I. L. Hofacker, M. A. Huynen, P. F. Stadler, and P. E. Stolorz. RNA folding on parallel computers. The minimum free energy structures of complete HIV genomes.

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

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

[46] J. Holland, K. Spindler, F. Horodyski, E. Grabau, S. Nichol, and S. V. Pol. Rapid evolution of RNA genomes. *Science*, 215:1577–1585, 1982.

[47] M. A. Huynen, R. Gutell, and D. A. M. Konings. Assessing the reliability of RNA folding using statistical mechanics. *J. Mol. Biol.*, 1997. in press.

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

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

[50] G. Kamer and P. Argos. Primary structural comparison of RNA-dependent polymerases from plant, animal and bacterial viruses. *Nucleic Acids Res.*, 12:7269–7282, 1984.

[51] A. A. Khromykh and E. G. Westaway. Completion of Kunjin virus RNA sequence and recovery of an infectious RNA transcribed from stably cloned full-length cDNA. *J. Virol.*, 68:4580–4588, 1994.

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

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

[54] E. V. Koonin and A. E. Gorbalenya. Evolution of RNA genomes: Does the high mutation rate necessiate high rate of evolution of viral proteins? *J. Mol. Evol.*, 28:524–527, 1989.

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

[56] Y.-S. Lee and B. L. Seong. Mutational analysis of influenza B virus RNA transcription in vitro. *J. Virol.*, 70:1232–1236, 1995.

[57] G. Luo, W. Luytjes, M. Enami, and P. Palese. The polyadenylation signal of influenza virus RNA involves a stretch of uridine followed by the RNA duplex of the panhandle structure. *J. Virol.*, 65:2861–2867, 1991.

[58] C. W. Mandl, M. Ecker, H. Holzmann, C. Kunz, and F. X. Heinz. Infectious cDNA clones of tick-borne encephalitis virus European subtype prototypic strain Neudoerfl and high virulence strain Hypr (submitted). Preprint, 1997.

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

[60] W. S. Mason, J. M. Taylor, and R. Hull. Retroid virus genome replication. *Adv. Virus Res.*, 32:35–96, 1987.

[61] R. E. F. Matthews, editor. *Plant Virology.* Academic Press, 3rd edition, 1991.

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

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

[64] F. A. Murphy, C. M. Fauquet, D. H. L. Bishop, S. A. Ghabrial, A. W. Jarvis, G. P. Martelli, M. A. Mayo, and M. D. Summers, editors. *Virus Taxonomy.* Springer-Verlag, 6th edition, 1995.

[65] R. C. L. Olsthoorn, G. Garde, T. Dayhuff, J. F. Atkins, and J. van Duin. Nucleotide sequence of a single-stranded RNA phage from *pseudomonas aeruginosa*: Kinship to coliphages and conservation of regulatory RNA structures. *Virology*, 206:611–625, 1995.

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

[67] H. D. Robertson. How did replicating and coding RNAs first get together? *Science*, 274:66–67, 1996.

[68] D. J. Robinson, W. D. O. Hamilton, B. D. Harrison, and D. C. Baulcombe. Two anomalous tobravirus isolates: Evidence for RNA recombination in nature. *J. Gen. Virol.*, 68:2551–2561, 1987.

[69] P. Schuster. How does complexity arise in evolution? *Complexity*, 2(1):22–30, 1996.

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

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

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

[73] K. G. Skryabin, S. Y. Morozov, A. S. Kraev, M. N. Rozanov, B. K. Chernov, L. I. Lukasheva, and J. G. Atabekov. Conserved and variable elements in RNA genomes of potexviruses. *FEBS Lett.*, 240:33–40, 1988.

[74] D. B. Smith and S. C. Inglis. The mutation rate and variability of eukaryotic viruses: An analytical review. *J. Gen. Virol.*, 68:2729–2740, 1987.

[75] C.-B. Stewart and A. C. Wilson. Sequence convergence and functional adaption of stomach lysozymes from foregut fermenters. *Cold Spring Harbor Symp. Quant. Biol.*, 52:891–899, 1987.

[76] L. Stryer, editor. *Biochemistry.* W. H. Freeman and Company, 3rd edition, 1988.

[77] H. Sumiyoshi, C. H. Hoke, and D. W. Trent. Infectious japanese encephalitis virus RNA can be synthesized from in vitro-ligated cDNA templates. *J. Virol.*, 66:5425–5431, 1992.

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

[79] J. D. Thompson, D. G. Higgs, and T. J. Gibson. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties, and weight matrix choice. *Nucleic Acids Res.*, 22:4673–4680, 1994.

[80] H. Toh, H. Hayashida, and T. Miyata. Sequence homology between retroviral reverse transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus. *Nature (London)*, 305:827–829, 1983.

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

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

[83] M. S. Waterman, editor. *Introducton to Computational Biology.* Chapman & Hall, 1st edition, 1995.

[84] M. S. Waterman and T. F. Smith. RNA secondary structure: A complete mathematical analysis. *Mathematical Biosciences*, 42:257–266, 1978.

[85] A. M. Weiner and N. Maizels. RNA-like structures at the 3' ends of genomic RNA molecules for replication: Implications for the origin of protein synthesis. *Proc. Natl. Acad. Sci., USA*, 84:7383–7387, 1987.

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

[87] F. H. Westheimer. Polyribonucleic acids as enzymes. *Nature*, 319:534–536, 1986.

[88] D. Zimmern. Evolution of RNA viruses. *RNA Genetics*, 2:211–240, 1988.

[89] M. Zuker. `mfold-2.0`. `pub/mfold.tar.Z @ nrcbsa.bio.nrc.ca`. (Public Domain Software).

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

[91] M. Zuker and A. B. Jacobson. "well-determined" regions in RNA secondary structure prediction: analysis of small subunit ribosomal RNA. *Nucleic Acids Res.*, 23:2791–2798, 1995.

[92] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bull. Math. Biol.*, 46(4):591–621, 1984.

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

Table of Contents

Curriculum vitae

Susanne Rauscher
$*$ 1967−03−06

1973−1977	:	Volksschule, 11th District, Vienna, Austria
1977−1981	:	Hauptschule 11th District, Vienna, Austria
1981−1986	:	Handelsakademie, 3rd District, Vienna, Austria
6/1986	:	Matura at the Handelsakademie
1986−1996	:	Studies of Biochemistry at the University of Vienna
9/1995−12/1996	:	Diploma thesis with Doz. Dr. Peter F. Stadler at the Institute of Theoretical Chemistry, University of Vienna

Publications:

- C. Flamm, S. Rauscher, C. W. Mandl and P. F. Stadler
  New Conserved Secondary Structure Motifs at the 3' End of RNA Genome of Tick-Borne Flaviviruses.
  Symposium on Modern Approaches to Flavivirus Vaccines, Vienna, Austria. Abstract volume B18 (1996).

- S. Rauscher, C. Flamm, C. W. Mandl, F. X. Heinz and P. F. Stadler
  Secondary Structure of the 3'-Non-Coding Region of Flavivirus Genomes.
  Submitted to RNA (1996).