Conserved Structure Elements in Viral RNA Genomes

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium

an der Formal- und Naturwissenschaftlichen Fakultät
der Universität Wien

eingereicht von

Mag. Susanne Rauscher

Institut für Theoretische Chemie und
Molekulare Strukturbiologie

Wien, im Juni 2000

Dank an alle,

die zum Gelingen dieser Arbeit beigetragen haben:

Peter Schuster, Peter Stadler, Christian Mandl, Ivo Hofacker, Aderonke Babajide, Dieter Blaas, Jan Cupal, Martin Fekete, Christoph Flamm, Walter Fontana, Dagmar Friede, Kurt Grünberger, Jörg Hackermüller, Christian Haslinger, Philipp Kobel, Michael Kospach, Ulli Mückstein, Stefan Müller, Andrea Reischl, Bärbel Stadler, Roman Stocsits, Andreas Svrcek-Seiler, Caroline Thurner, Günther Weberndorfer, Andreas Wernitznig, Christina Witwer, Michael Wolfinger, Gerlinde Aschauer, Daniela Dorigoni, Judith Jakubetz.
Berta Rauscher, Josef Rauscher, Renate Hammer, Manuela Hammer, Patrick Hammer, Helmut Rauscher.
Emre Tuncer, Dorothea Anrather, Birgit Halva-Barabasch, Akis Karakostas, Doris Pinka-Jilg, Reinhild Pürgy, Helga Smekal, Susanne Vorauer.

Abstract

The three-dimensional shape of RNA molecules plays a crucial role in processes such as protein synthesis and may exhibit a large variety of catalytic activities. The task of three-dimensional structure prediction for biopolymers like RNAs and proteins from sequence data can not be solved with current knowledge and methods. A simpler problem, namely the prediction of secondary structure is in principle tractable even for large molecules. Functional secondary structures can be thought of as an important descriptive component of the RNA molecules, leading to their application in the interpretation of molecular evolution data.

However, most RNA molecules — and particular viral RNA genomes and their subsequences — form distinctive secondary structures. The presence of secondary structure therefore does not indicate any functional significance. Computer simulations have shown that a small number of point mutations is very likely to cause large changes in the secondary structures. If selection acts to preserve a structural element then it must have some function. The detection of conserved structural motifs in related RNA sequences is therefore a first crucial step towards understanding their function.

We searched for functionally important structures that are conserved among a group of closely related viruses utilizing the information contained in a reliable multiple sequence alignment (`Ralign` and `ClustalW`) of the complete virus sequences to extract conserved secondary structures from a pool of plausible structures generated by thermodynamic prediction (McCaskill's algorithm).

We applied our technique to groups of RNA viruses for which a sufficient number of complete genomes have been sequenced. Three unrelated families of viruses, which contain a variety of human pathogens of global medical importance, served as examples, namely *Retroviridae*, *Picornaviridae*, and *Flaviviridae*.

A comprehensive survey of conserved RNA secondary structures in viral genomes is now feasible, and there is a collection of software available that allows a routine investigation of viral RNA sequences. The results prove that our methods are indeed capable of detecting the structural elements that were described in literature plus a large number of previously unknown conserved structural elements. The resulting data, in particular the "atlas" of conserved RNA structures for the virus families, provide a valuable basis for further investigations into viral evolution and phylogeny.

Zusammenfassung

Die dreidimensionale Struktur von RNA-Molekülen spielt eine wichtige Rolle in Prozessen wie z. B. der Proteinsynthese und zeigt häufig katalytische Aktivität. Die Vorhersage der dreidimensionalen Struktur von Biomolekülen, wie RNA und Proteinen, ist jedoch nach derzeitigem Wissensstand – im Gegensatz zur Sekundärstrukturvorhersage auch für große Moleküle – nicht möglich. Sekundärstrukturen sind gute Näherungen zur Beschreibung von RNA-Molekülen, und können zur Erklärung der molekularen Evolution herangezogen werden.

Da fast alle RNA-Moleküle, im besonderen virale Genome und auch ihre Teilsequenzen, Sekundärstrukturen ausbilden, gibt ein Vorhandensein von Strukturelementen noch keinen Aufschluss über ihre funktionelle Signifikanz. Umfangreiche Computersimulationen haben ergeben, dass schon eine kleine Anzahl von Punktmutationen ausreicht, um Sekundärstrukturen erheblich zu verändern. Die Selektion eines Sekundärstrukturelementes deutet also auf deren wichtige Funktion hin. Das Auffinden solcher Sekundärstrukturmotive in ähnlichen Sequenzen ist daher der erste Schritt, um Hinweise auf ihre mögliche Funktion zu erhalten.

Um diese funktionell wichtigen Strukturen, die innerhalb einer Gruppe nahverwandter Viren konserviert sind, zu finden, haben wir die Information aus einem verlässlichen multiplen Alignment der kompletten Virussequenzen (`Ralign` bzw. `ClustalW`) verwendeten, um aus einem Pool von möglichen Strukturen, berechnet mit einem thermodynamischen Algorithmus (McCaskill), die konservierten Sekundärstrukturen auszuwählen.

Wir haben unsere Methode an RNA-Viren getestet, von denen sich eine ausreichende Anzahl an kompletten Sequenzen in den Datenbanken befanden. Drei Virusfamilien *Retroviridae*, *Picornaviridae* und *Flaviviridae*, denen jeweils Humanpathogene von globaler Bedeutung angehören, dienten als Beispiele.

Ein umfangreicher Vergleich konservierter Sekundärstrukturen in Virusgenomen ist nun möglich, und die vorhandenen Programme erlauben eine routinemäßige Untersuchung von viralen Sequenzen. Die Ergebnisse beweisen, dass es mit unseren Methoden möglich ist, die in der Literatur beschriebenen Strukturelemente zu bestätigen und zusätzlich viele bislang unbekannte, konservierte Strukturelemente zu bestimmen. Die so erhaltenen Daten, im besonderen der "Atlas", der innerhalb einer Virusfamilie konservierten RNA-Strukturen, liefern eine wertvolle Basis zur weiteren Erforschung der Virusevolution und -phylogenie.

Contents

1 Introduction

RNA molecules are not only carriers of information, they can also be functionally active units. The three-dimensional structures of RNA molecules play a crucial role in the process of protein synthesis. Furthermore, they may exhibit a large variety of catalytic activities. One of the major problems facing computational molecular biology is the fact that the amount of sequence information surmounts the quantity of information about the three-dimensional structure of biopolymers by far. The task of three-dimensional structure prediction for biopolymers like RNAs and proteins from sequence data can not be solved with current knowledge and methods (see however [86] for a promising approach).

Presently, a simpler problem, namely the prediction of secondary structure is far more tractable even for large molecules. It has been observed, that functional secondary structures are conserved in evolution [48]. They can be thought of as an important descriptive component of the RNA molecules, leading to their application in the interpretation of molecular evolution data.

While long RNA molecules probably fold sequentially in nature, there are rearrangements between established and new helices during the folding process. Although there have been a number of different approaches to kinetic and/or sequential folding there is no consensus in the field and so far these approaches have not proved to be significantly superior to thermodynamic folding, which yields at least a controlled approximation of the real structure. It highlights possible global interactions that may or may not be accessible along kinetic folding pathways. As a consequence, thermodynamic predictions of base pairing probabilities are an ideal starting point for comparative approaches.

Extensive computer simulations [42, 123] with RNA sequences have shown that a small number of point mutations is very likely to cause large changes in the secondary structures. About 10% difference in the nucleic acid sequence almost certainly leads to unrelated structures if the mutated sequence positions are chosen at random. Secondary structure elements consistently present in a group of sequences with less than, say 95% average pairwise identity are therefore most likely the result of stabilizing selection, not a consequence of the high degree of sequence homology.

In this thesis we show that potentially functional RNA structures can be identified by a purely computational procedure that combines structure prediction and sequence comparison. RNA viruses are an ideal proving ground for developing novel approaches towards functional genome analysis for a variety of reasons:

- Distant groups of RNA viruses have very little or no detectable sequence homology and oftentimes very different genomic organization. Thus we can test our approach on essentially independent data sets.
- RNA viruses show an extremely high mutation rate, on the order to 10^{-5} to 10^{-3} mutations per nucleotide and replication. Due to this high mutation rate they form *quasispecies*, that is, diffuse “clouds” in sequence space [35] and their sequences evolve at a very high rate. In contrast functional secondary structures are strongly conserved. A substantial fraction of the mutations are likely to be almost neutral, implying a very rapid evolution at sequence level. Due to the high sequence variation, the application of classical methods of sequence analysis is therefore difficult or outright impossible. Indeed, except for the order *mononegavirales* (negative-stranded RNA viruses), there is no accepted taxonomy above the family level.
- The high mutation rate of RNA viruses also explains their short genomes of less than some 20,000 nucleotides [35]. A large number of *complete* genomic sequences is available in data bases. The non-coding regions are most likely functionally important since the high selection pressure acting on viral replication rates makes “junk RNA” very unlikely. 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 [148, 10, 62], RNA phages [17, 105], flaviviruses [124], pestiviruses [18, 32], picornaviruses [34, 64, 71, 81, 110, 116], hepatitis C viruses [130, 18], or hepatitis D viruses [141].

It was our goal to find functionally important structures in viral genomes. We used (i) thermodynamic folding procedures for structure prediction and (ii) structure based alignments for identifying those structural features conserved among a group of closely related viruses. The actual existence of such features is verified by the presence of compensatory mutations. Thermodynamic folding algorithms are collected in the `Vienna RNA Package` [61].

Our method of determing secondary structure elements depends crucially on the quality of the sequence alignment. We expect a significant improvement in alignment quality by using `Ralign` a protein sequence alignment for the coding regions and combining it with a nucleic acid based alignment.

A set of computer methods to scan moderate size samples of RNA sequences for conserved secondary structures was recently developed at the *Institute for Theoretical Chemistry and Molecular Structural Biology*. The program `pfrali` utilizes the information contained in a multiple alignment of related RNA sequences to extract conserved features from the pool of plausible structures calculated by thermodynamic prediction for each sequence.

We refined and extended these methods for determining conserved secondary structures and applied them to a wide variety of different RNA virus families. Our goals are threefold:

- Conserved secondary structures most likely have an important function in the viral life cycle. Our approach therefore immediately produces a list of promising targets for experimental investigations such as deletion studies.
- Functional secondary structures evolve much slower than the underlying sequences. Conserved secondary structures can therefore be used to extend the viral phylogeny to higher taxa.
- A list of conserved, and therefore evolved, RNA structures is a valuable data set in itself. Detailed analysis of such data can yield insights into general questions such as the evolution of robustness.

We applied our technique to families of RNA viruses for which a sufficient number of complete genomes have been sequenced. Three unrelated groups of viruses, which contain a variety of human pathogens of global medical importance, serve as examples:

The *Retroviridae* have received more attention from scientist in recent year than any other group of infectious agents. This reflects not only their importance as human and animal pathogens, but also their value as experimental objects. The Retroviridae comprise a large family of viruses, primarily of vertebrates. Despite the variety of host species and of interactions with the host, all retrovirus isolates are quite similar in virion structure, genome organization, and mode of replication. The genome consist of two, usually identical, molecules of single-stranded RNA, on the order of 10,000 nucleotides in length, modified like cell mRNAs, including capping at the 5' end and polyadenylation at the 3' end. The order of genes encoding structural proteins is invariable *gag-pro-pol-env*. The replication cycle includes a process known as reverse transcription. Complex exogenous viruses responsible for a variety of neurological and immunological diseases, like human immunodeficiency virus type 1 (HIV-1), HIV-2, and simian immunodeficiency virus (SIV), are members of the lentivirus genus.

The *Picornaviridae* are among the smallest ribonucleic acid-containing viruses known. Rhinovirus, poliovirus (genus enterovirus), human hepatitis A virus, and foot-and-mouth disease virus (FMDV) are members of the picornavirus family. Their genome consists of a single strand messenger-active (+) RNA of 7,000 to 8,000 nts. The genome is polyadenylated at the 3' terminus and carries a small protein (virion protein, genome; VPg) covalently attached to its 5' end. The translation mechanism is dependent on an ca. 450 nt *cis*-acting RNA element located within the 5'-untranslated region (UTR), originally designated as the internal ribosome entry site (IRES) [117].

The *Flaviviridae* are small enveloped particles with an unsegmented, plus-stranded RNA genome. This virus family contains the genera flavivirus (which includes the viruses causing Japanese encephalitis, dengue, yellow fever, and tick-borne encephalitis), pestivirus, hepatitis C, and the recently discovered hepatitis G virus [96].

The purpose of this work is to prove that a comprehensive survey of conserved RNA secondary structures in viral genomes is feasible, and there is now a collection of software available that allows a routine investigation of viral RNA sequences (Section 5, 6, and 7). The results reported here, in particular the "atlas" of conserved RNA structures for the virus families *Retroviridae*, *Picornaviridae*, and *Flaviviridae* prove that our methods are indeed capable of detecting a large number of previously unknown conserved structural elements, and that the resulting data provide a valuable basis for further investigations into viral evolution and phylogeny.

2 Folding of RNA

RNA polymers are macromolecules, consisting of a linear arrangement of building blocks, the monomers [50]. RNA polymers have the ability to fold back on themselves, due to interactions between individual base pairs. For biopolymers like proteins or RNA these interactions are specific, and can lead to the adaption of a unique compact conformation called 'native state'. During the structure formation process both, RNA and proteins, try to minimize the solvent exposure of hydrophobic residues by burying these residues in the interior of the structure [55]. But it is self-evident from the different chemical nature of RNA and proteins that the ways how these macromolecules achieve their compact conformation is different. For proteins the driving force of the collapse into compact conformations is the formation of a hydrophobic core. For RNA the formation of compact conformation is promoted by the tendency to maximize the stacking interaction between base pairs [56]. And it is essential for living cells that this formation of the correct and functional conformation is achieved in biologically relevant sufficiently short time [140, 51, 144, 142].

From a theoretical point of view, the problem of how biopolymers achieve their native state splits up into two aspects. The first aspect is the structure prediction problem. The second aspect deals with the dynamics of the folding process itself [21, 22, 53].

Since the sequence of a biopolymer specifies its three-dimensional structure, it should be possible, at least in principle, to predict its native structure solely from the knowledge of the sequence [65]. On the other hand, biopolymers like proteins or RNA are flexible and rapidly fluctuating molecules whose structural mobilities have functional significance [25, 91].

The native states of RNA consist of a large ensemble of closely related and rapidly inter-converting conformational substates of nearly equal stabilities. Theoretical methods for structure prediction require extensive computation. The secondary structure of RNA is defined as the pattern of base pairs, which is formed by hydrogen bonds between atoms of the four bases [104, 103].

RNA secondary structures provide a useful, though coarse grained, description of RNA molecules as exemplified by their frequent use in the literature. Several algorithms exist for the prediction of RNA secondary structures based on thermodynamic rules. The most widely used methods compute a single minimum energy structure through dynamic programming [61, 156, 157, 102, 101, 142, 143, 11]. Approaches to kinetic folding [47, 95] are also based on the thermodynamic rules. Because of inaccuracies of the energy model and the measured parameters [43, 52, 72, 134], the accuracy of these predictions is often insufficient. In cases where the correct structure is known from phylogenetic analysis it has been found that predicted structures contain only 30% to 80% of the correct base pairs. The correct structure can, however, be found within a relatively small energy interval above the ground state.

There are variants of the folding algorithm for computing a sample of suboptimal folds [155], or even *all* structures within a prescribed energy range [150]. Non-deterministic kinetic folding algorithms [47] can produce ensembles of structures by repeatedly running them with different random numbers.

A more elegant and efficient solution was suggested by McCaskill [92], which contains suitably weighted information about all possible secondary structures and therefore reduces the impact of inaccuracies in the structure prediction. The disadvantage of these methods is of course that they leave it up to the user to decide which of the proposed structures to believe. He proposed an algorithm to compute the partition function of the thermodynamic ensemble and the matrix of base pairing probabilities P_{hl} of an RNA molecule. The large size of, say, HIV genomes ($n \approx 9200$ nucleotides) implies that there is a huge number of low energy states. For example, the frequency of the minimum energy structure in the ensemble at thermodynamic equilibrium is in general smaller than 10^{-23} for RNAs of the size of a HIV viral genome. Hence one would need a huge number of different structures to adequately describe the ensemble. While such an approach is feasible for RNAs with up to some 100 nucleotides [150], the direct generation and analysis of the necessary amount of structure information for long sequences exceeds by far the capabilities of even the most modern computer systems. The pair probability matrix (P_{hl}) computed by McCaskill's algorithms is a much more suitable representation of such large structure ensembles. Thus McCaskill's approach provides a computationally feasible alternative, which is comparable to the requirements of the simple minimum free energy folding algorithm in terms of the required computational resources.

McCaskill's partition function folding algorithm has been implemented for message passing parallel computer architectures. The program is written in `C` and uses the `MPI` message passing interface. It is therefore easily portable to most currently available parallel computers.

In the following section we briefly recall the definition of RNA secondary structures and the standard energy model. A brief description of the partition function algorithm is provided in Section 2.2.

2.1 RNA Secondary Structures

Most RNA molecules are single stranded *in vivo*. The molecule folds back onto itself to form double helical regions stabilized by Watson-Crick G-C and A-U base pairs or the slightly less stable G-U pairs. Base stacking and base pairing are hence the major driving forces of structure formation in RNA. Other, usually weaker, intermolecular forces and the interaction with aqueous solvent shape its spatial structure. As opposed to the protein case, the secondary structure of RNA sequences is well defined, provides the major set of distance constraints that guide the formation of tertiary structure, and covers the dominant energy contribution to the 3D structure. Furthermore, secondary structures are conserved in evolutionary phylogeny [48] and therefore represent a qualitatively important description of the molecules.

A secondary structure consists of a set of vertices
$V = \{1, 2, ..., i, ..., N\}$ and a set of edges $S = \{i \cdot j, 1 \leq i < j \leq N\}$ fulfilling

(1) For $1 \leq i < n$, $i \cdot (i+1) \in S$.

(2) For each i there is at most one $h \neq i-1, i+1$ such that $i \cdot h \in S$.

(3) If $i \cdot j \in S$ and $h \cdot l \in S$ and $i < h < j$, then $i < l < j$.

The first condition simply states that RNA is a linear polymer, the second condition restricts each base to at most a single pairing partner, and the third forbids pseudoknots and knots. While pseudoknots are important structural elements in many RNA molecules [146], they are excluded from many studies mostly for a technical reason [143]. The folding problem for RNA can be solved efficiently by dynamic programming [156, 143] in their absence. In many cases pseudoknots

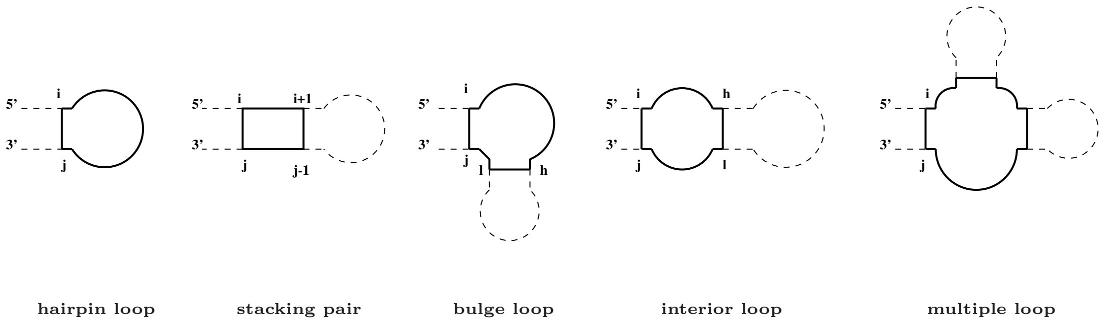


Figure 1: Secondary structures decompose into five distinct loop types, which form the basis of the additive energy model. One distinguishes three loop energy functions: $\mathcal{H}(i,j)$ for hairpin loops, $\mathcal{I}(i,j,h,l)$ for the three types of loops that are enclosed by base pairs $i \cdot j$ and $h \cdot l$ and the additive model for multi-loops described in the text. Stacked pairs ($h = i+1, l = j-1$) and bulges (either $h = i+1, l \neq j-1$ or $l = j-1, h \neq i+1$) are treated as special cases of interior loops. The energies depend on the types of closing base pairs indicated by $i \cdot j$ and interior base pairs as well as on the size of the loops.

can be "added" to a predicted secondary structure graph during a post-processing step.

A *base pair* $h \cdot l$ is called *interior* to the base pair $i \cdot j$, if $i < h < l < j$. It is *immediately interior* if there is no base pair $p \cdot q$ such that $i < p < h < l < q < j$. For each base pair $i \cdot j$ the corresponding *loop* is defined as consisting of $i \cdot j$ itself, the base pairs immediately interior to $i \cdot j$ and all unpaired regions connecting these base pairs. In graph theoretical terms, the loops form the unique minimal cycle basis of the secondary structure graph [82].

The standard energy model for RNA contains the following types of parameters: (i) *base pair stacking* energies depend explicitly on the types of the four nucleotides $i \cdot j$ and $(i+1) \cdot (j-1)$ that stack. For the purpose of the recursions in Table 1 it is useful to view stacked base pairs as a special type of interior loop, hence we denote the stacking energies $\mathcal{I}(i,j,i+1,j-1)$. (ii) *loop energies* depend on the type of the loop, its size, the closing pairs and the unpaired bases adjacent to them, see Figure 1. We write $\mathcal{H}(i,j)$ for hairpin loops and $\mathcal{I}(i,j,h,l)$ for interior loops. Multi-loops are assumed to have a linear contribution of the form $\mathcal{M} = \mathcal{M}_C + \mathcal{M}_I \cdot degree + \mathcal{M}_B \cdot unpaired$, in addition the so-called dangling end energies are taken into account which refer to mismatches next to the base pairs that delimit the loop. The implementation of the folding algorithms used in this contribution assumes the energy parameters summarized in [140] except that co-axial stacking of helices is neglected. (Co-axial stacking is, strictly speaking not part of the secondary structure graph as defined above.) The energy model is thus identical to Zucker's `mfold 2.3` [154].

2.2 McCaskill's Algorithm

McCaskill's partition function algorithm naturally decomposes into two parts, namely the computation of the partition function and the subsequent computation of the pairing probabilities. We will refer to the two parts as *folding* and *backtracking*, respectively. The logic of the folding part is essentially the same as for minimum energy folding [156] while the backtracking part is much more elaborate.

The partition function of the complete RNA molecule can be derived from the partition functions of all its subsequences. For the subsequence from i to j we have to distinguish whether $i \cdot j$ forms a base pair or not. We write Q^B_{ij} for the partition function of the substring subject to the constraint that $i \cdot j$ is paired and Q_{ij} for the unconstrained partition function. Consequently, the partition function of the entire molecule is $Q = Q_{1n}$.

If i to j are paired, this pair can close either a hairpin loop, an interior loop delimited by $i \cdot j$ and $h \cdot l$, or a multi-component loop. The three terms in Table 1 correspond to these possibilities. Multi-loops can be dealt with efficiently due to a linear approach for their energy contributions. This allows for a decomposition into three terms: one for unpaired substructures, one for substructures consisting of single component, and a multi-component reminder. The auxiliary variables Q^M and Q^{M1} are necessary for handling multi-loop contributions. Introducing Q^A and restricting the size of interior loops to $u \leq u_{\max}$ reduces the CPU requirements from $\mathcal{O}(n^4)$ to $O(n^3)$. Most programs set $u_{\max} = 30$. The restriction on the size of interior loops does not have a serious effect in practice, since long interior loops are energetically unfavorable and therefore very rare. For further details we refer to McCaskill's [92] original paper.

In the backtracking part of the algorithm, the pairing probabilities P_{ij} are obtained by comparing the partition functions Q^B_{ij} and Q_{ij} with and without an enforced pair $i \cdot j$. While the partition function for longer subsequences is computed from shorter ones during the folding part, the backtracking recursion pro-

Table 1: Recursion for Computing the Partition Function.
The parameter m is the minimum size of a hairpin loop, usually $m = 3$.

Folding	**Backtracking**
$Q^B_{ij} = e^{-\mathcal{H}(ij)/kT} + \sum_{\substack{h=i+1 \\ u \le u_{\max}}}^{j-m-2} \sum_{l=h+m+1}^{j-1} Q^B_{hl}\, e^{-[\mathcal{I}(i,j,h,l)]/kT} + \sum_{h=i+1}^{j-m-2} Q^M_{i+1,h-1} Q^{M1}_{h,j-1}\, e^{-\mathcal{M}_{\mathcal{C}}/kT}$	$P^c_{hl} = \frac{Q_{1,h-1} Q^B_{hl} Q_{l+1,n}}{Q_{1n}}$
$Q^{M1}_{ij} = \sum_{l=i+m+1}^{j} Q^B_{il}\, e^{-[\mathcal{M}_{\mathcal{I}} + \mathcal{M}_{\mathcal{B}}(j-l)]/kT}$	$P^i_{hl} = \sum_{\substack{i=1 \\ u<u_{\max}}}^{h-1} \sum_{j=l+1}^{n} P_{ij} \frac{Q^B_{hl}}{Q^B_{ij}} e^{-\mathcal{I}(i,j,h,l)/kT}$
$Q^M_{ij} = \sum_{h=i+m+1}^{j-m-1} Q^M_{i,h-1}\, Q^{M1}_{hj} + \sum_{h=i}^{j-m-1} Q^{M1}_{hj}\, e^{-\mathcal{M}_{\mathcal{B}}(h-i)/kT}$	$P^m_{hl} = Q^B_{hl} e^{-[(\mathcal{M}_{\mathcal{C}} + \mathcal{M}_{\mathcal{I}})/kT]} \times \sum_{i=1}^{h-1} \left(P^{M1}_{il} Q^M_{i+1,h-1} + P^M_{il}\, Q^M_{i+1,h-1} + P^M_{il}\, e^{-[(h-i-1)\mathcal{M}_{\mathcal{B}}/kT]}\right)$
$Q^A_{ij} = \sum_{l=i+m+1}^{j} Q^B_{il}$	$P^M_{il} = \sum_{j=l+2}^{n} \frac{P_{ij}}{Q^B_{ij}} Q^M_{l+1,j-1}$
$Q_{ij} = 1 + Q^A_{ij} + \sum_{h=i+1}^{j-m-1} Q_{i,h-1} Q^A_{hj}$	$P^{M1}_{il} = \sum_{j=l+1}^{n} \frac{P_{ij}}{Q^B_{ij}} e^{-[(j-l-1)\mathcal{M}_{\mathcal{B}}/kT]}$
	$P_{hl} = P^c_{hl} + P^i_{hl} + P^m_{hl}$

ceeds in the reverse direction. The probability P_{hl} of the pair $h \cdot l$ is the sum of three independent terms: (i) it closes a component with probability P^c_{hl}, (ii) it is an interior base pair of an interior loop, bulge, or stack with probability P^i_{hl}, or (iii) it is immediately interior to a multiloop with probability P^m_{hl}. Again, two auxiliary arrays are needed to handle the multi-loop contribution in cubic time. The complete recursion of McCaskill's algorithm is summarized in Table 1. An efficient implementation for serial machines is part of the `Vienna RNA Package` [135, 61].

For long (sub)sequences the partition functions Q_{ij} become very large since they are the products of a large number of exponential functions. In order to reduce the numerical problems we rescale the partition function of a subsequence of length ℓ by a factor $\tilde{Q}^{\ell/n}$, where $\tilde{Q}$ is an *a priory* estimate for the partition function. A sufficiently accurate estimate can be obtained from the ground state energy $E_{\min}$:

$$\ln \tilde{Q} \approx -1.04 \times \; E_{\min}/kT \tag{1}$$

We use the message passing implementation of the minimum energy folding algorithm, which is described in [60, 62] to compute $E_{\min}$.

2.3 Using Parallel Computers

RNA folding algorithms are quite demanding both in terms of memory and CPU time. For a sequence of length n, CPU time scales as $\mathcal{O}(n^3)$ and memory requirements are $\mathcal{O}(n^2)$. While this is not a problem for small RNA molecules, such as tRNAs, the requirements exceed the resources of most computers for large RNA molecules such as viral genomes. In most cases, memory, rather than computational speed, becomes the fundamental resource bottleneck. The use of modern parallel computers thus becomes unavoidable once the memory requirements exceed, say, 1 GByte.

For large RNA sequences we used McCaskill's partition function folding algorithm which has been implemented for message passing parallel computers. Since the folding and the backtracking part are independent of each other they were parallelized independently [36]. The folding part can be parallelized in a way that is very similar to the earlier message passing implementation of minimum energy folding algorithm [62, 61]. However, some of the intermediate results are required again during the backtracking stage. Storing these values in such a way that the backtracking recursion can efficiently be distributed among a large number of processors is the main difficulty of our task. For sequences longer than some 3,000 nucleotides it is necessary in general to use double precision reals. Hence Fekete's implementation needs some 2.5 GBytes to fold a HIV sequence.

Recently cost-effective workstation clusters have become widely available. We use a `Beowulf` architecture consisting of 9 two-processor PCs (Pentium II, 450 Mhz) with 512 MByte each, connected by 100 Mbit `Fast Ethernet`, running `Linux` and `LAM 6.1`. This setup is sufficient for the routine computation of base pairing probability matrices from complete RNA virus genomes. For comparison, folding the HIV LAI sequence, $n = 9229$, took about 77 min using 320 processors on the `Intel Delta` and 2 h on 16 Pentium II 450 MHz. The serial code took 42 h on a `DEC alpha` and 64 h on `Cray YMP` for the same sequence.

Despite the relatively slow network connection in the `Beowulf` workstation cluster we find efficiencies above 50% on 16 nodes already for the chain length $n = 4000$. Executing the parallel code on a single CPU shows that the overhead from the parallelization is about 20% to 25%. This is mainly because some parts of the algorithm can be implemented more efficiently in the serial version, where the memory organization is not constrained by requirements of easy message passing.

3 Aligning RNA Genomes

As functional important secondary structures are conserved in evolution, especially nucleotides that are important in maintaining these structures are subject to high selection pressure, e.g. mutation of a nucletide within a stack causes the exchange of the pairing base. For detecting these sequence covariations we need a reliable multiple sequence alignment, which allows us to find the locality of the correlated mutations within the different sequences. Our aim is to utilize the information contained in a multiple sequence alignment of related virus sequences to extract conserved secondary structures from a pool of plausible structures generated by thermodynamic prediction for each sequence, see Figure 9.

3.1 Background on Alignments

An alignment is the most basic sequence analysis task. It is used to tell whether two or more sequences are related, how close this relationship is, and which sequence positions are equivalent. To find the best possible alignment of sequences is of central importance for bioinformatics and data processing after routine laboratory procedures. In principle all alignment algorithms are based on two criteria, (i) maximum similarity or (ii) minimum distance [41, 45, 58]. For evaluating the difference between two sequences we have three possibilities of pairs of opposite symbols: (i) identity, (ii) substitution or mismatch and (iii) insertion or deletion.

Careful thought must be given to the scoring system used to evaluate an alignment by looking for evidence when sequences have diverged from a common ancestor by mutation and selection. As mentioned above, the basic mutational processes that are considered are substitutions, which change, and insertions and deletions, which add or remove residues in a sequence and are referred to as 'gaps'. The total score we assign to an alignment is a sum of terms for each aligned pair of residues, plus terms for each gap. Informally, using an additive scoring system we expect identities and conservative substitutions to be more likely in good (biologically relevant) alignments than we expect by chance, and so they should contribute positive score terms. On the other hand non-conservative changes are expected to be observed less frequently so they contribute negative score terms. This system also corresponds to the assumption that we can consider mutations at different sites in a sequence to have occurred independently (treating a gap of arbitrary length as a single event). All alignment algorithms depend crucially on such a scoring scheme and from a biological point of view the assumption of independence appears to be a reasonable approximation for DNA and protein sequences, although we know that intramolecular interactions between residues of a protein play a very important role in determining protein structure. Regarding the secondary structures of RNAs, where base pairing introduces very critical long range dependencies, the model of independent mutations is biologically inaccurate [66, 74, 79].

The most important dynamic programming algorithm in biological sequence analysis for obtaining the optimal global alignment between two sequences allowing gaps is the Needleman-Wunsch algorithm [98, 11].

The idea is to build up an optimal alignment using previous solutions for optimal alignments of smaller subsequences. A matrix F of the two sequences is constructed, indexed by i and j, one index for each sequence, where the value $F(i,j)$ is the score of the best alignment between the initial segment $x_{1\dots i}$ of x up to x_i and the initial segment $y_{1\dots j}$ of y up to y_j. The score value $F(i,j)$ is builded recursively and we start by initialising $F(0,0) = 0$.

As mentioned above we have three possibilities of pairs of opposite symbols: (i) identity, (ii) substitution or 'mismatch' and (iii) insertion or deletion. The values $F(i,0) = -id$ represent gaps in y and vice versa $F(0,j) = -jd$ for gaps in x. If $F(i-1,j-1), F(i-1,j)$ and $F(i,j-1)$ are known, it is possible to calculate $F(i,j)$. There are three possible ways that the best score $F(i,j)$ of an alignment up to x_i, y_j could be obtained: x_i could be aligned to y_j, in which case $F(i,j) = F(i-1,j-1) + s(x_i,y_j)$, where $s(x_i,y_j)$ is the individual score for this pair of amino acids or nucleotides; or x_i is aligned to a gap, in which case $F(i,j) = F(i-1,j) - d$, where d is the gap penalty; or y_j is aligned to a gap, in which case $F(i,j) = F(i,j-1) - d$. The best score up to (i,j) is the largest of these three options. Therefore, we have

$$F(i,j) = \sup \left\{ \begin{array}{c} F(i-1,j-1) + s(x_i,y_j), \\ F(i-1,j) - d, \\ F(i,j-1) - d. \end{array} \right\} \tag{2}$$

This equation is applied repeatedly to fill in the matrix of $F(i,j)$ values. To gain the alignment itself, we follow the pointers that we stored when building the matrix, a procedure called backtracking.

3.2 Multiple Alignments

Using dynamic programming in order to align just two sequences guarantees a mathematically optimal alignment. But attempts at generalising dynamic programming to multiple alignments are limited to small numbers of short sequences [83]. For much more than ten sequences the problem is infeasible given current computer power. Nowadays, the most widely used approach is to exploit the fact that homologous sequences are evolutionary related. Multiple alignment are produced progressively by a series of pairwise alignments, following the branching order in a phylogenetic tree [38]. First all possible pairs of sequences are aligned to derive a distance matrix in order to calculate the initial guide tree which is built up by the distances between the sequences. Then the most closely related sequences get aligned progressively according to the branching order in the guide tree, gradually adding in the more distant ones when we already have some information about the most basic mismatches or gaps.

This approach is fast enough to allow alignments of virtually any size. Further, in most (simple) cases, the quality of the alignments is very good, as judged by the ability to correctly align corresponding domains from sequences of known secondary or tertiary structures [8]. The placement of gaps in alignments between closely related sequences is much more accurate than between distantly related ones. Therefore, the positions of the gaps which were introduced during the early alignments of the closely related sequences are not changed as new sequences are added. One problem is that this approach becomes less reliable if all of the sequences are highly divergent. More specifically, any mistakes like misaligned regions made early in the alignment process cannot be corrected later as new information from other sequences is added. Thus, there is no guarantee that the global optimal solution has been found and the alignment is not captured in a local minimum. This risk increases with the divergence of the initially aligned sequences.

Furthermore, the parameter choice a weight matrix and two gap penalties (one for opening a new gap and one for extension of an existing gap) is very important. When the sequences are closely related identities dominate an alignment, almost any weight matrix will find approximately the correct solution. With very divergent sequences the scores given to non-identical residues will become critically important, because there are more mismatches than identities. The range of gap penalty values which will find the correct or best possible solution can be very broad for highly similar sequences, but the more divergent the sequences are, the more exact values of gap penalties have to be used [137].

3.3 `ClustalW`

A widely used multiple alignment program is `ClustalW` [132]. `ClustalW` addresses the alignment parameter choice problem and dynamically varies the gap penalties in a position- and residue-specific manner. As the alignment proceeds, `ClustalW` chooses different weight matrices depending on the estimated divergence of the sequences to be aligned at each stage. Some matrices are appropriate for aligning very closely related sequences where most weight by far is given to identities, with only the most frequent conservative substitutions receiving high scores. Other matrices work better at higher evolutionary distances where less importance is attached to identities. Besides, sequences are weighted by `ClustalW` to correct for unequal sampling across all evolutionary distances in the data set [136, 137]. This down-weights sequences which are very similar to other sequences in the data set and up-weights the most divergent ones. The weights are calculated directly from the branch lengths in the initial guide tree [133, 132]. In `ClustalW` the initial guide tree used to guide the multiple alignment, is calculated using the neighbour-joining method [118] which is quite robust against the effects of unequal evolutionary rates in different lineages and gives good estimates of individual branch lengths. These branch lengths are then used to derive the sequence weights. Furthermore, it is possible for the user to choose between fast approximate alignments [9] or full dynamic programming for the distance calculations used to make the guide tree.

The trees used to guide the final multiple alignment process are calculated from the distance matrix derived in the first step. This produces unrooted trees with branch lengths proportional to the estimated divergence (Figure 6). Then the root of one tree is established at a position where the means of the branch lengths on either side of the root are equal. These trees are then also used to derive a weight for each sequence.

The basic procedure of the progressive alignments is to use a series of pairwise alignments to align larger and larger groups of sequences, following the branching

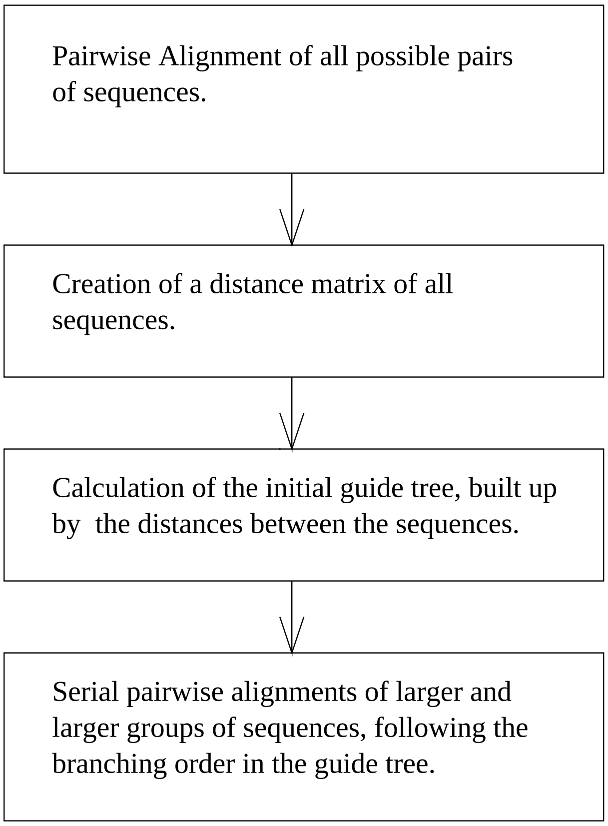


Figure 2: A flow chart showing the main steps of `ClustalW`.

order in the guide tree. First the most similar sequences at the tips of the tree get aligned. Then this alignment gets aligned with the third most similar sequence and so on. At each stage a full dynamic programming algorithm [97] is used with a residue weight matrix and penalties for opening and extending gaps. `ClustalW` varies gap penalties used with different weight (substitution) matrices to improve the accuracy of the sequence alignments. Further, the per cent identity of the two (groups of) sequences to be aligned is used to increase the gap opening penalty for closely related sequences and to decrease it for more divergent sequences. Also, if there are already gaps at a position, then the gap opening penalty is reduced in proportion to the number of sequences with a gap at this position and the gap extension penalty is lowered by a half.

Finally, `ClustalW` offers two main series of weight matrices for proteins to the user: the `Dayhoff PAM` series [31] and the `BLOSUM` series [54]. In each case there is a choice of matrices ranging from strict ones, useful for comparing very closely

```
SARGLSSTVSLGQFEHWSPR
+AR+LS+TVSL+QF+H SPR
NARNLSDTVSLSQFDHPSPR

AGTGCAAGAGGATTAAGTAGTACAGTAAGTTTAGGACAATTTGAACATTGGAGTCCAAGA
   GC  G G   T       AC G      T    CA TT GA CA       CC  G
GACGCCCGCGACCTCTCCGACACCGCTTCCCTCTCCCAGTTCGACCACCCCTCCCCCCGC
```

Figure 3: Example for the problem of higher sequence heterogenity on the level of nucleic acids. It shows an hypothetical amino acid alignment on top which represents a high degree of similarity between both protein sequences allowing for an unambiguous alignment. Below the same sequences are aligned on the level of nucleic acids. It is clearly visible that the sequences are much more heterogenous: the pairwise identity is only 33%. This is only slightly above the 25% identity expected for two random nucleic acid sequences.

related sequences, to less strict ones which are useful for aligning more divergent sequences. Depending on the distances between the two sequences or groups of sequences to be compared, `ClustalW` switches between four different matrices in each series. The distances are measured directly from the guide tree.

3.4 `Ralign`

Difficulties of Nucleic Acid Alignments

Alignments of nucleic acid sequences suffer from the small alphabet size and the fact that nucleic acid sequences evolve in generell faster than their proteins. The sequence heterogenity on the level of nucleic acid makes good alignments often infeasible. While protein sequences can still show substantial homology, the corresponding nucleic acid sequences are already essentially randomized. This is caused by the inherent redundancy of the genetic code: most amino acids have more than one codon on the level of nucleic acid. In a protein alignment these amino acids would match each other while the differences on the level of nucleic acids can produce gaps within coding regions in a nucleic acid alignment. Whereas, on the level of protein alignments many of this gaps could have been avoided.

Therefore, in most cases it is possible to obtain better alignments on the level of protein than on the level of nucleic acids (Figure 3). The scores (the per cent homologies) are higher and the number of gaps within the protein sequences is lower than in the case of nucleic acids. Reducing the gaps within an alignment improves the resulting alignment which may be used as input into other sequence data processing programs such as those searching for sequence covariations.

The Ralign Algorithm

A combined amino acid and nucleic acid based alignment procedure is made available in a program called `Ralign` developed by Roman Stocsits [127] and described in his diploma thesis.

The idea behind `Ralign` is that coding regions on the level of protein vary less than on the level of nucleic acid, because most amino acids are coded by more than one codon (base triplet) and some different nucleic acid sequences can produce the same protein sequence after translation. Thus, his approach was to improve the quality of sequence alignments of RNA viruses by creating and implementing a combined alignment algorithm.

One could argue that the quality of sequence alignments could be raised simply by translating the entire nucleic acid sequence into protein and processing on the level of proteins. The straight forward use of amino acid based alignments is, however, complicated by the overlapping reading frames that are often encountered in viral genomes. Overlapping open reading frames are possible, hence one part of the nucleic acid sequence codes for more than one protein in different frames. Theoretically, three open reading frames can be covered by the same nucleic acid sequence (e.g. realized in the hepatitis B virus). In addition, various non-coding regions can exist in a certain virus genome. These non-coding regions should, of course, be aligned as nucleic acids, and every open reading frame should be processed in the correct frame.

`Ralign` creates an output file which contains all data about the detected open reading frames including information about their length, start and stop, and the lengths of their proteins after translation. Also a second file is created: a `PostScript` output file which gives a graphical representation of the found open reading frames, see Figure 4.

In many cases we can see significant differences in the genetic structure regarding the number and order of various open reading frames even between very closely

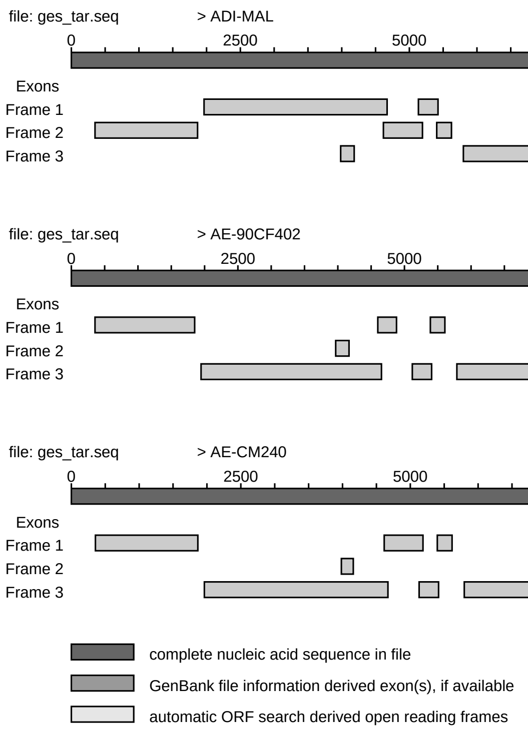


Figure 4: An example for the `PostScript` output of `Ralign`. The figure shows a graphical representation of the found open reading frames of HIV-1 sequences. The exon-intron organization in the is shown (as obtained through the `GenBank` file) and also the found open reading frames.

related sequences. This makes it difficult to decide which ORFs correspond to each other in the various sequences. Furthermore, if a certain part of the sequence is coding for two or three proteins, a decision has to be made which open reading frame is used for the protein alignment.

The proposed assignment of the open reading frames can be altered by the user. After the user has either manipulated or accepted the chosen open reading frames, `Ralign` uses `ClustalW` to align the homologous sequence parts, first the protein subsequences then the non-coding regions. The protein alignments are then reverse translated and for every gap of length n, a gap of $3n$ is inserted into the corresponding nucleic acid sequence at the corresponding site. Finally, all alignments, either on the level of protein or on the level of nucleic acid, are combined and a resulting output file of the complete nucleic acid sequence alignment is produced. The final alignment is checked for too divergent sequence which have to be removed.

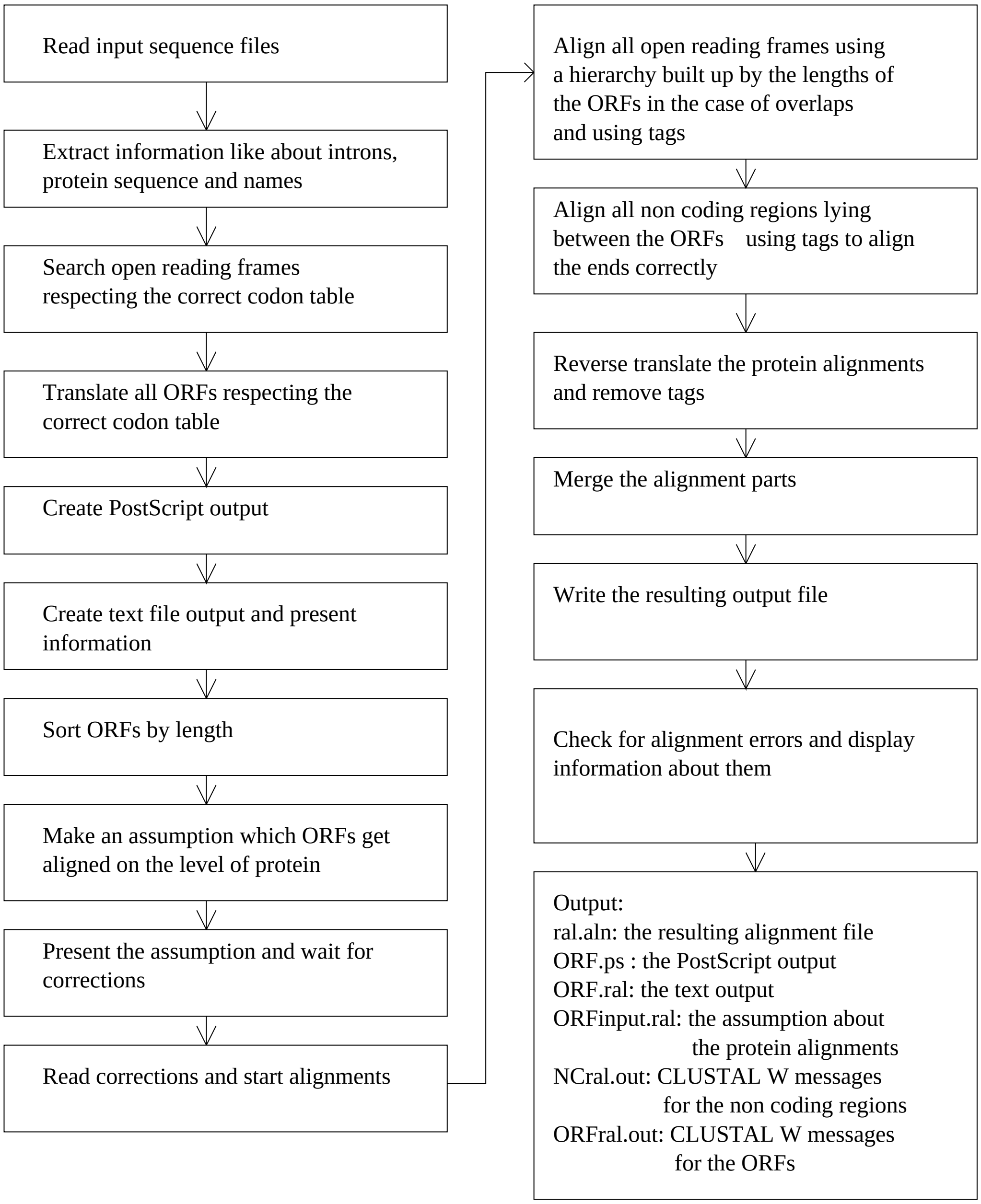


Figure 5: This flow chart shows the main steps of `Ralign` [128].

