Conserved and Consensus RNA Structures

DISSERTATION
zur Erlangung des akademischen Grades
Doctor rerum naturalium

Vorgelegt der
Fakultät für Formal- und Naturwissenschaften
der Universität Wien

von
Caroline Thurner

am Institut für Theoretische Chemie und Molekulare Strukturbiologie

im April 2004

Abstract

The function of a biomolecule is closely related to its structure, thus even in very different species, molecules with the same function exhibit very similar structure. For RNA molecules secondary structures are a useful coarse graining of the spatial structure, since they cover most of the free energy of folding, they are conserved during evolution and have been used successfully to interpret RNA function. Moreover they are computationally easy to handle, because only discrete base pair patterns (no coordinates) are considered.

We combined phylogenetic and thermodynamic information to search for evolutionarily conserved secondary structure motifs in the genomic RNAs of the family *Flaviviridae*. This family consists of the three genera *Flavivirus*, *Pestivirus*, *Hepacivirus*, and the group of GB virus C/hepatitis G virus with a currently uncertain taxonomic classification. The main findings of our survey are strong hints for the possibility of genome cyclization in hepatitis C virus and GB virus C, as it has been proposed previously for the members of the genus *Flaviviruses*, a surprisingly large number of conserved RNA motifs in the coding regions, and a lower level of detailed structural conservation of the motifs in the Internal Ribosomal Entry Sites and 3' untranslated regions than reported in the literature.

Pseudoknots are normally excluded from secondary structure by definition, yet they are occasionally functionally important. Therefore, in the second part of this thesis the secondary structure detection algorithm was extended to allow for prediction of some restricted kinds of pseudoknots. The algorithm is tested on three kinds of RNAs which are known to contain pseudoknots. These are Signal recognition particle RNA, ribonuclease P RNA and tmRNA.

In order to evaluate the predicted secondary structures we had to compare our data to experimental results contained in the literature. The search for literature containing the specific information resulted in a very time consuming task. Therefore we created an automated text categorization tool for bibliographic search in collaboration with computer scientists from the university of Leipzig. `Pubmed` result sets for two virus groups, *Picornaviridae* and *Flaviviridae*, have been manually labeled. We evaluated various classifiers from the `Weka` toolkit together with different feature selection methods to assess whether classifiers trained on documents dedicated to one virus group can be successfully applied to filter documents on other virus groups.

Zusammenfassung

Die Funktion von Biomolekülen hängt eng mit ihrer räumlichen Struktur zusammen. So haben Moleküle verschiedener Spezies, die aber gleiche Funktionen erfüllen, oft auch sehr ähnliche Strukturen. Für RNA Moleküle stellen Sekundärstrukturen eine sehr nützliche Vereinfachung der räumlichen Struktur dar, da sie den größten Teil der freien Faltungsenergie abdecken, in der Evolution erhalten geblieben sind und schon erfolgreich dazu verwendet werden konnten, RNA Funktionen zu interpretieren und leicht berechnet werden können, da nur diskrete Basenpaar-Kombinationen in Betracht gezogen werden müssen.

Um Sekundärstrukturen in genomischen RNAs von Viren der Familie *Flaviviridae* zu finden, die im Laufe der Evolution erhalten geblieben sind, verknüpften wir phylogenetische und thermodynamische Information miteinander. Die Familie *Flaviviridae* besteht aus drei Genera *Flavivirus*, *Pestivirus* und *Hepacivirus*, und die Gruppe GB Virus C/Hepatitis G, die derzeit noch unklassifiziert ist. Die wichtigsten Resultate unsere Untersuchungen ergaben, dass für die Genome der Viren Hepatitis C und GB Virus C eine Zyklisierung möglich sein könnte, ähnlich, wie das für die Mitglieder des Genus *Flaviviren* bereits beschrieben wurde. Weiters fanden wir eine Fülle von konservierten und potentiell funktionalen Strukturelementen in den kodierenden Regionen der Genome, und eine kleinere Anzahl von konservierten Strukturen in den nicht kodierenden Regionen der Virengenome, als in der Literatur allgemein angenommen wird.

Pseudoknoten werden im Allgemeinen nicht zu den Sekundärstrukturen gerechnet, ihnen kommt aber eine sehr wichtige Rolle in der Funktionsweise von RNA Molekülen zu. Im zweiten Teil dieser Dissertation erweitern wir deshlab unseren Algorithmus zur Sekundärstruktur-Vorhersage um eine eingeschränkte Gruppe von Pseudoknoten in die Vorhersage mitein zu beziehen. Der Algorithmus wurde an drei verschidenen RNAs mit Pseudoknoten getesten, das sind Signal recognition RNA, Ribonuklease P RNA und tmRNA.

Die Suche nach Literaturstellen, die experimentelle Informationen über unsere vorhergesagten Sekundärstrukturen enthalten, entpuppte sich als erstaunlich langwierig. Um eine ähnlich mühsame Literatursuche für andere Viren zu erleichtern, entwickelten wir in Zusammenarbeit mit Informatikern von der Universität Leipzig ein automatisches Textkathegorisierungs Programm. `Pubmed`-Ergebnis-Sets für zwei Virenfamilien wurden händisch gekennzeichnet. Wir werteten verschiedene Klassifizierer aus dem `Weka`-Toolkit gemeinsam mit verschiedenen Merkmal-Erkennungsmethoden aus, und stellten fest, dass Klassifizierer, die an Datensets von einem Virus trainiert wurden, erfolgreich dazu verwendet werden können, Dokumente eines anderen Viruses zu filtern.