3.5 Splitstree, Split Decomposition

Evolutionary relationships between taxa are generally represented by phylogenetic trees, and many different algorithms for tree construction have been developed. This is justified by the assumption that evolution is a tree-like or branching process. However, a set of real data often contains a number of different and sometimes conflicting signals and thus does not always clearly support an unique tree. To address this problem, H.-J. Bandelt and A.W.M. Dress (1992) developed the method of split decomposition.

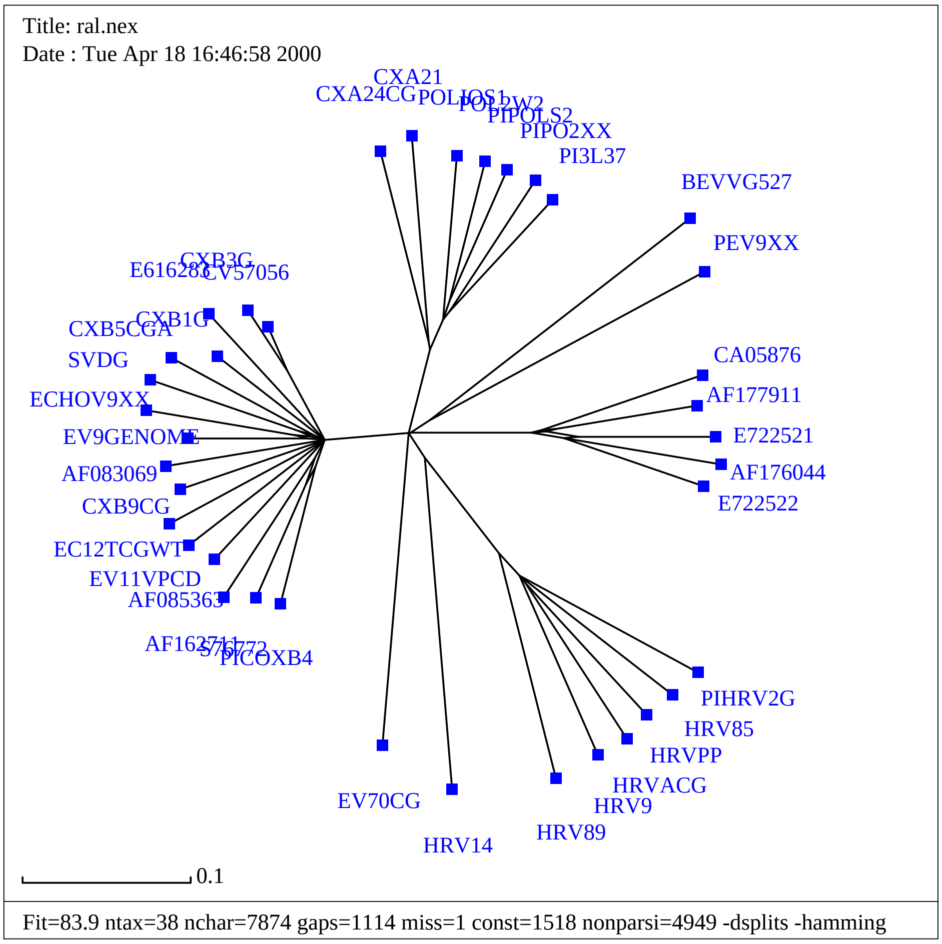


Figure 6: A distance matrix derived `Splitstree` plot for rhino- and enteroviruses. The branch lengths are proportional to the estimated divergences.

In contrast to methods such as maximum parsimony and maximum likelihood that reconstruct phylogenetic trees by optimizing parameters, split decomposition is a transformation-based approach. Essentially, evolutionary data is transformed or, more precisely, ”canonically decomposed”, into a sum of ”weakly compatible splits” and then represented by a so-called splits graph. It takes as input a distance matrix or a set of aligned sequences and produces as output a graph that represents the evolutionary relationships between the taxa, see Figure 6. For ideal data, this is a tree, whereas less ideal data will give rise to a tree-like graph that can be interpreted as possible evidence for different and conflicting phylogenies. Further, as split decomposition does not attempt to force data onto a tree, it can provide a good indication of how tree-like given data is [67].

4 Searching Conserved RNA Secondary Structures

Recently, a set of computer methods to scan moderate size samples of RNA sequences for conserved secondary structures was developed at our institute (described in Section 4.2). A brief overview of convenient ways to represent secondary structures is given in the following section.

4.1 Representing the Structure

Bracket Notation

The unique decomposition of secondary structures 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.

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 minimum free energy (mfe) structure for comparison, see Figure 7. Note, the mfe is the ground state structure, but not necessarily the most probable structure in the ensemble. Hair-

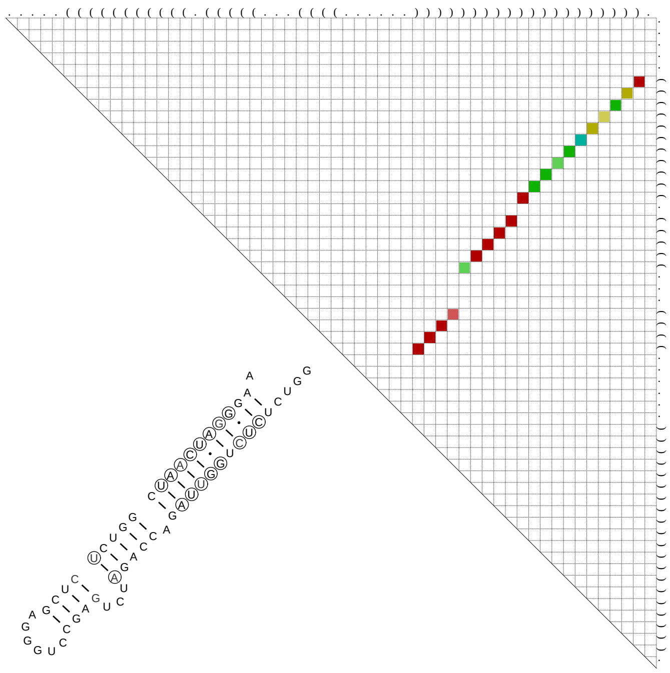


Figure 7: Colored dot plot and consensus structure the TAR structure of HIV-1. The upper right part shows the predicted base pair probabilities computed with the partition function algorithm of the `Vienna RNA Package`. (Here only the base pairs that occur in the mfe are indicated.) The area of the squares is proportional to the pairing probability. Colors indicate the number of consistent mutations 1, 2, 3 different types of base pairs. Saturated colors, , indicate that there are only compatible sequences. Decreasing saturation of the colors indicates an increasing number of non-compatible sequences: 1, 2 sequences that cannot form a (i, j). If there are more than 2 non-compatible sequences the entry is not displayed. The lower left part gives the computed consensus structure which matches the structure determined by probing and phylogenetic reconstruction [10]. A large number of compensatory mutations indicated by circles supports the thermodynamic predictions.

pin 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.

Mountain Representation

A convenient way of displaying the size and distribution of secondary structure elements is the *mountain representation* [65]. 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 8. 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 [76]. A modified version of the mountain representation [65] can be constructed easily from the base pairing probability matrix. The number

$$m(k) := \sum_{i<k} \sum_{j>k} p_{ij} \tag{3}$$

counts all base pairs enclosing[1] 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 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 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.

4.2 The Search Algorithm

`RNAfold` as part of the `Vienna RNA Package`[2] [61] reads RNA sequences from

[1] In the terminology of Zuker and Sankoff [156] these are all base pairs to which sequence position k is interior.

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

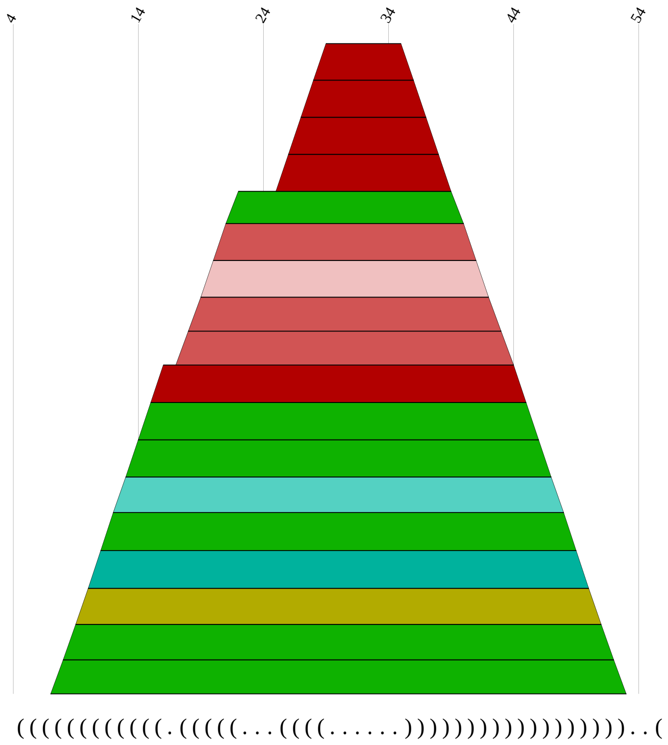


Figure 8: An example of a colored mountain plot of the TAR element in HIV-1. This representation is a convenient alternative to dot plots for larger structural motifs. Base pairs are represented by slabs connecting the two sequence positions. The width and color of a slab corresponds to size and color coding of the corresponding dot plot entry (Figure 7). The lower part shows the corresponding bracket notation.

`stdin` and calculates their mfe structure, partition function and base pairing probability matrix [135, 92]. 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` output files with plots of the resulting secondary structure graph and a dot plot of the base pairing matrix. The dot plot shows a matrix of squares with area proportional to the pairing probability in the upper half, and one square for each pair in the mfe structure in the lower half, see Figure 7. The results of `RNAfold` are used as an input for `alidot` and `pfrali` [59, 62, 61].

The programs `alidot` and `pfrali` for detecting conserved secondary structures [59, 63] aim at utilizing the information contained in a multiple alignment of a relatively small set of related RNA sequences to extract conserved features from the pool of plausible structures generated by thermodynamic prediction for each sequence. The basic two inputs for the algorithm `pfrali` are a multiple sequence alignment and the base pair probabilities from McCaskill's algorithm. A flow chart is shown in Figure 9. Our approach is different from efforts to simultaneously compute alignment and secondary structures [121, 129, 26]. One disadvantage of these methods is the much higher computational cost which makes them unsuitable for long sequences such as viral genomes. Furthermore, they assume implicitly that all sequences have a common structure, not just a few conserved structural features. The same is true for the related program `Construct` [85].

While the related `alidot` method [59] uses only mfe structures, i.e. one structure per sequence, `pfrali` uses base pairing probabilities as obtained from McCaskill's partition function algorithm. Since the base pairing probabilities contain information about a large number of plausible structures, this approach is less likely to miss parts of the correct structures. In both cases, we make explicit use of the sequence variation to select the credible parts of the predicted structures. Thus, we do not assume *a priori* that there is a conserved secondary structure for all (or even most) parts of the sequence.

We calculate the multiple sequence alignment using `Ralign` and `ClustalW` [132], respectively. No attempt is made to improve the alignment based on predicted secondary structures. While this might increase the number of predicted structural elements, it would also compromise the use of the sequence data for verifying these structures. Furthermore, we find that most regions that have functional secondary structure tend to align fairly well, at least locally.

The `pfrali` program reads the pair probabilities from dot plots files in `PostScript` format as well as a multiple sequence alignment in `ClustalW` format. The gaps in the alignment are inserted into the corresponding probability matrices. We can now superimpose the probability matrices of the individual sequences to produce a *combined dot plot*. To keep the number of base pairs manageable we keep only pairs that occur with a probability of at least $p^* = 10^{-3}$ for at least one sequence. Base pairs with even lower probabilities are very unlikely to be part of an important structure. In the combined dot plot the area of a dot at position i, j is proportional to the mean probability $\bar{p}_{i.j}$ (averaged over all sequences). In addition we use a color coding to represent the sequence variation. The number of non-compatible sequences, and the number $c_{i.j}$ of different pairing combinations is incorporated in the combined dot plot as color information. For details of the encoding scheme, see the caption of Figure 7.

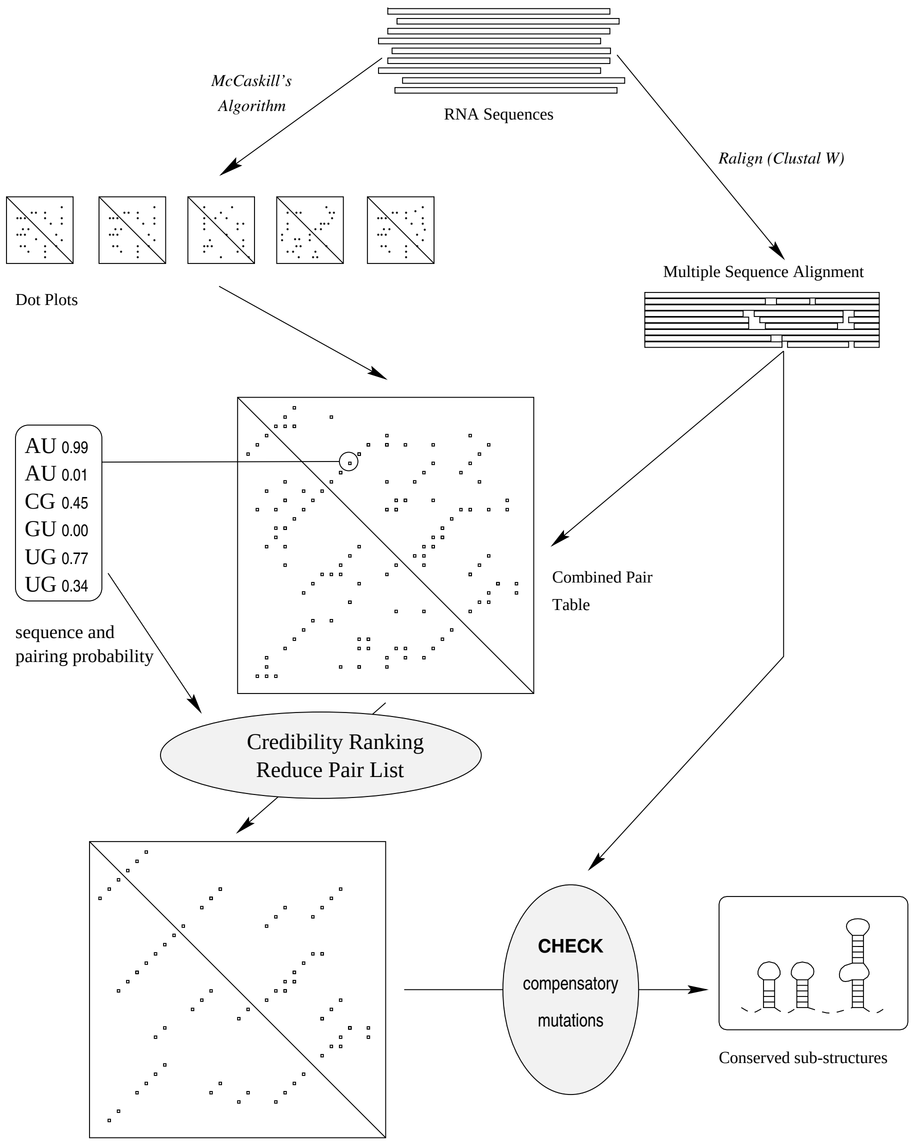


Figure 9: Flow diagram of the algorithm. A multiple sequence alignment is calculated using `Ralign`. RNA genomes are folded using McCaskill's partition function algorithm as implemented in the `Vienna RNA Package`. The sequence alignment is then used to align the predicted structures. From this structural alignment we extract putative conserved regions. In the final step the sequence information, in particular compensatory mutations, are used for validating or rejecting predicted structure elements.

A sequence is *compatible* with base pair $(i.j)$ if the two nucleotides at positions i and j of the multiple alignment can form either a Watson-Crick (**GC**, **CG**, **AU**, or **UA**) pair or a wobble (**GU**, **UG**) pair. When different pairing combinations are found for a particular base pair $(i.j)$ we speak of *consistent* mutations. If we find combinations such as **GC** and **CG** or **GU** and **UA**, where both positions are mutated at once we have *compensatory* mutations. The occurrence of consistent and, in particular, compensatory mutations strongly supports a predicted base pair, at least in the absence of non-consistent mutations.

Phylogenetic methods in general consider only compensatory mutations even though **GU** base pairs are clearly important as evidenced by the fact that **RY**→**YR** conversions are rare [57]. While compensatory mutations of the type **RY**→**RY**, such as **GC**→**AU**, can be obtained by two subsequent consistent point mutations, for instance **GC**→**GU**→**AU**, a double mutation is required for **RY**→**YR** mutations. We argue therefore that all consistent mutations, not only compensatory ones, should be seen as support for a proposed structure.

The base pairs contained in the combined dot plot will in general not be a valid secondary structure, i.e., they will violate one or both of the following two conditions: (i) No nucleotide takes part in more than one base pair. (ii) Base pairs never cross, that is, there may not be two base pairs $(i.j)$ and $(k.l)$ such that $i < k < j < l$. In the reminder of this section we describe how to extract credible secondary structures from the list of base pairs.

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

(i) The more sequences are non-compatible with $(i.j)$, the less credible is the base pair.

(ii) If the number of non-compatible sequences is the same, then the pairs are ranked by the product $\bar{p}_{i.j} \times c_{i.j}$ of the mean probability and the number of different pairing combinations.

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

The list now represents a valid secondary structure, albeit still containing ill-supported base pairs. Since our goal is to produce a list of well-supported secondary structure features that contains as few false positive as possible, we use a series of additional “filtering” steps: First, we remove all pairs with more than two non-compatible sequences, as well as pairs with two non-compatible sequences adjacent to a pair that also has non-compatible sequences. Helices with so many non-compatible sequences can hardly be called “conserved”. (For large samples these rules might have to be modified to tolerate somewhat larger numbers of non-compatible sequences.) Next, we omit all isolated base pairs. The remaining pairs are collected into helices and in the final filtering step only helices are retained that satisfy the following conditions: (i) the highest ranking base pair must not have non-compatible sequences. (ii) for the highest ranking base pair the product $\bar{p}_{i.j} \times c_{i.j}$ must be greater than 0.3. (iii) if the helix has length 2, it must not have more non-compatible sequences than consistent mutations. In general, these filtering steps only remove insignificant structural motifs that one would have disregarded upon visual inspection anyways. The remaining list of base pairs is the conserved structure predicted by the `pfrali` program.

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

Manual reconstruction of a consensus structure proved to be a time-consuming and error-prone task. In contrast, the structure in Figure 7 was produced without human intervention except for the layout of the structure drawing using the program `XRNA` [151].

4.3 Vienna RNA Viewer

Especially large virus genomes of several thousand nucleotides overwhelm the investigator with data. Therefore Ivo Hofacker and Martin Fekete at the *Institute for Theoretical Chemistry and Molecular Structural Biology* developed a graphical viewing tool in `perl` and `perlTk` called `Vienna RNA Viewer` [37]. This algorithm provides a more user friendly presentation of RNA secondary structures and substantially facilitates the analysis of large amounts of data. This graphical viewing tool with options for a semi-automatically selecting of conserved RNA secondary structures, accepts either output format produced by the

Vienna RNA Package and from our conserved structure searching algorithms, alidot and pfrali. The program uses as input dot plot files or the output file format from alidot and pfrali. Although several viewing tools are known for RNA secondary structures, e.g. RNAviz[3], XRNA[4] this new viewing tool made our search for conserved RNA secondary structure patterns feasible, since it allows us to process larger data sets and deal with dot plots.

The Vienna RNA Package produces a so-called dot plots in PostScript file format, containing the information of the secondary structure of the folded RNA sequence, either mfe and base pair probability, see Figure 10. In the main window of the Vienna RNA Viewer we see a typical dot plot output. The upper right triangle shows the base pairing probability matrix (p_{ij}), and the lower left part contains the mfe structure for comparison. Squares denote base pairs and their size the probability in the ensemble of structures. By *left-mouse* click on the colored squares information on this base pair can be obtained, e.g. the base pair position, the pairing nucleotides, and the probability, as well as the mfe of the enclosed structure. Several additional functions are available, e.g. zooming in and out the dot plot, centering an inserted base pair position, saving the screen as a PostScript screen shot of the main window, redrawing the main window. Special features are a Basepair List created in a new window with a list of all base pairs sorted by their credibility, where a base pair can be selected and centered in the main window. A Mountain Plot can be drawn for a region enclosed by a selected base pair. A Stack List allows to search for stacking regions, with respect to a minimal stack size and a minimum base pairing probability of the stacking base pairs. All stacks matching the search criteria are listed in the stack list window. PostScript output files of the RNA secondary structure and Xrna compatible structure files of a selected region can be created. The main functions implemented to the Vienna RNA Viewer are shown in Figure 11.

The Vienna RNA Viewer was designed to deal with dot plot files produced by either pfrali or alidot. The analysis of complete virus genomes with several thousand nucleotides, such as HIV could hardly be done in reasonable time without an investigation tool, that helps filtering the information.

Conserved RNA secondary structures are always present in RNA sequences. A

[3] http://www-rrna.uia.ac.be/rnaviz

[4] ftp://fangio.ucsc.edu/pub/XRNA

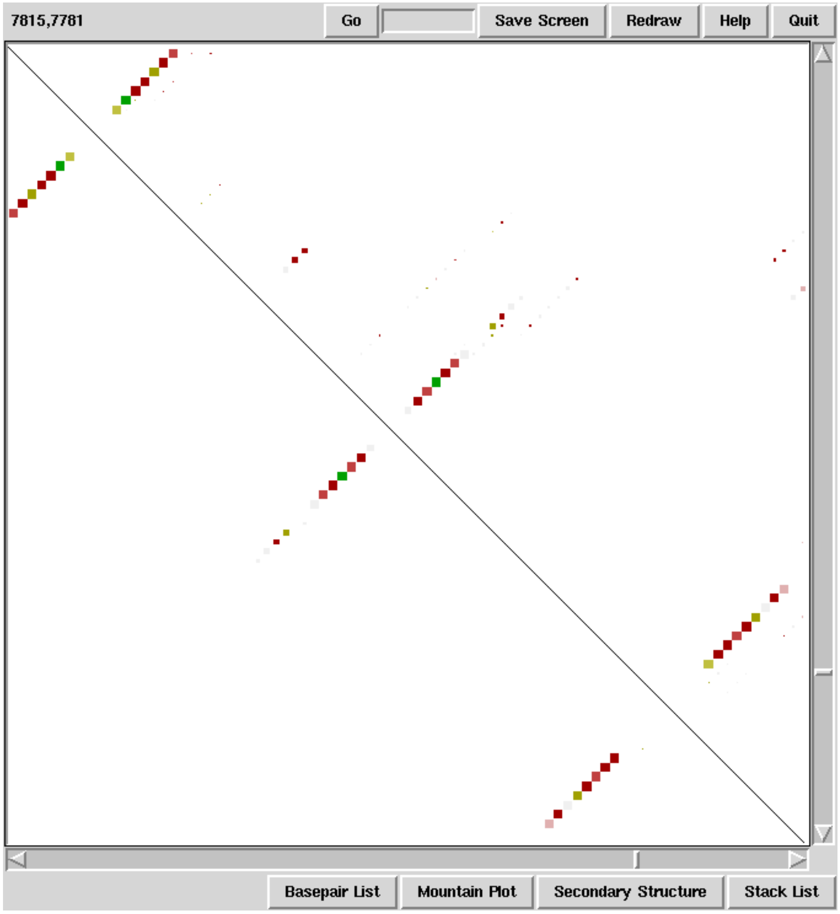
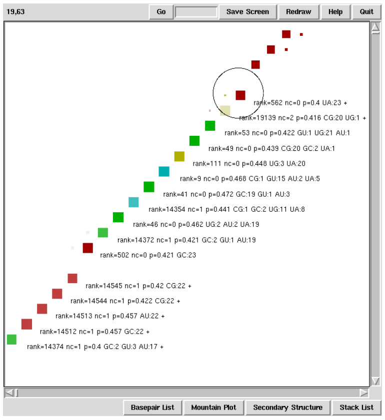
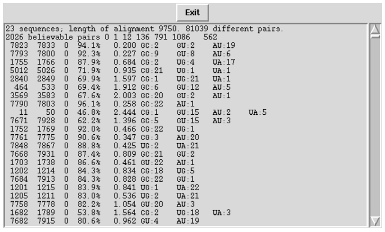


Figure 10: The main window of the `Vienna RNA Viewer` showing a dot plot output of 23 aligned sequences of the RRE in HIV-1. The upper left triangle displays the base pair probabilities. Consistent and compensatory base pairs are colored according to the color coding mentioned in Figure 7. The lower left triangle shows the mfe structure. Additional information on the particular base pairs can be obtained by a *left-mouse* click on the squares.

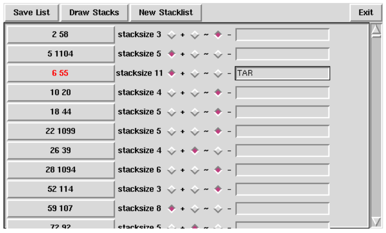
first overview of a large virus genome is provided by the mountain plot, which allows a qualitative analysis, whether conserved motifs can be found. As result of investigating these plots a number of structure files is created, either mountain plots and structure graphs of possibly conserved RNA secondary structures. The `Vienna RNA Viewer` enables to screen through even large data files from complete virus genomes of different families and to create a number of promising conserved RNA secondary structure output files. A first attempt to classify RNA virus species on basis of their conserved structural motifs can be made.



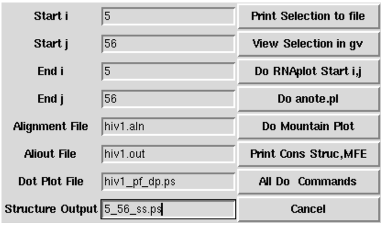
11.1 Main window



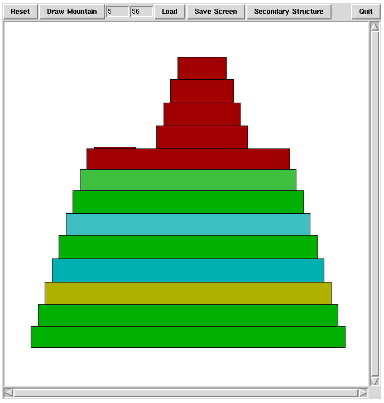
11.2 Base pair list



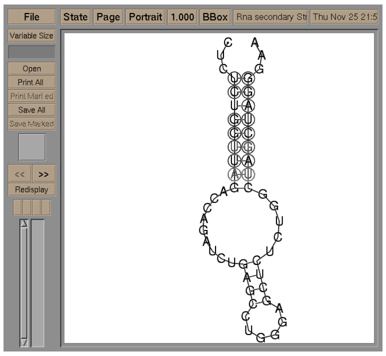
11.3 Selected stacks



11.4 Output selection



11.5 Mountain Plot



11.6 Secondary structure

Figure 11: `Vienna RNA Viewer` windows displaying the TAR element of HIV-1. 11.1 Dot plot (`pfrali` output) showing rank, number of non-compatible sequences, probability, and base pairs in the sequences. 11.2 List of base pairs sorted by their credibility. 11.3 List of stacks for a chosen minimum stack size and minimum base pair probability. 11.4 Window for selecting the region and the `PostScript` output files. 11.5 A colored Hogeweg mountain plot of the `pfrali` output. 11.6 Selected secondary structure shown in a `ghostview` window. Compensatory and consistent mutations are denoted by circles around the bases.

5 Retroviridae

5.1 Primate Lentiviruses

Knowledge gained over the last century has established that retroviruses cause a wide variety of diseases in many avian and mammalian species, including nonhuman primates. Although individual cases of unexplained immunosuppression accompanied by opportunistic infection, neurologic disorder, and unusual forms of cancer were recognized in industrialized societies during the 1960s and 1970s, 1981 proved to be a turning point in the recognition of a new epidemic of acquired immunodeficiency syndrome (AIDS). Outbreaks of immunodeficiency-associated conditions such as Kaposi's sarcoma, mucosal candidiasis or *Pneumocystis carinii* pneumonia were described. Epidemiologic studies implicated an infectious agent that was transmitted during sexual intercourse, through intravenous drug abuse, by therapies utilizing blood and blood products, and vertically from mother to child. The underlying infectious agent for this newly described immunodeficiency syndrome human retroviruses were identified, named *human immunodeficiency virus type 1* (HIV-1). In 1986, a second HIV (HIV-2) was isolated in West Africa and subsequently in Europe and North America.

The discovery of additional distinct lentiviruses in nonhuman primates as well as in humans has provided important insight into the biologic significance and evolutionary relationships of these viruses. A lentivirus was isolated from captive Asian macaques with an AIDS-like disease. The importance of the nonhuman primate lentiviruses is underscored by the fact that several HIV-2 isolates of West African origin are nearly indistinguishable at the nucleotide sequence level from certain strains of simian immunodeficiency virus (SIV) [40].

Over the course of infection, the virus an individual carries broadens in tropism and biologic variability. Small changes in the envelope glycoprotein amino acid composition can lead to large differences in phenotype. Sequence variation occurs rapidly. Although a predominant HIV species is maintained over time, swarms of quasispecies of subtly altered viruses emerge with broadened tropism and possibly increased cytopathic capacity.

Both HIV and SIV are genetically related members of the lentivirus genus of the Retroviridae. Lentivirus isolates from humans are grouped into one of two types, designated HIV-1 and HIV-2, on the basis of serological properties and sequence analysis. Independent isolates of each type display the greatest sequence variation in the *env* gene, which encodes the glycoprotein in the virion membrane. A classification scheme, based on *env* gene sequences, recognizes nine subtypes of HIV-1 (A through I); HIV-2 isolates have been classified into five subtypes (A through E). Accordingly, viral diversification, or speciation, is a feature of HIV-1 and HIV-2 phylogeny. Molecular epidemiology surveys indicate that the pattern of global variation and distribution is due to viral migration rather than to viral mutation. Sequence analysis reveals that the genomes of SIV_{SMM} from sooty mangabeys and HIV-2 exhibit a high degree of homology, and SIV_{CPZ} found in chimpanzees is most closely related to HIV-1 [40], see Figure 12.

Despite extensive sequence diversity, a unifying feature of human and nonhuman primate lentiviruses is that the main cell receptor for attachment is the CD4 antigen, a differentiation marker present on the surface of T-helper lymphocytes. But also other mechanisms for viral entry into certain cells are possible and the search for additional receptors (like the F_c receptor) that facilitate virus entry into cells has been intense.

Infectious virions of HIV and SIV contain two identical copies of single-stranded RNA, about 9.2kb long, that have positive polarity with respect to translation. In the early stage of infection, the virion RNA genome is converted into double-stranded linear DNA by the process of reverse transcription [via viral-encoded reverse transcriptase (RT)], which involves two strand-transfer steps to synthesize linear viral DNA with long terminal repeats (LTRs) flanking viral genes. This linear viral DNA is integrated into the host cell genome to produce provirus. Accordingly, HIV and SIV, like other retroviruses, have two genomic forms: single-stranded RNA in the extracellular phase of the viral life cycle (i.e., virions) and double-stranded DNA (i.e., provirus) within the cell. Genomic viral RNA is synthesized by cellular RNA polymerase II from proviral DNA and thus contains a cap structure at the 5' end and a poly(A) tail at the 3' end. The order of genes encoding structural proteins is invariable *gag-pro-pol-env*. The reverse transcription can be thought of as two phases. The first phase which is mediated by proteins found within the virion, includes entry of the virion core into the cytoplasm, synthesis of double-stranded DNA using the single-stranded genome as template, transfer of the core structure to the nucleus, and integration of the DNA into the host genome. The second phase includes synthesis and proceeding

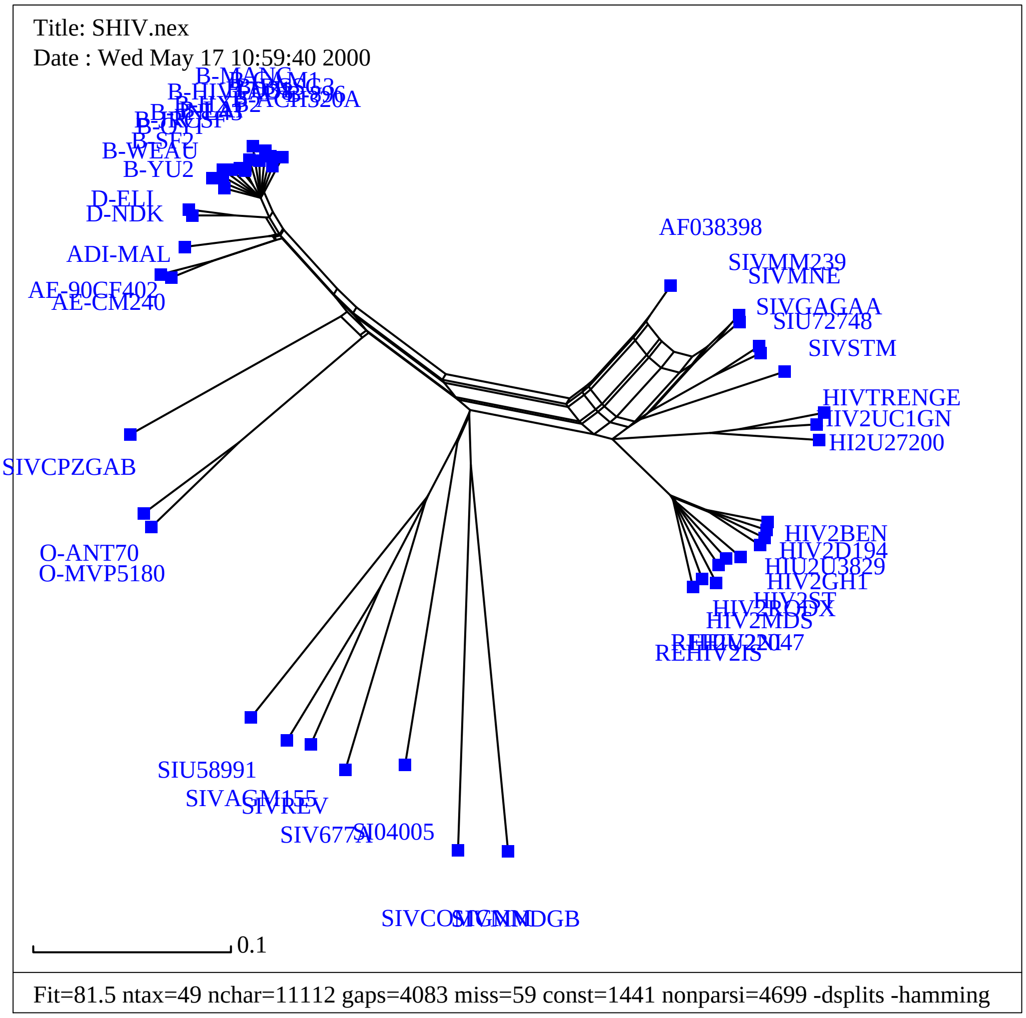


Figure 12: `Splitstree` plot of the aligned sequences of HIV-1, HIV-2, and SIV.

of viral genomes, mRNAs, and proteins using host cell systems including RNA polymerase. Virion assembly proceeds by encapsidation of the genome by unprocessed precursors of the *gag*, *pro*, and *pol* genes, association of the nucleocapsids with the cell membrane, release of the virion by budding, and finally processing of the precursors to the finished products [40].

Packaging of the genomic RNA into a retroviral particle is a highly specific process that achieves selection of a single RNA species from the total polyadenylated mRNA in the infected cell. Two copies of genomic RNA are encapsidated, linked usually at their 5' end through the dimer linkage site (DLS). Dimerisation is associated with encapsidation, but dimer stabilisation occurs post-capture by

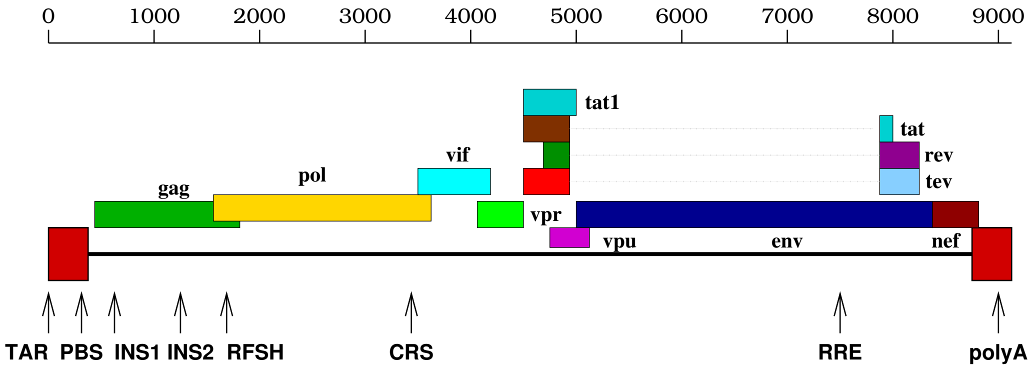


Figure 13: Organization of a retrovirus genome like HIV-1. Proteins are shown on top, known features of the RNA are indicated below. For details about the genes see the text.

the Gag protein. Viral RNAs to be specifically packaged are identified by the presence of an RNA sequence named the packaging signal [152].

Both HIV and SIV encode precursor polypeptides for virion proteins as well as several additional open reading frames (Figure 13). The *gag* gene encodes the precursor for structural proteins for the virion capsid, the *pol* gene encodes the precursor for several virion enzymes [protease (PR), RT, RNase H, and integrase (IN)], and the *env* gene encodes the precursor for envelope glycoprotein. The transcriptional transactivator (*tat*) and regulator of viral expression (*rev*) genes are encoded by two overlapping exons and produce small nonvirion regulatory proteins, Tat and Rev, that can bind to TAR and the RRE, respectively and are essential for viral replication. INS1, INS2, and CRS are RNA sequences that destabilize the transcript in the absence of the Rev protein. RFSH refers to the hairpin that is involved in the ribosomal frameshift from *gag* to *pol* during translation [70]. Poly(A) refers to the polyadenylation signal. PBS is the primer-binding site [13].

The untranslated leader region of a retroviral RNA genome contains multiple signals that control distinct steps of the viral life cycle. Motifs found in most retroviral species are signals that control mRNA splicing, dimerization and packaging, and reverse transcription. The 5' end of the leader RNA forms the R (repeat) region and seems particularly complex because it encodes additional regulatory motifs, an exact copy of which is present at the extreme 3' end of the viral genome [12, 28].

The transactivation response elements (TAR) for HIV-1 has been mapped to a stem-loop at the 5' end of the viral RNA [15, 29, 30], whereas TAR in HIV-2 and SIV is encompassed within the first 130 nucleotides of viral transcripts and appears to fold into a structure of two stem-loops. TAR interacts with the regulatory Tat protein. The binding of the Tat protein is responsible for the activation and/or elongation of transcription of the provirus [39, 75].

One of the best known regulatory elements is the Rev response element (RRE) which acts as a binding site for the Rev protein. The RRE structure is located within the *env* gene (Figure 13). Binding of the Rev protein to the RRE promotes the transport of unspliced HIV transcripts to the cytoplasm [24, 87]. RNA secondary structures play a role in both the entire genomic HIV sequence and in the separate HIV mRNAs, which are basically (combined) fragments of the entire genome [5]. The RRE region forms a well-defined structure on the outside of a large bulk of secondary structure. The consensus structure for the RRE region consists of five hairpins in a multiple branched conformation closed by a single stem structure [77] which separates it from the rest of the RNA molecule. An alternative structure of only four hairpins, in which the hairpins III and IV of the consensus model merge to form one hairpin, has been proposed by Mann *et. al.* [90]. Note, that this alternative structure matches the mfe structure obtained with the old energy parameters.

Extensive computer analysis has shown that the alignment of the RRE at the level of the sequence does not coincide with the alignment at the level of the secondary structure [77]. This has two important implications: 1) methods that predict secondary structure of RNA on the basis of co-variation of positions within the sequence [48] cannot provide unambiguous answers for this region, and 2) the RRE has intrinsic structural versatility and hence it is indispensable to consider ensembles of structures rather than only the single mfe structure.

5.2 Results

To decide whether there are common secondary structure elements within a group of viruses (in this case the lentivirus genus) we conducted the following experiments. In Figure 12 we see the results of a splits decomposition of all the aligned lentiviral sequences. For this analysis we used the complete sequences of all the different representatives of this genus available at the time. We had to edit some of the sequences that we found in the databases which did not start with the TAR element as they are supposed to. It is obvious that certain members of the genus are more closely related to each other than to the rest. For example we can see that HIV-2 and SIV have more common traits than HIV-1 and HIV-2.

For a more in depth analysis of the individual members of the lentivirus genus we performed alignments of the different strains of each member. We found high similarity between the strains of HIV-1 and HIV-2, but interestingly there seem to be two subgroups within the SIV strains (sub1 and sub2). We compared these two subgroups independently. Because we found that according to our prior experiments (Figure 12) HIV-2 and SIV are more closely related to each other we performed alignments between these two members. In detail, we separately compared HIV-2 to SIV_{sub1} and SIV_{sub2}.

In literature we find references [40] stating that the SIV_{sub1} is supposedly more closely related to HIV-1 than to HIV-2. To verify this assumption, we compared HIV-1 to SIV_{sub1}, but we could not find a significant amount of similarity, see Figure 14.

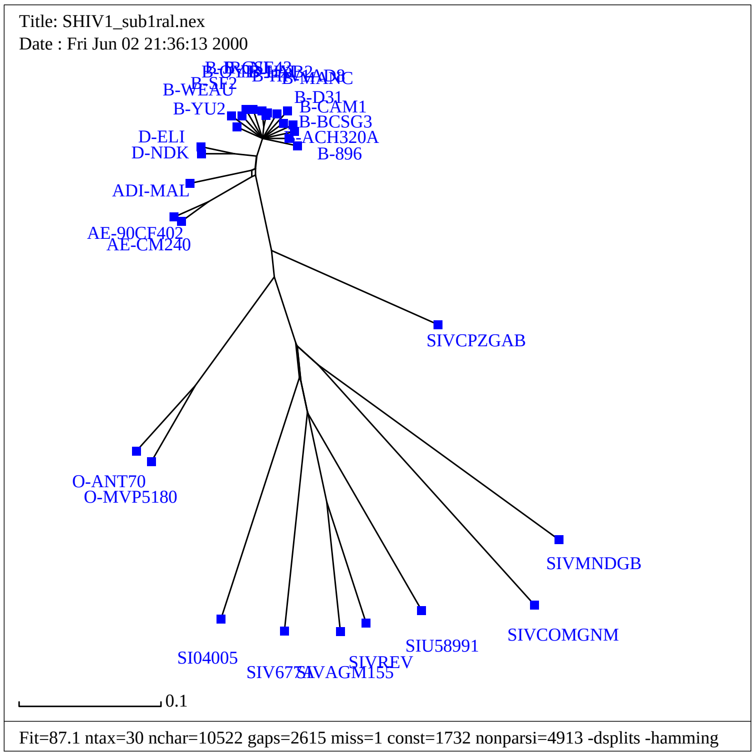


Figure 14: `Splitstree` plot of the aligned sequences of HIV-1 and SIV$_{sub1}$.

5.2.1 Human Immunodeficiency Virus Type 1

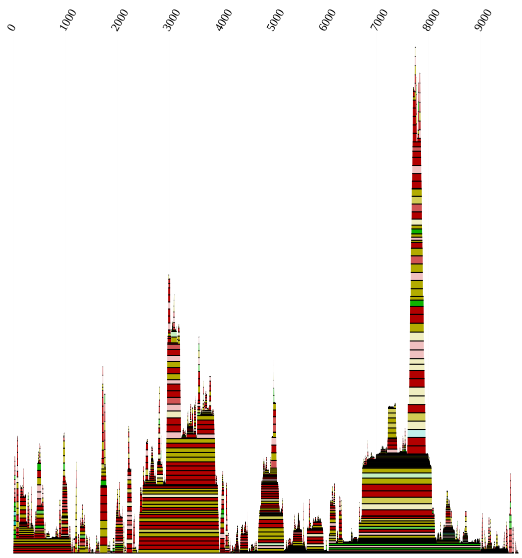
HIV-1 is a highly complex retrovirus with a single stranded RNA genome that is densely packed with information coding for proteins and for structural elements that regulate the viral life cycle [68].

The major genes of HIV-1 are *gag*, *pol*, *env*, *tat* and *rev*, nonessential for the viral replication are *vif*, *vpr*, *vpu*, and *nef* (Figure 13).

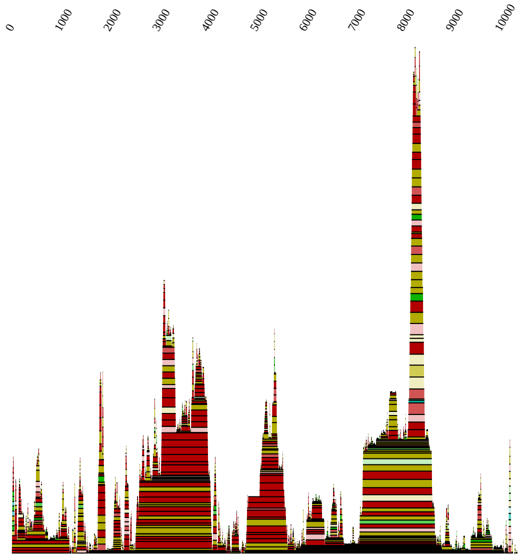
The untranslated leader region of the HIV-1 encodes multiple signals that regulate distinct steps of the viral replication cycle, see Figures 16.1 to 16.5 and Figure 19. The RNA secondary structure of several replicative signals in the leader are critical for function. Well-known examples include the TAR hairpin which is essential for upregulation of viral transcription by the Tat *trans*-activator protein [29, 30, 70]. On the basis of biochemical analysis [10] and computer prediction of the 5' end of the genome it is known that the TAR region in HIV-1 forms a single, isolated stem-loop structure of about 60 nucleotides with about 20 base pairs interrupted by two bulges. The immediately adjacent poly(A) hairpin contains the AAUAAA polyadenylation signal and is critical for repression of this signal in the 5' R of the leader [28], but the same signal is efficiently used in 3' R to provide the HIV-1 RNAs with a regular poly(A) tail. The primer-binding site (PBS) facilitates the binding of the tRNA primer and its subsequent elongation [13]. The dimer initiation site (DIS) [152, 73, 14] has been implicated both in RNA dimerization through base pairing of the palindromic loop sequences and in packaging of the viral RNA. The SD hairpin contains the major splice donor signal, and the Ψ hairpin is a key component of the packaging signal [12].

In Figure 15 we see the alignments of the different HIV-1 strains calculated with three different alignment algorithms (`Ralign, ClustalW`, and `Fasta`). The employed sequences can be found in Table 8. The sequences were found in the `HIV Sequence Database`[5] from Bette Korber *et. al.* Although we used three different algorithms the resulting secondary structure elements are very similar. The positions of the individual elements vary because of the differing alignment lengths of the respective algorithms. The major secondary structures, however are preserved. Due to the similarities between the `Ralign` and the `ClustalW` algorithm, the results are nearly identical. A more detailed analysis of these results can be found in Figures 16 to 18. These figures show the secondary structure elements with the highest base pairing probabilities. This is further underlined by the occurring compensatory and consistent mutations (highlighted by the color code) in these regions. Some of these structural elements have been described in literature [12, 28, 73, 70, 29, 24, 30, 14] like TAR, poly(A), DIS, SD, Ψ, RFSH, and RRE, but we found a substantial number of previously unknown features.