Contents

1 Introduction

1.1 Why RNA?

The fundamental biopolymers in molecular genetics are proteins and nucleic acid sequences, i.e. DNA and RNA. Genetic information in all eukaryotic and prokaryotic cells is stored in the form of DNA. RNA plays an important role in the expression of genes. DNA is copied to messenger RNA (mRNA) in a process called transcription by the help of the enzyme RNA-polymerase II. mRNA is subsequently decoded for protein synthesis. This process is called translation and is mediated by ribosomes, enzyme complexes composed by proteins and ribosomal RNA (rRNA), and transfer RNA (tRNA), which translates the codons of the mRNA into the amino acids of the protein. Proteins play a crucial role as enzymes in most biological processes. The remarkable scope of their activity includes catalysis of chemical reactions, transport of small molecules and ions through membranes, control of growth and differentiation of cells, and a key-role in immune protection, to mention just a few functions. Whereas all kind of living cells store their genetic information in the form of DNA, the genetic material of viruses can be either in the form of DNA or RNA.

For many years, proteins were assumed to be the only biomolecules with catalytic properties. Within the last two decades, this view has given way to a more detailed understanding due to several important discoveries. Various types of RNA molecules possessing catalytic properties have been found and in the following were called ribozymes. In the 1980s Cech *et al.* discovered the autocatalytic splicing of the precursor of rRNA [13, 81]. In the following year, the groups of Altmann and Pace revealed that RNase P, which processes the 5' ends of tRNA precursors in all organisms, was also a ribozyme [36].

Since then a large number of other ribozymes have been discovered. A partial list of such molecules includes small nuclear RNAs (snRNAs) [135] that compose the pre-mRNA splicing machinery, signal recognition particle (SRP)

RNA [71] necessary for protein translocation, and rRNA [105, 104]. Initially it was thought that the RNA component of the ribosome merely serves as a structural scaffold for the functional active ribosomal proteins, the current view is the reverse.

Recently more and more families of small RNAs have been discovered which carry important functions in regulatory processes of the cell. For example small nucleolar RNA (snoRNA) are stable RNA species localized in the eukaryotic nucleolus, where they are required for cleavage of precursor rRNA and modify many of its nucleotides [76]. An other most interesting class are microRNAs (miRNAs) which are involved in the regulation of translation and degradation of mRNAs [103, 145].

The conformation of messenger RNA, particularly at the 5'- and 3'-untranslated regions, determines the lifetime of the RNA and controls the efficiency of translation (see, for example [138]). Furthermore, it has been shown that pseudoknots in retroviral mRNAs cause programmed frame shifts that produce the correct ratios of proteins required for viral propagation [14]. Another example is presented by the highly conserved RNA secondary structure domains present in the 5'-non-translated region of, for instance, picornaviruses, hepatitis C viruses and pestiviruses [113, 67]. This so-called internal ribosome entry site (IRES) enables cap-independent initiation of translation. In addition, a number of IRES-containing eukaryotic mRNAs have been detected recently, reviewed in [48]. Functional important RNA structures are not restricted to non-coding RNA, as shown by the examples of the Rev response element (RRE) of HIV1, which is located within the *env* gene [89], and the cis-acting replication element located in the coding region of picornaviruses [97].

One criterion for the importance of RNA structure is the conservation in a set of homologous RNA sequences. Therefore it is of considerable practical interest to compute efficiently the consensus structure of a collection of such RNA molecules.

1.2 RNA Structure

RNA molecules are usually single stranded, except in some viruses. An RNA molecule can fold back onto itself to form double helical structures consisting mainly of Watson-Crick (GC and AU) base pairs or the slightly less stable Wobble pair GU. The stacking energy of these *allowed* base pairs is the major driving force for RNA structure formation. Other, usually weaker, intermolecular forces and the interaction with the aqueous solvent shape its spatial structure. The list of base pairs of an RNA structure which can be drawn as outerplanar graph, i.e. all base pairs can be drawn in the half-plane without intersections, forms the *secondary structure*. The three-dimensional configuration of the molecule is called the *tertiary structure*. The first such structure to be experimentally determined was the yeast tRNA$^{\text{phe}}$ [4] shown in Figure 1.

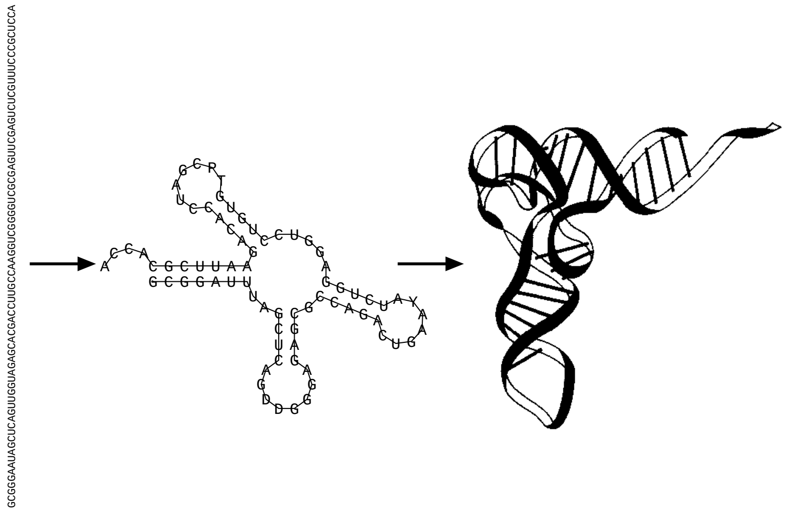

GCGGGAAUAGCUCAGUUGGUAGAGCACGACCUUGCCAAGGUCGGGGUCGCGAGUUCGAGUCUCGUUUCCCGCUCCA

Figure 1: Sequence, secondary structure and schematic representation of the tertiary structure of tRNA

Protein secondary and tertiary structure are highly coupled and difficult to predict accurately. The secondary structure of proteins is context dependent, their energies are comparable to the energies involved in tertiary interactions.

In contrast to proteins, RNA secondary structure covers the major part of the free energy of folding. Furthermore, secondary structures are used successfully in the interpretation of RNA function and reactivity, and secondary structures related to function are conserved in evolutionary phylogeny.

Extensive computer simulations [27, 129] 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.

The common theoretical secondary structure model comprises only a subset of all possible base pair patterns. The model excludes by definition all overlapping base pair interactions, subsequently called pseudoknots, mainly for computational reasons. It turns out that algorithms dealing with simple secondary structures based on a thermodynamic energy model can be implemented in a very elegant way, with the help of a method called dynamic programming. In contrast the prediction of RNA structure including pseudoknots based on the same model has been proven to be NP-complete [88, 3].

1.3 Objectives of this Work

When a sufficiently large number of sequences is available, phylogenetic methods can be used to faithfully determine conserved structural elements or consensus structures, while the accuracy of purely thermodynamic structure prediction is often not satisfactory.

But in a large majority of cases the number of available sequences is small, and it is desirable to gain structural information out of smaller sets of related sequences. In this field mainly algorithms are successful that combine thermodynamic and phylogenetic information for structure prediction [85, 51]. Consensus structures are unsuitable when a significant part of the whole molecule has no conserved structures. RNA virus genomes, for instance, contain only local structural patterns. Such features can be identified with a related approach, the algorithm `alidot` developed by Hofacker *et al.* [54, 58]. One part of this work is concerned with the identification of potentially functional important structures in the genomes of the virus family *Flaviviridae*, using `alidot`. These are RNA containing viruses, including important human pathogens, such as Hepatitis C virus, Tick-borne encephalitis virus, Japanese encephalitis virus and the four serotypes of dengue virus, and the members of the genus *Pestiviruses*, which comprise venerinerian pathogens, such as Bovine viral diarrhea virus and Hog cholera virus. Research has been concentrated mainly on the 5'-non-translated regions of the genome, because of the particular interest in the IRES region. Here we describe a comprehensive computational survey of conserved structural elements, including the coding region, in the genera of the family *Flaviviridae*.

Another important class of structural elements in RNA sequences are pseudoknots, as they cover a broad diversity of molecular biological functions. These are translational control at 5'ends of mRNAs, forming reaction centers of ribozymes, and replicational control at the 3'end of RNA molecules. Thus it is desirable to be able to predict the consensus structure including pseudoknots based on a smaller set of sequences.

In the second part of this thesis we adapt the algorithm `alidot` in a way that it allows the prediction of some kinds of pseudoknots, on the basis of a combination of both thermodynamic and phylogenetic information.

As an important part of our work, we had to compare our computational results with experimental data. These are, unfortunately, often hidden in the vast body of molecular biology literature. More often than not, the data that are of interest for a particular computational study are mentioned only in passing and in a different context in the experimental literature. To our surprise, the bibliographic search for experimental evidence of and further information on “RNA secondary structures” in a given group of virus (a seemingly rather straightforward task) turned out to be more tedious than the work on the actual sequence and structure data. Therefore, in collaboration with Lukas Faulsitch from the Institut für Informatik, Universität Leipzig, Germany, we created an automated text categorization tool (`litsift`). The elaborate imput data supply for the virus family *Picornaviridae* was provided by Christina Witwer and for the virus family *Flaviviridae* as part of this thesis.

2 General Concepts

In this chapter basic concepts that are fundamental for later discussion will be established. These include the definition and representation of secondary and bi-secondary RNA structures. Since secondary structures are conveniently described in terms of a graph, we therefore introduce some basic definitions from graph theory, and some basic notation. These can be found e.g. [2]. We follow the presentation of Haslinger in [45].

2.1 Graphs

A *graph* $G = (V, E)$ consists of a finite set V of vertices (nodes, in the case of RNA: bases) and a finite set E of edges (arcs, for RNA: backbone or hydrogen bonds).

The edge e containing vertices u and v is often denoted uv, vertices u and v are said to be *adjacent*, and the edge e is *incident* to u and v.

The *degree* of a vertex v is the number of edges incident to v.