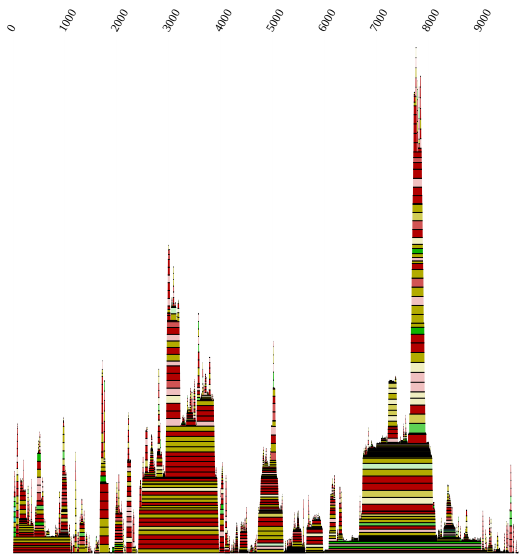
[5] `http://hiv-web.lanl.gov/cgi-bin/get_align.cgi`



15.1 Mountain plot `ClustalW`

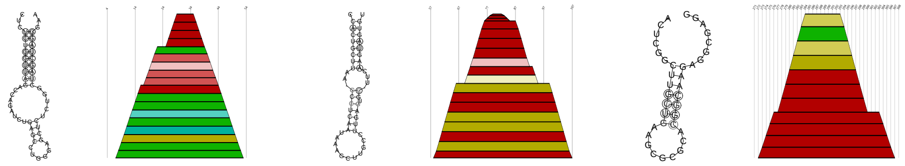


15.2 Mountain plot `Fasta`

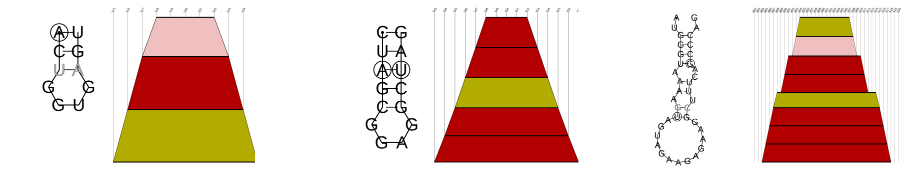


15.3 Mountain plot `Ralign`

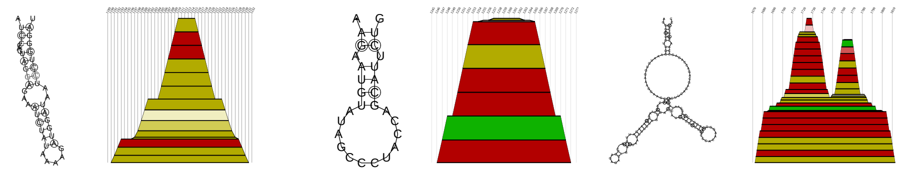
Figure 15: Mountain plots of HIV-1: For information on sequences see Table 8.
23 sequences aligned by `ClustalW`. Alignment length is 9731 bases, conserved 4244 and the mean pairwise homology is 84.9 %.
22 sequences aligned by `Fasta`. Alignment length is 10338 bases, conserved 4441 and the mean pairwise homology is 84.8 %.
23 sequences aligned by `Ralign`. Alignment length is 9750 bases, conserved 4215 and the mean pairwise homology is 84.6 %.



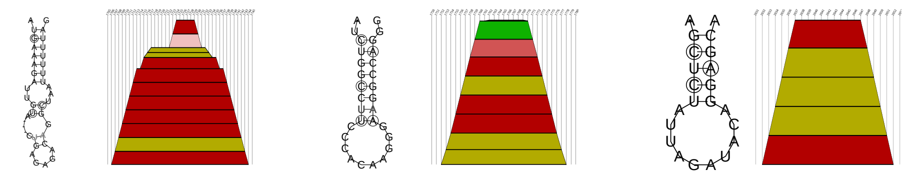
16.1 5-56 TAR   16.2 58-108 poly(A)   16.3 272-308 DIS



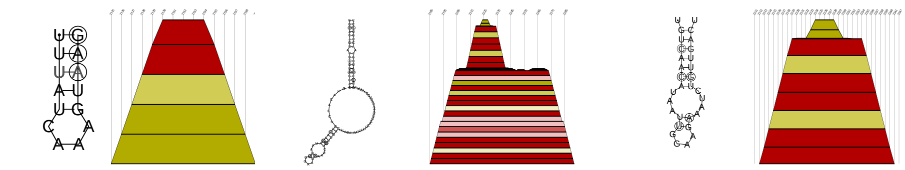
16.4 316-325 SD   16.5 344-357 Ψ   16.6 883-920



16.7 1190-1232   16.8 1246-1273   16.9 1671-1815

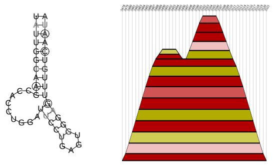


16.10 1706-1745   16.11 1751-1780 RFSH   16.12 2032-2053

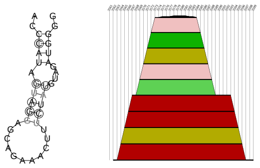


16.13 2126-2139   16.14 2186-2292   16.15 2212-2242

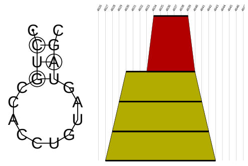
Figure 16: Detected conserved secondary structures of HIV-1 sequences. The numbers denote the base pair range in the `Ralign` alignment. (partI)



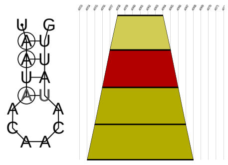
17.1 3479-3521



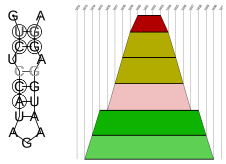
17.2 3562-3599



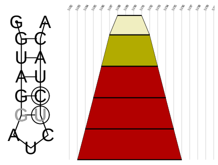
17.3 4027-4044



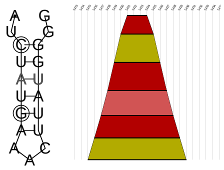
17.4 4954-4969



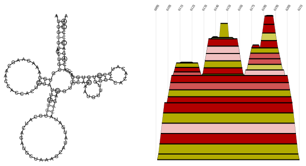
17.5 5013-5031



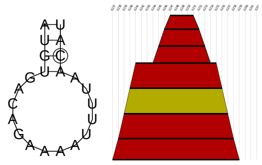
17.6 5183-5197



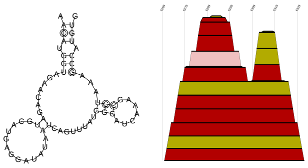
17.7 5414-5432



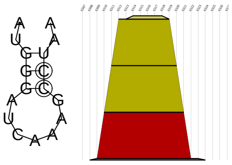
17.8 6094-6214



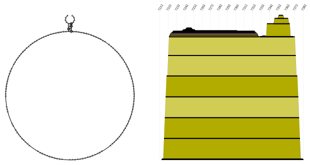
17.9 6238-6259



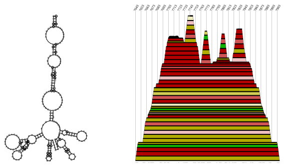
17.10 6270-6333



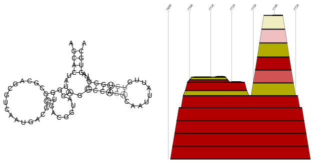
17.11 6308-6324



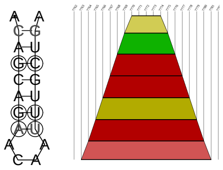
17.12 7213-7383



17.13 7644-7907 RRE

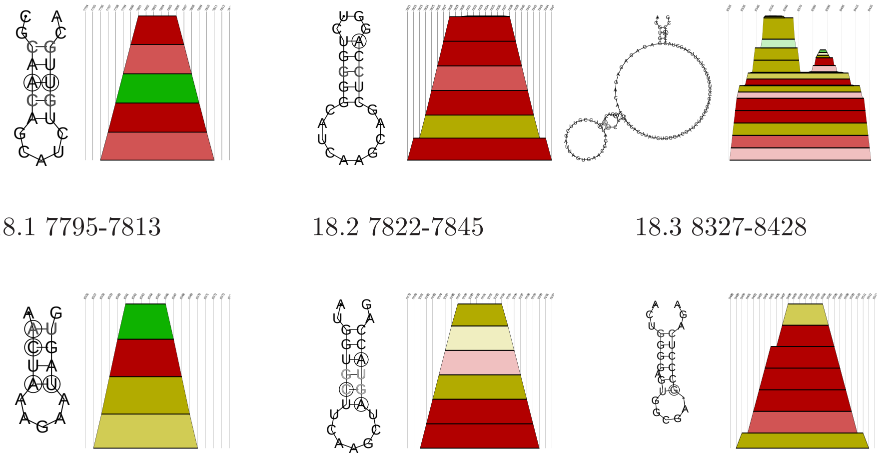


17.14 7695-7762



17.15 7763-7782

Figure 17: Detected conserved secondary structures of HIV-1 sequences. The numbers denote the base pair range in the `Ralign` alignment. (partII)

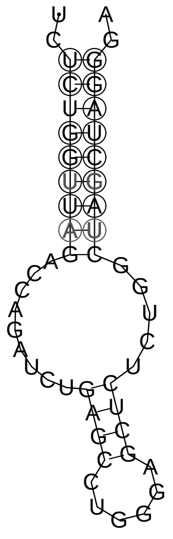


18.1 7795-7813 18.2 7822-7845 18.3 8327-8428

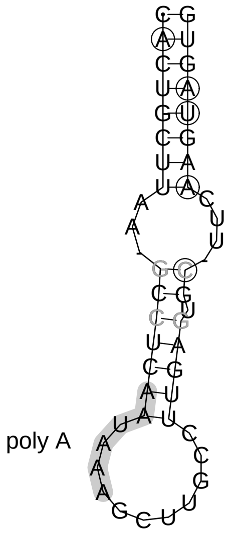
18.4 8557-8571 18.5 9180-9202 18.6 9489-9513

Figure 18: Detected conserved secondary structures of HIV-1 sequences. The numbers denote the base pair range in the `Ralign` alignment. (partIII)

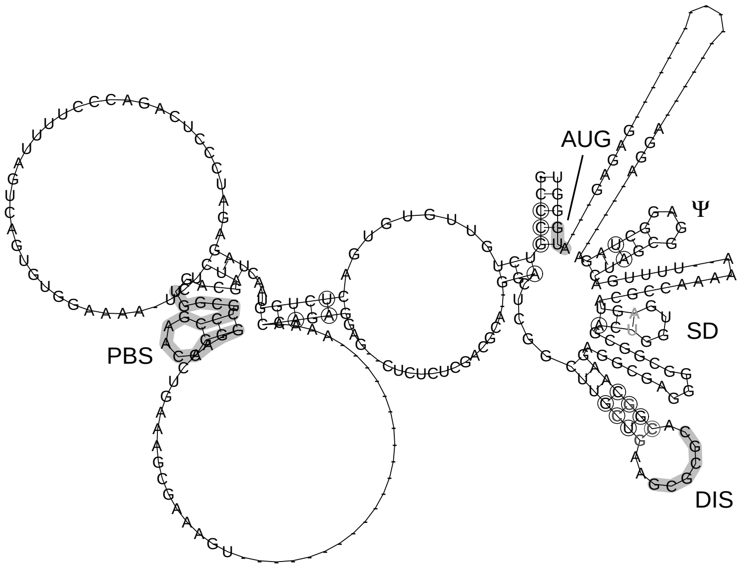
Figure 19 shows the structural elements contained in the leader region described in literature [12, 152, 28, 29, 14]. With our calculations we verified the elements mentioned in this paper. Especially the TAR, poly(A), DIS, and Ψ hairpin are highly conserved, whereas the SD hairpin seems to be shorter, only the PBS region forms a completely different structure [13].



19.1 5-56 TAR



19.2 58-108



19.3 110-399 PBS, DIS, SD, Ψ, AUG

Figure 19: The leader of the HIV-1 RNA genome contains RNA secondary structure elements that are important for viral replication. The numbers denote the base pair range in the `Ralign` alignment.

5.2.2 Human Immunodeficiency Virus Type 2

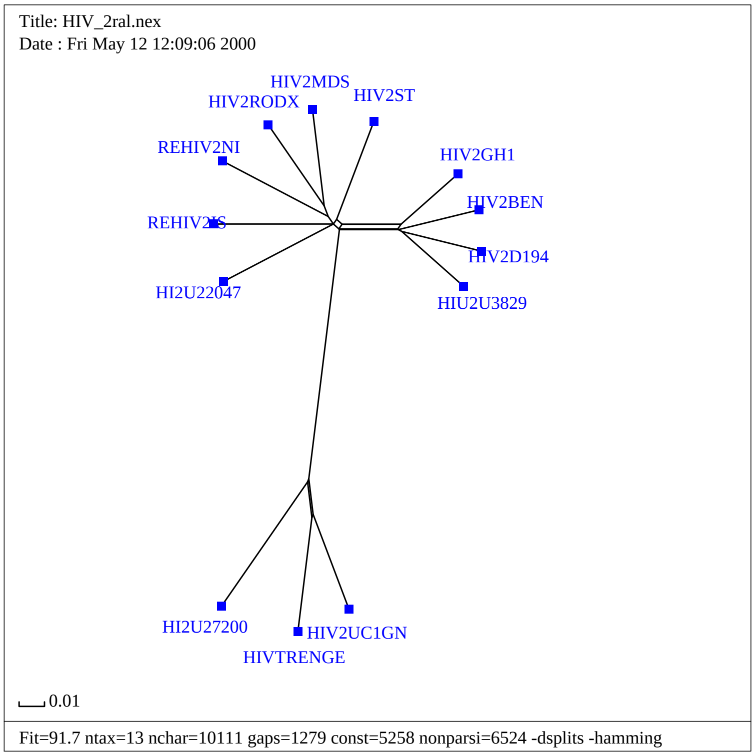


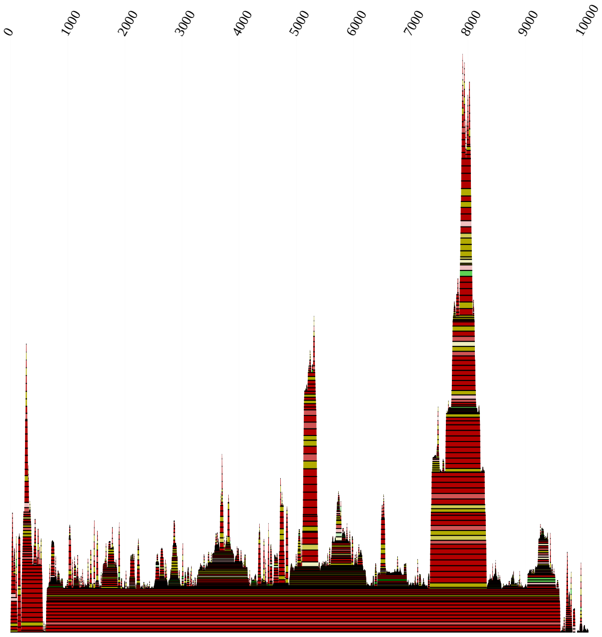
Figure 20: `Splitstree` plot of the aligned HIV-2 sequences.

HIV-2 causes AIDS, but generally after a much longer period of clinical latency and lower morbidity than that which follows infection with HIV-1. At the molecular level, HIV-2 is much more closely related to the simian immunodeficiency viruses (SIV) than to HIV-1 [19]. While the TAR in HIV-1 is a single stem-loop, the HIV-2/SIV TAR is more complex, consisting of as many as three stem-loops. Nonessential genes for viral replication in HIV-2 and SIV are *vif*, *vpx*, and/or *vpr*, and *nef*.

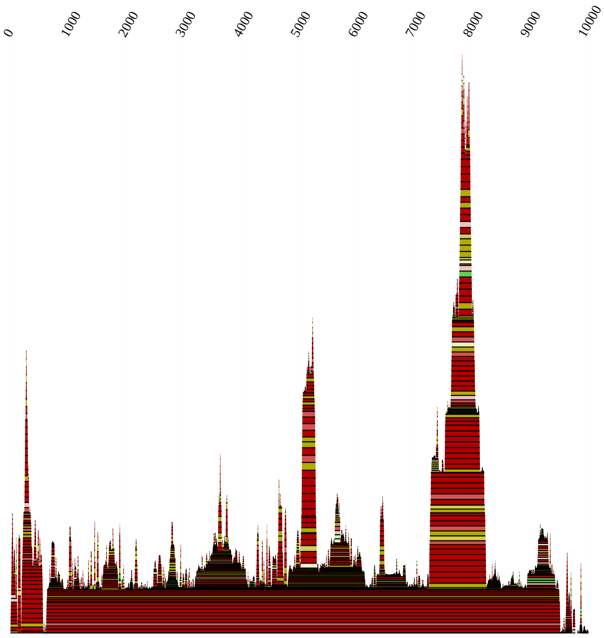
Figure 20 shows the split decomposition for all the HIV-2 sequences that we used for this experiment. We can see that the different strains are closely related. The sequences were taken from the `GenBank`[6], details can be found in Table 9. In Figure 21 we see the alignments of the different HIV-2 strains calculated with two different alignment algorithms (`Ralign and ClustalW`). Although we used two different algorithms the resulting secondary structure elements are very similar.

[6] `http://www2.ncbi.nlm.nih.gov/genbank/`

The positions of the individual elements are consistent because of the similar alignment lengths of the respective algorithms. The major secondary structures are preserved. Due to the similarities between the `Ralign` and the `ClustalW` algorithm, the results are nearly identical. A more detailed analysis of these results can be found in Figures 22 to 27. These figures show the secondary structure elements with the highest base pairing probabilities. This is further underlined by the occurring compensatory and consistent mutations (highlighted by the color code) in these regions. The secondary structure elements which were found only with the `ClustalW` algorithm can be seen in Figure 28. Some of these structural elements have been described in literature [19, 14, 15] like TAR , DIS, and RRE, but we again found a large number of previously unknown features. While the TAR in HIV-1 is a single stem-loop, the HIV-2 TAR is more complex, consisting of as many as three stem-loops. In Figure 27.4 and 27.5 we found hairpin structures which are idenical with two hairpin in the TAR element (Figure 22.1). We also found a further secondary structure element 27.6 at the far 3' end which occurs only in 5 of 13 sequences, leading to the conclusion that some of the sequences in the `GenBank` might have been subject to sequencing errors at the 3' end.

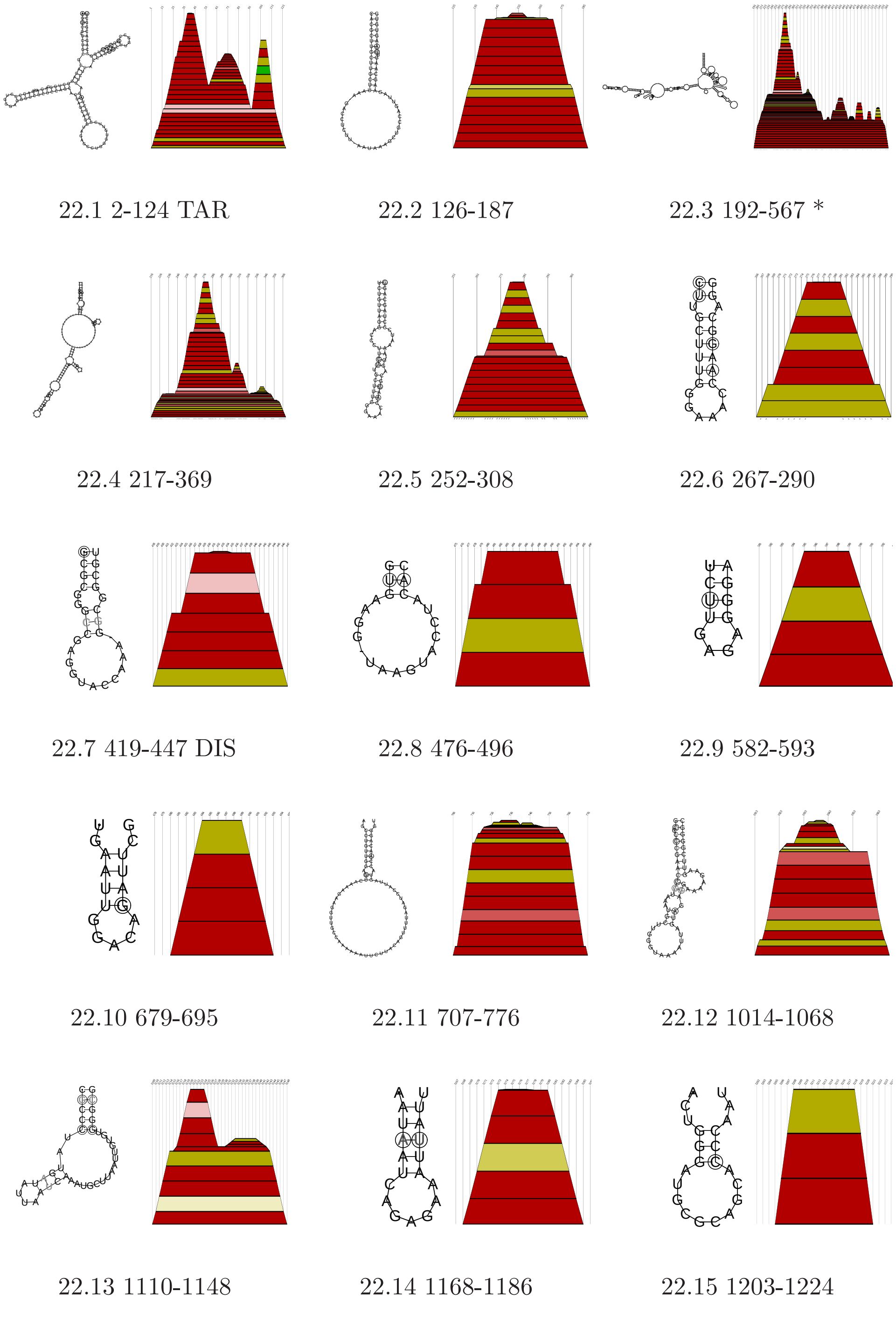


21.1 Mountain plot `Ralign`



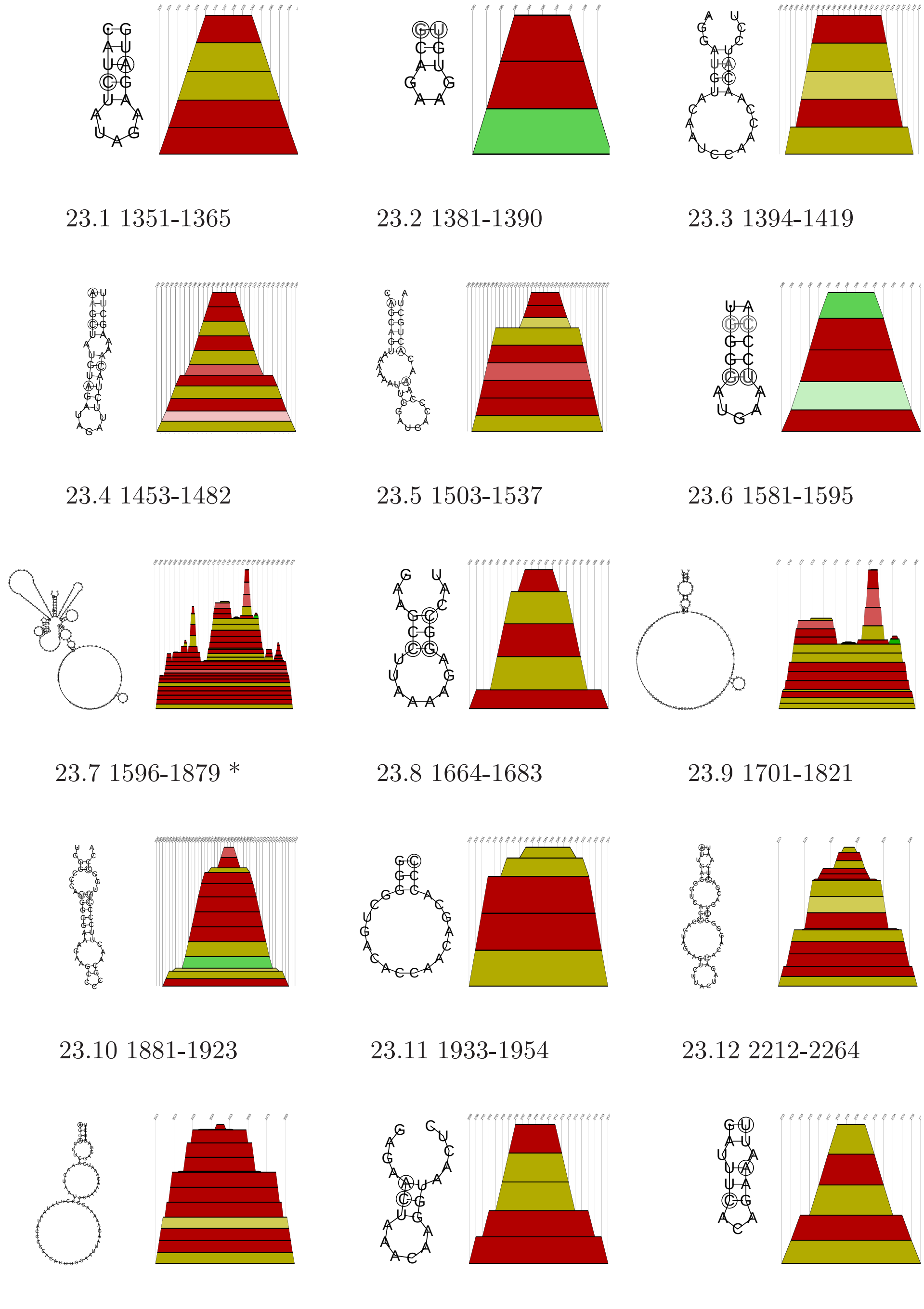
21.2 Mountain plot `ClustalW`

Figure 21: Mountain plots of HIV-2:
13 sequences aligned by `Ralign`. Alignment length is 10111 bases, conserved 5258 and the mean pairwise homology is 82.0%.
13 sequences aligned by `ClustalW`. Alignment length is 10063 bases, conserved 5285 and the mean pairwise homology is 82.2%.



22.1 2-124 TAR
22.2 126-187
22.3 192-567 *
22.4 217-369
22.5 252-308
22.6 267-290
22.7 419-447 DIS
22.8 476-496
22.9 582-593
22.10 679-695
22.11 707-776
22.12 1014-1068
22.13 1110-1148
22.14 1168-1186
22.15 1203-1224

Figure 22: Detected conserved secondary structures of HIV-2 sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. The structure labeled by (*) has long range interactions, see also Table 2. (partI)



23.1 1351-1365 | 23.2 1381-1390 | 23.3 1394-1419

23.4 1453-1482 | 23.5 1503-1537 | 23.6 1581-1595

23.7 1596-1879 * | 23.8 1664-1683 | 23.9 1701-1821

23.10 1881-1923 | 23.11 1933-1954 | 23.12 2212-2264

23.13 2614-2688 | 23.14 2700-2720 | 23.15 2723-2737

Figure 23: Detected conserved secondary structures of HIV-2 sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. The structure labeled by (*) has long range interactions, see also Table 2. (partII)

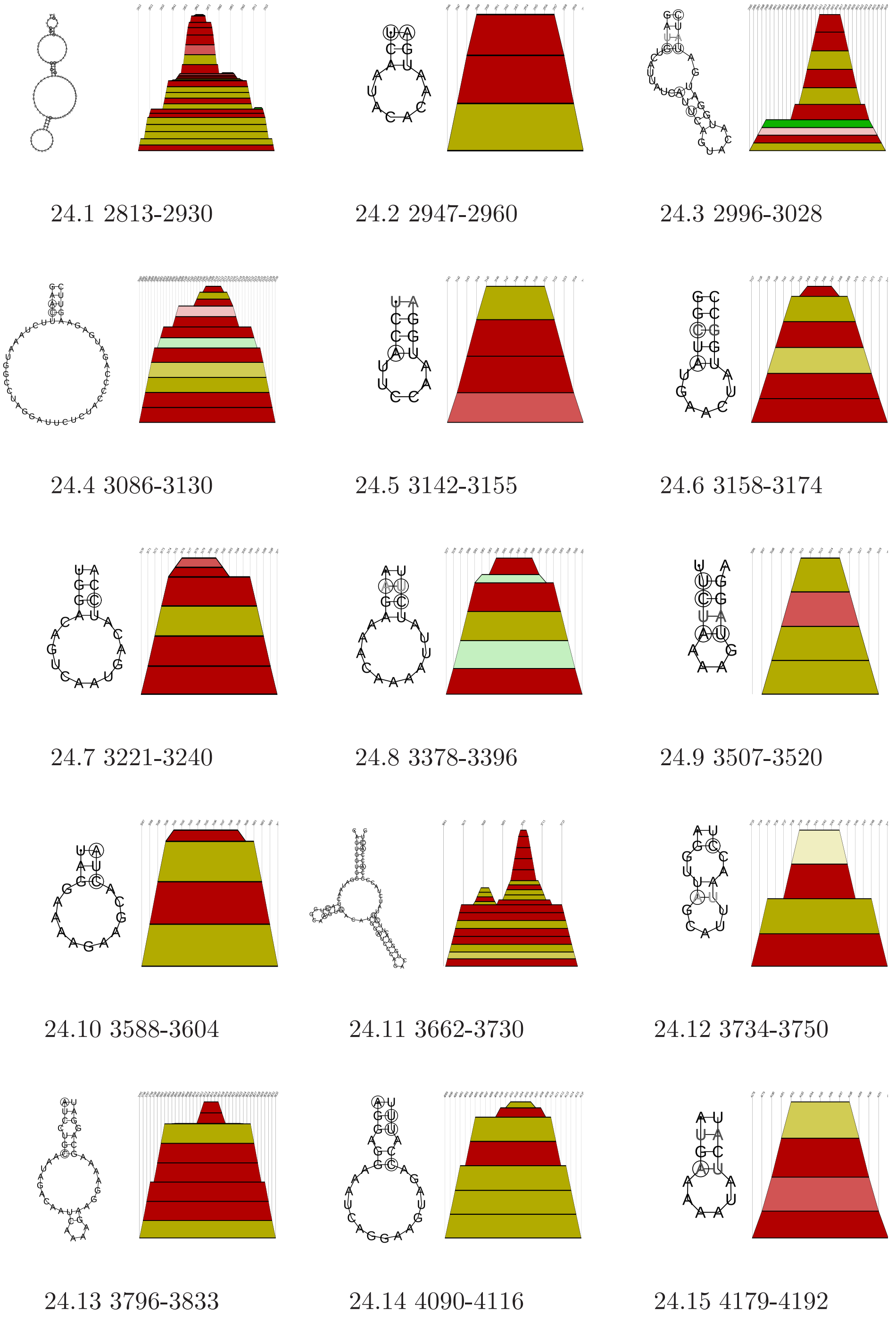
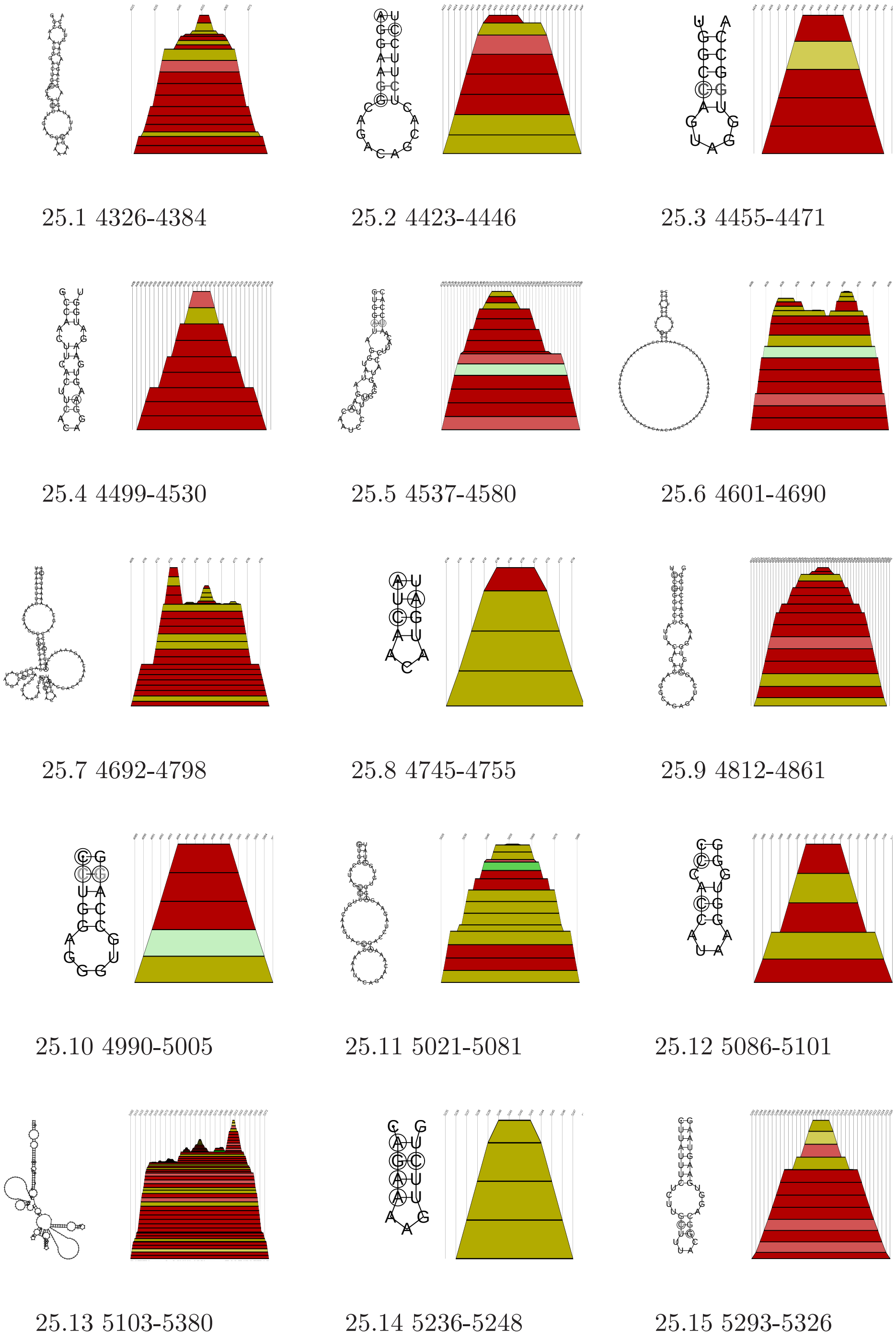


Figure 24: Detected conserved secondary structures of HIV-2 sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. (partIII)



25.1 4326-4384 | 25.2 4423-4446 | 25.3 4455-4471
25.4 4499-4530 | 25.5 4537-4580 | 25.6 4601-4690
25.7 4692-4798 | 25.8 4745-4755 | 25.9 4812-4861
25.10 4990-5005 | 25.11 5021-5081 | 25.12 5086-5101
25.13 5103-5380 | 25.14 5236-5248 | 25.15 5293-5326

Figure 25: Detected conserved secondary structures of HIV-2 sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. (partIV)

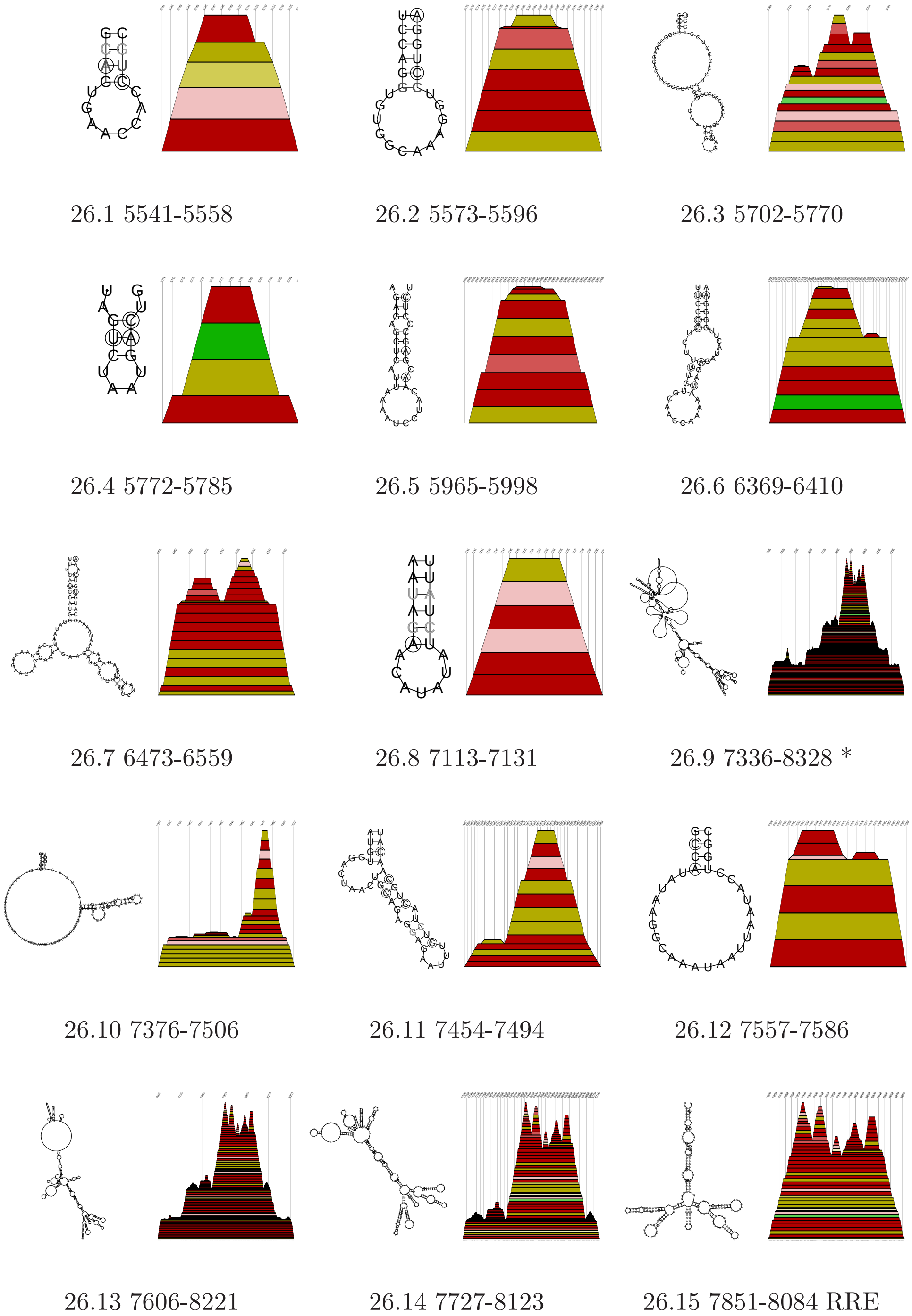
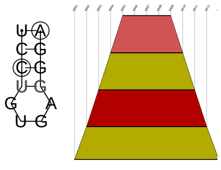
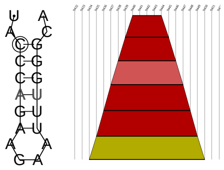


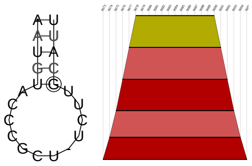
Figure 26: Detected conserved secondary structures of HIV-2 sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. The structure labeled by (*) has long range interactions, see also Table 2. (partV)



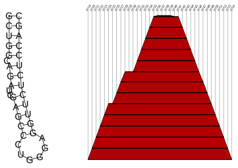
27.1 8902-8913



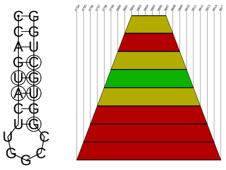
27.2 9433-9452



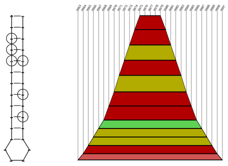
27.3 9674-9695



27.4 9720-9754

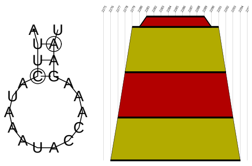


27.5 9795-9815

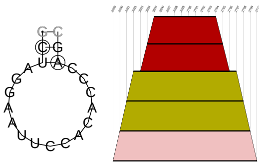


27.6 9964-9991

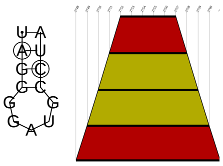
Figure 27: Detected conserved secondary structures of HIV-2 sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. (partVI)



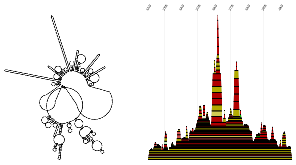
28.1 2276-2295



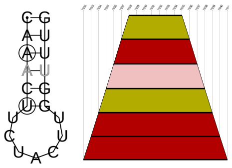
28.2 2690-2710



28.3 2749-2761



28.4 3230-4103 *



28.5 7023-7041

Figure 28: Detected conserved secondary structures of HIV-2 sequences after aligning the sequences with `ClustalW`, but not found in the `Ralign` alignment. The numbers denote the base pair range of the aligned sequences in which the structural element is found. The structure labeled by (*) has long range interactions, see also Table 2.

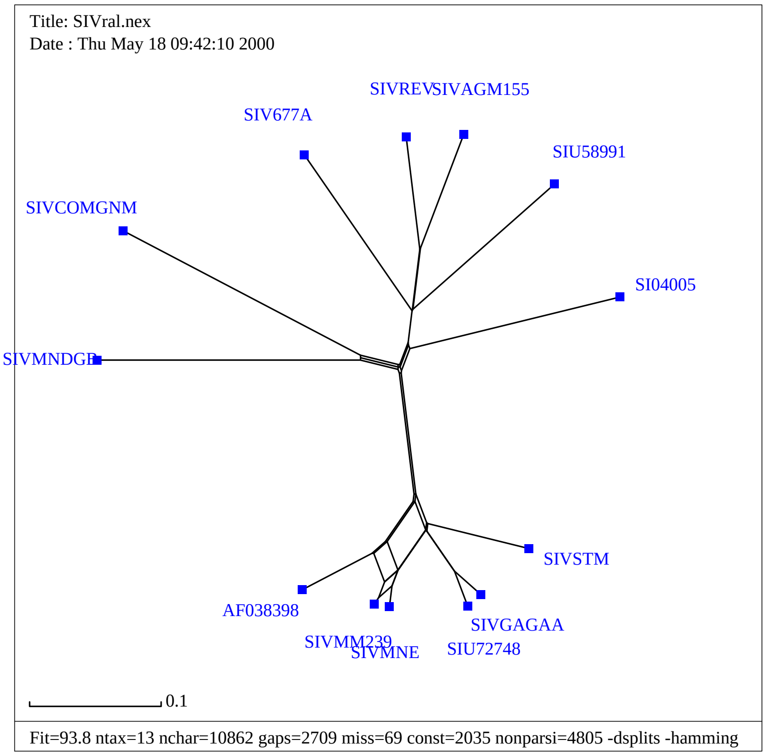


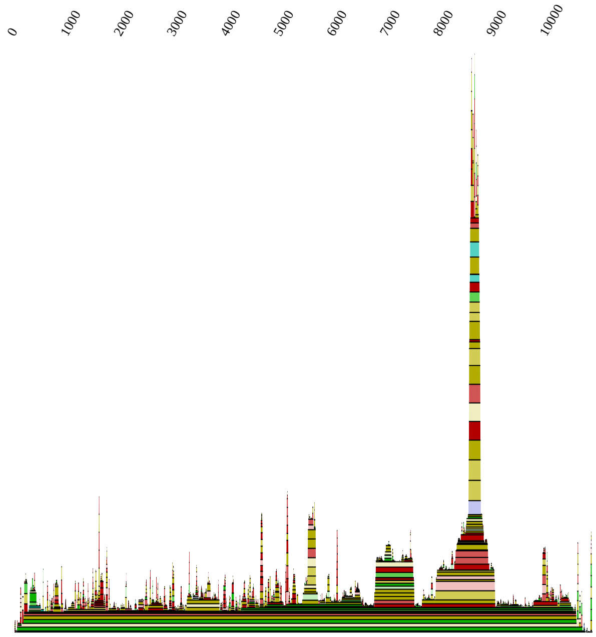
Figure 29: `Splitstree` plot of the aligned sequences of SIV.

5.2.3 Simian Immunodeficiency Virus

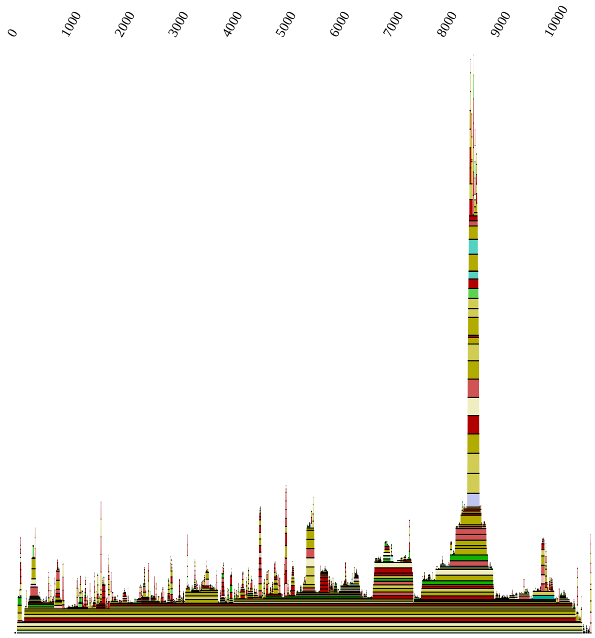
Simian immunodeficiency viruses (SIV), are very close relatives to HIV which have virtually the same genetic composition and regulatory capabilities. At the molecular level, HIV-2 is much more closely related to the simian immunodeficiency viruses (SIV) than to HIV-1 [19]. While the TAR in HIV-1 is a single stem-loop, the SIV TAR similar to the one in HIV-2 is more complex, consisting of three stem-loops. Nonessential genes for viral replication in HIV-2 and SIV are *vif*, *vpx*, and/or *vpr*, and *nef*.

Figure 29 shows the split decomposition for all the SIV sequences that we used for this experiment. The sequences were taken from the `GenBank`, details can be found in Table 10. We can see that the strains are split in two different subgroups. Due to this we performed separate split decompositions for the two proposed groups. The results of this calculations can be found in Figure 32 and 36. As expected the strains within the subgroups are closely related.

In Figure 30 we see the alignments of the different SIV strains calculated with two different alignment algorithms (`Ralign` and `ClustalW`). Although we used two different algorithms the resulting secondary structure elements are very similar. The positions of the individual elements are very consistent because of the similar alignment lengths of the respective algorithms. The major secondary structures are preserved. But we found a much smaller number, because only some elements are present in both subgroups. Due to the similarities between the `Ralign` and the `ClustalW` algorithm, the results are nearly identical.



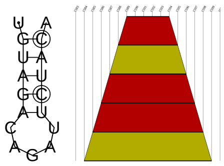
30.1 Mountain plot `Ralign`



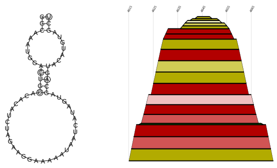
30.2 Mountain plot `ClustalW`

Figure 30: Mountain plots of SIV:
13 sequences aligned by `Ralign`. Alignment length is 10862 bases, conserved 2035 and the mean pairwise homology is 62.4%.
13 sequences aligned by `ClustalW`. Alignment length is 10761 bases, conserved 2094 and the mean pairwise homology is 62.5%.

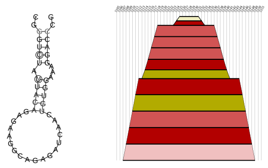
A more detailed analysis of these results can be found in Figure 31. This figure shows the secondary structure elements with the highest base pairing probabilities. This is further underlined by the occurring compensatory and consistent mutations (highlighted by the color code) in these regions.



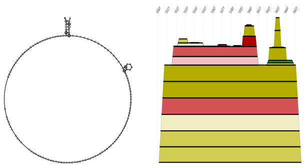
31.1 1584-1600



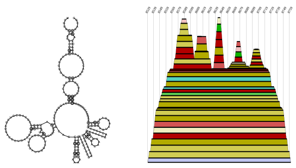
31.2 4616-4674



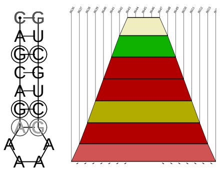
31.3 5106-5151



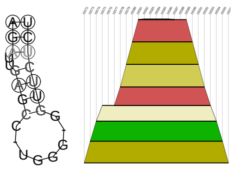
31.4 5508-5665



31.5 8530-8762 RRE



31.6 8637-8654



31.7 **10573-10596**

Figure 31: Detected conserved secondary structures of SIV sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. The structure labeled in bold was not predicted in the `ClustalW` alignment.