The *adjacency matrix* $\mathbf{A}$ of a graph G with n vertices is a $n \times n$ matrix, whose rows and columns correspond to vertices, with $\mathbf{A}_{uv} = 1$ if $uv \in E$, and $\mathbf{A}_{uv} = 0$ otherwise.

A graph is *bipartite* if the vertices partition into sets V_1 and V_2, such that for each edge $uv \in E$ either $u \in V_1$ and $v \in V_2$, or $u \in V_2$ and $v \in V_1$. A graph $G' = (V', E')$ is a *subgraph* of $G = (V, E)$, if $V' \subseteq V$ and $E' \subseteq E$.

A *walk* is a sequence of vertices, $(v_1, v_2, \dots v_n)$, such that for $1 \leq i < n$, $v_i v_{i+1}$ is an edge. A *path* is a walk where no vertex occurs more than once in the sequence. A *cycle* is a path that starts and ends at the same vertex. Two nodes u and v are *connected* if the graph contains at least one path from u to v. A graph is *connected* if every pair of its nodes is connected, otherwise, the graph is *disconnected.* A *component* of a graph is a maximal connected subgraph.

The drawing of a graph is *planar* if no two distinct edges intersect. A graph is *planar* if it admits a planar drawing.

2.2 Contact Structures

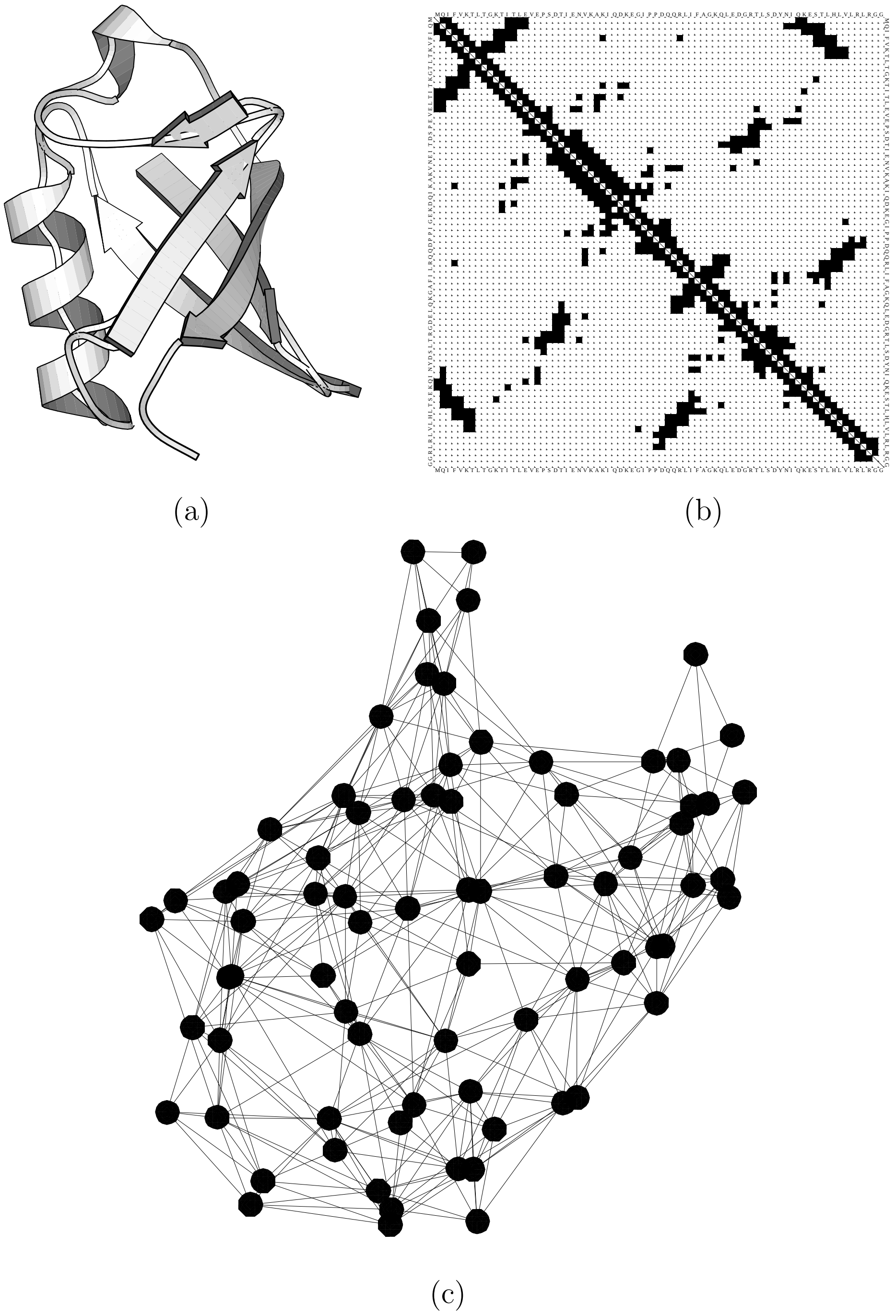
The three-dimensional structure of a linear biopolymer, such as RNA, DNA, or a protein can be approximated by their *contact structure*, i.e., by the list of all pairs of monomers that are spatial neighbors. Contact structures of polypeptides have been introduced by Ken Dill and co-workers in the context of lattice models of protein folding [15, 18]. The secondary structures of single stranded RNA and DNA form a special class of contact structures.

We assume that the monomers, aminoacids and nucleotides alike, are numbered from 1 to n along the backbone. For simplicity we shall write $[n] = \{1, \ldots, n\}$. The adjacency matrix of the backbone $\mathbf{B}$ has the entries $\mathbf{B}_{i,i+1} = \mathbf{B}_{i+1,i} = 1$, $i \in [n-1]$. In a more general context, polymers with cyclic or branched backbones can be considered, see e.g. [44].

A contact structure is faithfully represented by the *contact matrix* $\mathbf{C}$ with the entries $\mathbf{C}_{ij} = 1$ if the monomers i and j are spatial neighbors without being adjacent along the backbone, and $\mathbf{C}_{ij} = 0$ otherwise. Hence $\mathbf{C}_{ij} = 0$ if $|i-j| \leq 1$. Note that both $\mathbf{B}$ and $\mathbf{C}$ are symmetric matrices.

Definition 1 *A* (contact) diagram $([n], \Omega)$ *consists of n vertices labeled* 1 *to n and a set Ω of* arcs *that connect non-consecutive vertices.*

The diagram is simply a graphical representation of the contact matrix. As an example, the conventional ribbon diagram of the protein ubiquitin together with its discretized structure represented by contact matrix and contact graph is shown in Fig.2.



(a)

(b)

(c)

Figure 2: **The structure of the ubiquitin molecule**, `pdb` entry `1ubq`. (a) Conventional ribbon diagram, (b) contact matrix, (c) contact graph. The figure is adapted from [130]

2.3 RNA Contact Structures

The "classical" definition of RNA secondary structure [152] cannot be extended easily to include pseudoknots without allowing overly involved knotted structures or nested pseudoknots. Therefore we use an alternative definition of secondary structure which is generalized to so-called bi-secondary structures [46]. Bi-secondary structures include almost all known pseudoknotted RNA structures, with the exception of the *Escherichia coli* αmRNA. This chapter mostly follows the definitions given by Haslinger in [46].

2.3.1 Secondary Structures

The classical definition of an RNA secondary structure [152] consists of two conditions: (i) No nucleotide takes part in more than one base pair. Thus RNA secondary structures are special types of 1-diagrams. (ii) Base pairs must not cross, that is, there may not be two base pairs $(i.j)$ and $(k.l)$ such that $i < k < j < l$. In terms of the contact matrix this means, if $\mathbf{C}_{ij} = \mathbf{C}_{kl} = 1$ and $i < k < j$ then $i < l < j$. With the following notation we will find an alternative formulation of condition (ii):

Let $\alpha = \{i, j\}$ with $i < j$ be an arc of a diagram. We write $\bar{\alpha} \overset{\text{def}}{=\!=} [i, j] \subset \mathbb{R}$ for the associated interval. Two arcs of a diagram are *consistent* if they can be drawn in the same half-plane without crossing each other. Equivalently, two arcs $\alpha, \beta \in \Omega$ of a diagram are consistent if either one of the following four conditions is satisfied:

(i) $\bar{\alpha} \cap \bar{\beta} = \emptyset$.

(ii) $\bar{\alpha} \subseteq \bar{\beta}$.

(iii) $\bar{\beta} \subseteq \bar{\alpha}$.

(iv) $\bar{\alpha} \cap \bar{\beta} = \{k\}$, a single vertex.

Case (iv) is ruled out by definition in 1-diagrams. The non-crossing condition thus may be expressed as follows: Whenever the intervals of two arcs $\{i, j\}$ and $\{k, l\}$ have non-empty intersection then one is contained in the other [128]. This leads to the following definition:

Definition 2 *A secondary structure is a 1-diagram in which any two arcs are consistent.*

Thus secondary structure graphs are planar, i.e., they can be drawn in such a way that the backbone forms a circle and all base pairs are represented by chords that must not cross each other, see the example of tRNA in Fig.4.

A base i is said to be interior to the base pair (k, l) if $k < i < l$. If, in addition, there is no base pair (p, q) $k < p < q < l$ such that $p < i < q$ we will say that i is immediately interior to the base pair (k, l). A base pair (p, q) is said to be (immediately) interior if p and q are (immediately) interior to (k, l).

Definition 3 *A secondary structure consists of the following structure elements*

(i) A stem consists of subsequent base pairs $(p - k, q + k), (p - k + 1, q + k - 1), ..., (p, q)$ such that neither $(p - k - 1, q + k + 1)$ nor $(p + 1, q - 1)$ is a base pair. $(k + 1)$ is the length of the stem, $(p - k, q + k)$ is the terminal base pair of the stem. Isolated single base pairs are considered as stems $(length = 1)$ as well.

(ii) A loop consists of all unpaired bases which are immediately interior to some base pair (p, q), the “closing” pair of the loop. The number of these bases is called the size of the loop.

(iii) An external base is an unpaired base which does not belong to a loop. A collection of adjacent external bases is called an external element. If it contains the base 1 or n it is a free end, otherwise it is called joint.

Any secondary structure S can be uniquely decomposed into stems, loops, and external elements.