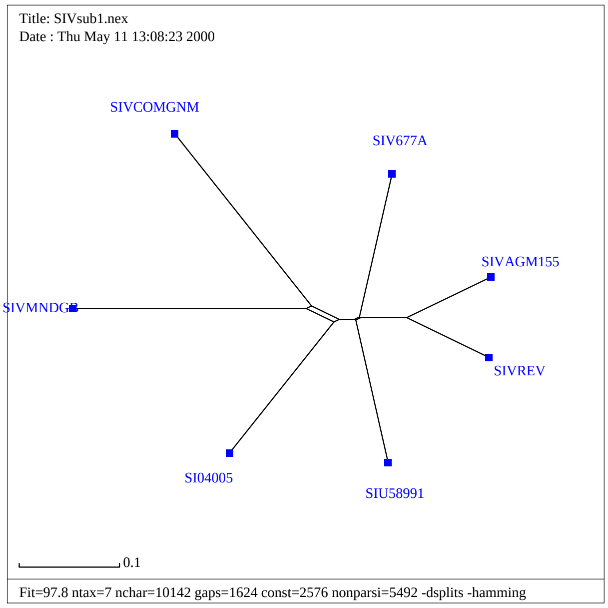
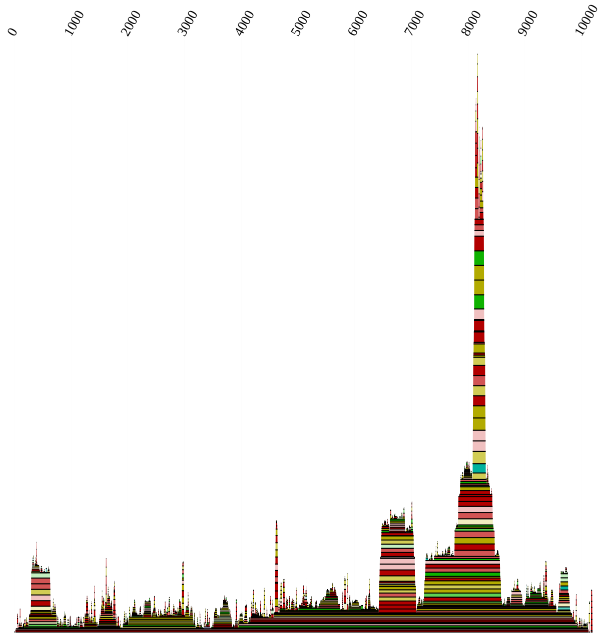


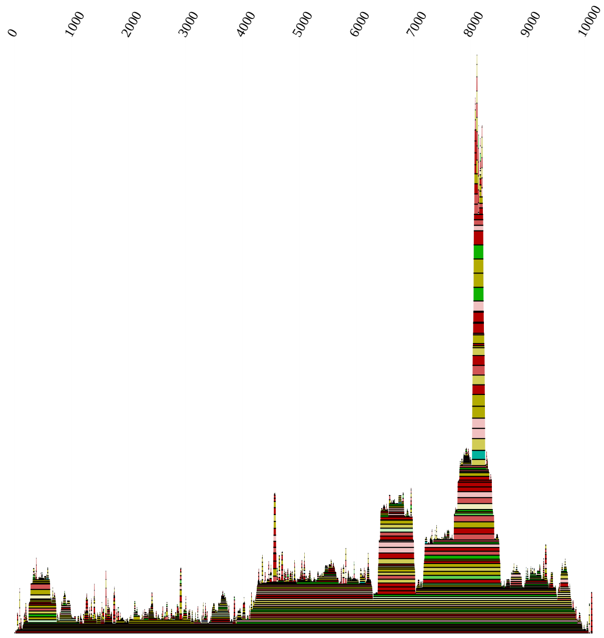
Figure 32: `Splitstree` plot of the aligned sequences of SIV$_{sub1}$.

In Figure 33 we see the alignments of the different SIV$_{sub1}$ strains calculated with two different alignment algorithms (`Ralign and ClustalW`). We found slight differences between the two mountain plots. Although the positions of the individual elements are very consistent, because of the similar alignment length of the respective algorithms, we detected long range interactions in the `ClustalW` alignment. The major secondary structures are preserved. A more detailed analysis of these results can be found in Figure 34. This figure shows the secondary structure elements with the highest base pairing probabilities. This is further underlined by the occurrence of compensatory and consistent mutations (highlighted by the color code) in these regions. We again found a number of previously unknown features.

In Figure 37 we see the alignments of the different SIV$_{sub2}$ strains calculated with two different alignment algorithms (`Ralign and ClustalW`). Although we used two different algorithms the resulting secondary structure elements are very similar. The positions of the individual elements are very consistent because of the similar alignment lengths of the respective algorithms. The major secondary structures are preserved. In the `Ralign` mountain plot we detected less long range interactions. A more detailed analysis of these results can be found in Figures 38 to 43. These figures show the secondary structure elements with the highest base pairing probabilities. This is further underlined by the occurring compensatory and consistent mutations (highlighted by the color code) in these regions. Some of these structural elements, TAR and RRE, have been described in literature [15], but we again found a large number of previously unknown features. We found a second TAR element at position 9924 to 10047 (Figure 43.7) which is at the 3' end of the sequence.

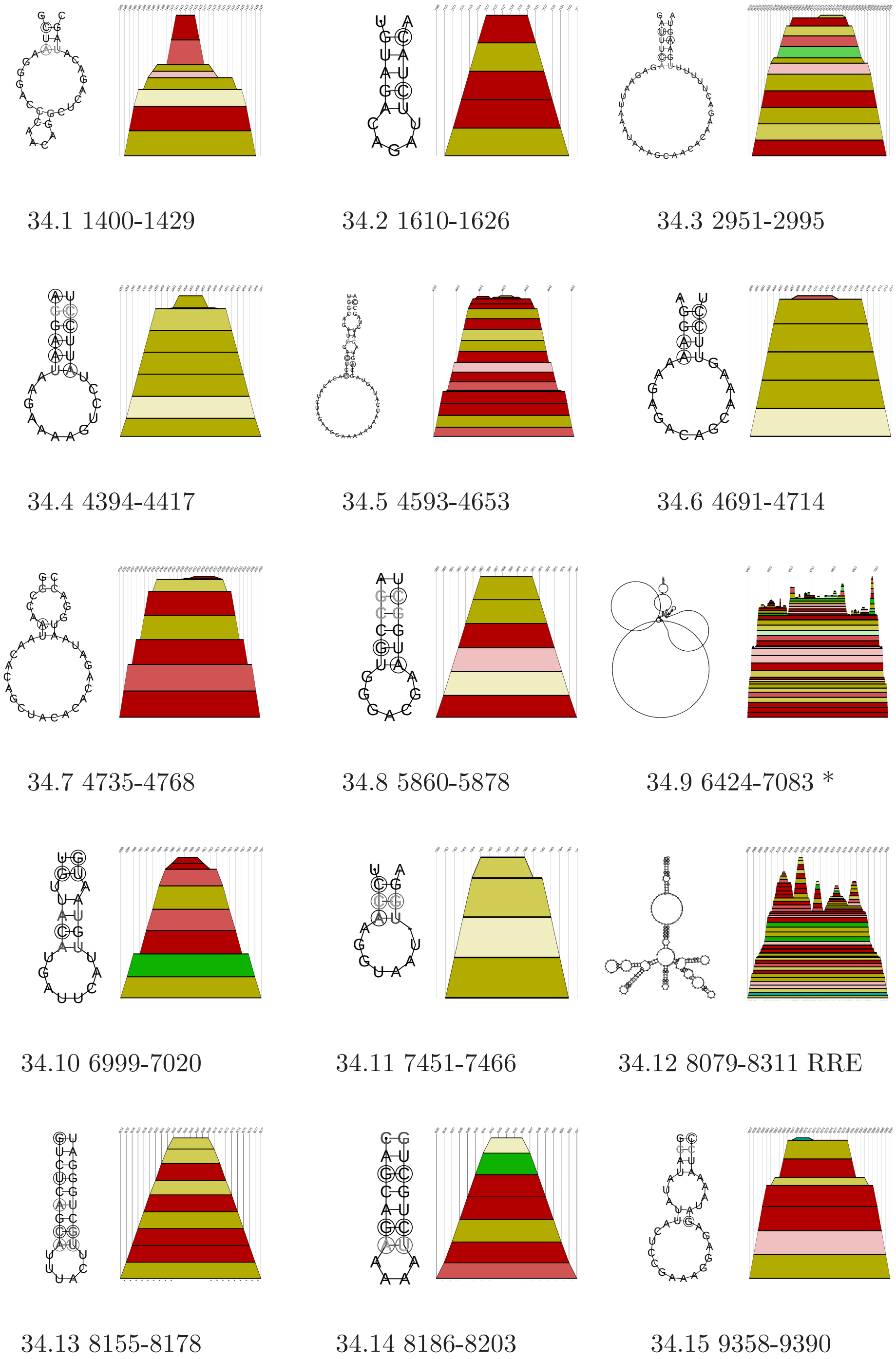


33.1 Mountain plot `Ralign`



33.2 Mountain plot `ClustalW`

Figure 33: Mountain plots of SIV$_{sub1}$:
7 sequences aligned by `Ralign`. Alignment length is 10195 bases, conserved 2572 and the mean pairwise homology is 61.1%.
7 sequences aligned by `ClustalW`. Alignment length is 10142 bases, conserved 2576 and the mean pairwise homology is 68.4%.



34.1 1400-1429

34.2 1610-1626

34.3 2951-2995

34.4 4394-4417

34.5 4593-4653

34.6 4691-4714

34.7 4735-4768

34.8 5860-5878

34.9 6424-7083 *

34.10 6999-7020

34.11 7451-7466

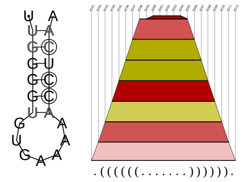
34.12 8079-8311 RRE

34.13 8155-8178

34.14 8186-8203

34.15 9358-9390

Figure 34: Detected conserved secondary structures of SIV$_{sub1}$ sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. The structure labeled by (*) has long range interactions, see also Table 2.



35.1 8192-8212

Figure 35: Additional conserved secondary structure of SIV$_{sub1}$ sequences after aligning the sequences with `ClustalW`. The numbers denote the base pair range in the alignment.

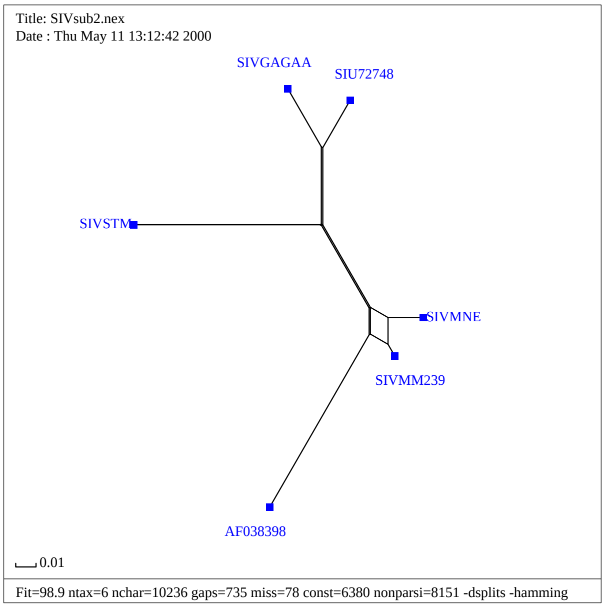
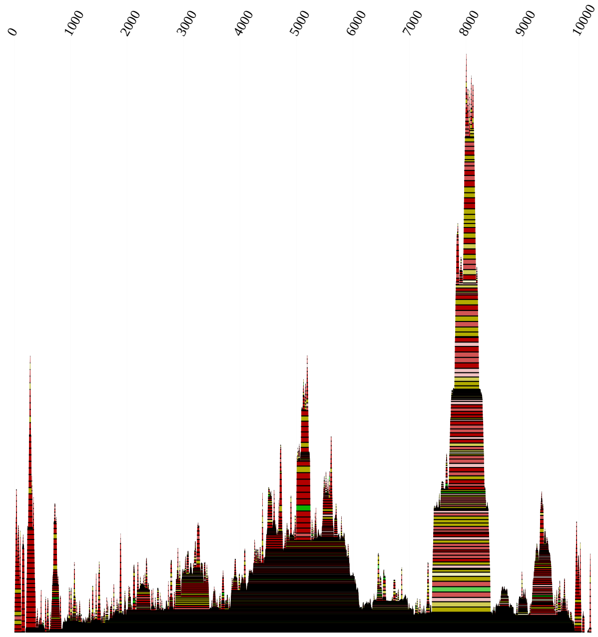
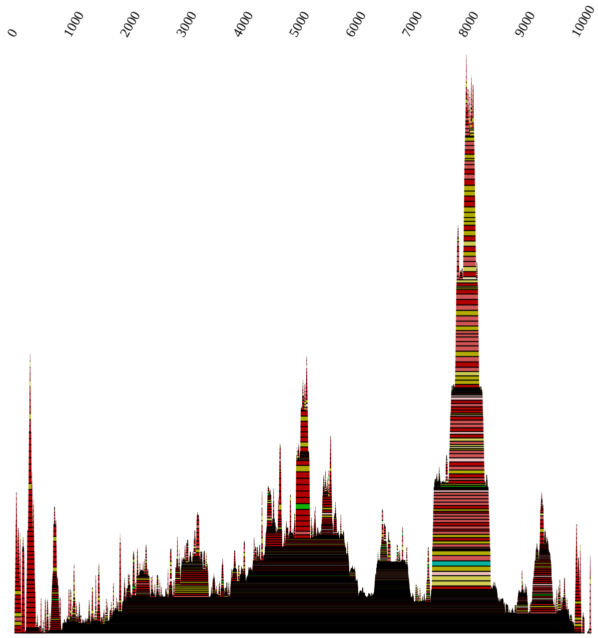


Figure 36: `Splitstree` plot of the aligned sequences of SIV$_{sub2}$.

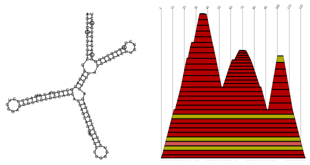


37.1 Mountain plot `Ralign`

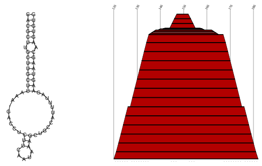


37.2 Mountain plot `ClustalW`

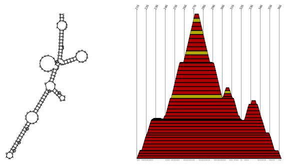
Figure 37: Mountain plots of SIV_{sub2}:
6 sequences aligned by `Ralign`. Alignment length is 10226 bases, conserved 6380 and the mean pairwise homology is 82.9%.
6 sequences aligned by `ClustalW`. Alignment length is 10236 bases, conserved 6380 and the mean pairwise homology is 82.9%.



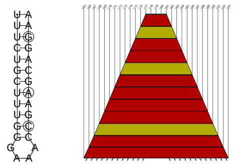
38.1 2-125 TAR



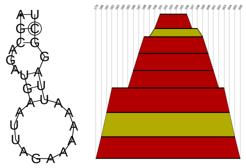
38.2 127-188



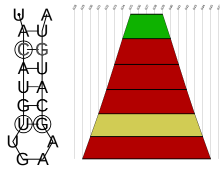
38.3 217-368



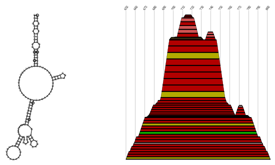
38.4 266-293



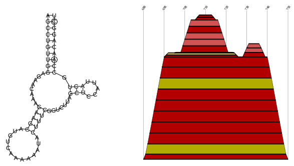
38.5 580-606



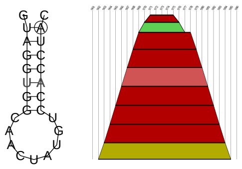
38.6 629-646



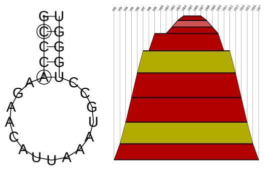
38.7 653-805



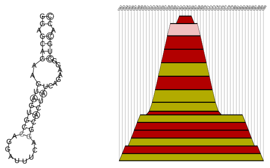
38.8 689-758



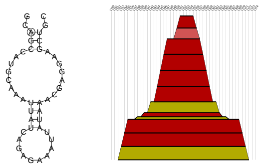
38.9 962-986



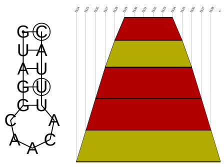
38.10 993-1017



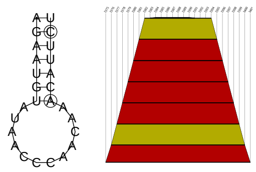
38.11 1043-1090



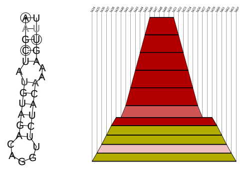
38.12 1136-1179



38.13 1325-1339

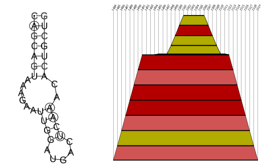


38.14 1376-1401

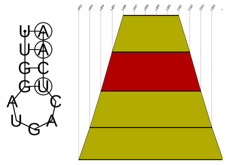


38.15 1435-1464

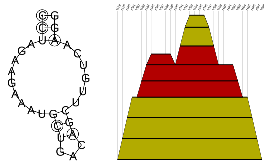
Figure 38: Detected conserved secondary structures of SIV$_{sub2}$ sequences after aligning the sequences with `Ralign`. The numbers denote the base pair range in the `Ralign` alignment. (partI)



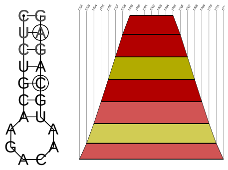
39.1 1485-1519



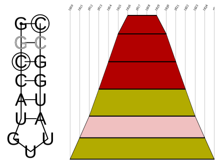
39.2 1498-1510



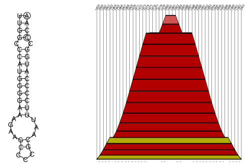
39.3 1579-1608



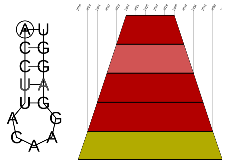
39.4 1753-1772



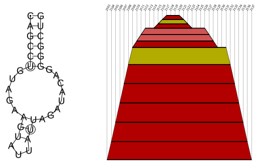
39.5 1811-1825



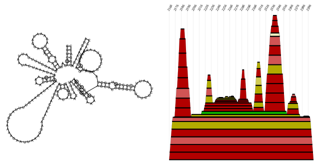
39.6 1860-1903



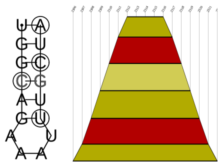
39.7 2020-2034



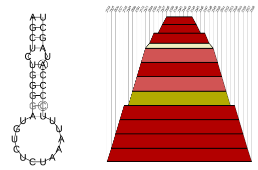
39.8 2104-2137



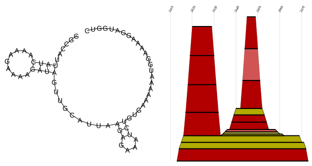
39.9 2167-2405 *



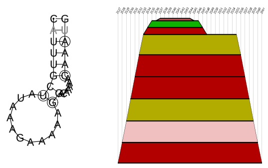
39.10 2307-2322



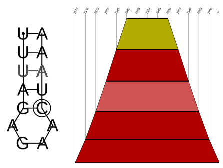
39.11 2325-2358



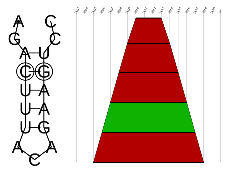
39.12 2420-2485



39.13 2528-2561

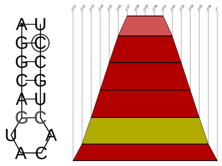


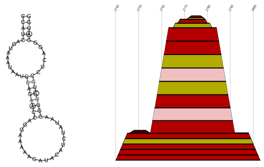
39.14 2578-2591

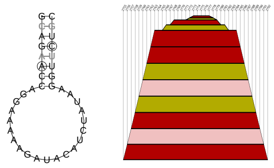


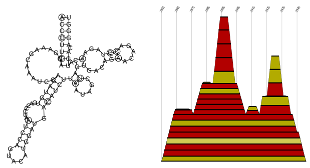
39.15 2604-2620

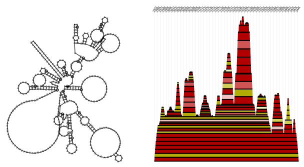
Figure 39: Detected conserved secondary structures of SIV_{sub2} sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. The structure labeled by (*) has long range interactions, see also Table 2. (partII)

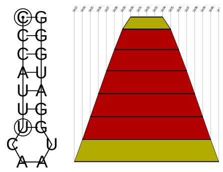

40.1 2722-2737

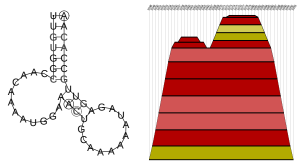

40.2 2746-2808

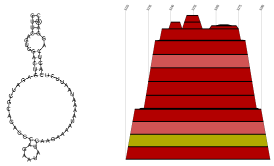

40.3 2761-2797

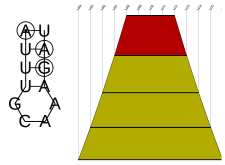

40.4 2856-2951

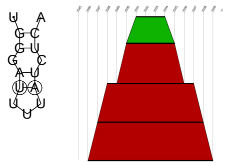

40.5 2952-3446 *

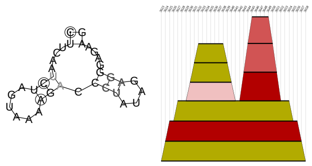

40.6 3024-3041

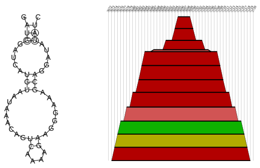

40.7 3049-3093

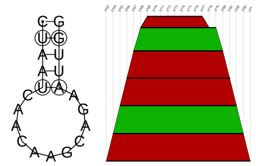

40.8 3226-3289

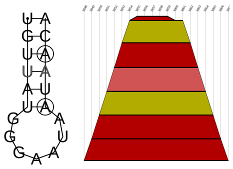

40.9 3405-3416


40.10 3586-3600


40.11 3624-3658


40.12 3673-3720


40.13 3764-3784


40.14 3849-3867

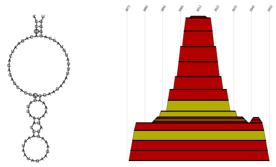
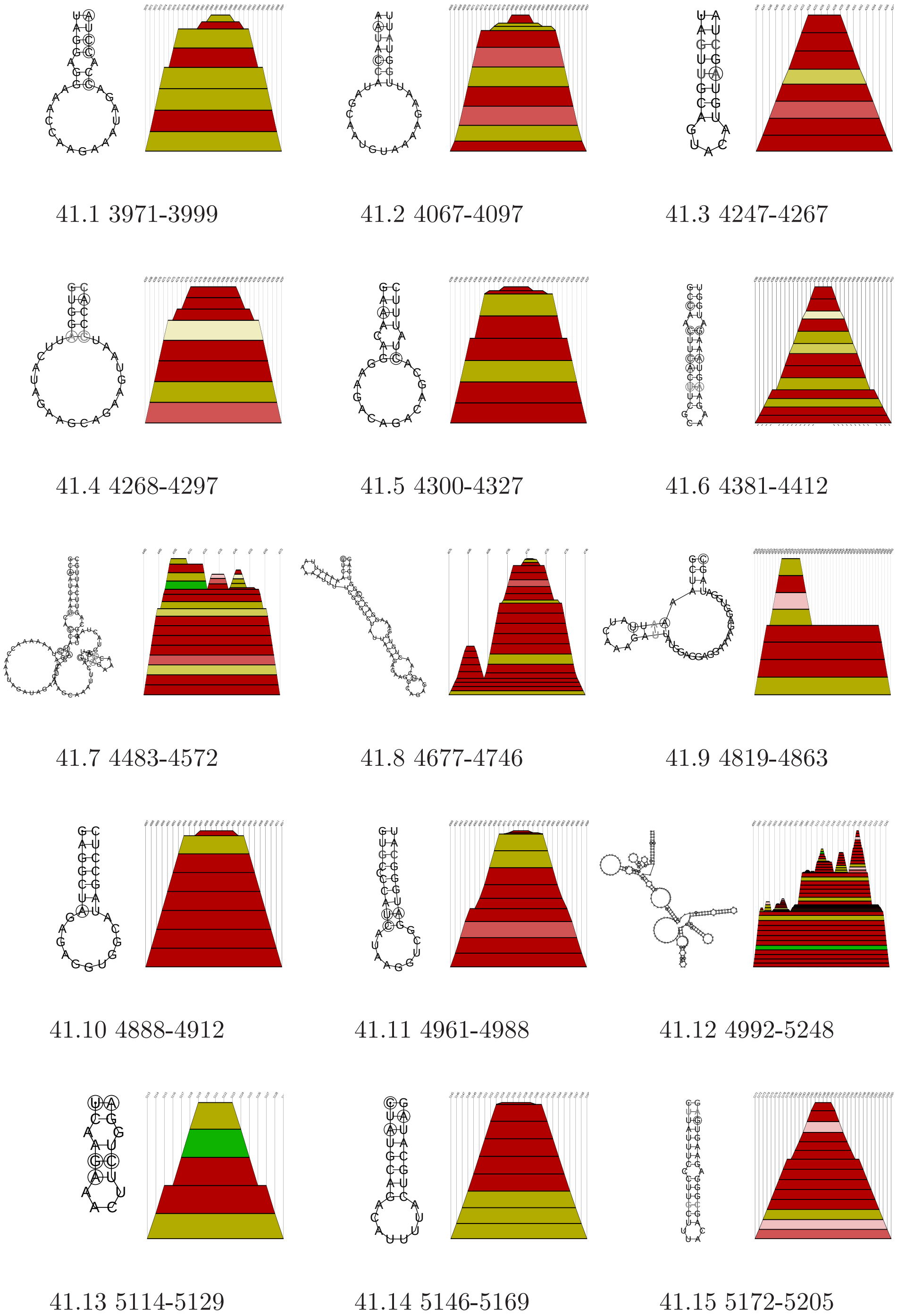
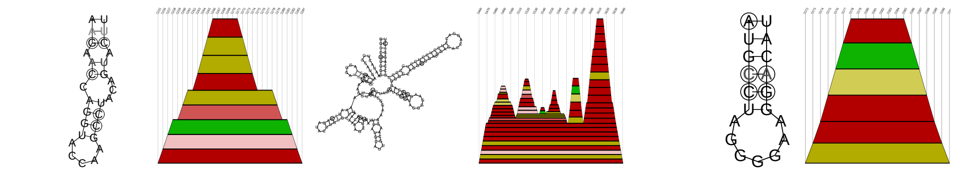

40.15 3873-3953

Figure 40: Detected conserved secondary structures of SIV_{sub2} sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. The structure labeled by (*) has long range interactions, see also Table 2. (partIII)

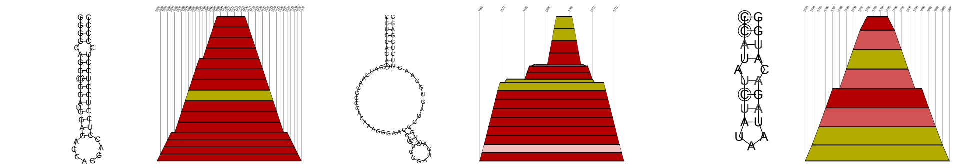


41.1 3971-3999
41.2 4067-4097
41.3 4247-4267
41.4 4268-4297
41.5 4300-4327
41.6 4381-4412
41.7 4483-4572
41.8 4677-4746
41.9 4819-4863
41.10 4888-4912
41.11 4961-4988
41.12 4992-5248
41.13 5114-5129
41.14 5146-5169
41.15 5172-5205

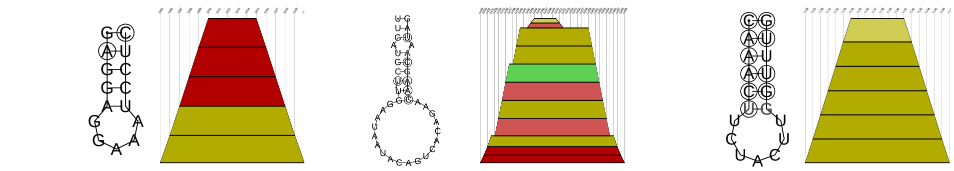
Figure 41: Detected conserved secondary structures of SIV$_{sub2}$ sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. (partIV)



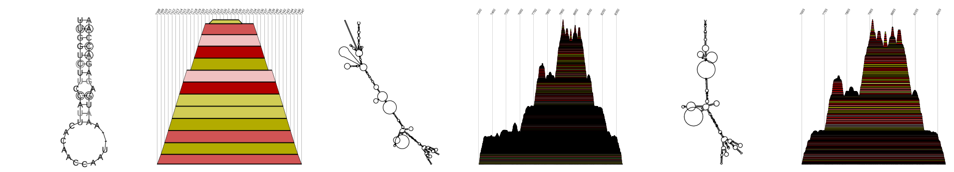
42.1 5261-5289 42.2 5461-5641 42.3 5573-5591



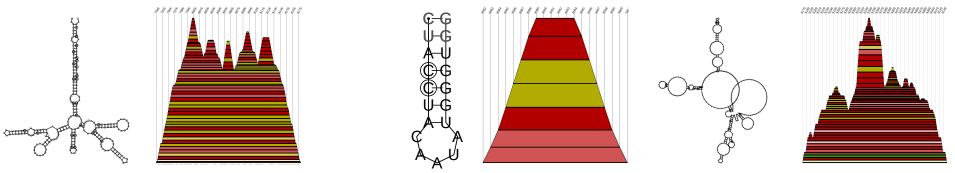
42.4 5592-5632 42.5 5662-5725 42.6 5784-5804



42.7 5906-5920 42.8 **6430-6469** 42.9 7129-7147

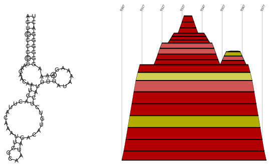


42.10 **7309-7347** 42.11 7393-8439 * 42.12 7694-8328

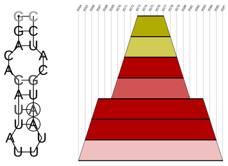


42.13 7947-8179 RRE 42.14 8983-9001 42.15 9171-9539 *

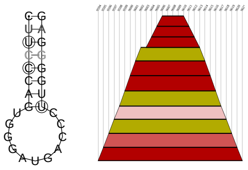
Figure 42: Detected conserved secondary structures of SIV$_{sub2}$ sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. Structures labeled in bold were not predicted by `ClustalW`. Structures labeled by (*) have long range interactions, see also Table 2. (partV)



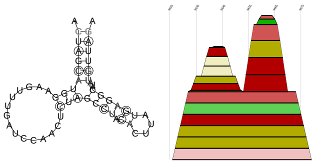
43.1 9308-9379



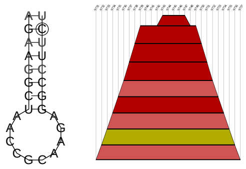
43.2 9565-9586



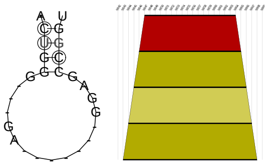
43.3 9595-9621



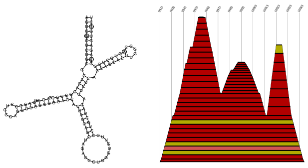
43.4 9626-9680



43.5 9732-9757



43.6 **9843-9869**



43.7 9924-10047 TAR

Figure 43: Detected conserved secondary structures of SIV$_{sub2}$ sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. Structures labeled in bold were not predicted by `ClustalW`. (partVI)

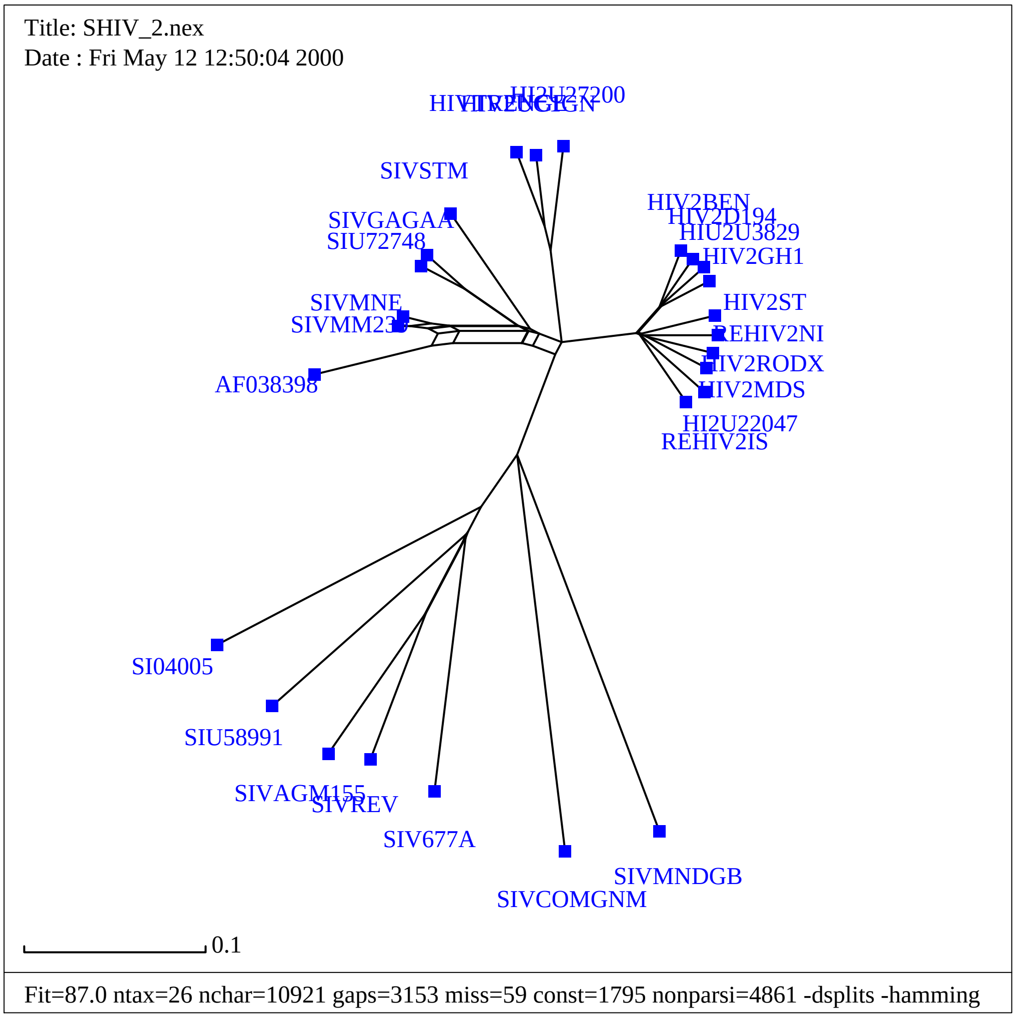
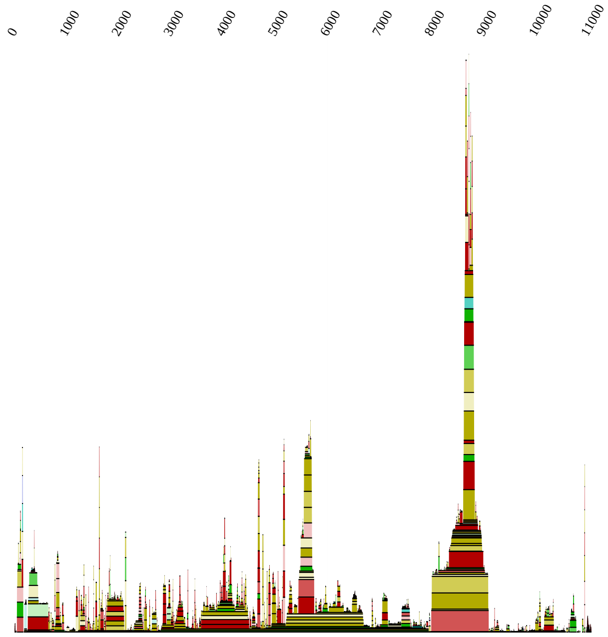
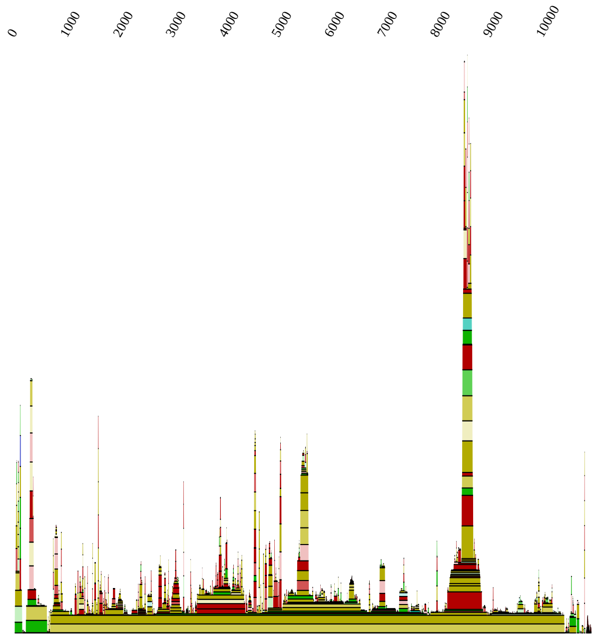


Figure 44: `Splitstree` plot of the aligned sequences of HIV-2 and SIV.

Figure 44 shows the split decomposition for all the HIV-2 and SIV sequences that we used for this experiment. The sequences were taken from the `GenBank`, details can be found in Table 9 and 10. We clearly see that the SIV strains are split in two different subgroups and that the sequences in subgroup 2 are more similar to the HIV-2 strains. Due to this we performed separate split decompositions for the two proposed groups with HIV-2 respectively. The results of this calculations can be found in Figure 48 and 51. As expected the strains of subgroup 2 are more closely related to HIV-2 than those of subgroup 1.

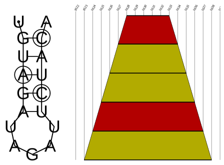


45.1 Mountain plot `Ralign`

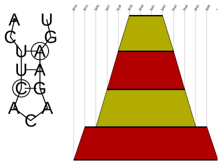


45.2 Mountain plot `ClustalW`

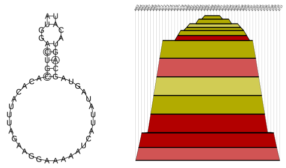
Figure 45: Mountain plots of HIV-2 and SIV:
26 sequences aligned by `Ralign`. Alignment length is 11071 bases, conserved 1773 and the mean pairwise homology is 69.0%.
26 sequences aligned by `ClustalW`. Alignment length is 10921 bases, conserved 1795 and the mean pairwise homology is 68.8%.



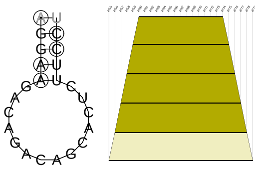
46.1 1623-1639



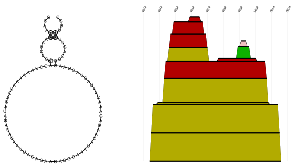
46.2 3035-3047



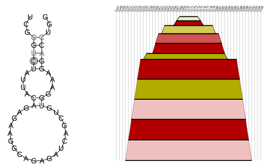
46.3 4664-4711



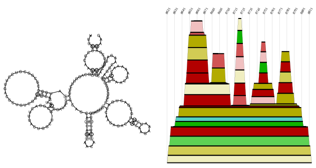
46.4 4756-4779



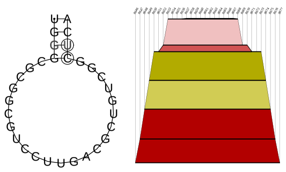
46.5 4935-5024



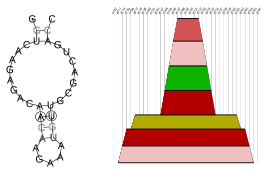
46.6 5148-5195



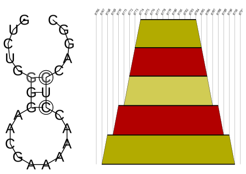
46.7 8622-8814 RRE*



46.8 8647-8677

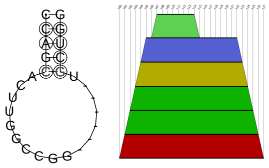


46.9 8733-8766



46.10 8767-8792

Figure 46: Detected conserved secondary structures of HIV-2 and SIV sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. The structure labeled by (*) has long range interactions, see also Table 2.



47.1 101-127

Figure 47: Additional conserved secondary structures of HIV-2 and SIV sequences detected after aligning the sequences with `ClustalW`. The numbers denote the base pair range in the alignment.

In Figure 45 we see the mountain plots of the alignment for the HIV-2 and SIV strains calculated with two different alignment algorithms (`Ralign and ClustalW`). We found slight differences between the two mountain plots. Although the positions of the individual elements are very consistent because of the similar alignment lengths of the respective algorithms, we detected long range interactions in the `ClustalW` alignment. However, the major secondary structures are preserved and the long range interactions are not very convincing because of the fact that only two base pairs are involved. A more detailed analysis of these results can be found in Figure 46. This figure shows the secondary structure elements with the highest base pairing probabilities. This is further underlined by the occurring compensatory and consistent mutations (highlighted by the color code) in these regions. Figure 47 shows a secondary structure element located near the 5' end which was only found with `ClustalW`.

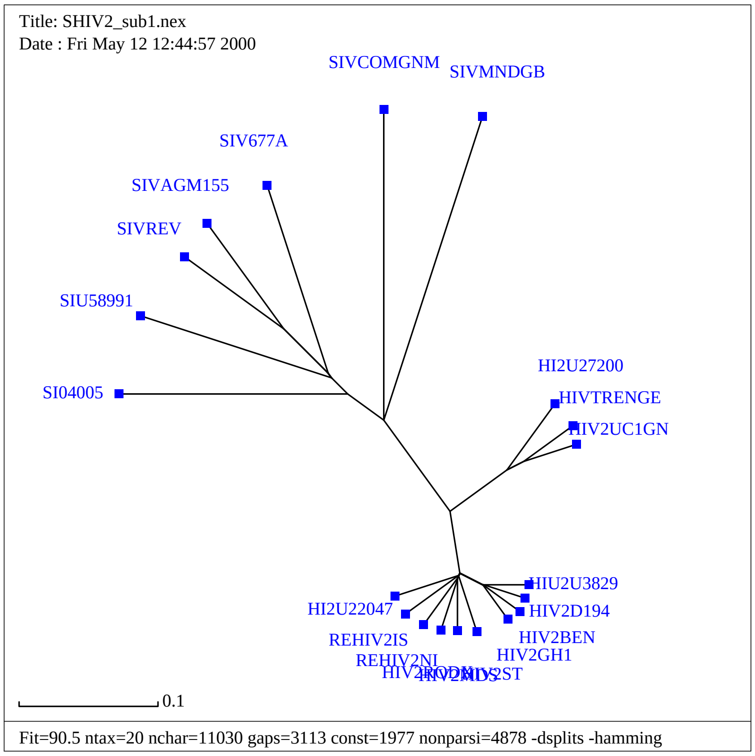
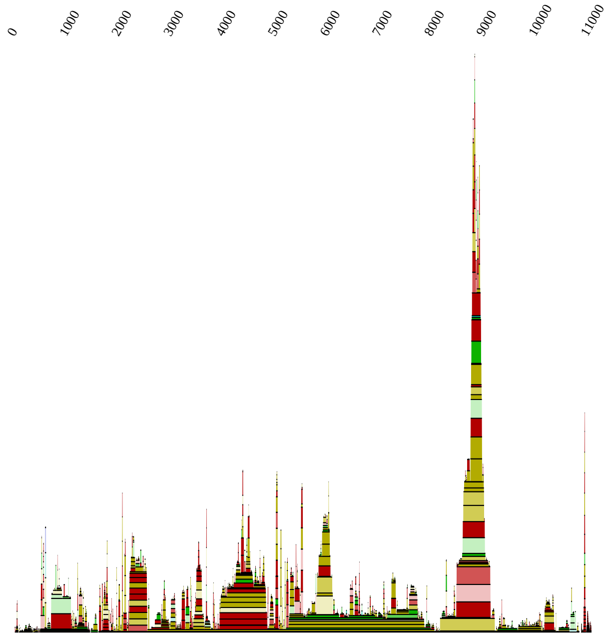
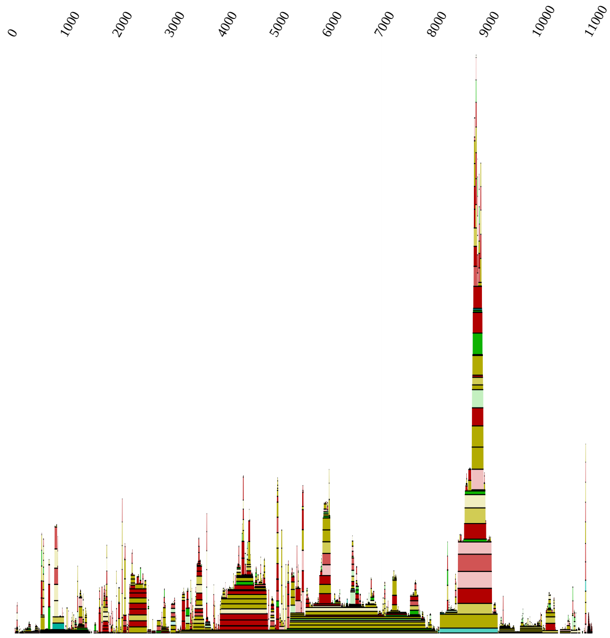


Figure 48: `Splitstree` plot of the aligned sequences of HIV-2 and SIV_{sub1}.

In Figure 49 we see the mountain plots of the alignment for the HIV-2 and SIV_{sub1} strains calculated with the two different alignment algorithms (`Ralign and ClustalW`). Although we used two different algorithms the resulting secondary structure elements are very similar. The positions of the individual elements are very consistent because of the similar alignment lengths of the respective algorithms. The major secondary structures are preserved. A more detailed analysis of these results can be found in Figure 50. This figure shows the secondary structure elements with the highest base pairing probabilities. This is further underlined by the occurring compensatory and consistent mutations (highlighted by the color code) in these regions.

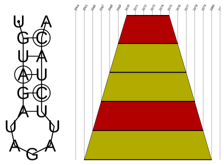


49.1 Mountain plot `Ralign`

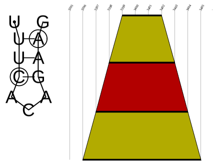


49.2 Mountain plot `ClustalW`

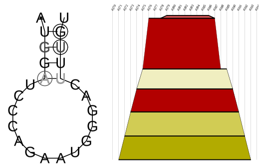
Figure 49: Mountain plots of HIV-2 and SIV$_{sub1}$:
20 sequences aligned by `Ralign`. Alignment length is 11074 bases, conserved 1905 and the mean pairwise homology is 66.0%.
20 sequences aligned by `ClustalW`. Alignment length is 11030 bases, conserved 1977 and the mean pairwise homology is 66.6%.



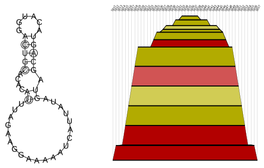
50.1 2065-2081



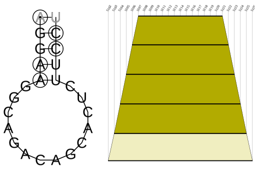
50.2 3396-3406



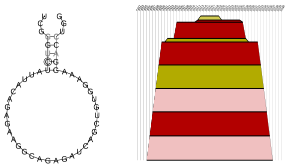
50.3 4371-4394



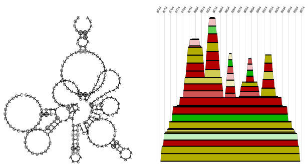
50.4 5012-5057



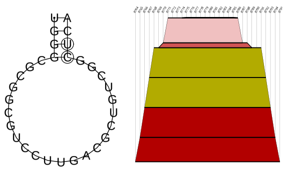
50.5 5103-5126



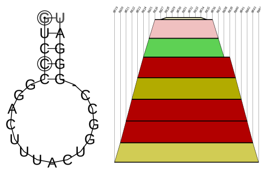
50.6 5493-5540



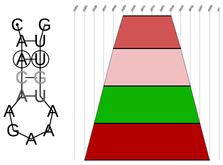
50.7 8745-8976 RRE*



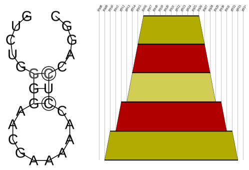
50.8 8789-8851



50.9 8820-8844



50.10 8885-8899



50.11 8909-8934

Figure 50: Detected conserved secondary structures of HIV-2 and SIV$_{sub1}$ sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. The structure labeled by (*) has long range interactions, see also Table 2.

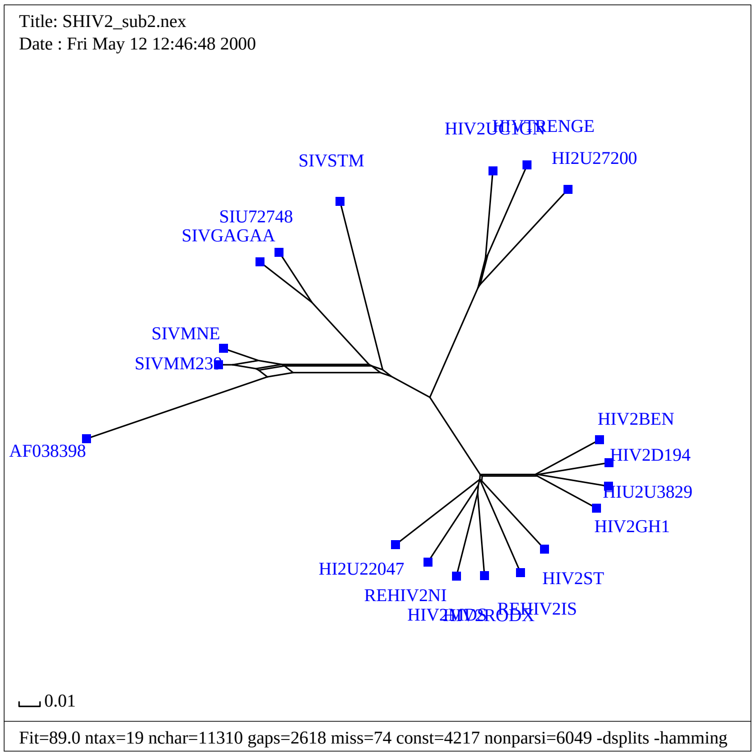
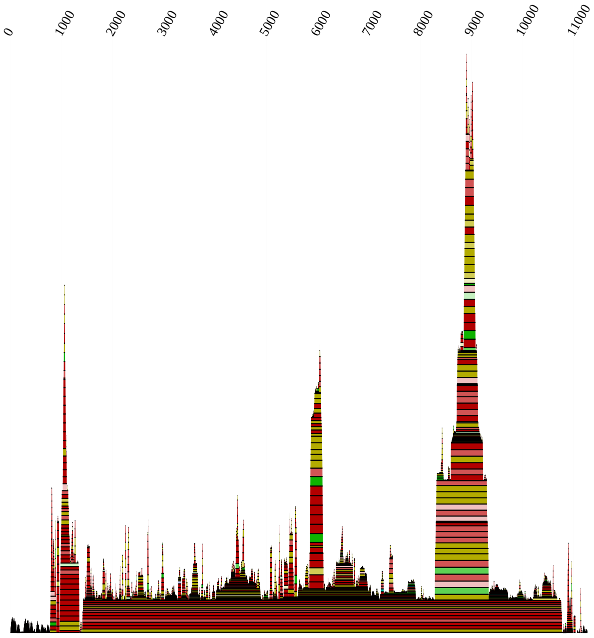


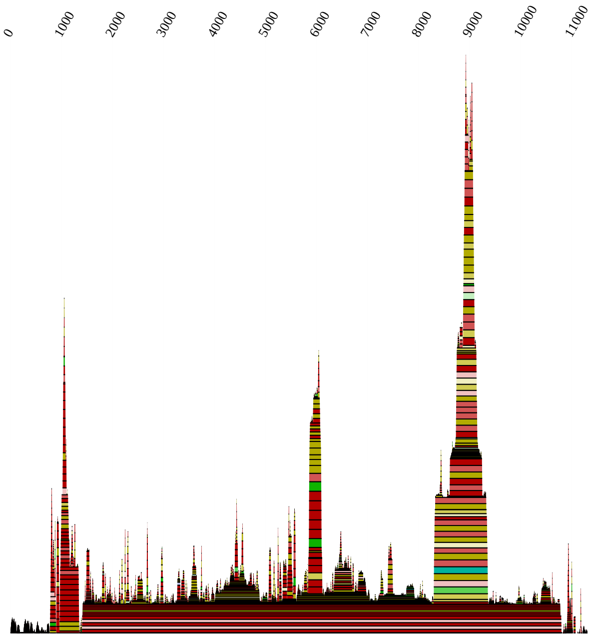
Figure 51: `Splitstree` plot of the aligned sequences of HIV-2 and SIV$_{sub2}$.

In Figure 52 we see the mountain plots of the alignment for the HIV-2 and SIV$_{sub2}$ strains calculated with the two different alignment algorithms (`Ralign and ClustalW`). Although we used two different algorithms the resulting secondary structure elements are very similar. In this case we found a number of long range interactions were the results shown in the `Ralign` mountain plot are more convincing as more base pairs with a higher probability are involved. Still, the positions of the individual elements are very consistent because of the similar alignment lengths of the respective algorithms. The major secondary structures are preserved. A more detailed analysis of these results can be found in Figures 53 to 56. These figures show the secondary structure elements with the highest base pairing probabilities. This is further underlined by the occurring compensatory and consistent mutations (highlighted by the color code) in these regions. Some of these structural elements have been described in literature like TAR (Figure 53.1) and RRE (Figure 56.3), but we again found a large number of previously unknown features. Figure 56.5 und 56.6 show hairpins which are identical with those in the 5' TAR element. We again found a further secondary structure element (Figure 56.8) at the far 3' end which occurs only in 7 of 19 sequences, which is probably due to sequencing errors.

The alignment of HIV-2 with SIV_{sub2} shows a larger homology that with SIV_{sub1} (79% vs. 66%). As a consequence we find significantly more conserved secondary structure elements for HIV-2 with SIV_{sub2}. The question remains, however, whether elements that are conserved between HIV-2 with SIV are not detected because of alignment errors in the case of HIV-2/SIV_{sub1}, or whether HIV-2 truly shares more structure elements with SIV_{sub2} than SIV_{sub1}.

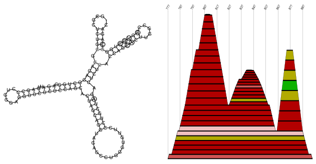


52.1 Mountain plot `Ralign`

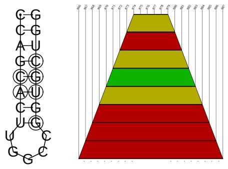


52.2 Mountain plot `ClustalW`

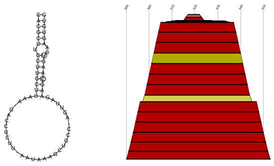
Figure 52: Mountain plots of HIV-2 and SIV$_{sub2}$:
19 sequences aligned by `Ralign`. Alignment length is 11267 bases, conserved 4177 and the mean pairwise homology is 79.1%.
19 sequences aligned by `ClustalW`. Alignment length is 11310 bases, conserved 4217 and the mean pairwise homology is 79.1%.



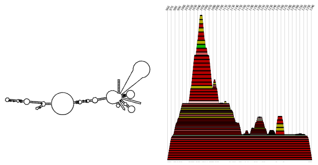
53.1 778-895 TAR



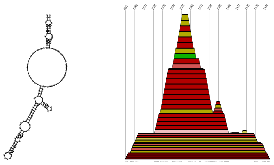
53.2 867-887



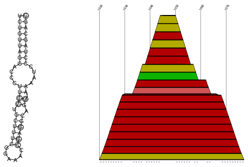
53.3 900-962



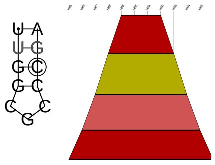
53.4 967-1348 *



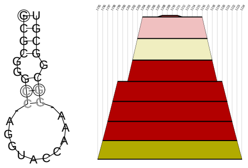
53.5 992-1146



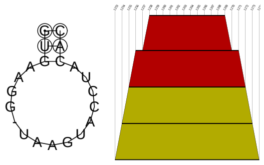
53.6 1029-1085



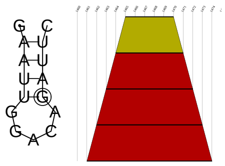
53.7 1086-1096



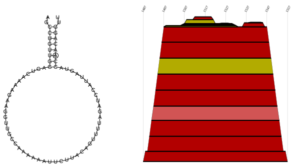
53.8 1196-1224



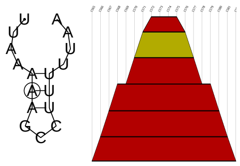
53.9 1254-1274



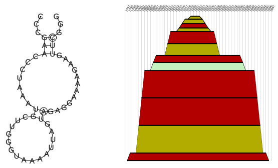
53.10 1461-1475



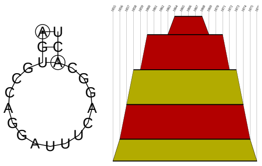
53.11 1488-1557



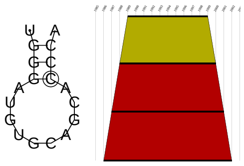
53.12 1566-1582



53.13 1798-1847

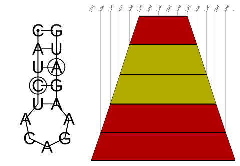


53.14 1856-1876

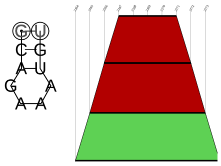


53.15 1986-2003

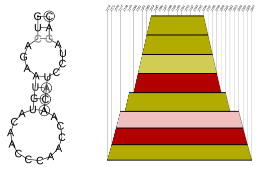
Figure 53: Detected conserved secondary structures of HIV-2 and SIV$_{sub2}$ sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. The structure labeled by (*) has long range interactions, see also Table 2. (partI)



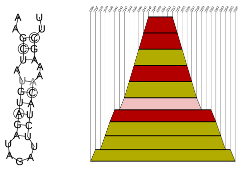
54.1 2135-2149



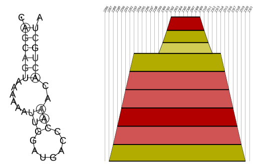
54.2 2165-2174



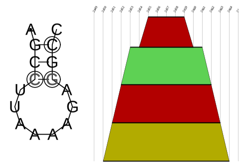
54.3 2175-2207



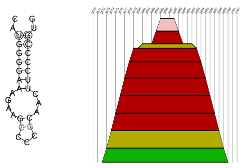
54.4 2237-2266



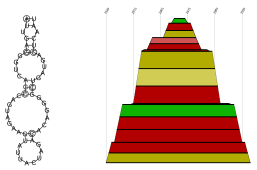
54.5 2287-2321



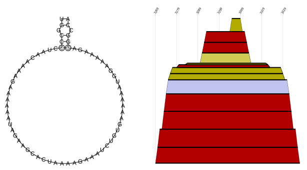
54.6 2450-2465



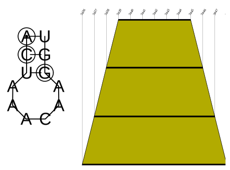
54.7 2671-2702



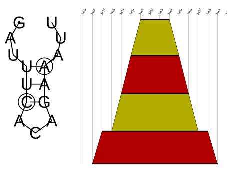
54.8 2946-2998



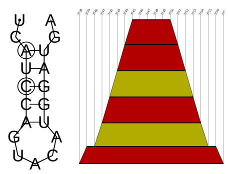
54.9 3270-3337



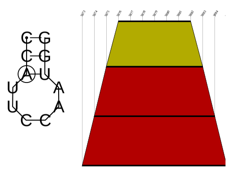
54.10 3437-3448



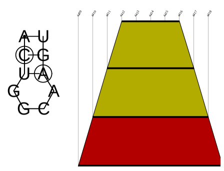
54.11 3456-3470



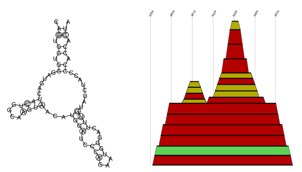
54.12 3739-3757



54.13 3876-3887

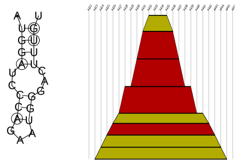


54.14 4411-4420

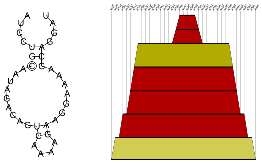


54.15 4395-4463

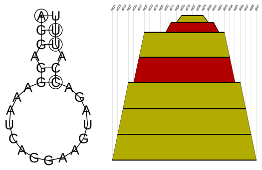
Figure 54: Detected conserved secondary structures of HIV-2 and SIV$_{sub2}$ sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. (partII)



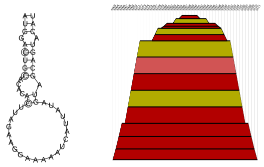
55.1 4425-4446



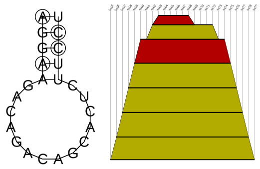
55.2 4529-4566



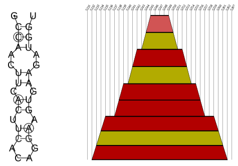
55.3 4823-4849



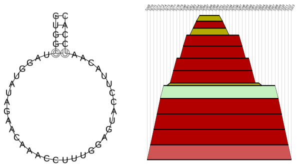
55.4 5064-5111



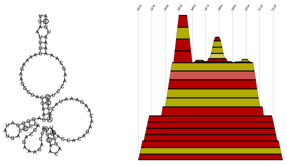
55.5 5156-5179



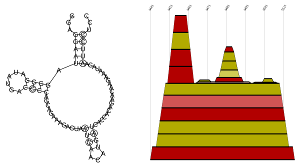
55.6 5232-5263



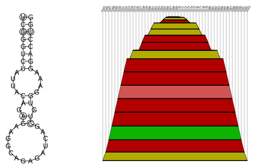
55.7 5270-5313



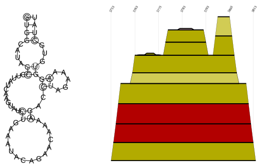
55.8 5425-5531



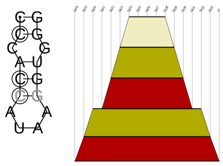
55.9 5442-5517



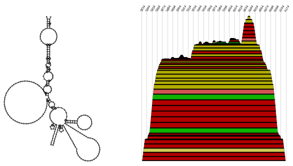
55.10 5546-5593



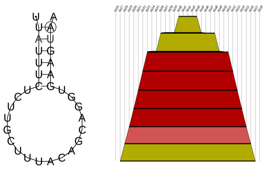
55.11 5754-5814



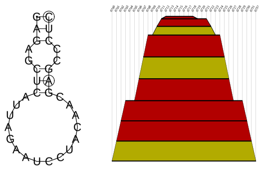
55.12 5819-5834



55.13 5835-6114

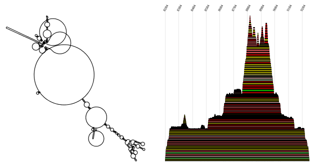


55.14 6027-6058

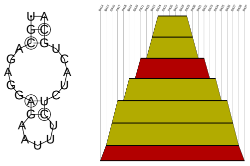


55.15 6701-6732

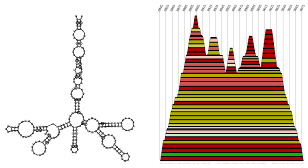
Figure 55: Detected conserved secondary structures of HIV-2 and SIV$_{sub2}$ sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. (partIII)



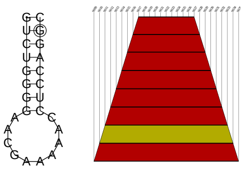
56.1 8285-9336 *



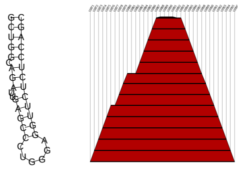
56.2 8415-8439



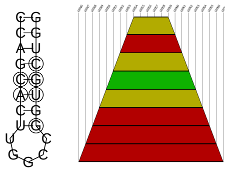
56.3 8846-9079 RRE



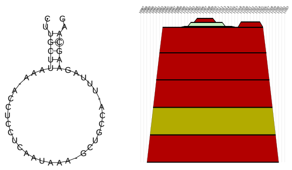
56.4 9010-9035



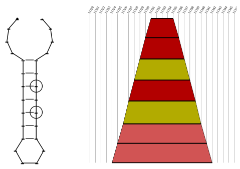
56.5 10872-10906



56.6 10947-10967



56.7 10988-11033



56.8 11121-11146

Figure 56: Detected conserved secondary structures of HIV-2 and SIV$_{sub2}$ sequences after aligning the sequences with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. The structure labeled by (*) has long range interactions, see also Table 2. (partIV)

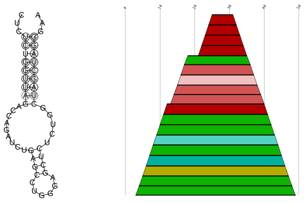
5.3 Discussion

Our calculations verify the fact that HIV-2 is more closely related to SIV than to HIV-1, especially to the so-called subgroup 2. We also confirmed the already published secondary structural elements in the 5' end of the viral sequences, RFSH and RRE.

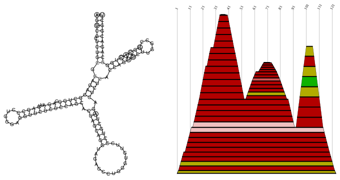
As we can see in Figure 57.1 the TAR in HIV-1 is a single stem-loop, whereas the HIV-2/SIV TAR consists of three stem-loops (Figure 57.2 and 57.5). In HIV-2 we found two haipins see Figure 57.3 and 57.4, near the 3' end which are identical to the corresponding ones in the TAR element at the 5' end (Figure 57.2). The TAR element in SIV$_{sub2}$ seems to be present at the 5' end as well as at the 3' end with slight differences in one hairpin, see Figure 57.5 and 57.6, respectively. As for HIV-2 alone we again found two hairpins (Figures 57.8 and 57.9) at the 3' end corresponding to the 5' TAR (Figures 57.7) by aligning HIV-2 together with SIV_{sub2}. We did not find a TAR element in SIV sequences belonging to subtype 1, this could be due to either the divergence of the sequences and/or to the fact that the sequences had to be cut prior to the analysis which might have been at the wrong site.

In Figure 58 we listed the RRE elements found in the primate lentiviruses. The RRE element in HIV-2 and SIV_{sub2} (Figure 58.2 and 58.5, respectively) are nearly identical, whereas there are differences in the sizes of the outermost hairpins in the RRE in HIV-1 (Figure 58.1) and in hairpin V in the RRE in SIV_{sub1} (Figure 58.4). The differences in SIV_{sub1} are due to the fact that in some sequences the stacks are shifted compared to the consensus structure. Especially sequence SIVMNDGB exhibits structure variation in the area of hairpin V and VI. However, the general structural pattern is conserved. As we can see the RRE is not convincing in all SIV sequences, see Figure 58.3, because of shifted structural patterns.

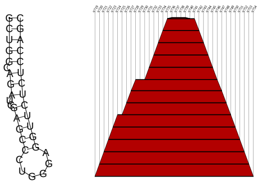
The folding of long RNA molecules as a single piece as opposed to the folding of short subsequence allows us to observe long range interactions, see Table 2. It seems that stacks with a high probabilities enclose rather flexible regions. We further found many hairpin loops with relatively short stacks, but the fact that they occurred in all sequences gives credence to their existence, e.g. see SD element in Figure 16.4. Furthermore, we found a large number of promising secondary structure elements, but unfortunately, it is impossible to specify their functions with our methods.



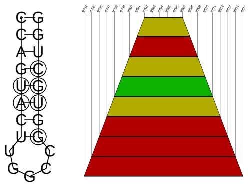
57.1 TAR element HIV-1



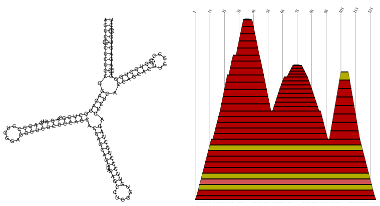
57.2 5'TAR in HIV-2



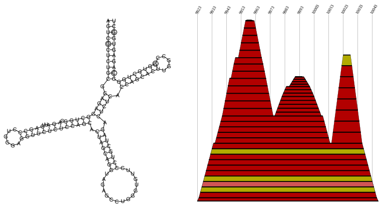
57.3 3'TAR in HIV-2



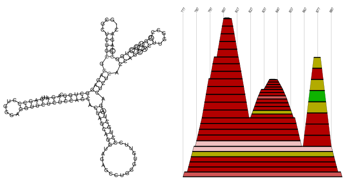
57.4 3'TAR in HIV-2



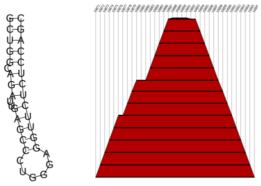
57.5 5'TAR in SIV$_{sub2}$



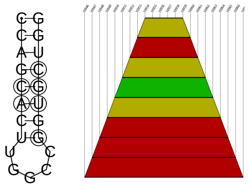
57.6 3'TAR in SIV$_{sub2}$



57.7 5'TAR in HIV-2 and SIV$_{sub2}$

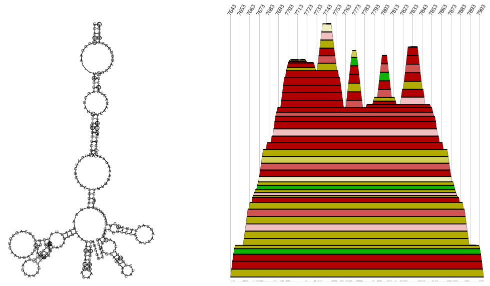


57.8 3'TAR in HIV-2 and SIV$_{sub2}$

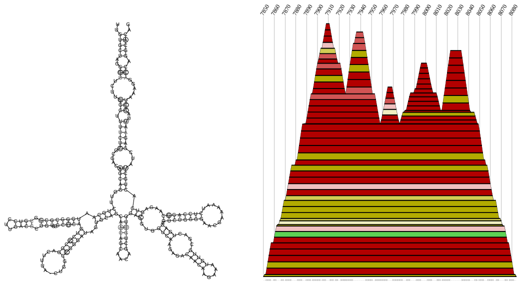


57.9 3'TAR in HIV-2 and SIV$_{sub2}$

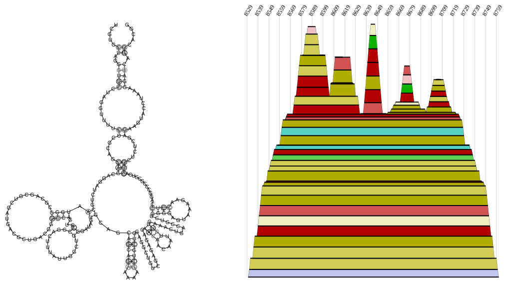
Figure 57: The TAR elements in HIV-1, HIV-2, and SIV$_{sub2}$



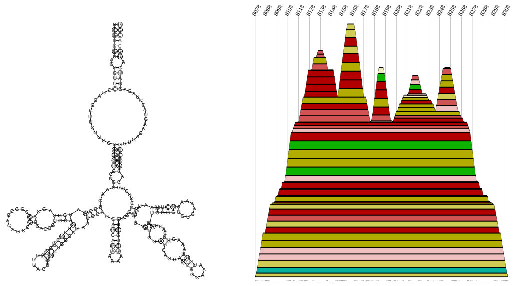
58.1 RRE in HIV-1



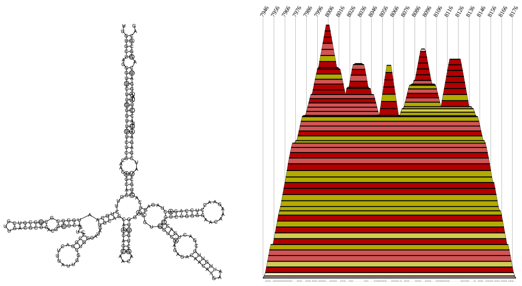
58.2 RRE in HIV-2



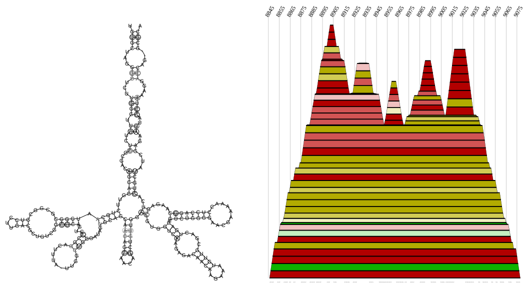
58.3 RRE in SIV



58.4 RRE in SIV_{sub1}



58.5 RRE in SIV_{sub2}



58.6 RRE in HIV-2 and SIV_{sub2}

Figure 58: The RRE elements in HIV-1, HIV-2, and SIV (SIV_{sub2})

Table 2: List of secondary structure elements with long range interactions in primate lentivirus sequences.

Figure	Position (nt)	`Ralign`	`ClustalW`
HIV-2			
23.3	192-567	+	+
23.7	1596-1879	+	+
28.4	3230-4103	-	+
26.9	7336-8328	+	+
SIV_{sub1}			
34.9	6424-7083	+	+
SIV_{sub2}			
39.9	2167-2405	+	+
40.5	2952-3446	+	+
42.11	7393-8439	+	+
42.15	9171-9539	+	+
HIV-2 and SIV			
46.7	8622-8814	+	+
HIV-2 and SIV_{sub1}			
50.7	8745-8976	+	+
HIV-2 and SIV_{sub2}			
53.4	967-1348	+	+
56.1	8285-9336	+	+

6 Picornaviridae

6.1 Rhinoviruses and Enteroviruses

The *Picornaviridae* are among the smallest ribonucleic acid-containing viruses known. Rhinovirus, poliovirus (genus enterovirus), human hepatitis A virus and foot-and-mouth disease virus (FMDV) are members of the picornavirus family. Their genome consists of a single strand messenger-active (+) RNA of 7,000 to 8,000 nts is polyadenylated at the 3' terminus and carries a small protein (virion protein, genome; VPg) covalently attached to its 5' end [2, 3].

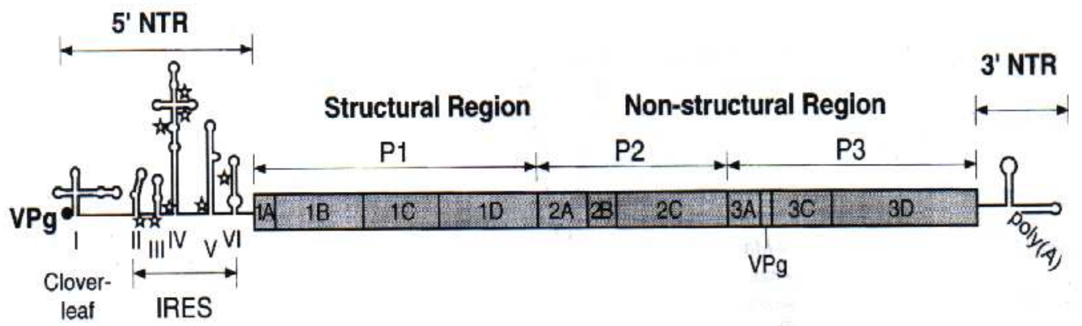


Figure 59: Structure of the poliovirus genome. The single-stranded RNA is shown with the genome-linked protein VPg (3B) at the 5' end of the non-translated region (5'NTR) and the 3'NTR connected to the poly(A) tail. The boxed region shows the polyprotein and vertical lines within the box indicate proteinase cleavage sites. The locations of the structural (P1) and non-structural (P2,P3) region are shown on top. RNA structural domains within the 5' non-translated region are shown by Roman numerals, cloverleaf (I), IRES (II-VI). Stars indicate the positions of non-initiating AUG triplets. Figure adapted from M. Gromeier *et. al.* [33].

At the 5' nontranslated region (NTR) of picornavirus RNA two characteristic folding patterns were found, see Figure 59, 65.1, and 67. The cloverleaf at the far 5' end is found in genomes of all rhinoviruses and enteroviruses. Furthermore, the internal ribosomal entry site (IRES) element which is a highly structured *cis*-acting element and plays a keyrole in initiating synthesis of viral protein [106, 139]. The primary structure of the poliovirus 5'NTR is strongly conserved between different serotypes and isolates, particularly within the first 650 nts. The 5'NTRs of other enteroviruses and rhinoviruses share a high degree of sequence identity with poliovirus. The influence of small mutations within the poliovirus

5'NTR on the attenuation phenotype and cap-independent translation suggests that the structure of the 5'NTR might be important for its function. Various models for the secondary structure of the 5'NTRs of members of the enterovirus and rhinovirus genera have been derived by biochemical probing in solution, computational methods, and comparative sequence analysis. They share many secondary structure elements referred to as type 1 IRES element. The boundaries of the poliovirus IRES have been defined by deletion analysis: 5' and 3' borders are at about nt 134 and nt 556, respectively [149].

6.2 Results

In our first experiment we compared all the members of the picornavirus family with each other. Figure 60 shows the split decomposition for all the picornavirus sequences that we used for this experiment. We can see that the different members are not closely related. The split between rhinoviruses and enteroviruses is the only significant one, as e.g. the coxsackievirus types A21 and A24 form a group with the polioviruses. The sequences were taken from the `GenBank`, details can be found in Table 11, in the Appendix.

In Figure 61 we see the alignments of the different genera calculated with two different alignment algorithms (`Ralign and ClustalW`). Due to the fact that the rhinovirus and the enterovirus sequences have little in common, we found very few secondary structure elements that are present in both genera. Parts of the IRES seem to be highly conserved in all sequences (Figure 62.4 and 62.7). A more detailed analysis of these results can be found in Figure 62. These figures show the secondary structure elements with the highest base pairing probabilities. This is further underlined by the occurring compensatory and consistent mutations (highlighted by the color code) in these regions.

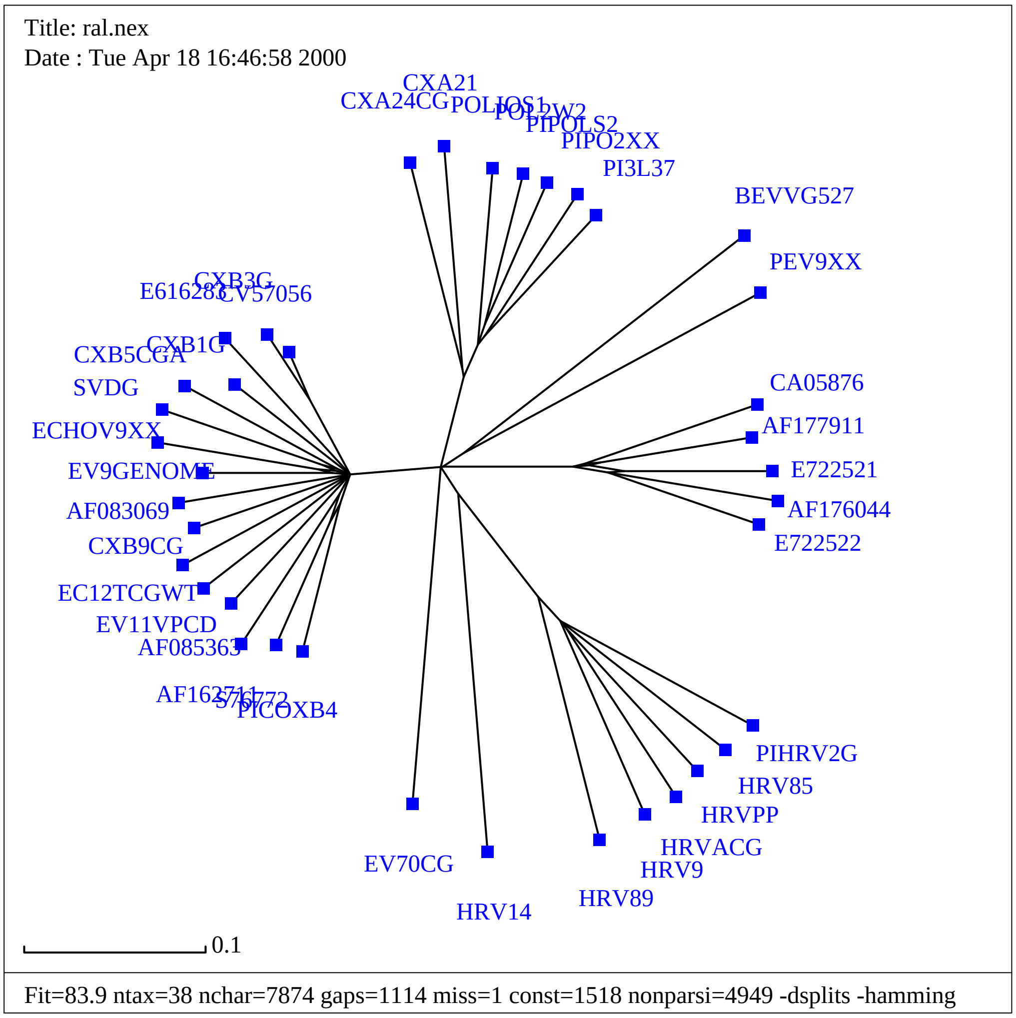
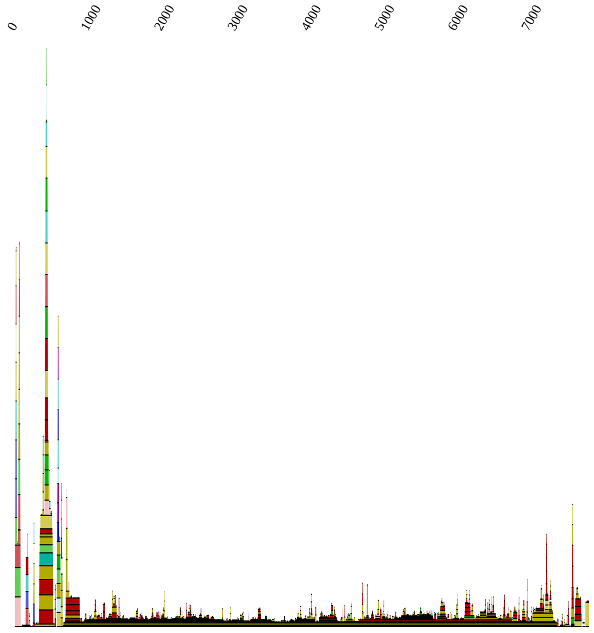
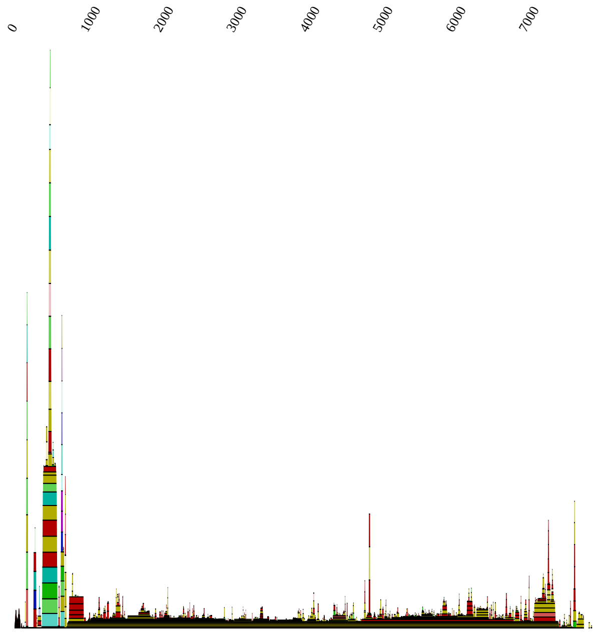


Figure 60: `Splitstree` plot of the aligned rhinovirus and enterovirus sequences.

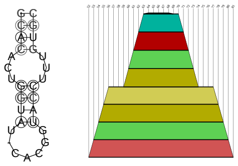


61.1 Mountain plot `Ralign`

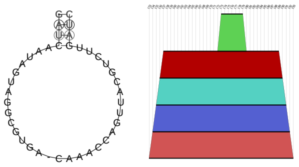


61.2 Mountain plot `ClustalW`

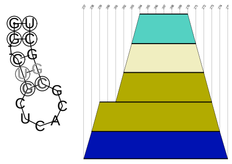
Figure 61: Mountain plots of rhino- and enteroviruses: For information on sequences see Table 11.
38 sequences aligned by `Ralign`. Alignment length is 7874 bases, conserved 1518 and the mean pairwise homology is 61.7%.
38 sequences aligned by `ClustalW`. Alignment length is 7911 bases, conserved 1513 and the mean pairwise homology is 61.6%.



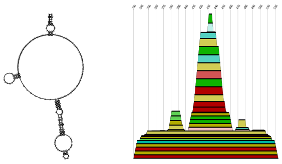
62.1 53-81 I



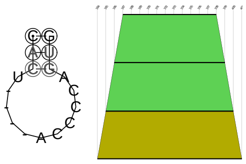
62.2 154-193



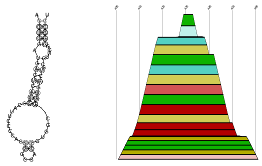
62.3 **258-275**



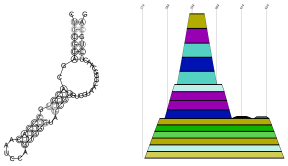
62.4 337-530 IV



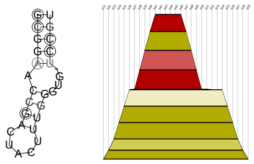
62.5 **385-401**



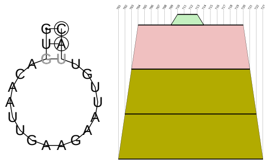
62.6 409-470



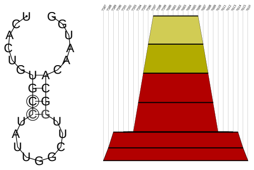
62.7 575-632 VI



62.8 632-659



62.9 **702-723**



62.10 7588-7616

Figure 62: Detected conserved secondary structures of rhino- and enterovirus sequences aligned with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. I indicates a hairpin of the cloverleaf structure, the elements IV and VI are part of the IRES. Secondary structure features after `ClustalW` alignment were found shiftet downstream in the aligned sequence. Position of secondary features found in the `Ralign` but not in the `ClustalW` alignment are given in bold.

6.2.1 Rhinovirus

The human rhinoviruses are the most important etiological agents of the common cold in adults and children and comprise as much as 102 serotypes. Most of the picornaviruses isolated from the human respiratory system are acid labil, and this lability has become a defining characteristic of rhinoviruses [40].

In this experiment we compared all the members of the rhinovirus genus with each other including strain 14 which is the most divergent strain. Figure 63 shows the split decomposition for all the rhinovirus sequences that we used for this experiment. We can see that most of the different strains are closely related. The sequences were taken from the `GenBank`, details can be found in Table 11.

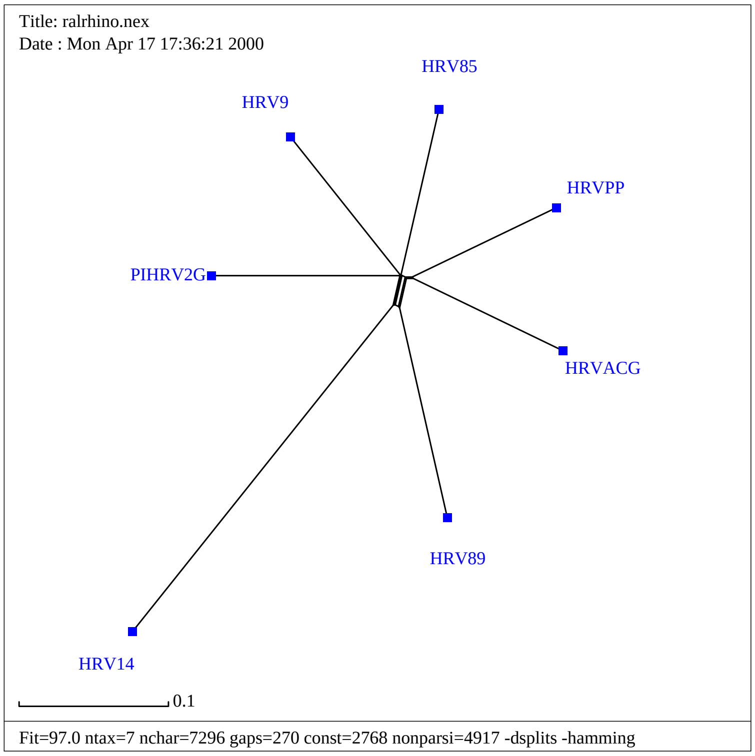
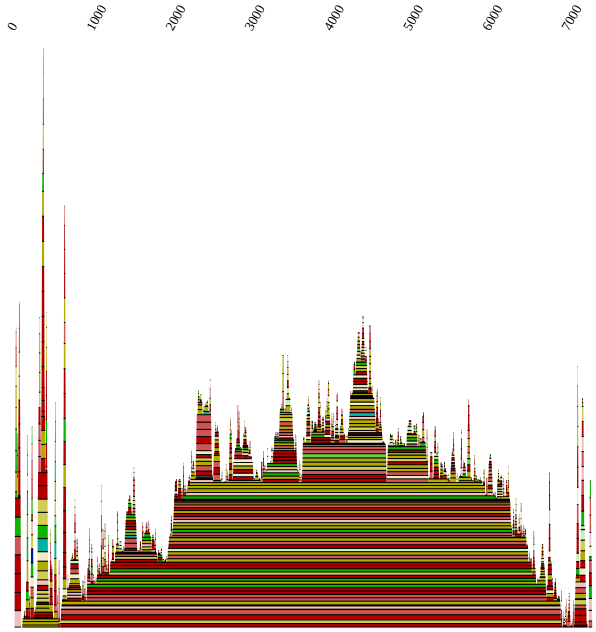
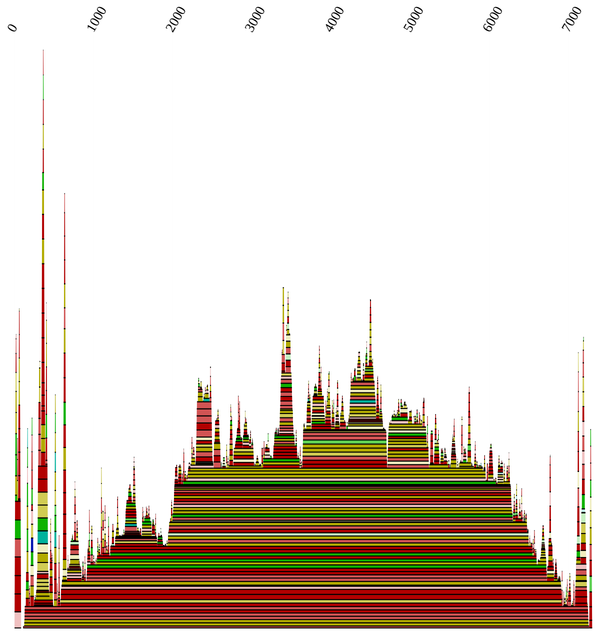


Figure 63: `Splitstree` plot of the aligned rhinovirus sequences.

In Figure 64 we see the alignments of the different strains calculated with two different alignment algorithms (`Ralign` and `ClustalW`). It seems that in the `Ralign` plot the 5'NCR and 3'NCR form distinct units, whereas in the `ClustalW` alignment we find long range interactions between the 5' and the 3' end. But the secondary structure elements are nearly identical in both alignments. A more detailed analysis of these results can be found in Figures 65 to 66. These figures show the secondary structure elements with the highest base pairing probabilities. This is further underlined by the occurring compensatory and consistent mutations (highlighted by the color code) in these regions. The secondary structure elements which were found with the `Ralign` algorithm only, are emphasised by bold numbers. Some of these structural elements have been described in literature [80] like cloverleaf (CL), see Figure 65.1 and IRES (Figure 67), but we found a large number of previously unknown features. In this case `Ralign` is more convincing than `ClustalW`, because it detects hairpin V of the IRES, see Figure 65.5. However, the most promising secondary structure features are located at the 5' and 3' end of the sequences.

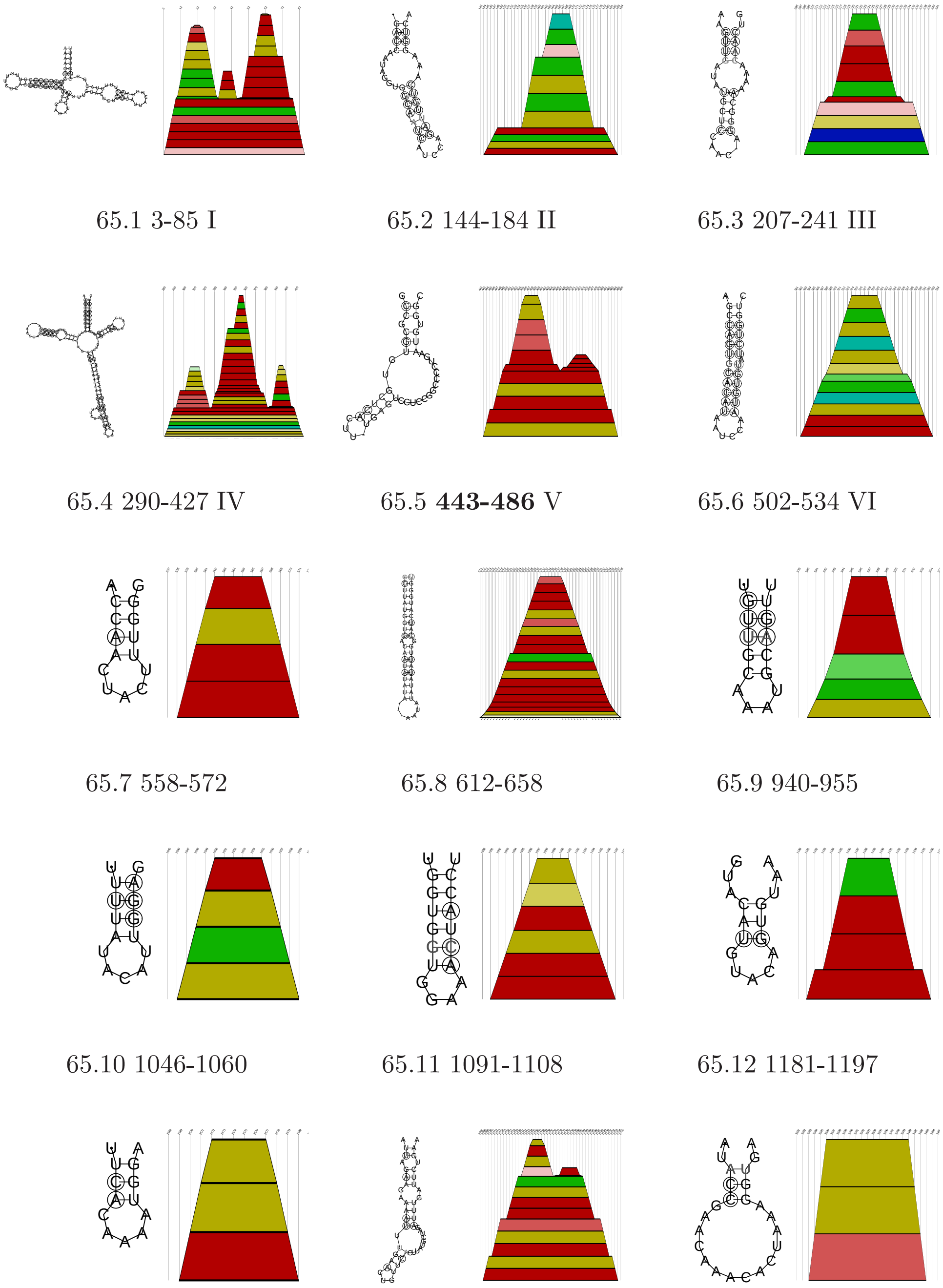


64.1 Mountain plot `Ralign`



64.2 Mountain plot `ClustalW`

Figure 64: Mountain plots of rhinovirus: For information on sequences see Table 11.
7 sequences aligned by `Ralign`. Alignment length is 7296 bases.
7 sequences aligned by `ClustalW`. Alignment length is 7293 bases, conserved 2819 and the mean pairwise homology is 69.0 %.