Definition 4 *A stem $[(p,q),...,(p+k,q-k)]$ is called terminal if $p-1=0$ or $q+1=n+1$ or if the two bases $p-1$ and $q+1$ are not interior to any base pair. The sub-structure enclosed by the terminal base pair (p,q) of a terminal stem will be called a component of S.*

Definition 5 *The degree of a loop is given by 1 plus the number of terminal base pairs of stems which are interior to the closing bond of the loop. A loop of degree 1 is called hairpin (loop), a loop of a degree larger than 2 is called multi-loop. A loop of degree 2 is called bulge if the closing pair of the loop and the unique base pair immediately interior to it are adjacent; otherwise a loop of degree 2 is termed interior loop. Two stacked base pairs form an interior loop with size 0.*

2.3.2 Representation of Secondary Structures

Figure 4 shows a variety of different representation forms for RNA secondary structure. Beside the conventional drawing as a planar graph a secondary structure can be represented as a *dot plot*. A square in row i and column j in the upper right side of the dot plot indicates a base pair (i,j) which is predicted by McCaskill's algorithm, the area of the square is proportional to the predicted base-pairing probability. A square in row j and column i in the lower left side of the dot plot indicates a base pair (i,j) which is part of the minimum-free-energy structure of the sequence.

As a consequence of definition 2 each secondary structure can be encoded as a string s of length n in the following way: If the base i is unpaired, then s_i = '`.`'. Each base pair $\alpha = \{p,q\}$ with $p<q$ translates to s_p = '`(`' and s_q = '`)`'. Since any base is allowed to pair only once and base pairs

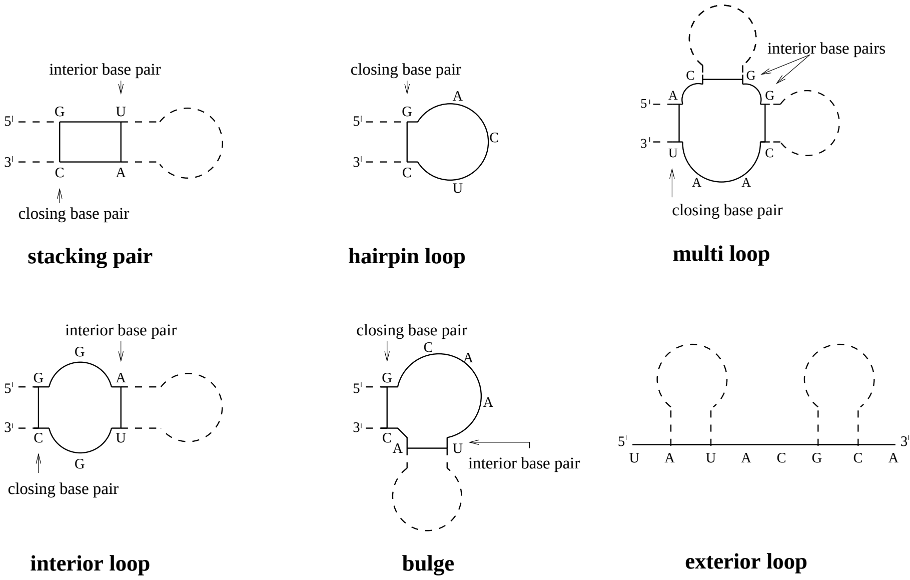


Figure 3: Basic loop types

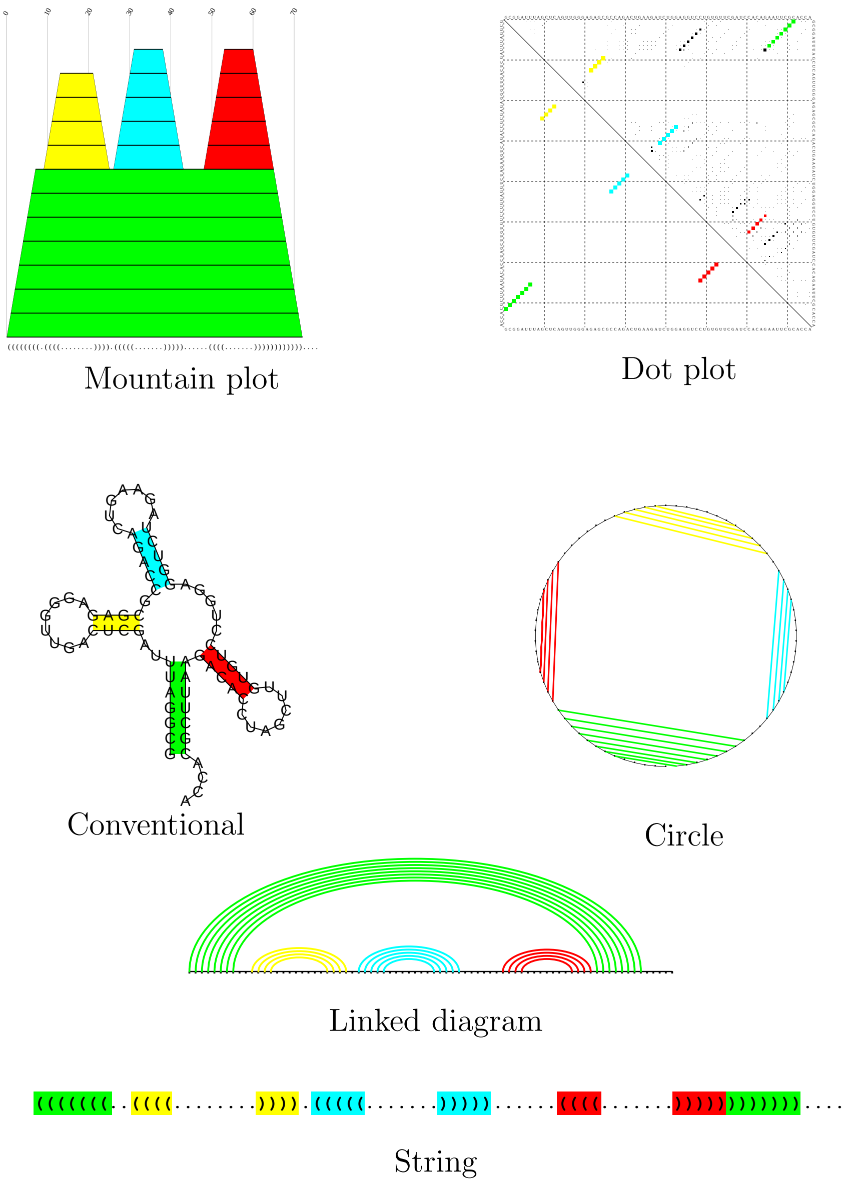
must not cross their corresponding parentheses are either nested, `(( ))`, or next to each other, `()()`. As there are no base pairs between neighboring bases there is at least one dot contained within each parenthesis. The "dot-bracket" notation is used as a convenient notation in input and output of the `Vienna RNA Package`, a piece of free software for folding and comparing RNA molecules [56].