65.1 3-85 I

65.2 144-184 II

65.3 207-241 III

65.4 290-427 IV

65.5 **443-486** V

65.6 502-534 VI

65.7 558-572

65.8 612-658

65.9 940-955

65.10 1046-1060

65.11 1091-1108

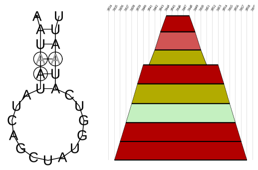
65.12 1181-1197

65.13 2669-2681

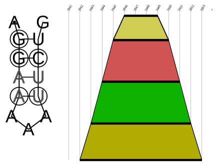
65.14 2708-2755

65.15 3382-3405

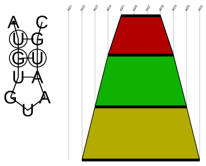
Figure 65: Detected conserved secondary structures of rhinovirus sequences aligned by `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. I indicates the cloverleaf structure, the elements II to VI form the IRES. The position of the secondary structure feature found in the `Ralign` but not in the `ClustalW` alignment is given in bold. (part I)



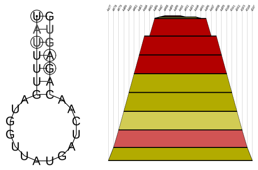
66.1 3835-3859



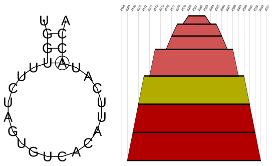
66.2 3942-3954



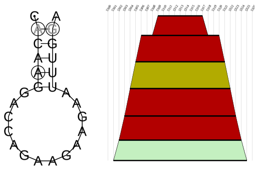
66.3 4022-4032



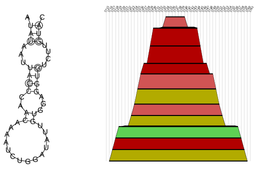
66.4 4478-4505



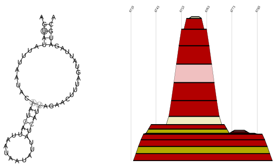
66.5 **4569-4594**



66.6 5301-5326

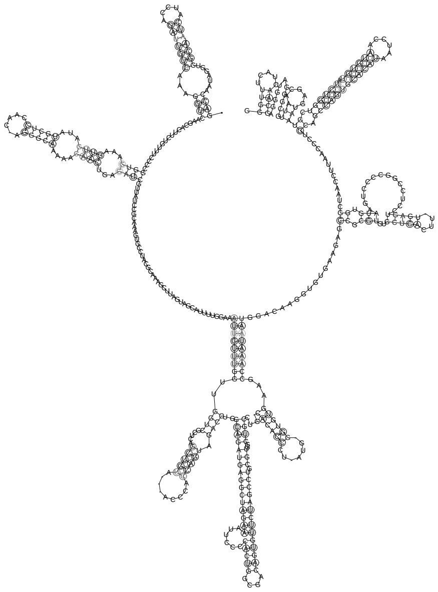


66.7 **5716-5761**

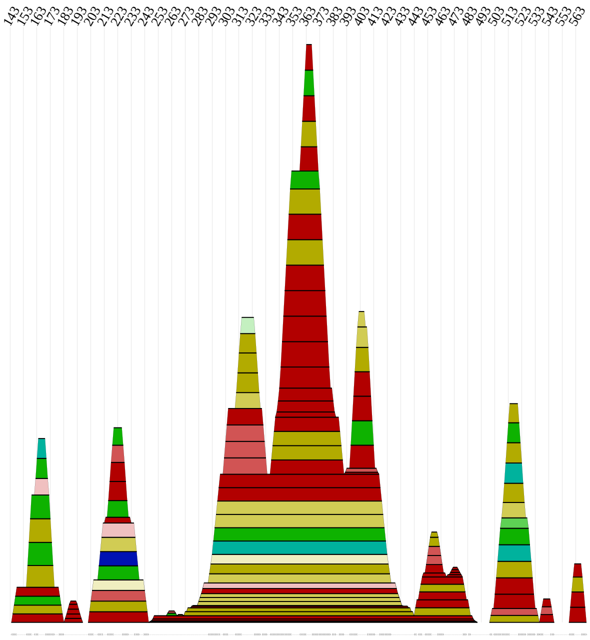


66.8 6734-6790

Figure 66: Detected conserved secondary structures of rhinovirus sequences aligned by `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. Positions of secondary features found in the `Ralign` but not in the `ClustalW` alignment are given bold. (part II)



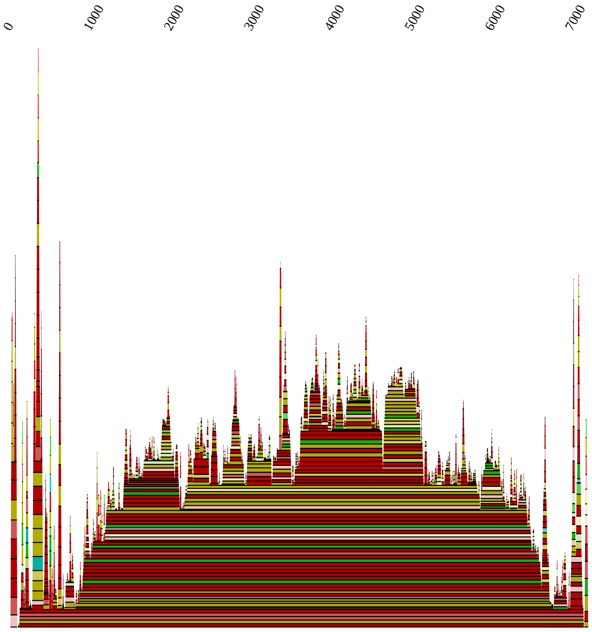
67.1 Secondary structure plot of the IRES



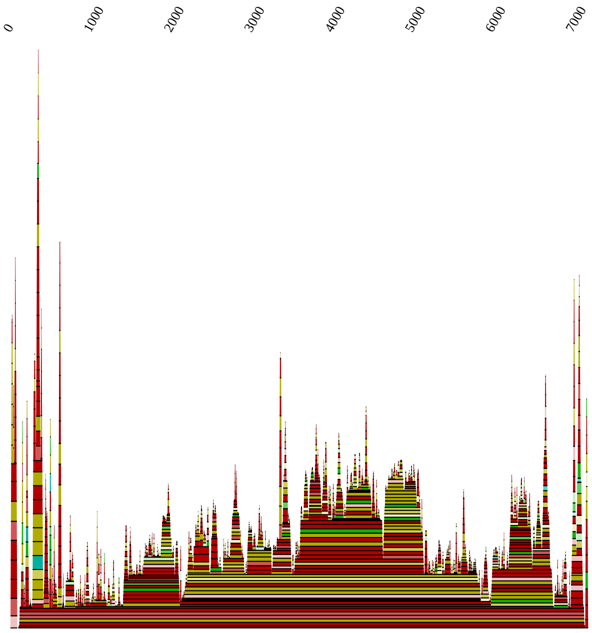
67.2 Mountain plot of the IRES

Figure 67: Rhinoviruses sequences at position nt 144 to 572 of the `Ralign` alignment containing the IRES region.

Since strain 14 is very divergent from the other strains [93] we decided to perform a further experiment excluding it. In Figure 68 we see the alignments again calculated with the two different alignment algorithms (`Ralign` and `ClustalW`). We found the same secondary structure elements in both alignments. However, the `Ralign` mountain plot shows long range interactions not present in the `ClustalW` plot. A more detailed analysis of these results can be found in Figures 69 to 70 and Figure 71. These figures show the secondary structure elements with the highest base pairing probabilities. This is further underlined by the occurring compensatory and consistent mutations (highlighted by the color code) in these regions. Some of these structural elements have been described in literature [80] like cloverleaf (CL) and IRES, but we found a large number of previously unknown features. The secondary structure elements which were found only with the `ClustalW` algorithm can be seen in Figure 71. However, by excluding strain 14 there are more promising secondary structure features located at the 3' end of the sequences, see Figures 70.13, 70.14, and 70.15.

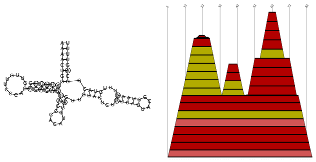


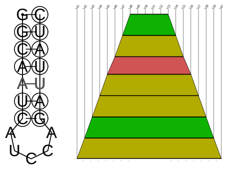
68.1 Mountain plot `Ralign`

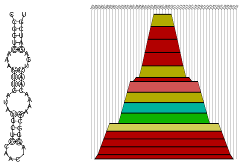


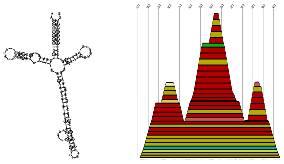
68.2 Mountain plot `ClustalW`

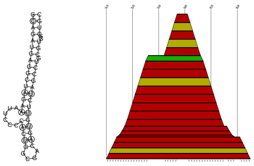
Figure 68: Mountain plots of rhinoviruses excluding strain 14:
6 sequences aligned by `Ralign`. Alignment length is 7198 bases, conserved 3690 and the mean pairwise homology is 73.4%.
6 sequences aligned by `ClustalW`. Alignment length is 7189 bases, conserved 3705 and the mean pairwise homology is 73.6%.

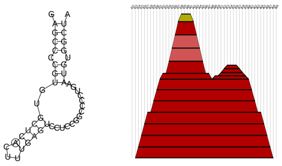

69.1 3-85 I

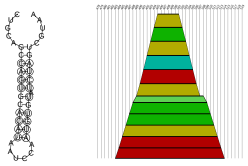

69.2 142-160 II

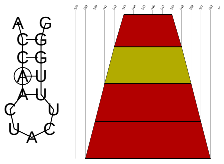

69.3 184-232 III

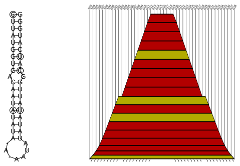

69.4 273-410 IV

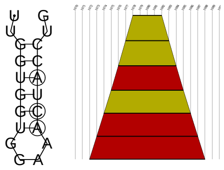

69.5 319-373

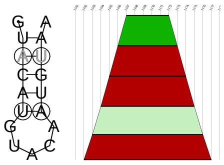

69.6 423-469 V

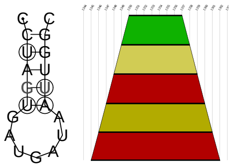

69.7 479-519 VI

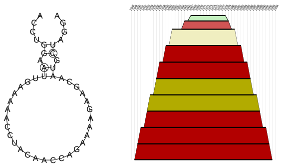

69.8 539-553

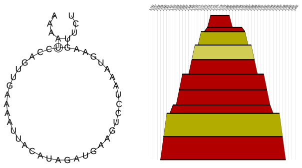

69.9 594-638


69.10 1071-1090


69.11 1162-1178


69.12 1345-1363


69.13 2049-2094


69.14 **2357-2402**

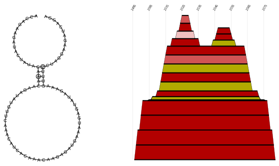
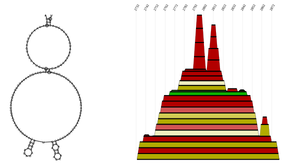
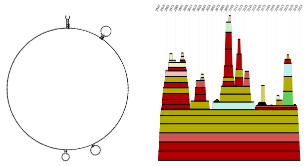

69.15 2496-2583

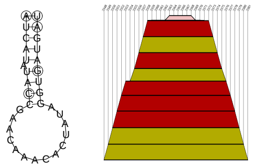
Figure 69: Detected conserved secondary structures of rhinovirus sequences excluding strain 14 aligned with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. The numbers denote the base pair range in the alignment. I indicates the cloverleaf structure, the elements II to VI form the IRES. Position of secondary features found in the `Ralign` but not in the `ClustalW` alignment are given in bold. (part I)



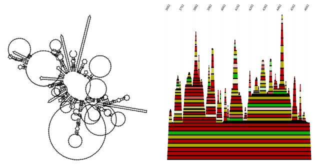
70.1 2733-2881



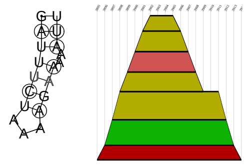
70.2 2944-3257



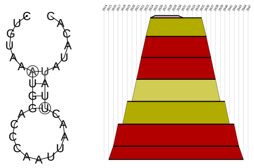
70.3 3349-3381



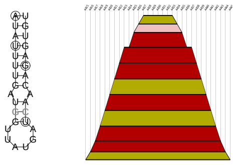
70.4 3603-4641 *



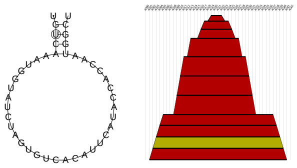
70.5 3896-3914



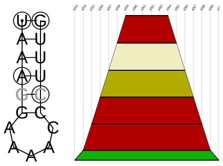
70.6 3915-3945



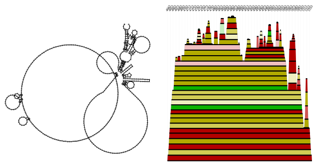
70.7 4416-4445



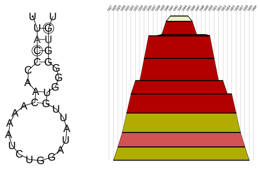
70.8 4501-4542



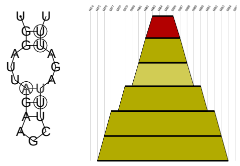
70.9 4554-4570



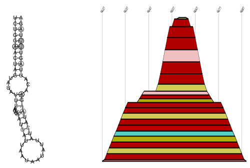
70.10 4644-5141 *



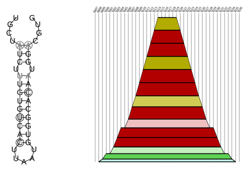
70.11 5628-5660



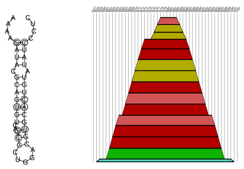
70.12 6075-6095



70.13 6628-6689



70.14 6998-7036



70.15 7058-7102

Figure 70: Detected conserved secondary structures of rhinovirus sequences excluding strain 14 aligned with `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. Structures labeled by (*) have long range interactions, see also Table 3. (part II)

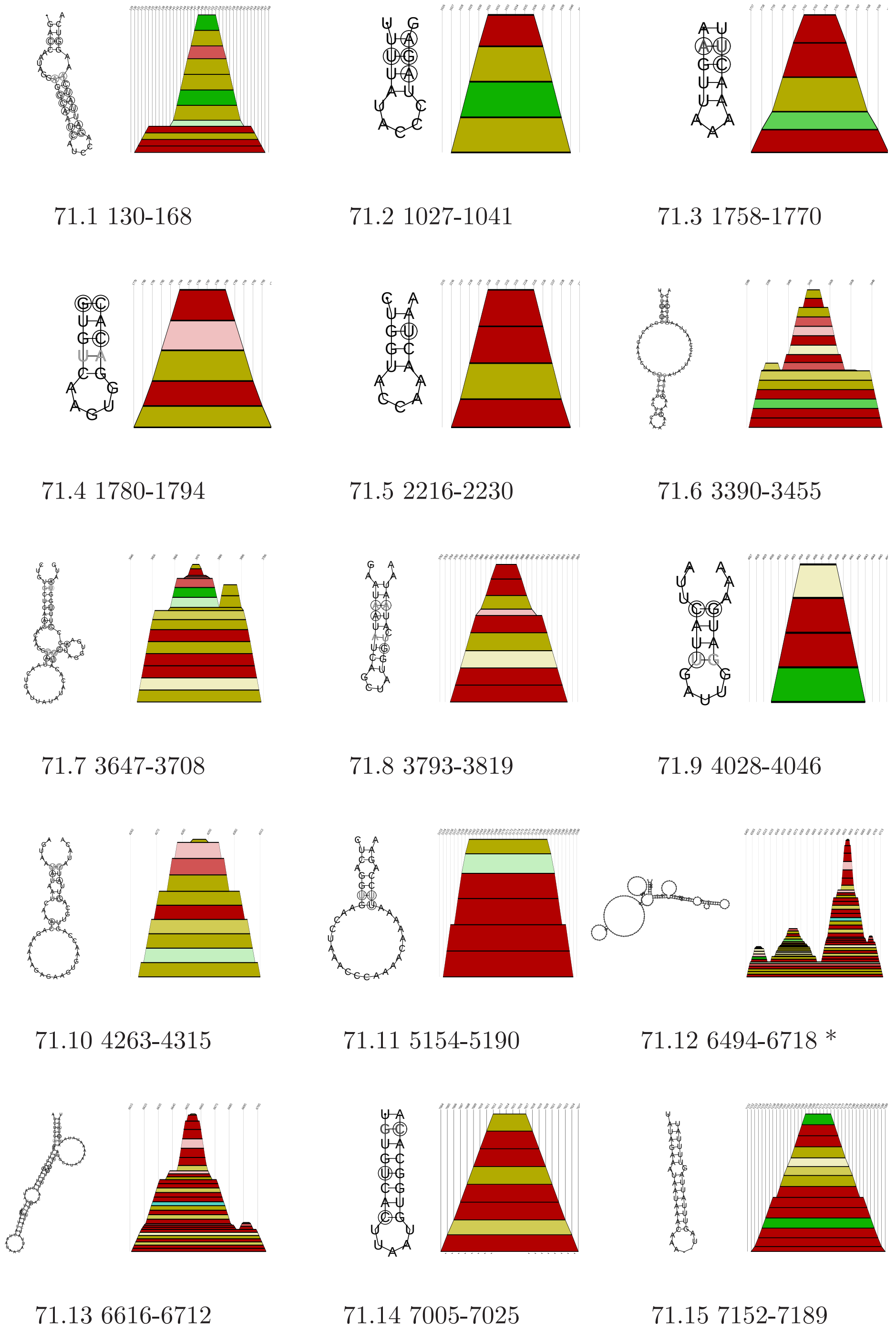


Figure 71: Detected conserved secondary structures of rhinovirus sequences excluding strain 14 aligned by `ClustalW` not found in the `Ralign` alignment. The numbers denote the base pair range in the alignment. The structure labeled by (*) has long range interactions, see also Table 3.

6.2.2 Enterovirus

The Enterovirus genus, so-called because most inhabit the alimentary (enteric) tract, include not only the polioviruses [149], but also the coxsackieviruses, the echoviruses, human enteroviruses, and a number of nonhuman enteric viruses.

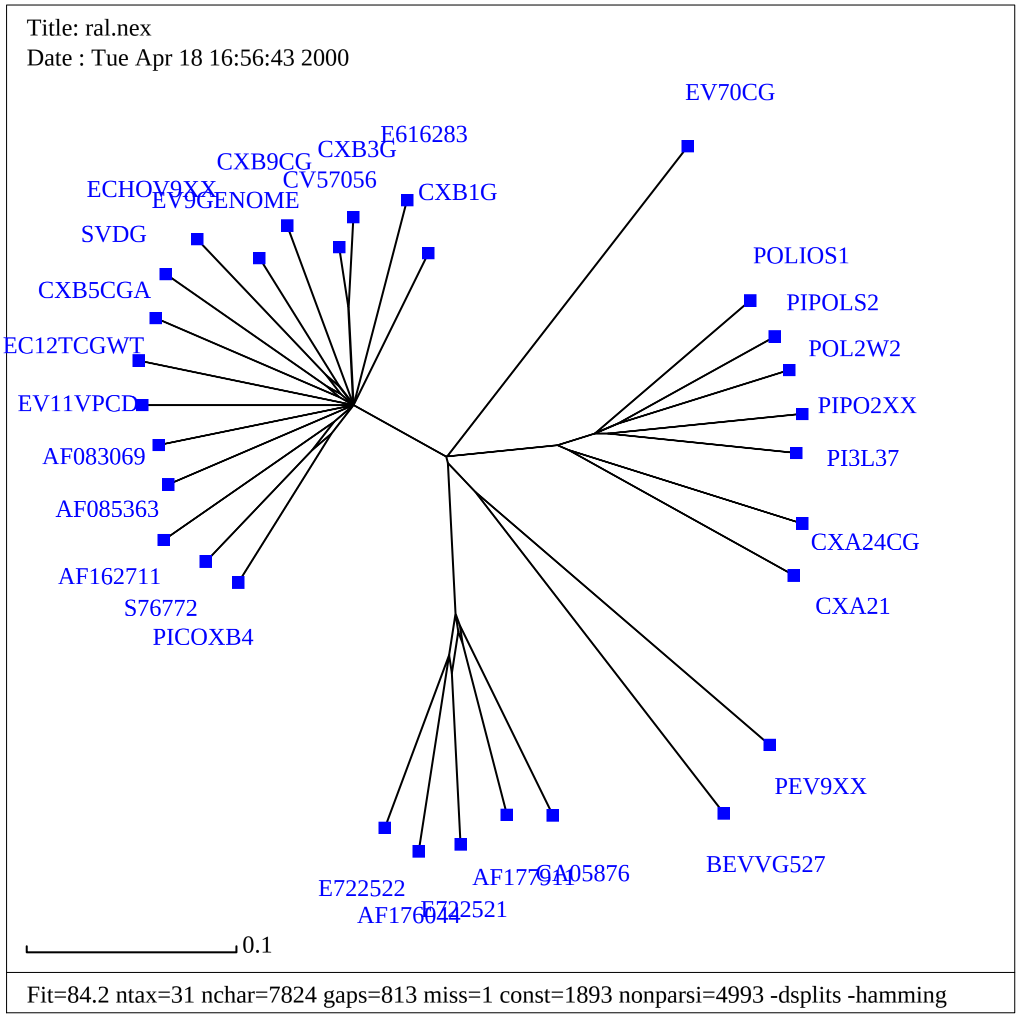
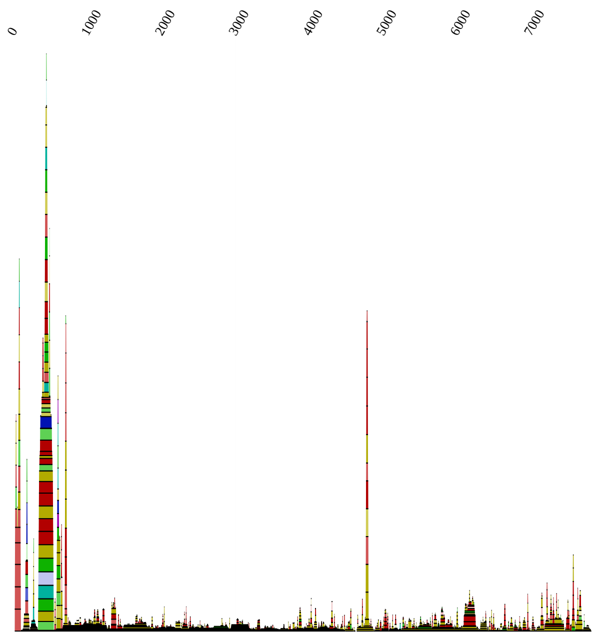


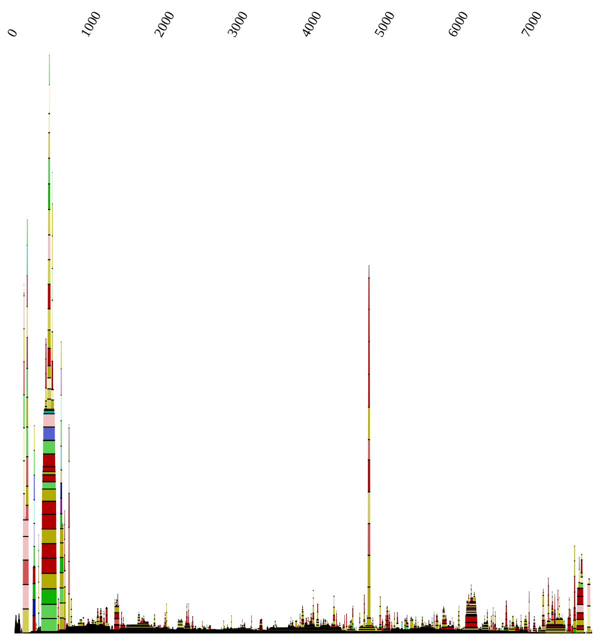
Figure 72: `Splitstree` plot of the aligned enterovirus sequences.

In this experiment we compared all the members of the enterovirus genus with each other. Figure 72 shows the split decomposition for all the enterovirus sequences that we used for this experiment. We can see that most of the different strains are not closely related. As expected, two of the coxsackievirus strains, namely coxsackievirus A21 and A24 are similar to poliovirus, whereas the other are far apart. The sequences were taken from the `GenBank`, details can be found in Table 11.

In Figure 73 we see the alignments of the different strains calculated with the algorithms `Ralign` and `ClustalW`. Although we found very few secondary structure elements, they are the same in both alignments. A more detailed analysis of these results can be found in Figure 74. This figure shows the secondary structure elements with the highest base pairing probabilities. This is further underlined by the occurring compensatory and consistent mutations (highlighted by the color code) in these regions. The secondary structure elements which were found only with the `Ralign` algorithm, are emphasized by bold numbers. The most significant secondary structures seem to be the IRES and a second element at position nt 4757 to 4797, see Figure 74.9.



73.1 Mountain plot `Ralign`



73.2 Mountain plot `ClustalW`

Figure 73: Mountain plots of enterovirus: For information on sequences see Table 11.
31 sequences aligned by `Ralign`. Alignment length is 7824 bases, conserved 1893 and the mean pairwise homology is 65.1%.
31 sequences aligned by `ClustalW`. Alignment length is 7872 bases, conserved 1908 and the mean pairwise homology is 65.3%.

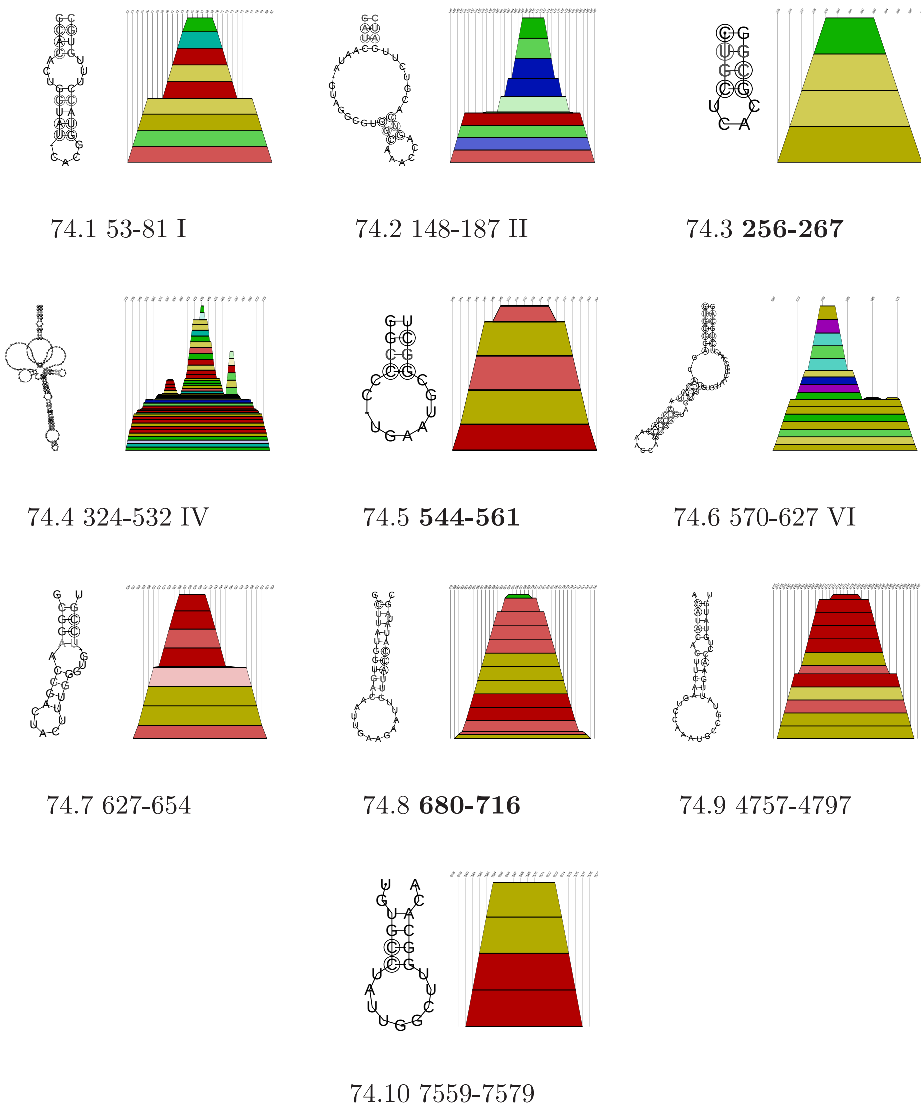


Figure 74: Detected conserved secondary structures of enterovirus sequences aligned by `Ralign` and `ClustalW`. The numbers denote the base pair range in the `Ralign` alignment. I indicates a hairpin of the cloverleaf structure, the elements II, IV, and VI are part of the IRES. Positions of secondary features found in the `Ralign` but not in the `ClustalW` alignment are given in bold.

6.3 Discussion

By aligning both rhinoviruses and enteroviruses which are not closely related according to their sequences we found just a few secondary structures present in both serocomplexes. Our calculations for the rhinoviruses confirmed the already known secondary structure elements at the 5' end of the sequences, e.g. the IRES region. However, we could not verify the *cis*-acting replication element *cre* at position nt 2318 to 2413 in rhinovirus sequences, see Figure 69.14, suggested for HRV-14 in literature by using the `mfold` algorithm [93]. Most secondary structure predictions in the literature have so far only considered a fairly small sample of suboptimal structures, as provided, e.g. by Zuker's `mfold` package [156, 155]. McCaskill's partition function approach [92], 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. The dot plot of the region nt 2282 to 2509 in rhinovirus sequences (Figure 75) shows that their is no convincing evidence for base pairing in this area.

Folding the entire RNA sequences allows us to detect long range interactions, see Table 3, where it seems that stacks with high probabilities enclose rather flexible regions.

Table 3: List of secondary structure elements with long range interactions in rhinovirus sequences excluding strain 14.

Figure	Position (nt)	`Ralign`	`ClustalW`
70.4	3603-4641	+	+
70.10	4644-5141	+	+
71.12	6494-6718	-	+

The enterovirus serocomplex is very divergent, therefore we could not find many secondary structural features. A very interesting feature can be seen in Figure 74.8, every single basepair of this element is underlined by a compensatory mutation. In the mountain plots of the enterovirus sequence alignments (Figure 73) a further promising element occurred within the open reading frame at ca. nt 4800, see Figure 74.9. Unfortunately, it is impossible to specify the function of a particular secondary structure element with our methods.

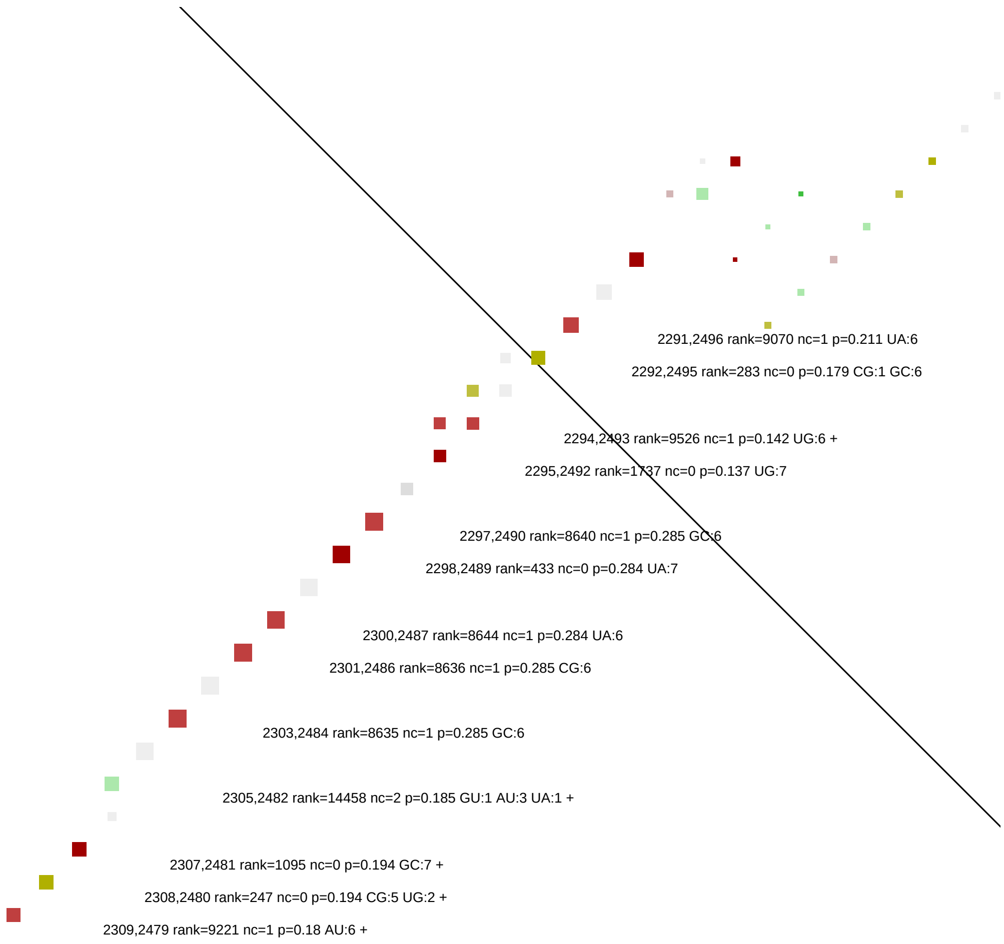


Figure 75: Dot plot of the region nt 2282 to 2509 in rhinovirus sequences. (+) indicates pairs which conflict with another higher ranking (more plausible) pair.

7 Flaviviridae

7.1 Flaviviruses

The *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. Virions contain one molecule of linear positive-sense single stranded RNA. Total genome length is 9,500 to 12,500 nts. The 5' end of the genome has a cap, or a genome-linked protein (VPg).

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 [96]. 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 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. 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 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.

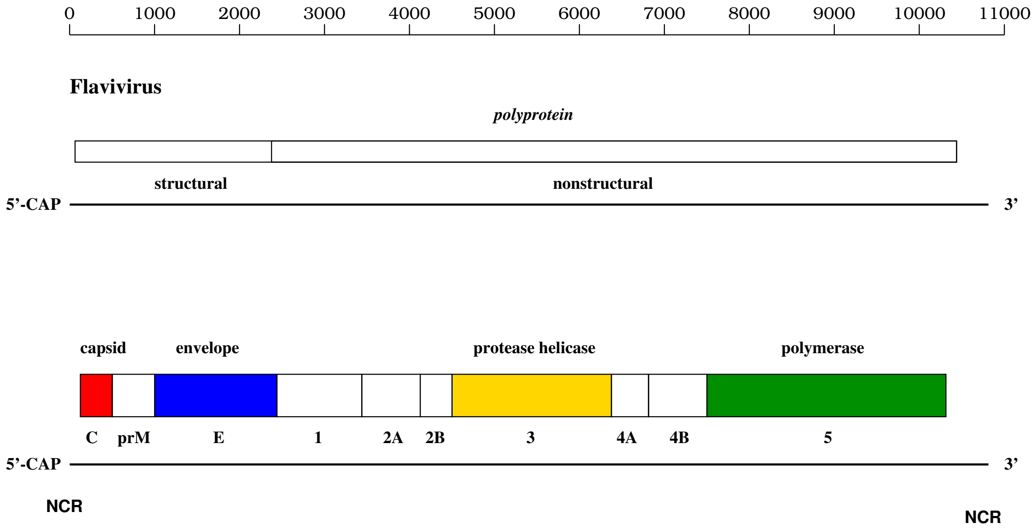


Figure 76: Flavivirus genome map. Translation and processing of the flavivirus polyprotein. At the top is the viral genome with structural and non-structural protein coding region. Boxes below indicate mature proteins generated by the proteolytic processing cascade [40, 131].

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 [40, 131, 23], see Figure 76. 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 [124]. Genome-length RNA appear to be the only virus specific messenger RNA (mRNA) molecule in flavivirus infected cells. The genomic RNA has a type I cap at its 5' end (mGpppAmp). Genomic RNAs of mosquito-borne and tick-borne flaviviruses lack a 3' terminal poly(A) tract and terminate with the conserved dinucleotide CU.

For our analysis we searched sequence data bases for all available complete flavivirus sequences, and found 84 complete sequences. After sorting out similar sequences (pairwise homology greater than 98%) 24 sequences remained , see Table 12. The length of alignment `ClustalW` is 11435 bases and the mean pairwise homology 62.5%. The data set of flavivirus sequences was too diverse to detect conserved structures with our method. In this thesis we focussed on the dengue viruses as well as on the 3'NCR of dengue, yellow fever, and Japanese encephalitis viruses.

7.2 Results

7.2.1 Dengue Virus

The dengue virus group contains 9 different complete sequences. Two sequences belong to the dengue-1, five to the dengue-2, one constitutes the dengue-3 and dengue-4 type. For information on sequences see Table 12. All 9 sequences were aligned by `Ralign`. Alignment length is 10777 bases, the number of conserved bases is 5229 and the mean pairwise homology is 76.3%.

Figure 77 shows the split decomposition for all the dengue strain sequences that we used for this experiment. We can see that although the strains belonging to the respective types are closely related, there are great dissimilarities among the different types.

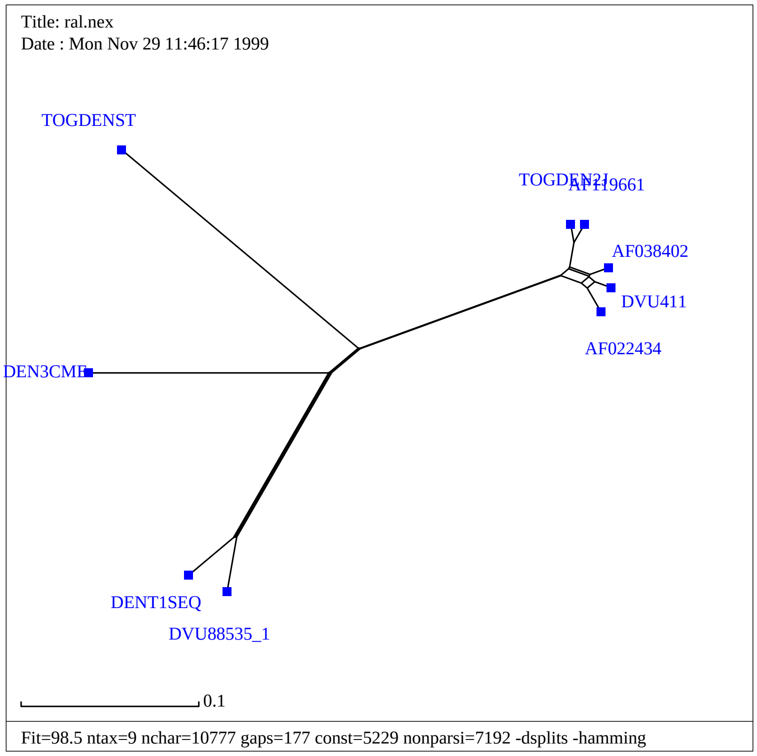


Figure 77: `Splitstree` plot of the aligned dengue sequences.

In Figure 78 we see the alignment of the different dengue strains calculated with the `Ralign` algorithm. The employed sequences (see Table 12) were found in the `GenBank`. A more detailed analysis of these results can be found in Figures 79 to

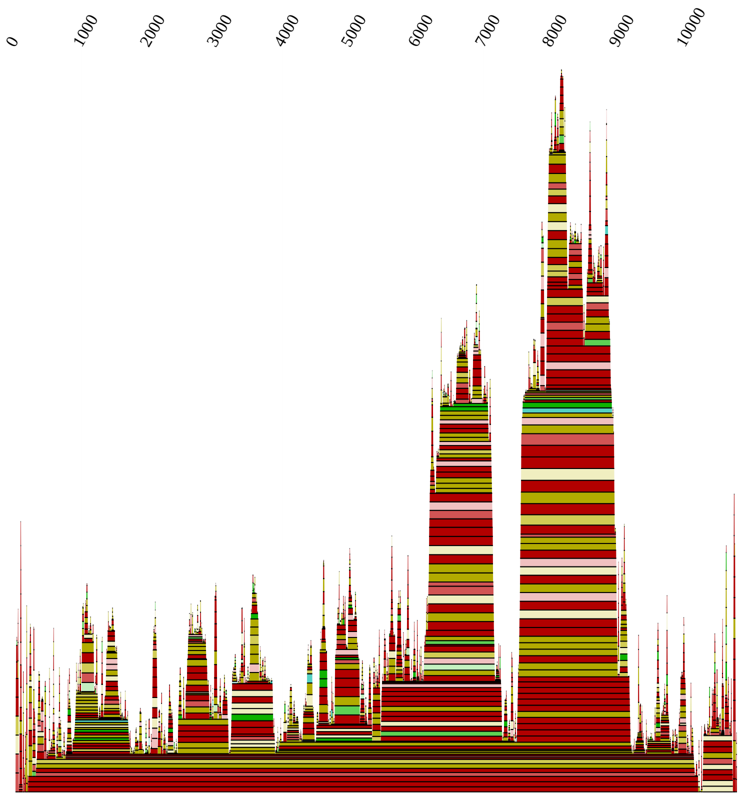
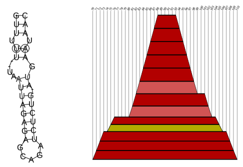
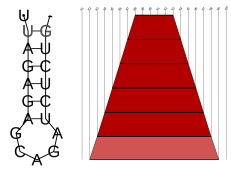


Figure 78: Mountain plot of dengue virus sequences aligned by `Ralign`.

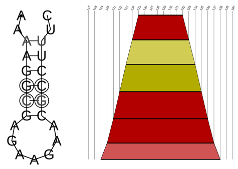
83. These figures show the secondary structure elements with the highest base pairing probabilities. This is further underlined by the occurring compensatory and consistent mutations (highlighted by the color code) in these regions.



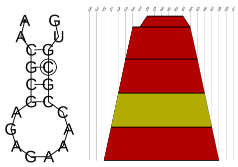
79.1 71-110



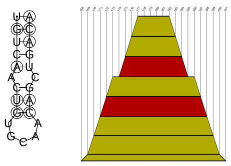
79.2 82-100



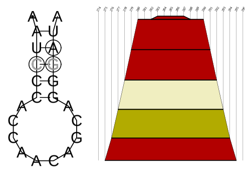
79.3 118-140



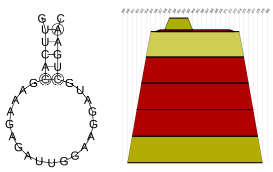
79.4 **151-170**



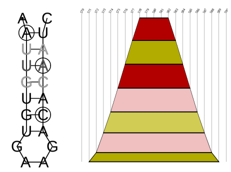
79.5 169-191



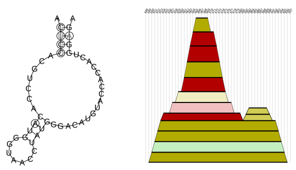
79.6 **275-296**



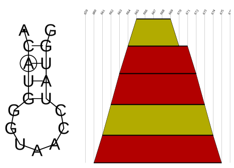
79.7 350-380



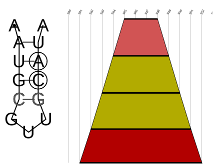
79.8 571-590



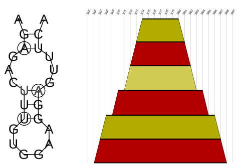
79.9 649-697



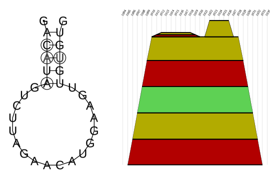
79.10 **660-676**



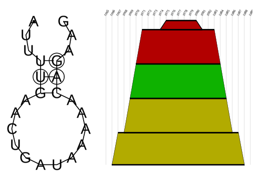
79.11 941-953



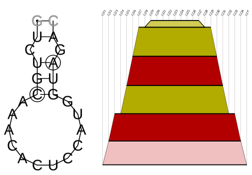
79.12 966-989



79.13 1005-1034



79.14 1066-1089



79.15 1212-1235

Figure 79: Detected conserved secondary structures of dengue virus. The numbers denote the base pair range in the `Ralign` alignment. Structural elements which were not found in the DEN-2/DEN-4 alignment are given in bold. (partI)

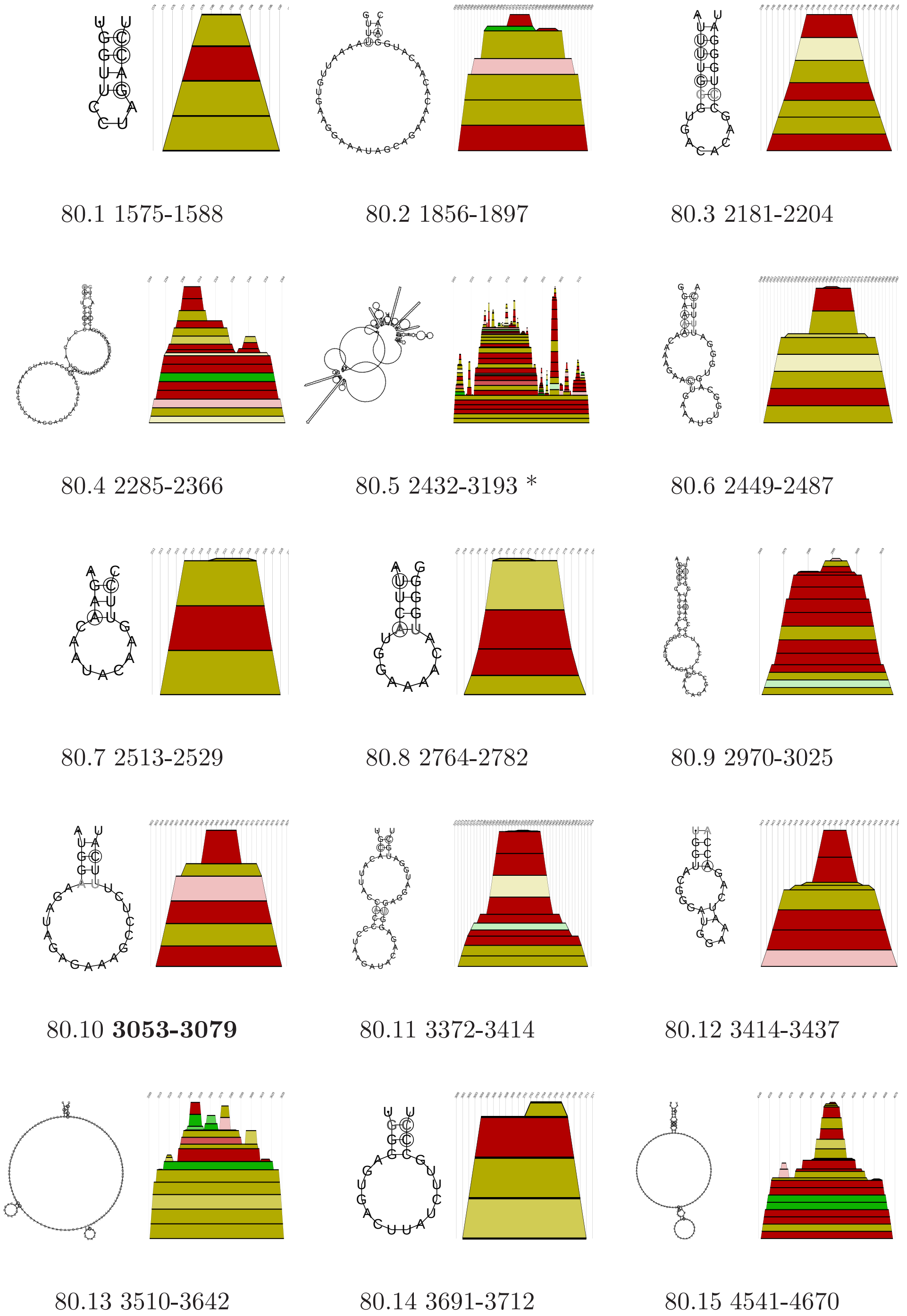


Figure 80: Detected conserved secondary structures of dengue virus. The numbers denote the base pair range in the `Ralign` alignment. The structural element which was not found in the DEN-2/DEN-4 alignment is given in bold. The structure labeled by (*) has long range interactions, see also Table 4. (partII)

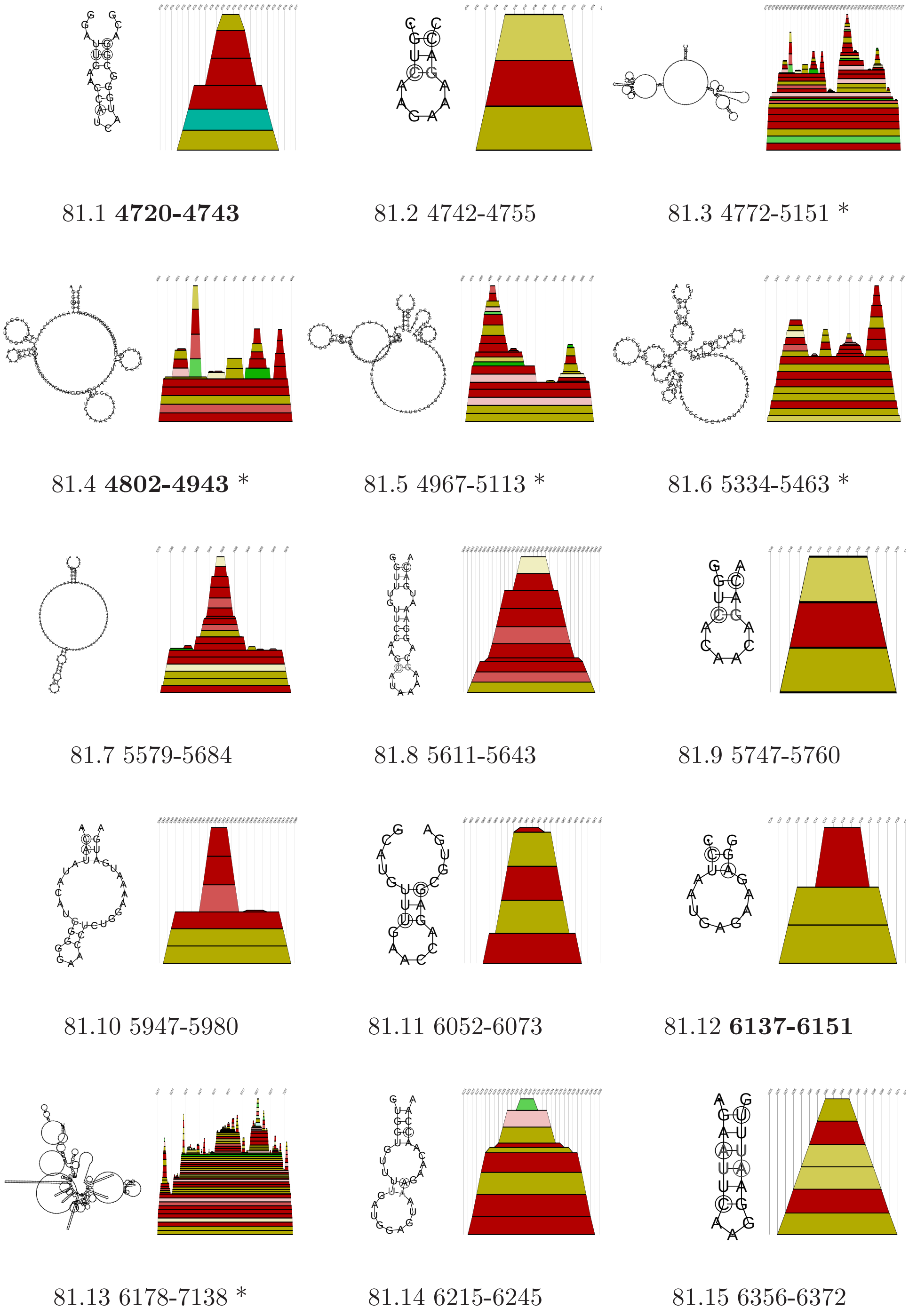
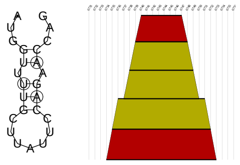
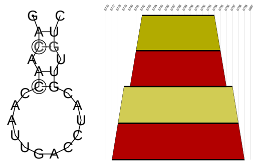


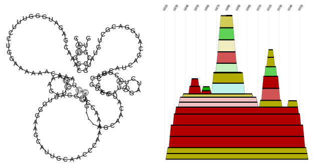
Figure 81: Detected conserved secondary structures of dengue virus. The numbers denote the base pair range in the `Ralign` alignment. Structural elements which were not found in the DEN-2/DEN-4 alignment are given in bold. Structures labeled by (*) have long range interactions, see also Table 4. (partIII)



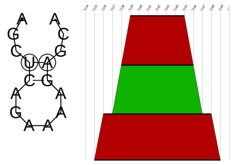
82.1 6732-6756



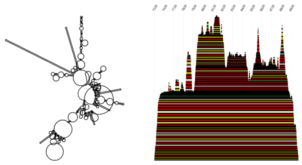
82.2 **6777-6800**



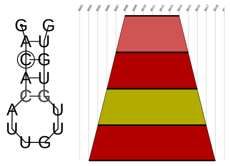
82.3 6830-6968 *



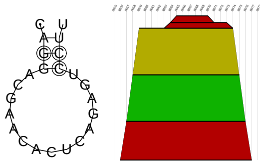
82.4 7235-7250



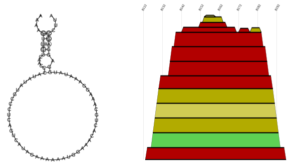
82.5 7521-8997 *



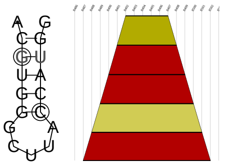
82.6 8004-8019



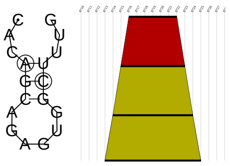
82.7 8056-8078



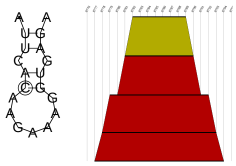
82.8 8123-8198



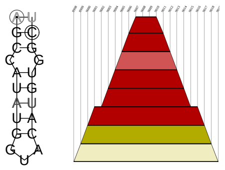
82.9 8487-8503



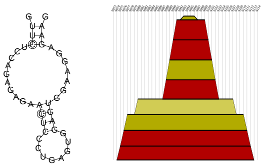
82.10 **8711-8728**



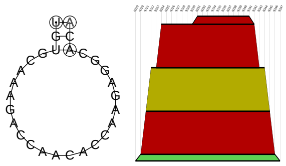
82.11 8777-8795



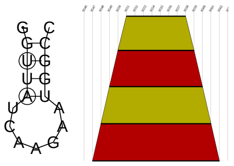
82.12 **8999-9019**



82.13 9074-9114

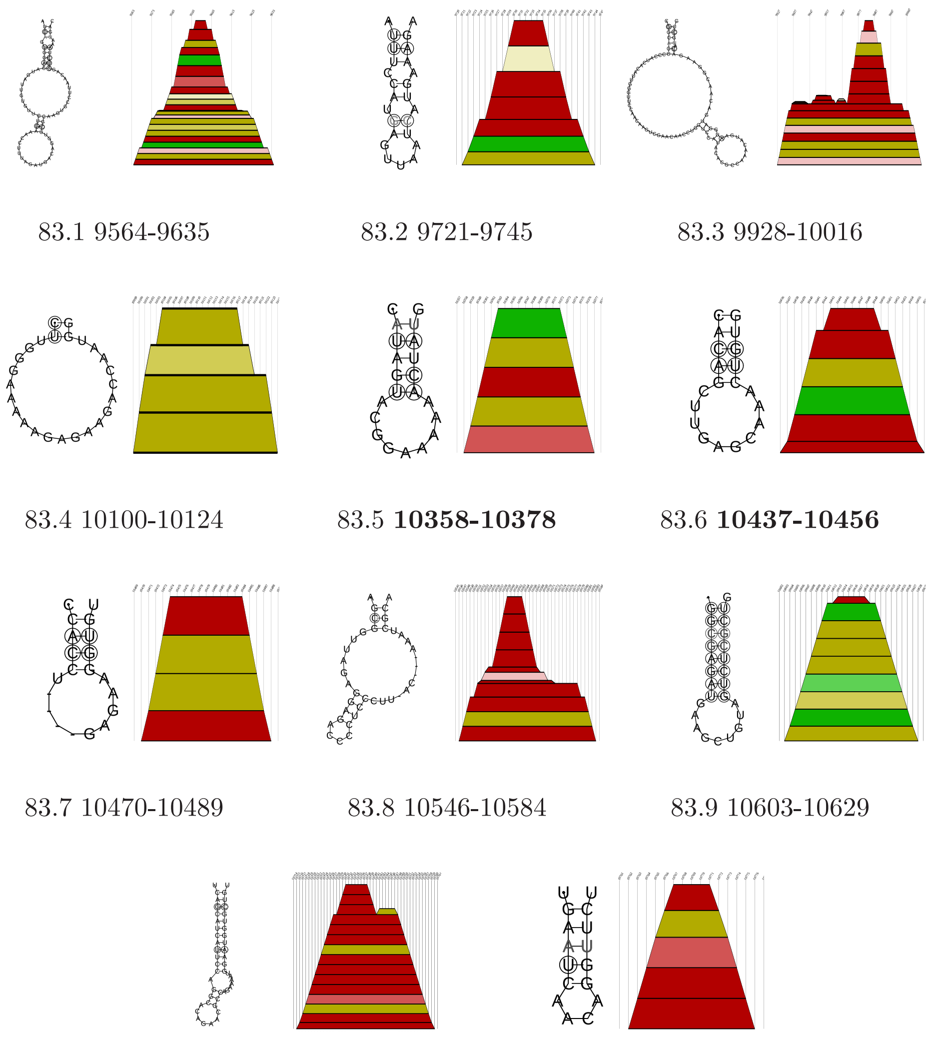


82.14 9320-9347



82.15 9547-9563

Figure 82: Detected conserved secondary structures of dengue virus. The numbers denote the base pair range in the `Ralign` alignment. Structural elements which were not found in the DEN-2/DEN-4 alignment are given in bold. Structures labeled by (*) have long range interactions, see also Table 4. (partIV)



83.1 9564-9635 | 83.2 9721-9745 | 83.3 9928-10016

83.4 10100-10124 | 83.5 **10358-10378** | 83.6 **10437-10456**

83.7 10470-10489 | 83.8 10546-10584 | 83.9 10603-10629

83.10 10715-10761 | 83.11 10762-10777

Figure 83: Detected conserved secondary structures of dengue virus. The numbers denote the base pair range in the `Ralign` alignment. Structural elements which were not found in the DEN-2/DEN-4 alignment are given in bold. (partV)

Dengue-2 and Dengue-4 Virus

Dengue virus types are very divergent. The largest group (sequenced genomes) is dengue-2. Five complete sequences were found and aligned, which gives an alignment length of 10726 and a mean pairwise homology of 95,2%. In order to reduce the mean pairwise homology and to make the results more meaningful, we added the dengue-4 sequence. Aligning dengue-2 together with dengue-4 sequences gives almost twice the number of convincing secondary structure elements compared to aligning all nine dengue sequences. All six sequences were aligned by `Ralign`. Alignment length is 10750 bases, the number of conserved bases is 6795 and the mean pairwise homology is 85.7%.

Figure 84 shows the split decomposition for the dengue-2 and dengue-4 sequences that we used for this experiment. We can see that dengue-2 and dengue-4 are not closely related.

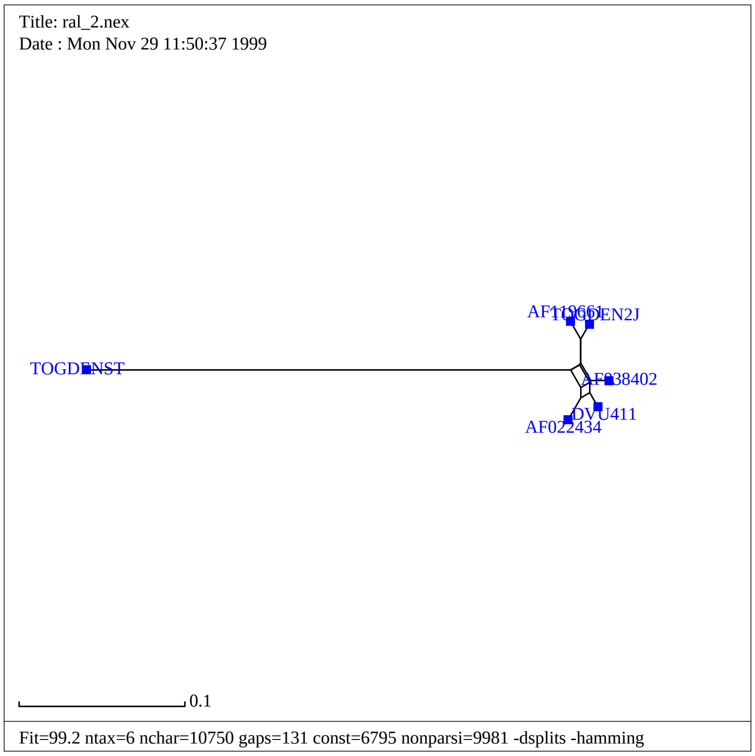


Figure 84: `Splitstree` plot of the aligned dengue-2 and dengue-4 sequences.

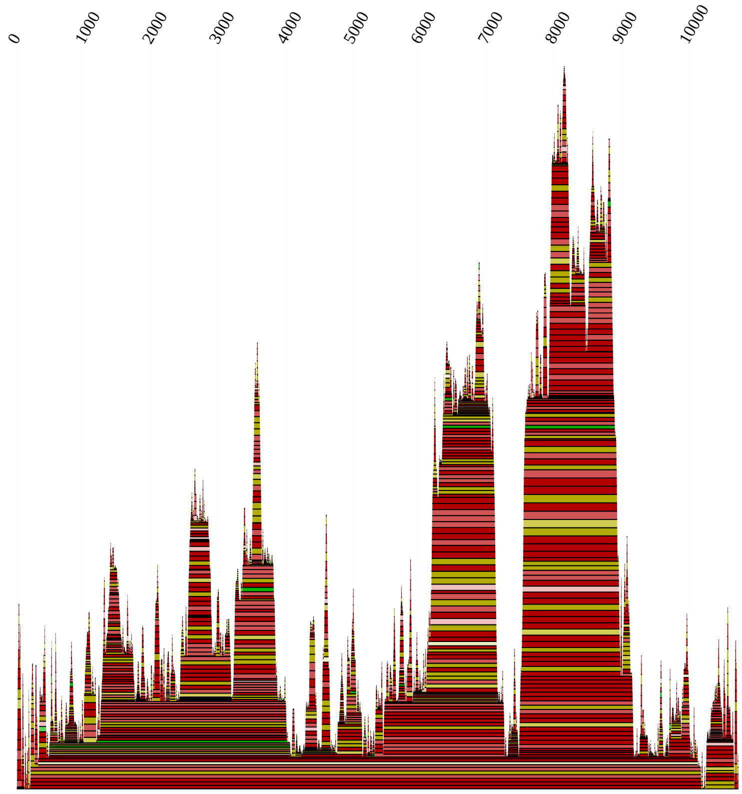
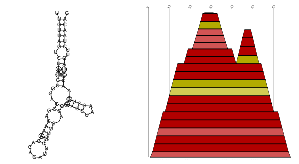
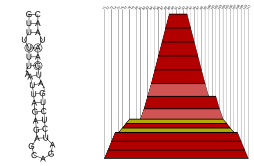


Figure 85: Mountain plot of dengue-2 and dengue-4 sequences.

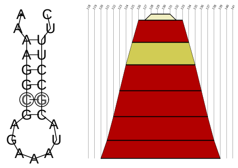
In Figure 85 we see the alignment of dengue-2 and dengue-4 strains calculated with the `Ralign` algorithm. The employed sequences (see Table 12) were found in the `GenBank`. A more detailed analysis of the results can be found in Figures 86 to 94. These figures show the secondary structure elements with the highest base pairing probabilities. This is further underlined by the occurring compensatory and consistent mutations (highlighted by the color code) in these regions.



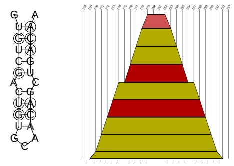
86.1 **4-72**



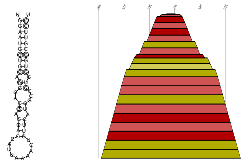
86.2 72-111



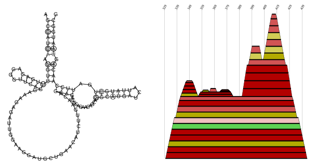
86.3 119-141



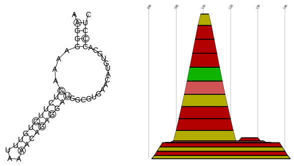
86.4 169-193



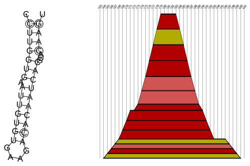
86.5 **207-263**



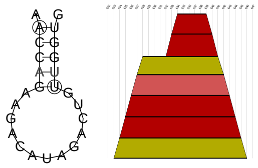
86.6 **330-444**



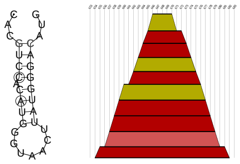
86.7 **500-552**



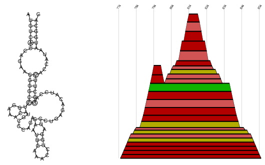
86.8 564-601



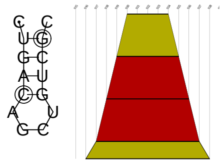
86.9 **623-647**



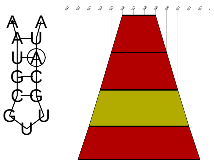
86.10 654-683



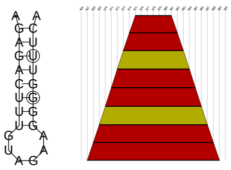
86.11 **775-856**



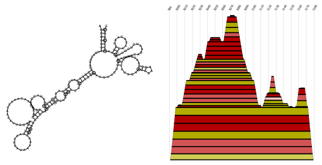
86.12 **916-929**



86.13 942-954

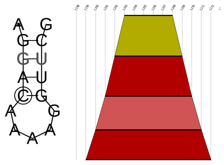


86.14 967-990

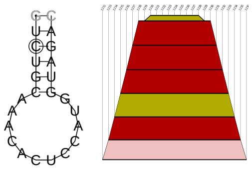


86.15 **997-1186**

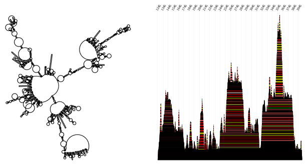
Figure 86: Detected conserved secondary structures of dengue-2 and dengue-4 virus. Structural elements which were not found in all dengue sequences are given in bold. The numbers denote the base pair range in the `Ralign` alignment. (partI)



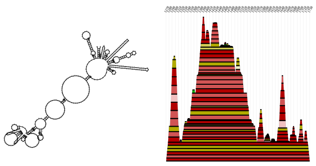
87.1 **1199-1213**



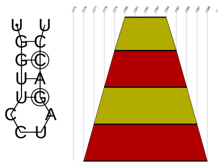
87.2 1213-1236



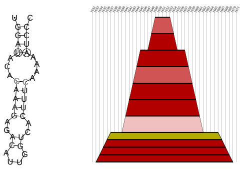
87.3 **1246-4017** *



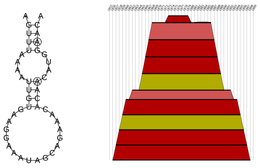
87.4 **1279-1739** *



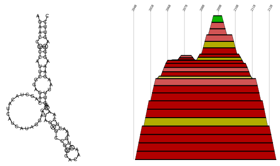
87.5 1576-1589



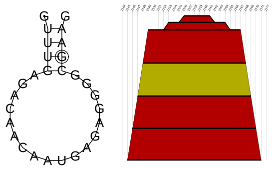
87.6 **1633-1671**



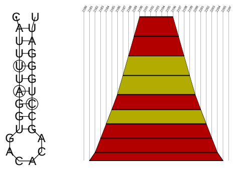
87.7 1856-1899



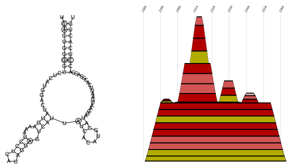
87.8 **2049-2133**



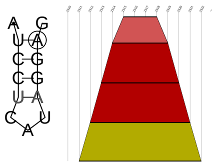
87.9 **2145-2172**



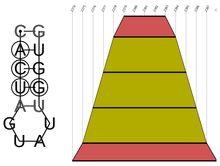
87.10 2181-2206



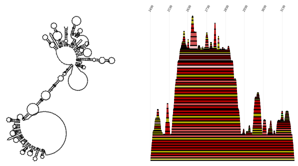
87.11 2285-2368



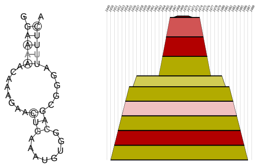
87.12 **2311-2323**



87.13 2375-2388



87.14 2431-3196 *



87.15 2450-2488

Figure 87: Detected conserved secondary structures of dengue-2 and dengue-4 virus. The numbers denote the base pair range in the `Ralign` alignment. Structural elements which were not found in all dengue sequences are given in bold. Structures labeled by (*) have long range interactions, see also Table 4. (partII)

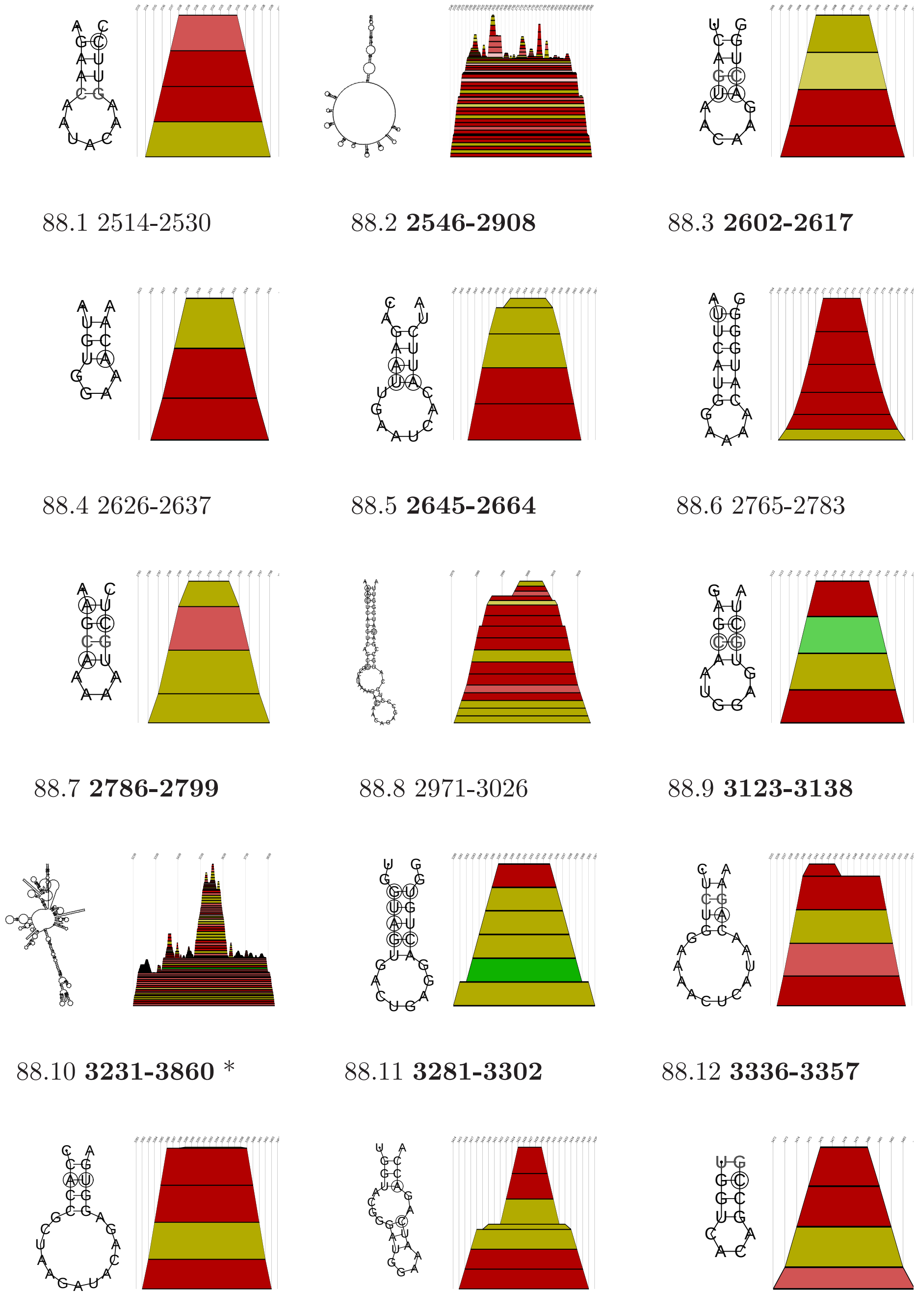


Figure 88: Detected conserved secondary structures of dengue-2 and dengue-4 virus. The numbers denote the base pair range in the `Ralign` alignment. Structural elements which were not found in all dengue sequences are given in bold. The structure labeled by (*) has long range interactions, see also Table 4. (partIII)

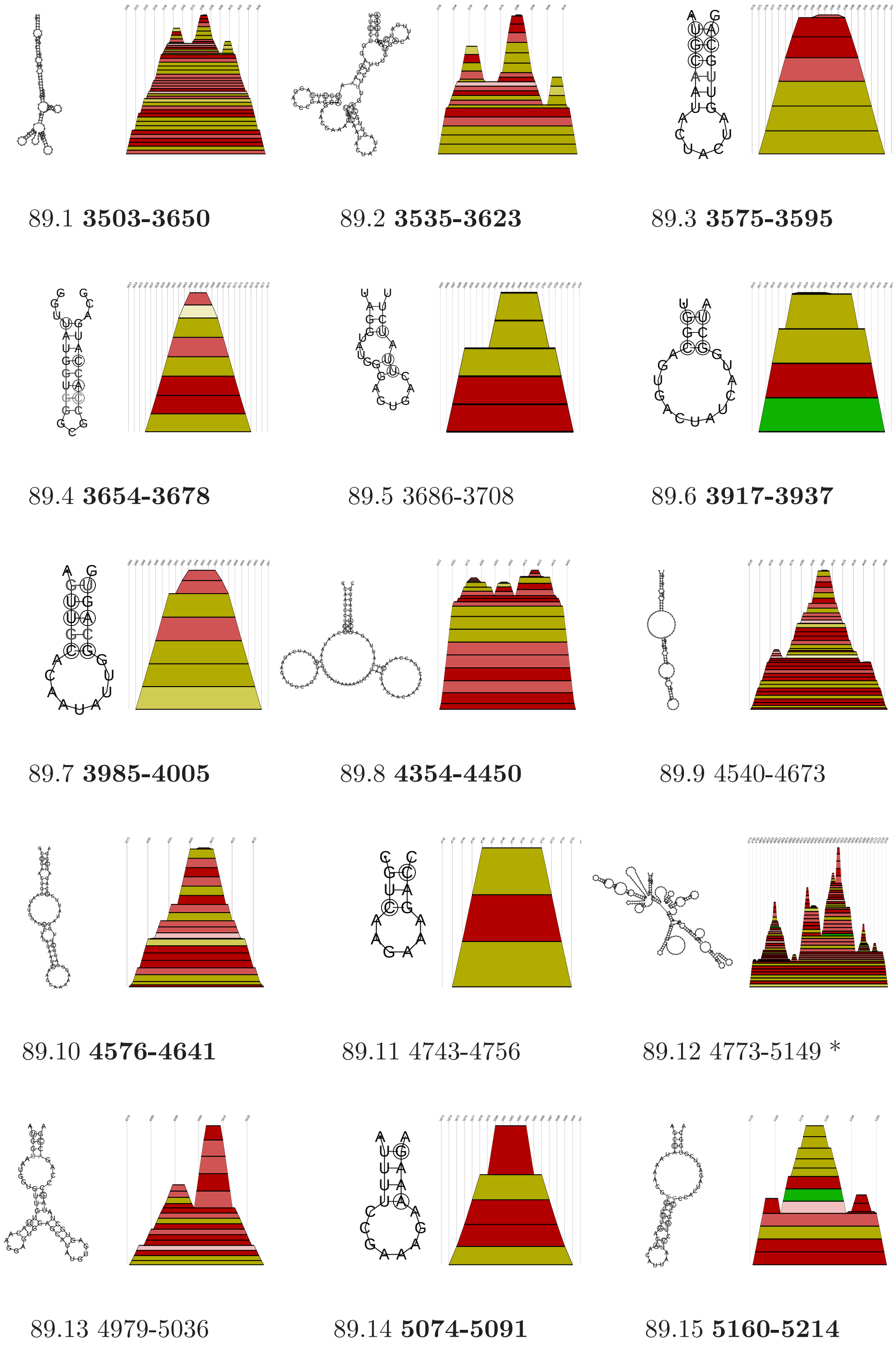
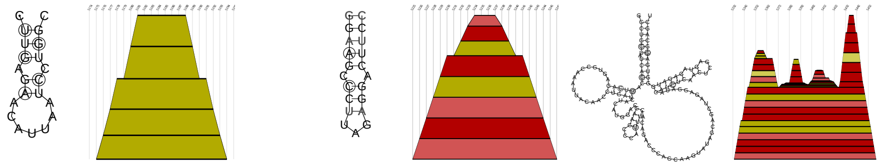
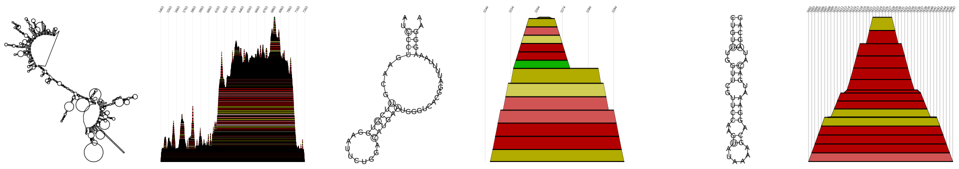


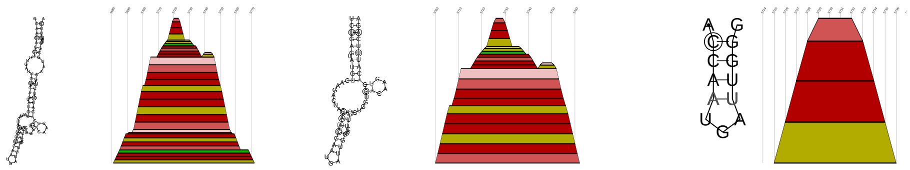
Figure 89: Detected conserved secondary structures of dengue-2 and dengue-4 virus. The numbers denote the base pair range in the `Ralign` alignment. Structural elements which were not found in all dengue sequences are given in bold. The structure labeled by (*) has long range interactions, see also Table 4. (partIV)



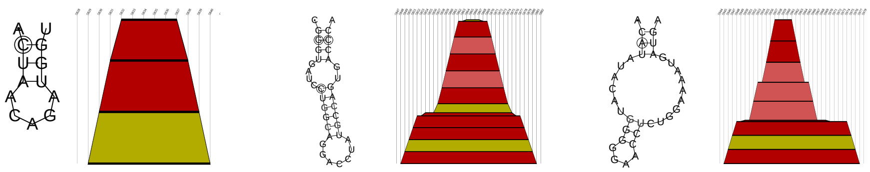
90.1 **5175-5195** 90.2 **5226-5246** 90.3 5333-5460



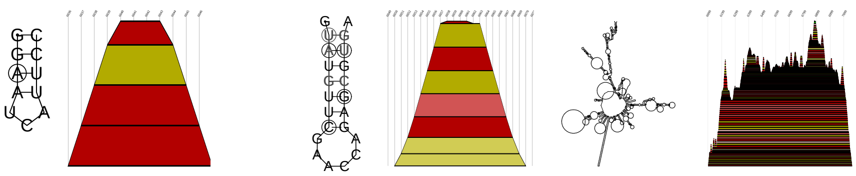
90.4 **5464-7265** * 90.5 5545-5600 90.6 5603-5647



90.7 **5690-5782** 90.8 5704-5765 90.9 **5725-5737**

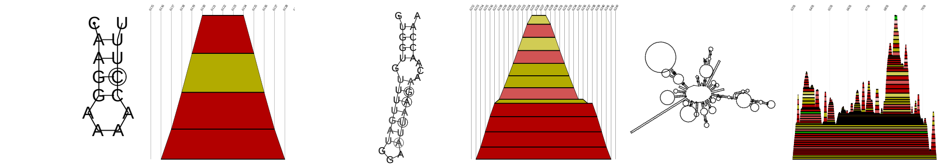


90.10 **5829-5841** 90.11 **5848-5882** 90.12 5945-5978

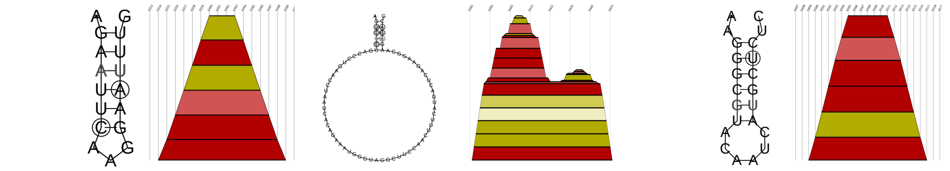


90.13 **6037-6047** 90.14 6050-6071 90.15 **6100-7161** *

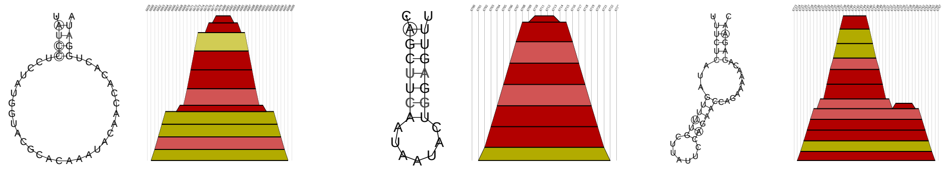
Figure 90: Detected conserved secondary structures of dengue-2 and dengue-4 virus. The numbers denote the base pair range in the `Ralign` alignment. Structural elements which were not found in all dengue sequences are given in bold. Structures labeled by (*) have long range interactions, see also Table 4. (partV)



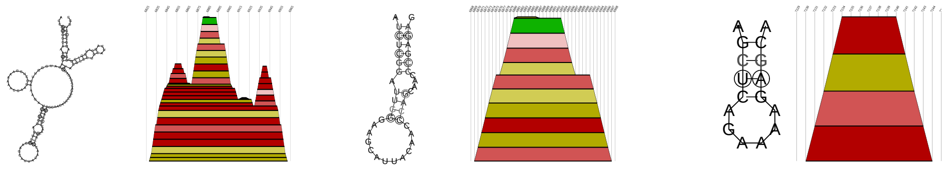
91.1 **6116-6129** 91.2 6213-6243 91.3 6332-7106 *



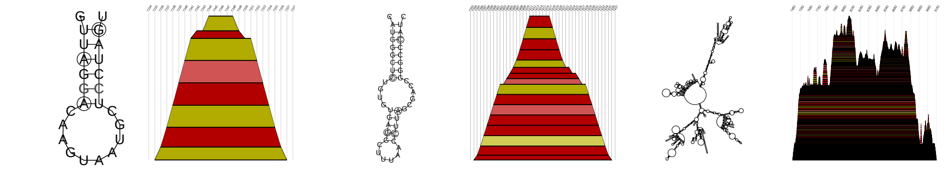
91.4 6354-6370 91.5 **6383-6454** 91.6 **6498-6519**



91.7 **6660-6699** 91.8 **6701-6723** 91.9 6724-6767

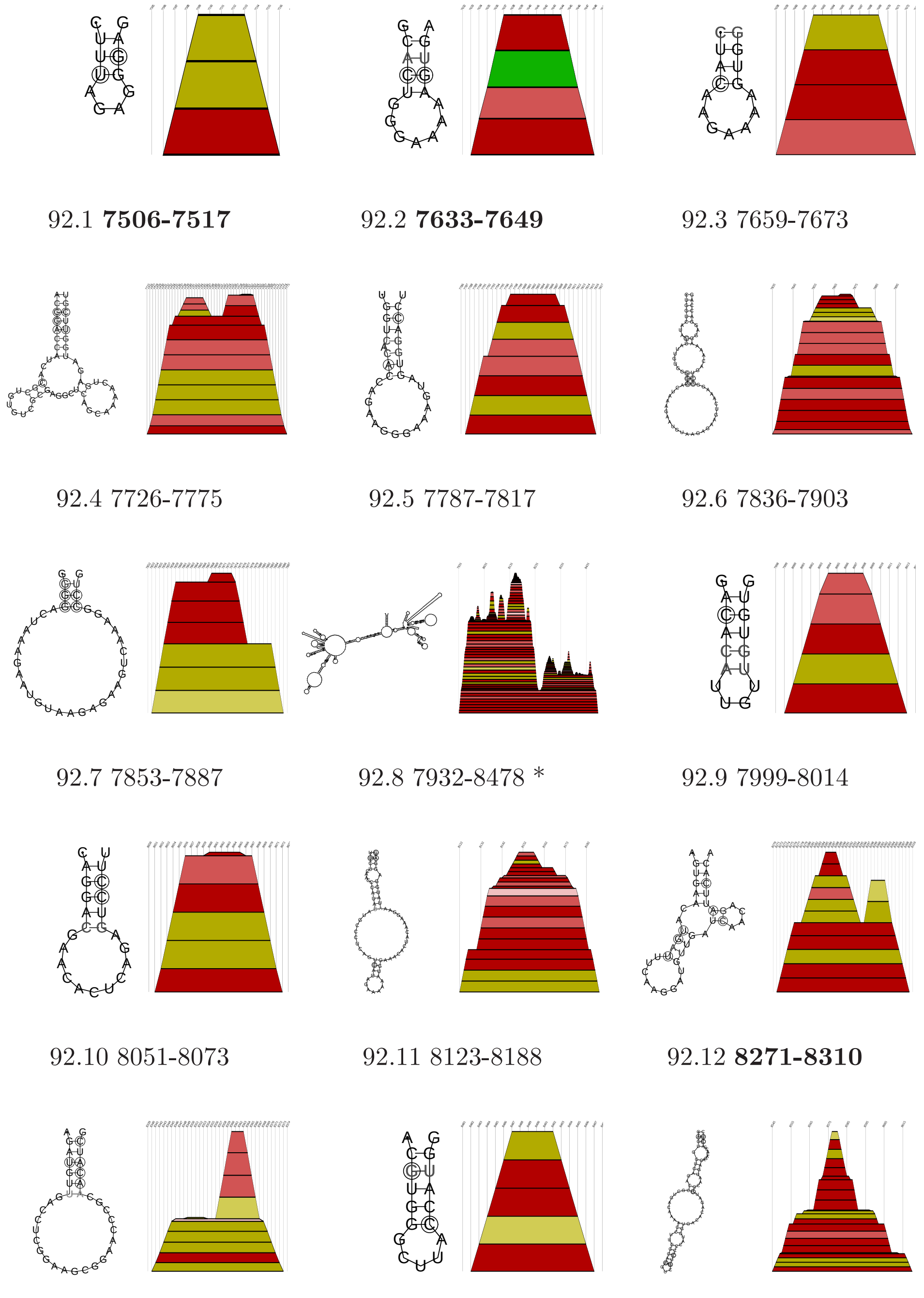


91.10 6826-6965 91.11 6869-6908 91.12 7230-7245



91.13 **7335-7358** 91.14 **7394-7437** 91.15 **7483-9188** *

Figure 91: Detected conserved secondary structures of dengue-2 and dengue-4 virus. The numbers denote the base pair range in the `Ralign` alignment. Structural elements which were not found in all dengue sequences are given in bold. Structures labeled by (*) have long range interactions, see also Table 4. (partVI)



92.1 **7506-7517** 92.2 **7633-7649** 92.3 7659-7673

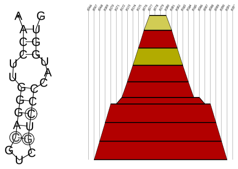
92.4 7726-7775 92.5 7787-7817 92.6 7836-7903

92.7 7853-7887 92.8 7932-8478 * 92.9 7999-8014

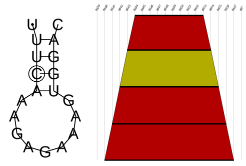
92.10 8051-8073 92.11 8123-8188 92.12 **8271-8310**

92.13 **8340-8374** 92.14 8482-8498 92.15 **8546-8620**

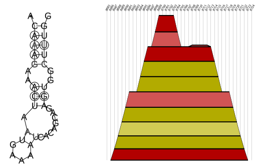
Figure 92: Detected conserved secondary structures of dengue-2 and dengue-4 virus. The numbers denote the base pair range in the `Ralign` alignment. Structural elements which were not found in all dengue sequences are given in bold. Structures labeled by (*) have long range interactions, see also Table 4. (partVII)



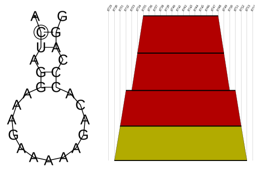
93.1 **8567-8592**



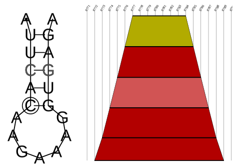
93.2 **8640-8658**



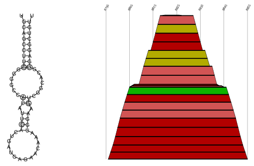
93.3 **8686-8724**



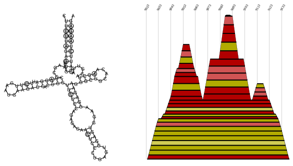
93.4 **8730-8754**



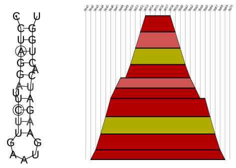
93.5 8772-8790



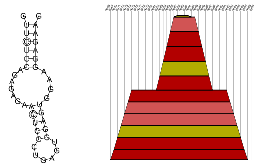
93.6 **8796-8856**



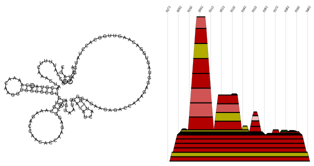
93.7 **9024-9140**



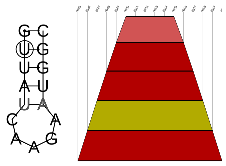
93.8 **9042-9070**



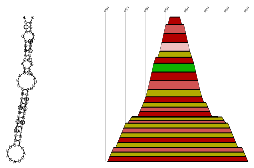
93.9 9069-9109



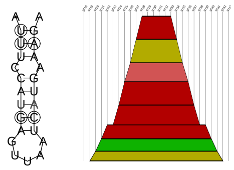
93.10 9273-9406 *



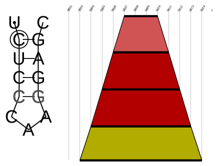
93.11 9546-9560



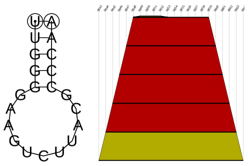
93.12 9562-9633



93.13 9719-9743

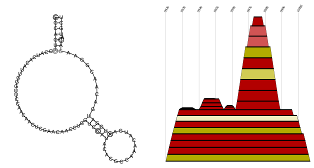


93.14 **9803-9815**

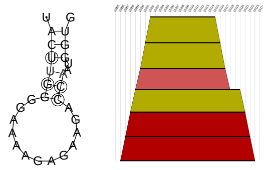


93.15 **9844-9864**

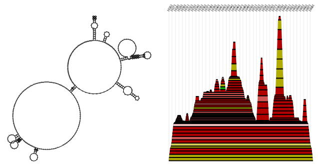
Figure 93: Detected conserved secondary structures of dengue-2 and dengue-4 virus. The numbers denote the base pair range in the `Ralign` alignment. Structural elements which were not found in all dengue sequences are given in bold. The structure labeled by (*) has long range interactions, see also Table 4. (partVIII)



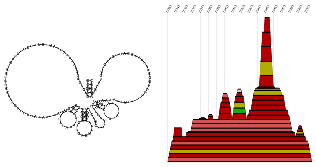
94.1 9927-10013



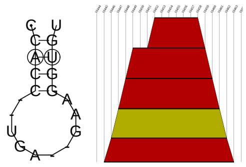
94.2 10096-10124



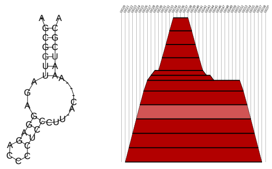
94.3 **10262-10688** *



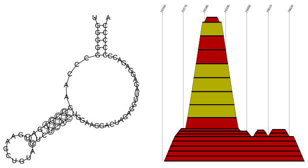
94.4 **10334-10510** *



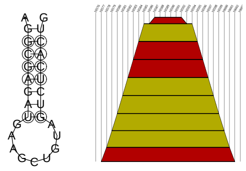
94.5 10445-10464



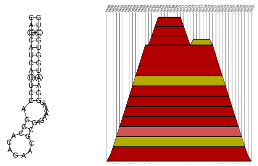
94.6 10521-10559



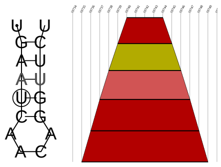
94.7 **10567-10634**



94.8 10577-10603



94.9 10689-10733



94.10 10735-10750

Figure 94: Detected conserved secondary structures of dengue-2 and dengue-4 virus. The numbers denote the base pair range in the `Ralign` alignment. Structural elements which were not found in all dengue sequences are given in bold. Structures labeled by (*) have long range interactions, see also Table 4. (partIX)

7.2.2 Discussion

The alignment of all dengue strains produced a large number of highly probable secondary structure elements. We found many hairpin loops with relatively short stacks, but the fact that they occurred in all sequences gives credence to their existence. Furthermore, we found a number of multi-loops e.g. see Figure 81.5, 81.6, and 82.3, some with special long range interactions e.g. see Figure 80.5, 81.13, and 82.5, all of them occurring with high probability, see Table 4. We conclude that although the different types of dengue virus are not closely related (see Figure 77), a large number of secondary structures are conserved throughout the serocomplex. Especially structures in the 5' and 3' region seem to be highly conserved see Figure 79.1, 83.9, and 83.10, leading to the conclusion that these regions are crucial for preserved viral functions.

A comparison of only two types, namely dengue-2 and dengue-4, reveals a number of other interesting details. Apart from the secondary structures we had already found in the general alignment of all strains (region nt 5,500 to 10,000), the mountain plot (Figure 85) shows a new multi-loop element with long range interactions (region nt 1,200 to 4,000), which were not found in Figure 78. We found almost double the number of secondary structures in this alignment, see Figures 86 to 94. However, some of these elements were also present in the general alignment. Those only found in this alignment are emphasized by bold numbers in the above mentioned figures. We found an interesting structure element in the nt 3503 to 3650 region see Figure 89.1 which contains many consistent and compensatory mutations. It was not present in the general alignment, where we found a very unconvincing multi-loop with very short stacks (Figure 80.13). The structures that were particular to the dengue-2/dengue-4 alignment also contained many multi-loops, some with long range interactions Figure 87.3, 88.10, and 91.15. Elements with long range interactions, see Table 4 can be found by folding the sequence in its entirety.

Table 4: List of secondary structure elements with long range interactions in dengue virus sequences.

all dengue viruses		DEN-2/DEN-4	
Figure	Position (nt)	Figure	Position (nt)
		87.3 87.4	1246-4017 1279-1739
80.5	2432-3193	87.14	2431-3196
		88.10	3231-3860
81.3	4772-5151	89.12	4773-5149
81.4 81.5 81.6	4802-4943 4967-5113 5334-5463		
		90.4	5464-7265
81.13	6178-7138	90.15	6100-7161
82.3	6830-6968	91.3	6332-7106
82.5	7521-8997	91.15	7483-9188
		92.8 93.10 94.3 94.4	7932-8478 9273-9406 10262-10688 10334-10510

7.2.3 The 3'NCR of Flaviviruses Folds as Distinct Unit

In our paper in 1997 [111] and in my master's thesis [112] we focussed our attention on the 3'NCR of flaviviruses. Short conserved primary sequence motifs were identified in the 3'NCRs of mosquito-borne flavivirus genomes [49], but these were found to be absent in tick-borne flaviviruses. Sequence analysis of a number of TBE virus strains revealed a surprising heterogeneity in the length of the 3'NCRs even among closely related strains [138]. A secondary structure was proposed for the 3' terminal 106 nucleotides of this core element, which is also found in the sequence of Powassan virus [88]. Very similar structures were reported for the sequences of mosquito-borne flaviviruses [46, 145] 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 [16].

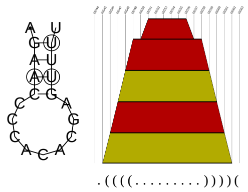
The folding of long RNA molecules as a single piece as opposed to the folding of short subsequences allows us to observe the effects inherent in the non-locality of RNA folding. It is well known that the fold of an RNA subsequence depends strongly on its size and exact location. This is because subsequences fold independently of the rest of the sequence only if they form isolated components 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, to fold the sequence in its entirety. By folding the complete sequences we checked whether the 3'NCR folds as a distinct unit, i.e., that the terminal nucleotides form the same structure irrespective of long range interaction.

Yellow Fever Virus

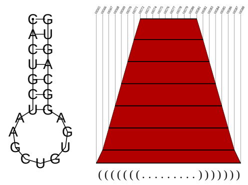
Now that we were able to fold the complete genomes we compared the conserved secondary structures in the 3'NCR folded as subsequences and of the entirely folded sequences. In Figure 95.8 we see the secondary structure elements that were found by folding only the 3'NCR region as described in our paper [111]. Folding of the entire sequences revealed the same elements see Figure 95.1 to 95.7. Obviously long range interaction have no influence on the folding of the 3'NCR of the yellow fever viruses.

Table 5: List of found structure elements in yellow fever virus sequences.

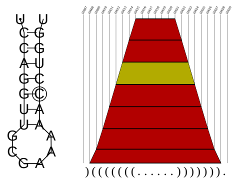
position in the 3'NCR (Fig. 95.8)	position in resent calculation	Figure
10543-10563	10545-10563	95.1
10565-10587	10566-10588	95.2
10608-10627	10608-10629	95.3
10629-10651	10630-10652	95.4
10662-10672	10663-10673	95.5
10673-10697	10673-10699	95.6
10789-10861	10791-10849	95.7



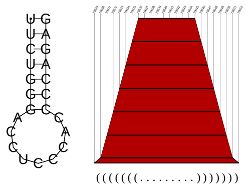
95.1 10545-10563



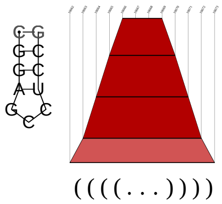
95.2 10566-10588



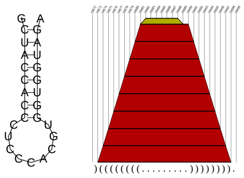
95.3 10608-10629



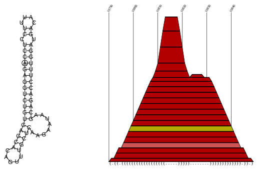
95.4 10630-10652



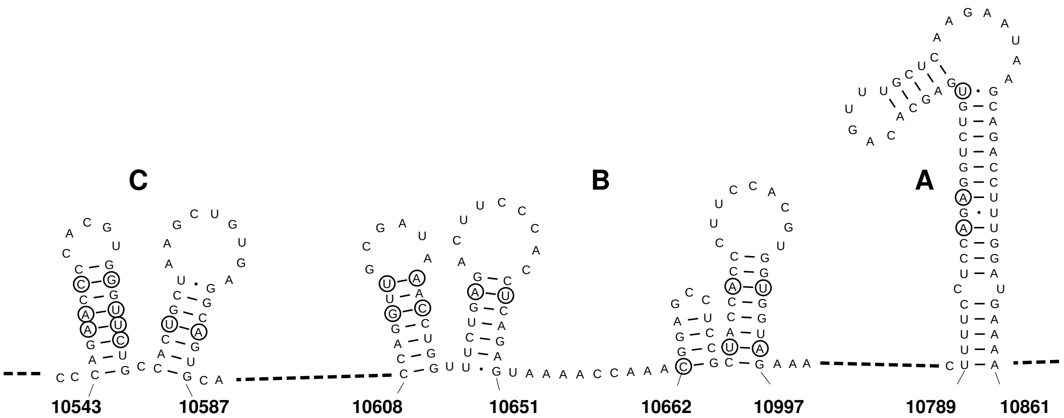
95.5 10663-10673



95.6 10673-10699



95.7 10791-10849



95.8 YF virus 3'NCR

Figure 95: Detected secondary structures of 3'NCR in yellow fever virus genomes. Figure 95.8 shows the published structural elements in the 3'NCR of dengue viruses [111]. In Figures 95.1 to 95.7 we see the recently found which correspond to those in Figure 95.8 (from left to right). The numbers in the upper part denote the base pair range of the `Ralign` alignment.