Especially useful to compare even large structures is the *mountain*-representation (or *mountain plot*) [59]. The three symbols of the string representation '.', '(' and ')' are assigned to three directions "horizontal', 'up' and 'down' in the plot. The structural elements match certain secondary structure features.

- *Peaks* correspond to hairpins. The symmetric slopes represent the stems 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.
- *Valleys* indicate the unpaired regions between the branches of a multi loop or, when their height is zero, they indicate external vertices.

In the *linked diagram* representation the sequence is arranged along the x-axis and the base pairs are drawn as arcs confined to the upper half-plane. The *circle* representation places the sequence along a circle and the base pairs are represented by the arcs.



Mountain plot

Dot plot

Conventional

Circle

Linked diagram

String

Figure 4: Different representations of RNA secondary structure. All drawings show the same structure and use the same colors to mark the different stems.

2.3.3 Bi-secondary Structures

Bi-secondary structures can be understood as superpositions of two disjoint secondary structures. Their contact graphs are still planar, but now the chords may be drawn on the inside and on the outside of the circle that represents the backbone.

Definition 6 *A bi-secondary structure is a 1-diagram that can be drawn in the plane without intersections of arcs.*

We may draw the arcs in the upper or lower half-plane, but they are not allowed to intersect the x-axis. Bi-secondary structures are therefore "superpositions" of two secondary structures.

The virtue of bi-secondary structures is that they capture a wide variety of RNA pseudo-knots, while at the same time they exclude true knots. Knotted RNAs could in principle arise either from parallel stranded helices (Fig 5), or in very large molecules from sufficiently complicated cross-linking patterns. Parallel-stranded RNA has not been observed (so far), see however ref. [28] on parallel-stranded DNA. Castor *et al.* [12] have searched unsuccessfully for knots in large RNAs. The definition of bi-secondary structures, by allowing a planar drawing of the structure, rules out both possibilities. Almost all known RNA pseudoknots fall into the class of bi-secondary structures, with the exception of *Escherichia coli* α-operon mRNA.

2.3.4 The Inconsistency Graph of a Diagram

Definition 7 *Let* $\Delta = ([n], \Omega)$ *be a diagram. The inconsistency graph* $\Theta(\Delta)$ *of the diagram has vertex set* Ω *and* $\{\alpha, \beta\}$ *is an edge of* $\Theta(\Delta)$ *if and only if the arcs* α *and* β *are inconsistent in* Δ*.*