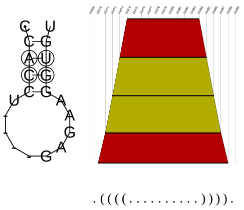
Dengue Virus

In Figure 96.8 we see the secondary structure elements that were found by folding only the 3'NCR region as described in our paper [111]. Folding of the entire sequences revealed most of the elements see Figure 96.1 to 96.7. However, we could not verify the closing stack of the CS2 element descibed in our paper. In Table 6 the first column shows the position of the secondary element by folding the 3'NCR as a distinct unit. The second gives the position after aligning the nine dengue sequences with `Ralign` and folding the complete sequences.

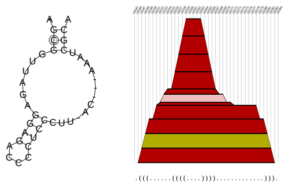
However, by aligning just the six dengue-2 and dengue-4 sequences an additional stack in element B can be found nt 10567 to 10634 (Figure 96.4) shifted downstream because of the different alignment length. This element contains several conserved base pairs with low probability to occur. Even in the dot plot of all dengue virus sequences there is evidence for this closing stack, see Figure 98. The mfe structure of the CS2 element when folding the entire sequence is almost identical to the published structure, but differs when folded with a gap as constraint, see Figure 97.

Table 6: List of found structure elements in dengue virus sequences.

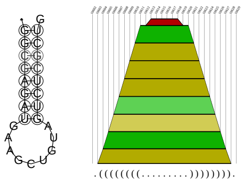
position in the 3'NCR (Fig. 96.8)	position in resent calculation	Figure
10429-10442	10470-10489	96.1
10499-10531	10546-10584	96.2
10540-10607	10603-10629	96.3
10617-10656	found in DEN-2/DEN-4	96.4
10631-10644	10685-10698	96.5
10662-10706	10715-10761	96.6
10709-10722	10762-10777	96.7



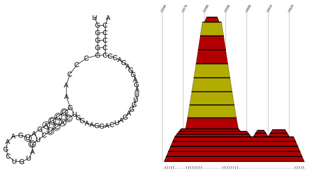
96.1 10470-10489



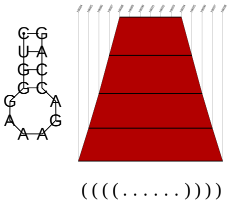
96.2 10546-10584



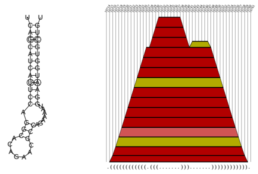
96.3 10603-10629



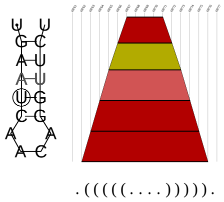
96.4 **10567-10634** in DEN-2/DEN-4



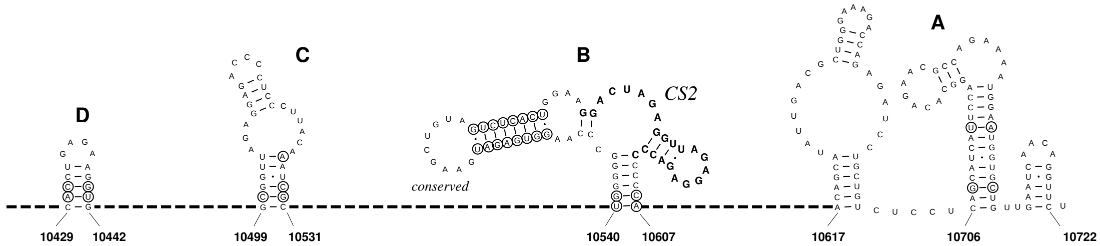
96.5 10685-10698



96.6 10715-10761

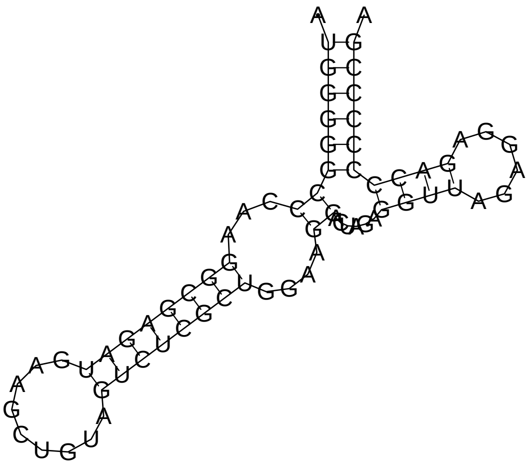


96.7 10762-10777

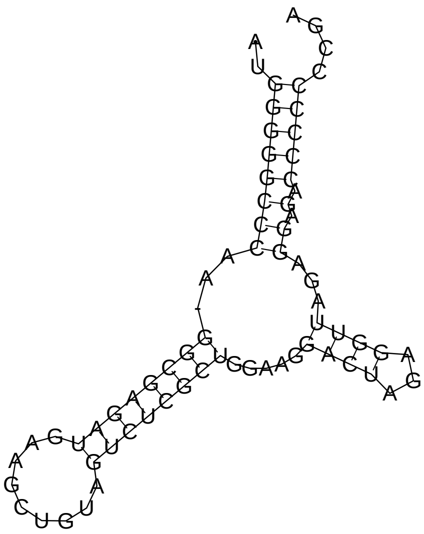


96.8 DEN 3'NCR

Figure 96: Detected secondary structures of 3'NCR in dengue virus genomes. We confirmed the closing stack of the CS2 element only in dengue-2/dengue-4 viruses. Figure 96.8 shows the published structural elements in the 3'NCR of dengue viruses [111]. In Figures 96.1 to 96.7 we see the recently found which correspond to those in Figure 96.8 (from left to right). The numbers in the upper part denote the base pair range of the `Ralign` alignment.



97.1 Mfe structure of the CS2 element without gap



97.2 Mfe structure of the CS2 element with gap

Figure 97: The mfe structure of the CS2 element found when folding the entire genome. The two alternative mfe structures can be obtained by folding the sequence either with a gap after aligning or without this gap.

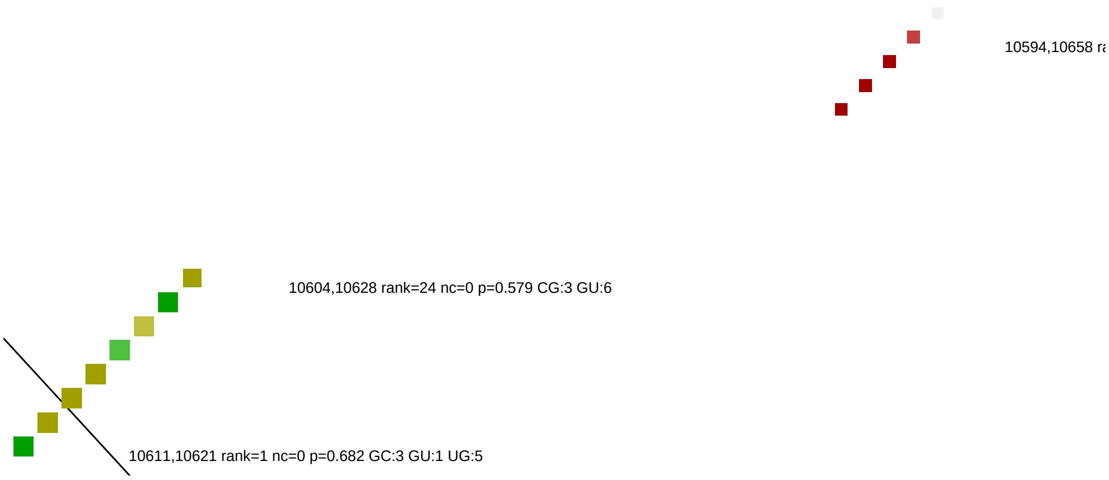


Figure 98: The dot plot of the CS2 element in dengue virus sequences shows that there is evidence for the closing stack (upper right red squares), consisting of only 3 conserved base pairs with low probability.

Japanese Encephalitis Virus

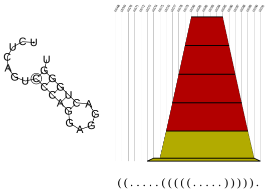
In Figure 99.4 we see the secondary structure elements that were found by folding only the 3'NCR region as described in our paper [111]. Folding of the entire sequences revealed the same elements see Figure 99.1 to 99.3. Obviously, long range interactions have no influence on the folding of the 3'NCR of the Japanese encephalitis viruses.

Table 7: List of found structure elements in Japanese encephalitis virus sequences.

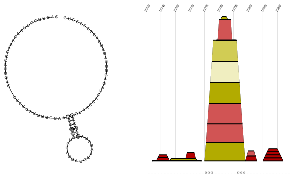
position in the 3'NCR (Fig. 99.4)	position in resent calculation	see Figure
10511-10531	10569-10591	99.1
10704-10774	10740-10838	99.2
10893-10976	10965-11049	99.3

7.2.4 Discussion

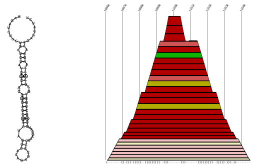
Our experiments verified the results of our previous work. On the whole we found the same secondary structures for all 3'NCR regions. Even the closing stack of the CS2 element of the dengue virus sequences occurs in all strains but with low probability. It seems that the 3'NCR region of the flaviviruses forms a distinct unit which is not influenced by long range interactions.



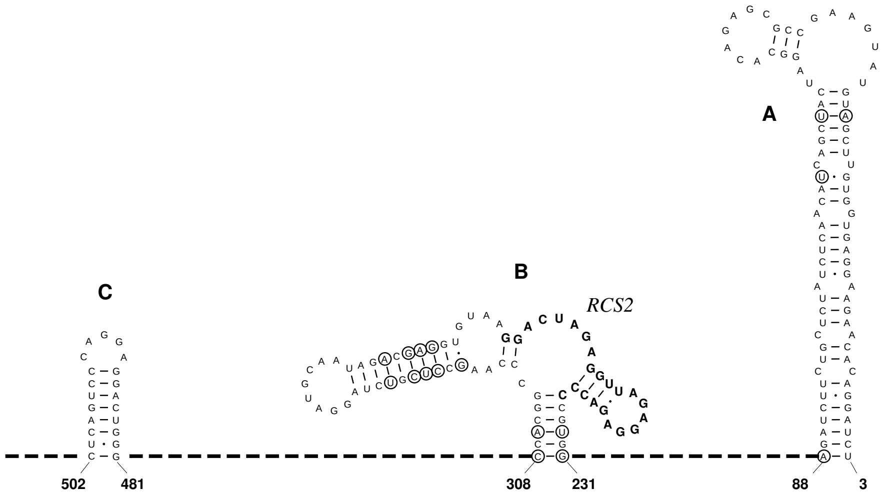
99.1 10569-10591



99.2 10740-10838



99.3 10965-11049



99.4 JE 3'NCR

Figure 99: Detected secondary structures of 3'NCR in Japanese encephalitis virus genomes. Figure 99.4 shows the published structural elements in the 3'NCR of dengue viruses [111]. In Figures 99.1 to 99.3 we see the recently found secondary structures which correspond to those in Figure 99.4 (from left to right). The numbers in the upper part denote the base pair range of the `Ralign` alignment.

8 Conclusion and Outlook

The purpose of this work was to find preserved secondary structure elements in different families of RNA viruses. It can be assumed that the conservation of certain structural elements by evolution is linked to an important function in the viral life cycle. We set out to find such functionally important secondary structures in the virus families *Retroviridae*, *Picornaviridae*, and *Flaviviridae* (Section 5, 6, and 7) by utilizing the information contained in a reliable multiple sequence alignment (`Ralign` and `ClustalW`) of the complete virus sequences to extract conserved secondary structures from a pool of plausible structures generated by thermodynamic prediction. The efficiency of our parallel implementation of McCaskill's algorithm (`RNAfold`) allows us to routinely fold complete RNA virus genomes.

Our experiments yielded the following results:

In *Retroviridae* we were able to confirm that HIV-2 is more closely related to SIV than HIV-1, especially to the so-called subgroup 2 as described in literature [40, 19]. We found that they have more structural elements in common with each other than with HIV-1 and SIV_{sub1}. For example the TAR elements in HIV-2 and SIV_{sub2} are nearly identical, see Figure 57. We could not find a TAR element in SIV_{sub1}, this could be due to either the divergence of the sequences and/or to the fact that the sequences had to be cut prior to the analysis which might have been at the wrong site. The RRE element was found in all members of primate lentiviruses we examined. However, the RRE elements in HIV-1 and SIV_{sub1} show slight differences. Again this finding underlines the closer relationship between HIV-2 and SIV_{sub2}. We also found a number of other secondary structure elements common to the two members, see Figure 53 to 56.

In *Picornaviridae* we were able to find the IRES region described in literature [106, 40, 139] in rhinoviruses and enteroviruses. However, only parts of the IRES could be found when both members were aligned together. This is probably due to the high sequential diversity. We also confirmed the cloverleaf structure in rhinoviruses downstream of the IRES. The most promising structural elements conserved throughout the family were at the 5' end. We could not confirm the cre element in rhinoviruses discribed in literature [93]. It was interesting to see that we found much more conserved secondary structures when we removed the most divergent strain (strain 14) from the sample of rhinoviruses.

The only members of the *Flaviviridae* family we examined were the dengue viruses because these were the only ones with a sufficient amount of fully sequenced genomes. We found a substantial number of promising secondary structure elements throughout the entire genome. The alignment of dengue-2 and dengue-4 produced almost twice the number of structures than the alignment of all members together.

Now that we were able to fold the entire sequences of RNA viruses, we verfied the results of a previous experiment. As a further test of the method we examined the folding of the 3'NCR region in the flaviviruses (YF, DEN, and JE viruses) and found the same conserved structures as in an earlier analysis performed only with partial sequences [111]. This shows that we chose the correct sites for the cutting of the sequences in our earlier experiments and that the 3'NCR folds as a distinct unit, i.e., there are no base pairs to the outside of the sequence window.

One major problem of our experiments was the necessity to preselect the sequences from the data bases because they were sometimes subject to sequencing errors. In order to be able to compare the sequences, we had to process e.g. to truncate some of the sequences of HIV before the TAR hairpin. When comparing the sequences, both incomplete and too long sequences cause severe problems.

A further difficulty is a too diverse sample of sequences. Although the sequence alignment has been improved, our method is still crucially dependent on the quality of the data set. Alignment errors oftentimes significantly decrease the quality of our results. Which sequences to select remains a somewhat subjective decision. In general, a large data set with sequences at a homology from about 60 to 90 % would be ideal for our investigation. On the other hand, the sorting procedure in `pfrali` has been optimized for a rather small sample of sequences.

Still, from our finding we deduce that the method introduced here is capable of reliably predicting functional structural elements on the basis of sequence information alone. We showed that a routine investigation of viral RNA sequences is possible and applied these techniques to the virus families. In all cases we were able to find the structural elements that were previously described in literature plus a large number of additional promising candidates. We found some hairpins with relatively short stacks, but the fact that they occurred in all sequences gives credence to their existence. 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 [156, 155]. McCaskill's partition function approach [92], 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 are able to predict the structure and estimate the reliability of the prediction at the same time. Due to the differences and the more comprehensive nature of our approach it is difficult to compare our findings to those of other groups, but we believe that inclusion of more complete sequences and larger numbers of candidates into our experiments make our findings more substantial. It is impossible for us to specify the function of the conserved structural features with our methods. This would be a task for experimental groups who could use our data to perform highly directed analysis of viral genomes.

Our program currently does not support the detection of pseudoknots and non-standard base pairs, like **GA** or **UU**. While pseudoknots are important structural elements (e.g. kissing hairpin motifs) in many RNA molecules [146], they are excluded from many studies mostly for technical reasons [143]. Further, in some cases well-predicted stacked regions are interrupted by individual "holes" or show a single base pair with a few non-compatible sequences. While in many cases these features reflect structural variability or the existence of an internal loop, they can be attributed to non-standard base pairs that do not necessarily disrupt the helix [84] in other cases. In order to make our predictions even more meaningful pseudoknots and non-standard base pairs should be incorporated.

Another interesting goal would be to develop a method for comparing virus genomes which are unrelated according to their sequences. These diverse sequences can not be aligned, but may exhibit similar secondary structure elements, which are suspect to evolutionary relationship.

However, in this work we have proven that a comprehensive survey of conserved RNA secondary structures in viral genomes is feasible and that our methods are indeed capable of detecting a large number of previously unknown conserved structural elements. The functional importance of the secondary structures described in this thesis will have to be verified by direct biological testing. The im-mediate objective of our approach is to analyze all complete RNA virus genomes available, which leads to a data bank based on conserved RNA structure elements. This will provide benefits for the understanding of the relationship between sequence and structure and may help in tasks such as drug discovery, as well as in the study of molecular evolution. The results, in particular the "atlas" of conserved RNA structures for the virus families, provide a valuable basis for further investigations into viral evolution and phylogeny.

9 Appendix

Table 8: List of human immunodeficiency virus type 1 sequences

REM	ID	Accession No	Length/Bases	Organism	
1	HIVMALCG	X04415 K03456	9229	HIV-1	ADI-MAL
2			9365	HIV-1	AE-90CF402
3			9203	HIV-1	AE-CM240
4			9257	HIV-1	B-896
5			9176	HIV-1	B-ACH320A
6	HIV1SG3X	L02317	9168	HIV-1	B-BCSG3
7			9135	HIV-1	B-CAM1
8	HIVU43096	U43096	9074	HIV-1	B-D31
9			9255	HIV-1	B-HIV1AD8
10			9265	HIV-1	B-HXB2
11			9085	HIV-1	B-JRCSF
12	HIVBRUCG	K02013	9229	HIV-1	B-LAI
13			9255	HIV-1	B-MANC
14	HIVOYI	M26727	9190	HIV-1	B-OYI
15			9282	HIV-1	B-SF2
16			9265	HIV-1	B-WEAU
17			9255	HIV-1	B-YU2
18			9254	HIV-1	B-pNL43
19	HIVELICG	K03454 X04414	9176	HIV-1	D-ELI
20	A34828	A34828	9143	HIV-1	D-NDK
21			9265	HIV-1	O-ANT70
22			9331	HIV-1	O-MVP5180
23			9338	HIV-1	SIVCPZGAB

Table 9: List of human immunodeficiency virus type 2 sequences

REM	ID	Accession No	Length/Bases	Organism
1	HI2U22047	U22047	10172	HIV-2
2	HIU2U3829	U38293	10312	HIV-2
3	HIV2BEN	M30502	10359	HIV-2
4	HIV2UC1GN	L07625	10271	HIV-2
5	HIVTRENGE	X61240 X16109	10269	HIV-2
6	HI2U27200	U27200 L14545	10351	HIV-2
7	HIV2ST	M31113	9672	HIV-2
8	REHIV2IS	J04498	9636	HIV-2
9	HIV2MDS	Z48731	9525	HIV-2
10	HIV2RODX	X05291	9671	HIV-2
11	REHIV2NI	J03654	9431	HIV-2
12	HIV2D194	J04542	9472	HIV-2
13	HIV2GH1	M30895 D00477	9480	HIV-2

Table 10: List of simian immunodeficiency virus sequences

REM	ID	Accession No	Length/Bases	Organism
1	SIVAGM155	M29975	9794	SIV
2	SIVREV	L40990	9815	SIV
3	SIV677A	M58410	9623	SIV
4	SIU58991	U58991	9784	SIVtan
5	SI04005	U04005	10036	SIVagmSAB-1
6	SIVCOMGNM	L06042	9597	SIV
7	SIVMNDGB	M27470 X15781	9215	SIV
8	SIU72748	U72748	10289	SIVsmE543
9	SIVGAGAA	M80193	9675	SIV
10	SIVMM239	M33262	10535	SIV
11	AF038398	AF038398	10000	SIV
12	SIVMNE	M32741	9628	SIV
13	SIVSTM	M83293	9892	SIV

Table 11: List of Picornaviridae sequences

Rhinovirus				
REM	ID	Accession No	Length/Bases	Organism
1	HRV14 PI	K02121 X01087	7212	Human rhinovirus type 14
2	HRV85		7140	Human rhinovirus type 85
3	HRV89	M16248 A10937	7152	Human rhinovirus type 89
4	HRV9		7128	Human rhinovirus type 9
5	HRVACG	D00239	7133	Human rhinovirus type 1B
6	HRVPP	L24917	7124	Human rhinovirus type 16
7	HRV2 PIHRV2G	X02316	7102	Human rhinovirus type 2

Enterovirus				
REM	ID	Accession No	Length/Bases	Organism
1	AF083069	AF083069	7433	Echovirus 5
2	AF085363	AF085363	7411	Coxsackievirus B2
3	AF162711	AF162711	7440	Echovirus 30
4	AF176044	AF176044	7433	Enterovirus 71
5	AF177911	AF177911	7410	Coxsackievirus A16
6	BEVVG527	D00214	7414	Bovine enterovirus
7	CA05876	U05876	7413	Coxsackievirus A16
8	CV57056	U57056	7400	Coxsackievirus B3
9	CXA21	D00538	7401	Coxsackievirus A
10	CXA24CG	D90457	7461	Coxsackievirus A24
11	CXB1G	M16560	7389	Coxsackievirus B1
12	CXB3G	M16572	7396	Coxsackievirus B3
13	CXB5CGA	X67706	7402	Coxsackievirus B5
14	CXB9CG	D00627	7452	Coxsackievirus A9
15	E616283	U16283 U05851	7417	Echovirus 6
16	E722521	U22521	7408	Enterovirus 71
17	E722522	U22522	7411	Enterovirus 71
18	EC12TCGWT	X79047	7501	Echovirus 12
19	ECHOV9XX	X92886	7451	Echovirus 9
20	EV11VPCD	X80059	7438	Echovirus 11
21	EV70CG	D00820	7390	Enterovirus 70
22	EV9GENOME	X84981	7420	Echovirus 9
23	PEV9XX	Y14459	7351	Porcine enterovirus type 9
24	PI3L37	K01392	7431	Poliovirus P3
25	PICOXB4	X05690 D00149	7395	Coxsackievirus B4
26	PIPO3XX	X04468	7435	Poliovirus 3
27	PIPOLS2	X00595	7439	Poliovirus 2
28	POL2W2	D00625	7434	Human poliovirus 2
29	POLIOS1	V01150 J02282 J02285 J02286 V01133	7441	Poliovirus
30	S76772	S76772	7397	Coxsackievirus B4
31	SVDG	D00435	7401	Coxsackievirus B

Table 12: List of flavivirus sequences

Dengue virus				
REM	ID	Accession No	Length/Bases	Organism
1	DENT1SEQ	M87512	10717	Dengue virus type 1
2	DVU88535_1	U88536	10735	Dengue virus type 1
3	AF022434	AF022434	10724	Dengue virus type 2
4	AF038402	AF038402	10724	Dengue virus type 2
5	AF119661	AF119661	10723	Dengue virus type 2
6	DVU411	U87411	10723	Dengue virus type 2
7	TOGDEN2J	M20558	10723	Dengue virus type 2
8	DENCME	M93130	10696	Dengue virus type 3
9	TOGDENST	M14931	10644	Dengue virus type 4
Yellow fever virus				
REM	ID	Accession No	Length/Bases	Organism
1	AF094612	AF094612	10760	Yellow fever virus
2	YF21055	U21055	10862	Yellow fever virus
3	YF21056	U21056	10862	Yellow fever virus
4	YFU54798	U54798	10862	Yellow fever virus
5	YFVCG	K02749	10862	Yellow fever virus
Japanese encephalitis virus				
REM	ID	Accession No	Length/Bases	Organism
1	AF045551	AF045551	10963	Japanese encephalitis virus
2	AF075723	AF075723	10976	Japanese encephalitis virus
3	AF080251	AF080251	10977	Japanese encephalitis virus
4	AF098736	AF098736	10976	Japanese encephalitis virus
5	AF098737	AF098737	10976	Japanese encephalitis virus
6	JEBEICG	L48961	10976	Japanese encephalitis virus
7	JEU47032	U47032	10976	Japanese encephalitis virus
8	KUNCG	D00246	10664	Kunjin virus
9	TOGJEV01	D90195	10976	Japanese encephalitis virus
10	TOGWNFCG	M12294	10960	West Nile virus

References

[1] J. P. Abrahams, M. van den Berg, E. van Batenburg, and C. Pleij. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucl. Acids Res.*, 18:3035–3044, 1990.

[2] V. I. Agol, A. V. Paul, and E. Wimmer. Paradoxes of the replication of picornaviral genomes. *Vir. Res.*, 62:129–147, 1999.

[3] V. I. Agol. Recombination and other genomic rearrangements in picornaviruses. *Semin. Virol.*, 8:77–84, 1997.

[4] V. V. Anshelevich, A. V. Vologodskii, A. V. Lukashin, and M. D. Frank-Kamenetskii. Slow relaxational processes in the melting of linear biopolymers: A theory and its application to nucleic acids. *Biopolymers*, 23:39–58, 1984.

[5] G. Awang and D. Sen. Mode of dimerization of HIV-1 genomic RNA. *Biochemistry*, 32:11453–11457, 1993.

[6] H. J. Bandelt and A. W.M. Dress. A canonical decomposition theory for metrics on a finite set. *Advances in mathematics*, 92(1):47, 1992.

[7] A. R. Banerjee, J. A. Jaeger, and D. H. Turner. Thermal unfolding of a group I ribozyme: The low-temperature transition is primarily disruption of tertiary structure. *Biochemistry*, 32:153–163, 1993.

[8] G. J. Barton and M. J. E. Sternberg. A strategy for the rapid multiple alignment of protein sequences. confidence levels from tertiary structure comparisons. *J. Mol. Biol.*, 198:327–337, 1987.

[9] D. Bashford, C. Chothia, and A. M. Lesk. Determinants of a protein fold. unique features of the globin amino acid sequences. *J. Mol. Biol.*, 196:199–216, 1987.

[10] 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.

[11] R. Bellman. On the theory of dynamic programming. *Proc. Natl. Acad. Sci. USA*, 38:716–719, 1952.

[12] B. Berkhout and J. L. B. Van Wamel. The leader of the HIV-1 RNA genome forms a compactly folded tertiary structure. *RNA*, 6:282–295, 2000.

[13] B. Berkhout, A. T. Das, and J. L. B. vanWamel. The native structure of the human immunodeficiency virus type 1 RNA genome is required for the first strand transfer of reverse transcription. *Virol.*, 249(2):211–218, 1998.

[14] B. Berkhout and J. L. B. vanWamel. Role of the DIS hairpin in replication of human immunodeficiency virus type 1. *J. Virol.*, 70(10):6723–6732, 1996.

[15] B. Berkhout. Structural features in TAR RNA of human and simian immunodeficiency viruses: a phylogenetic analysis. *Nucl. Acids Res.*, 20:27–31, 1992.

[16] 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.

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

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

[19] C. Browning, J. M. Hilfinger, S. Rainier, V. Lin, S. Hedderwick, M. Smith, and D. M. Markovitz. The sequence and structure of the 3' arm of the first stem-loop of the human immunodeficiency virus type 2 *trans*-activation responsive region mediate tat-2 transactivation. *J. Virol.*, 71:8048–8055, 1997.

[20] Ch. Calisher. Classification and Nomenclature of Viruses - Fifth Report of The International Committee on Taxonomy of Viruses. In DL Knudson RIB Francki, CM Fauquet and F Brown, editors, *Archives of Virology Supplementum 2*, Bunyaviridae, pages 273–274. Springer-Verlag, Wien, 1991.

[21] J. H. Cate, A. R. Gooding, E. Podell, K. Zhou, B. L. Golden, A. A. Szewczak, C. D. Kundrot, T. R. Cech, and J. H. Doudna. Crystal structure of a group I ribozyme domain: Principles of RNA packing. *Science*, 273:1678–1685, 1996.

[22] T. R. Cech and B. L. Bass. Biological catalysis by RNA. *Annu. Rev. Biochem.*, 55:599–630, 1986.

[23] 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.

[24] B. Charpentier, F. Stutz, and M. Rosbash. A dynamic *in vivo* view of the HIV/1 rev-RRE interaction. *J. Mol. Biol.*, 266:950–962, 1997.

[25] M. Chastain and I. Tinoco. Nucleoside triples from the group I intron. *Biochemistry*, 32:14220–14228, 1993.

[26] J. Corodkin, L. J. Heyer, and G. D. Stormo. Finding common sequences and structure motifs in a set of RNA molecules. In T. Gaasterland, P. Karp, K. Karplus, Ch. Ouzounis, Ch. Sander, and A. Valencia, editors, *Proceedings of the ISMB-97*, pages 120–123, Menlo Park, CA, 1997. AAAI Press.

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

[28] A. T. Das, B. Klaver, and B. Berkhout. A hairpin structure in the R region of the human immunodeficiency virus type 1 RNA genome is instrumental in polyadenylation site selection. *J. Virol.*, 73(1):81–91, 1999.

[29] A. T. Das, B. Klaver, and B. Berkhout. The 5' and 3' TAR elements of human immunodeficiency virus exert effects at several points in the viral life cycle. *J. Virol.*, 72(11):9217–9223, 1998.

[30] A. T. Das, B. Klaver, B. I. F. Klasens, J. L. B. vanWamel, and B. Berkhout. A conserved hairpin motif in the R-U5 region of the human immunodeficiency virus type 1 RNA genome is essential for replication. *J. Virol.*, 71(3):2346–2356, 1997.

[31] M. O. Dayhoff, R. M. Schwartz, and B. C. Orcutt. *Atlas of Protein Sequence and Structure*, volume 5 of *3*, chapter 3, pages 345–352. NBRF Washington, 1978.

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

[33] E. Domingo, R. Webster, and J. Holland, editors. *Origin and Evolution of Viruses.* Academic Press, 1999.

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

[35] M. Eigen, J. McCaskill, and P. Schuster. The molecular quasispecies. *Adv. Chem. Phys.*, 75:149–263, 1989.

[36] M. Fekete, I. L. Hofacker, and P. F. Stadler. Prediction of RNA base pairing probabilities using massively parallel computers. *J. Comp. Biol.*, 1999. in press, Santa Fe Institute preprint 98-06-057.

[37] M. Fekete. Scanning RNA virus genomes for functional secondary structures. PhD. thesis, Faculty of Sciences, University of Vienna, Austria, 2000.

[38] D. F. Feng and R. F. Doolittle. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J. Mol. Evol.*, 25:351–360, 1987.

[39] S. Feng and E.C. Holland. HIV-1 tat trans-activation requires the loop sequence within tar. *Nature*, 334:165–167, 1988.

[40] 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.

[41] W. M. Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155:279–284, 1967.

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

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

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

[45] W. B. Goad and M. I. Kanehisa. Pattern recognition in nucleic acid sequences. A general method for finding local homologies and symmetries. *Nucl. Acids Res.*, 10:247–263, 1982.

[46] 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.

[47] A. P. Gultyaev, F. H. D. vanBatenburg, and C. W. A. Pleij. The computer simulation of RNA folding pathways using a genetic algorithm. *J. Mol. Biol.*, 250:37–51, 1995.

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

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

[50] R. W. Hamming. *Coding and Information Theory*, pages 44–47. Prentice-Hall, Englewood Cliffs, 2 edition, 1989.

[51] T. P. Hausner, J. Atmadja, and K. H. Nierhaus. Evidence that the G2661 region of 23S rRNA is located at the ribosomal binding site of both elongation factors. *Biochemie*, 69:911–923, 1987.

[52] L. He, R. Kierzek, J. SantaLucia, A. E. Walter, and D. H. Turner. Nearest-neighbour parameters for G-U mismatches. *Biochemistry*, 30:11124–11132, 1991.

[53] R. Hecker, Z. Wang, G. Riesner, and D. Steger. Analysis of RNA structures by temperature-gradient gel electrophoresis: Viroid replication and processing. *Gene*, 72:59–74, 1988.

[54] S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA*, 89:10915–10919, 1992.

[55] D. Herschlag. RNA chaperones and the RNA folding problem. *J. Biol. Chem.*, 270:20871–20874, 1995.

[56] P. G. Higgs and S. R. Morgan. Thermodynamics of RNA folding. when is an RNA molecule in equilibrium. In F. Morán, A. Moreno, J. J. Merelo, and Chacón, editors, *Advances in Artificial Life*, pages 852–861, Berlin, 1995. ECAL 95, Springer Verlag.

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

[58] D. S. Hirschberg. A linear space algorithm for computing maximal common subsequences. *Comm. Assoc. Comp. Mach.*, 18:341–343, 1975.

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

[60] I. L. Hofacker, M. A. Huynen, P. F. Stadler, and P. E. Stolorz. RNA folding and parallel computers: The minimum free energy structures of complete HIV genomes. Technical report SFI Santa Fe, New Mexico, 1996. # 95-10-089.

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

[62] 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*, 20–25. AAAI Press, Portland, OR, 1996.

[63] I. L. Hofacker and P. F. Stadler. Automatic detection of conserved base pairing patterns in rna virus genome. *Comp & Chem.*, 23:401–414, 1999.

[64] M. A. Hoffman and A. C.Palmenberg. Mutational analysis of the J-K stem-loop region of the encephalomyocarditisvirus IRES. *J. Virol.* 69:4399–406, 1995.

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

[66] J. W. Hunt and M. D. McIlroy. An algorithm for differential file comparison. Technical Report Comp. Sci. 41, Bell Laboratories, 1976.

[67] D. H. Huson. SplitsTree: a program for analyzing and visualizing evolutionary data. *Bioinformatics*, 14(1):68–73, 1998.

[68] 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.

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

[70] M. A. Huynen, and D. A. M. Konings. Viral regulatory structures and their degeneracy. In: *Questions about RNA structures in HIV and HPV*, (Gerald Myers editor) pp. 143–162, 1998.

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

[72] 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.

[73] F. Jossinet, J.-Ch. Paillart, E. Westhof, Th. Hermann, E. Skripkin, J. St. Lodmell, C. Ehresmann, B. Ehresmann, and R. Marquet. Dimerization of HIV-1 genomic RNA of subtypes A and B: RNA loop structure and magnesium binding. *RNA*, 5:1222–1234, 1999.

[74] M. I. Kanehisa and W. B. Goad. Pattern recognition in nucleic acid sequences. An efficient method for finding locally stable secondary structures. *Nucl. Acids Res.*, 10:265–277, 1982.

[75] B. Klaver and B. Berkhout. Evolution of a disrupted TAR RNA hairpin structure in the HIV-1 virus. *EMBO J.*, 13:2650–2659, 1994.

[76] 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.

[77] D. A. M. Konings. Coexistence of multiple codes in messenger RNA molecules. *Comp. & Chem.*, 16:153–163, 1992.

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

[79] M. Kunze and G. Thierrin. Maximal common subsequences of pairs of strings. *Congr. Num.*, 34:299–311, 1982.

[80] S. Y. Le, and M. Zuker. Common structures of the 5' non-coding RNA in enteroviruses and rhinoviruses. *J. Mol. Bol.*, 216:729–741, 1990.

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

[82] J. Leydold and P. F. Stadler. Minimal cycle basis, outerplanar graphs. *Elec. J. Comb.* 5, R16, 1998. See `http://www.combinatorics.org`.

[83] D. J. Lipman, S. F. Altschul, and J. D. Kececioglu. A tool for multiple sequence alignment. *Proc. Natl. Acad. Sci. USA*, 86:4412–4415, 1989.

[84] S. Limmer. Mismatch base pairs in RNA. *Prog. Nucleic Acid Res. Mol. Biol.*, 57:1–39, 1997.

[85] R. Lück, G. Steger, and D. Riesner. Thermodynamic prediction of conserved secondary structure: Application to the RRE element of HIV, the tRNA-like element of CMV, and the mRNA of prion protein. *J. Mol. Biol.*, 258:813–826, 1996.

[86] F. Major, M. Turcotte, D. Gautheret, G. Lapalme, E. Fillion, and R. Cedergren. The combination of symbolic and numerical computation for three-dimensional modeling of RNA. *Science*, 253:1255–1260, 1991.

[87] M. H. Malim, J. Hauber, S. Y. Le, J. V. Maizel, and B. R. Cullen. The HIV-1 Rev transactivator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. *Nature*, 338:254–257, 1989.

[88] 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.

[89] C. W. Mandl, H. Holzmann, T. Meixner, S. Rauscher, P. F. Stadler, St. L. Allison, and F. X. Heinz. Spontaneous and engineered deletions in the 3'-noncoding region of tick-borne encephalitis virus: Construction of highly attenuated mutants of a flavivirus. *J. Virology*, 72:2132–2140, 1998.

[90] D. A. Mann, I. Mikaelian, R.W. Zemmel, S. M. Green, A. D. Lowe, T. Kimura, M. Singh, P. J. Butler, M. J. Gait, and J. Karn. A molecular rheostat. Co-operative rev binding to stem I of the rev-response element modulates human immunodeficiency virus type-1 late gene expression. *J. Mol. Biol.*, 241:193–207, 1994.

[91] H. M. Martinez. An RNA folding rule. *Nucl. Acids Res.*, 12:323–324, 1984.

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

[93] K. L. McKnight and S. M. Lemon. The rhinovirus type 14 genome contains an internally located RNA structure that is required for viral replication. *RNA*, 4:1569–1584, 1998.

[94] D. L. Mills, editor. *A new algorithm to determine the Levenshtein distance between two strings*, Conference on Sequence Comparison. University of Montreal, 1978.

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

[96] T. P. Monath and F. X. Heinz. Flaviviruses. In: *Fields Virology*, (Fields, B. N., Knipe, D. M., Howley, P. M., Chanock, R. M., Melnick, J. L., Monath, T. P., Roizmann, B., & Straus, S. E., eds) pp. 961–1034. Lippincott-Raven Philadelphia 3rd edition996.

[97] E. W. Myers and W. Miller. Optimal alignments in linear space. *CABIOS*, 4:11–17, 1988.

[98] S. B. Needleman and C.D. Wunsch. A general method applicable to the search for similarities in the amino acid sequences of two proteins. *J. Mol. Biol.*, 48:443–453, 1970.

[99] J. M. Norman. *Elementary Dynamic Programming.* Crane, Russak and Co., New York, 1975.

[100] C. I. Nugent, K. L. Johnson, P. Sarnow, and K. Kirkegaard. Functional coupling between replication and packaging of poliovirus replicon RNA. *J. Virol.*, 73(1):427–435, 1999.

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

[102] R. Nussinov and A. B. Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proc. Natl. Acad. Sci. USA*, 77:6309–6313, 1980.

[103] R. Nussinov, I. Tinoco, and A. Jacobsen. Secondary structure model for the complete simian virus 50 late precursor RNA. *Nucl. Acids Res.*, 10:351–363, 1982.

[104] R. Nussinov, I. Tinoco, and A. B. Jacobson. Small changes in free energy assignments for unpaired bases do not affect predicted secondary structures in single stranded RNA. *Nucl. Acids Res.*, 10:341–349, 1982.

[105] 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.

[106] A. C. Palmenberg and P. Argos. Topological organization of picornaviral genomes: Statistical prediction of RNA structural signals. *Semin. Virol.*, 8:231–241, 1997.

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

[108] S. Pascarella and P. Argos. Analysis of insertions/deletions in protein structures. *J. Mol. Biol.*, 224:461–471, 1992.

[109] T. Pfister, L. Pasamontes, M. Troxler, D. Egger, and K. Bienz. Immunocytochemical localization of capsid-related particles in subcellular fractions of poliovirus-infected cells. *J. Virol.*, 188:676–684, 1992.

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

[111] S. Rauscher, C. Flamm, C. Mandl, F. X. Heinz, and P. F. Stadler. Secondary structure of the 3' noncoding regions of flavivirus genomes: Comparative analysis of base pairing probabilities. *RNA*, 3:779–791, 1997.

[112] S. Rauscher. Exploring secondary structure features of RNA virus genomes. Master's thesis, Faculty of Sciences, University of Vienna, Austria, 1997.

[113] C. Reidys, P. F. Stadler, and P. Schuster. Generic properties of combinatory maps: Neural networks of RNA secondary structures. *Bull. Math. Biol.*, 59:339–397, 1997.

[114] C. M. Rice. Flaviviridae: the viruses and their replication. 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 931–959. Lippincott-Raven, Philadelphia, 3rd edition, 1996.

[115] A. Rieger and M. Nassal. Distinct requirements for primary sequence in the 5'-and 3'-part of a bulge in the hepatitis b virus rna encapsidation signal revealed by a combined in vivo selection/in vitro amplification system. *Nucl. Acids Res.*, 23:3909–3915, 1995.

[116] V. M. Rivera, J. D. Welsh, and J. V. Maizel. Comparative sequence analysis of the 5' noncoding region of the enteroviruses and rhinoviruses. *Virology*, 165:42–50, 1988.

[117] R. R. Rueckert 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.

[118] N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.*, 4:406–425, 1987.

[119] D. Sankoff, R. J. Cedergren, and G. Lapalme. Frequency of insertion-deletion, transversion and transition in the evolution of 5S ribosomal RNA. *J. Mol. Evol.*, 7,:133–149, 1976.

[120] D. Sankoff, C. Morel, and R. J. Cedergren. Evolution of 5S RNA and the nonrandomness of base replacement. *Nature New Biology*, 245:232–234, 1973.

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

[122] C. S. Schmaljohn, G. B. Jennings, J. Hay, and J. M. Dalrymple. Coding strtegy of the S genome segment of hantaan virus. *Virology*, 155:633–643, 1986.

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

[124] 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.

[125] D. B. Smith, J. Mellor, L. M. Jarvis, F. Davidson, J. Kolberg, M. Urdea, P. Yap, P. Simmonds, and The International HCV Collaborative Study Group. Variation of the hepatitis C virus 5' non-coding region: implications for secondary structure, virus detection and typing. *J. Gen. Virol.*, 76:1749–1761, 1995.

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

[127] R. Stocsits, I. L. Hofacker, and P. F. Stadler. Conserved secondary structures in hepatitis B virus RNA. *Computer Science in Biology*, pages 73–79, Bielefeld, D, 1999. Univ. Bielefeld. Proceedings of the GCB'99, Hannover, D.

[128] R. Stocsits. Improved alignments based on a combination of amino acid and nucleic acid information. Master's thesis, Faculty of Sciences, University of Vienna, Austria, 1999.

[129] J. E. Tabaska and G. D. Stormo. Automated alignment of RNA sequences to pseudoknotted structures. In T. Gaasterland, P. Karp, K. Karplus, Ch. Ouzounis, Ch. Sander, and A. Valencia, editors, *Proceedings of the ISMB-97*, pages 311–318, Menlo Park, CA, 1997. AAAI Press.

[130] 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.

[131] 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. Virus Taxonomy. *Springer-Verlag*, 6th, 1995.

[132] J. D. Thompson, D. G. Higgins, and T. J. Gibson. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignments through sequence weighting, position specific gap penalties and weight matrix choice. *Nucl. Acids Res.*, 22:4673–4680, 1994.

[133] J. D. Thompson, D. G. Higgins, and T. J. Gibson. Improved sensitivity of profile searches through the use of sequence weights and gap excision. *CABIOS*, 10:19–29, 1994.

[134] D. H. Turner, N. Sugimoto, and S. Freier. RNA structure prediction. *Ann. Rev. Biophys. Chem.*, 17:167–192, 1988.

[135] I. L. Hofacker, W. Fontana, P. F. Stadler, and P. Schuster. `Vienna RNA Package`. `http://www.tbi.univie.ac.at/~ivo/RNA/`, 1994. (Free Software).

[136] M. Vingron and P. R. Sibbald. Weighting in sequence space: A comparison of methods in terms of generalized sequences. *Proc. Natl. Acad. Sci. USA*, 90:8777–8781, 1993.

[137] M. Vingron and M. S. Waterman. Sequence alignment and penalty choice. Review of concepts, case studies and implications. *J. Mol. Biol.*, 235:1–12, 1994.

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

[139] P. A. Walker, L. E.-C. Leong, and A. G. Porter. Sequence and structural determinants of the interaction between the 5'-noncoding region of picornavirus RNA and rhinovirus protease 3C. *J. Biol. Chem.*, 270(24):14510–14516, 1995.

[140] A. E. Walter, D. H. Turner, J. Kim, M. H. Lyttle, P. Müller, D. H. Mathews, and M. Zuker. Co-axial stacking of helixes enhances binding of oligoribonucleotides and improves predicions of RNA folding. *Proc. Natl. Acad. Sci. USA*, 91:9218–9222, 1994.

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

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

[143] M. Waterman and T. F. Smith. RNA secondary structure: A complete mathematical analysis. *Math. Biosc.* 42, 257–266, 1978.

[144] M. S. Waterman and T. Byers. A dynamic programming algorithm to find all solutions in the neighborhood of the optimum. *Math. Biosci.*, 77:179–188, 1985.

[145] 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.

[146] E. Westhof and L. Jaeger. RNA pseudoknots. *Current Opinion Struct. Biol.* 2, 327–333, 1992.

[147] W. J. Wilbur and D. J. Lipman. Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA*, 80:726–730, 1983.

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

[149] E. Wimmer, C. U. T. Hellen, and X. Cao. Genetics of poliovirus. *Annu. Rev. Genet.*, 27:353–436, 1993.

[150] St. Wuchty, I. L. Hofacker W. Fontana, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 1998. in press, Santa Fe Institute Preprint 98-05-040.

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

[152] A. Zeffman, S. Hassard, G. Varani, and A. Lever. The major HIV-1 packaging signal is an extended bulged stem-loop whose structure is altered on interaction with the Gag polyprotein. *J. Mol. Biol.*, 297:877–893, 2000.

[153] R. W. Zemmel, A. C. Kelley, J. Karn, and P. J. G. Butler. Flexible regions of RNA structure facilitate cooperative rev assembly of the rev-response element. *J. Mol. Biol.*, 258:763–777, 1996.

[154] M. Zuker. `mfold-2.3`. `ftp://snark.wustl.edu/`. (Free Software).

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

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

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

List of Publications

- S. Rauscher, C. Flamm, C.W. Mandl, F.X. Heinz, and P.F. Stadler
  Secondary Structure of the 3'-Noncoding Region of Flavivirus Genomes: Comparative Analysis of Base Pairing Probabilities.
  RNA, 3:779-791 (1997).

- C. W. Mandl, H. Holzmann, T. Meixner, S. Rauscher, P.F. Stadler, and F.X. Heinz
  Spontaneous and Engineered Deletions in the 3'-Noncoding Region of Tick-Borne Encephalitis Virus: Construction of Highly Attenuated Flavivirus Mutants.
  J. Virology 72:2132-2140 (1998)

- I. L. Hofacker, M. Fekete, C. Flamm, M. A. Huynen, S. Rauscher, P. E. Stolorz, and P. F. Stadler
  Automatic Detection of Conserved RNA Structure Elements in Complete RNA Virus Genomes.
  Nucl. Acids Res. 26:3825-2836 (1998)

Curriculum vitae

Susanne Rauscher
* 6. März 1967 in Wien

Ausbildung

1981 – 1986	Handelsakademie Wien III
1986 – 1997	Biochemiestudium an der Universität Wien Diplomarbeit am Institut für Theoretische Chemie bei Prof. Peter F. Stadler: *Exploring Secondary Structure Features of RNA Virus Genomes*
17/3/1997	Sponsion zur *Magistra rerum naturalium*
9/1997	Sixth International Summer School on Biophysics *Supramolecular Structure and Function*, Rovinj/Kroatien
6/1999	*Complex Systems* Summer School, Santa Fe/New Mexico
seit 1997	Doktoratsstudium am Institut für Theoretische Chemie und Molekulare Strukturbiologie, Universität Wien Dissertation bei Prof. Peter F. Stadler: *Conserved Structure Elements in Viral RNA Genomes*

Tagungen

19 – 21/6/1996	Posterpräsentation: *New Conserved Secondary Structure Motifs at the 3'-End of the RNA Genome of Tick-Borne Flaviviruses*, Symposium on Modern Approaches to Flavivirus Vaccines, Wien
29/6/ – 3/7/1998	Tagung der Nobelpreisträger für Chemie, Lindau/Deutschland
26 – 28/4/1999	Meeting on RNA Structure Prediction, Wien

Tutoriumsaufträge

1993 – 1998	*Chemische Übungen für BiologInnen und ErnährungswissenschaftlerInnen*, Institut für Organische Chemie, Universität Wien
1998 – 1999	*Computerübungen zu Biochemie IV*, Institut für Theoretische Chemie, Universität Wien