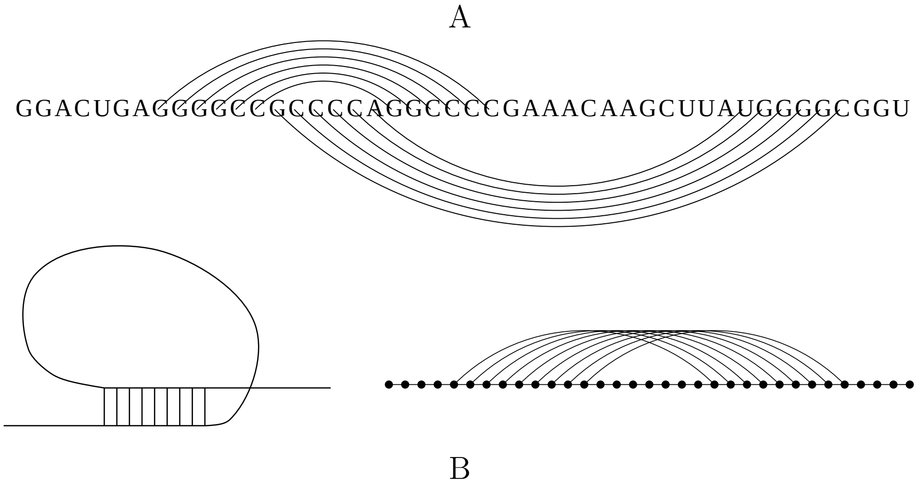


Figure 5: The contact structure of the proposed SRV-1 frame-shift signal contains a pseudo-knot, see reference [146]. (A) Pseudo-knot which belongs to the class of bi-secondary structures. (B) Knots do not belong to the class of bi-secondary structures. Knots, in contrast to pseudo-knots, may contain parallel stranded helices which so far have not been described for RNA.

The following notation will be useful: Two arcs $\alpha = \{i, j\}$ and β are *stacked* if $\beta = \{i-1, j+1\}$ or $\beta = \{i+1, j-1\}$. A *stem* is a subset Ψ of arcs α_0 through α_h such that α_p and α_{p+1} are stacked for $p = 0, \dots, h-1$. It is easy to show that the arcs of a stem Ψ of a 1-diagram are either all isolated vertices or they are contained in the same component of the inconsistency graph $\Theta(\Delta)$. Furthermore, all arcs of a stem have the same adjacent vertices in $\Theta(\Delta)$. We may therefore use a reduced intersection graph $\hat{\Theta}(\Delta)$ the vertices of which are the stems. Examples of reduced intersection graphs are given in Figures 6 and 7.

The following example shows that there are natural RNA structures that are more complicated than bi-secondary structures. The *Escherichia coli* α-operon mRNA folds into a structure that is required for allosteric control of translational initiation [143]. Compensatory mutations have defined an

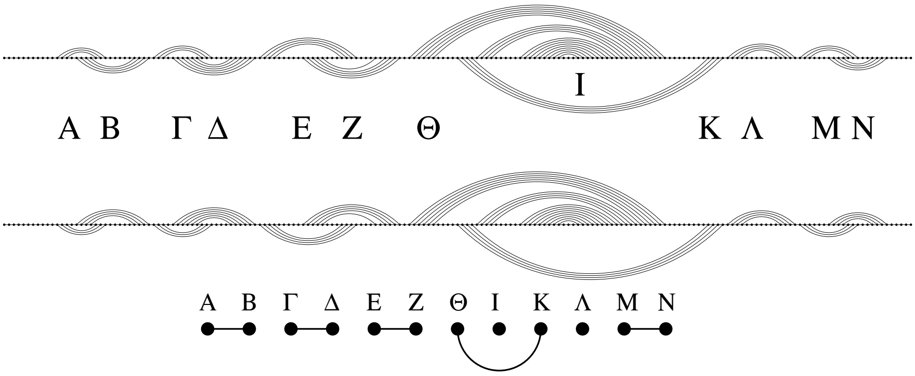


Figure 6: Two diagrams encoding the 3' non-coding region of tobacco mosaic virus RNA [1]. The upper diagram corresponds to the normal form, the lower diagram maximizes the number of upper arcs. Stems are labeled by uppercase Greek letters. The third line shows the inconsistency graph of the tmvRNA structure.

unusual pseudo-knotted structure [142], the thermodynamics of which were subsequently investigated in detail [32]. The diagram of its contact structure cannot be drawn without intersections, see Figure 7.

Being the union of the two secondary structures $([n], \Omega_U)$ and $([n], \Omega_L)$ we can represent each bi-secondary structure as a string s using two types of parentheses: As in a secondary structure we write a dot `.' for all unpaired vertices. A pair $\{p, q\} \in \Omega_U$ becomes $s_p =$ `(' and $s_q =$ `)', while an arc $\{p, q\} \in \Omega_L$ becomes $s_p =$ `[' and $s_q =$ `]'. Unfortunately, the decomposition of a bi-secondary structure into two secondary structures in general is not unique, see Figure 6. However it is possible to define the *normal form* of a bi-secondary structure by means of the following rule: The leftmost arc of each connected component of $\Theta(\Delta)$ belongs to Ω_U. In particular, all isolated vertices of $\Theta(\Delta)$ are contained in Ω_U.

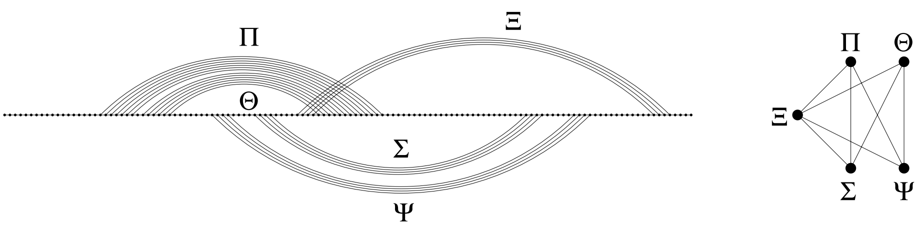


Figure 7: Diagram of the contact structure of E. coli α-mRNA. The structure contains 5 stems, labeled by uppercase Greek letters.

2.3.5 Isambert-Siggia Decomposition of Secondary Structures

In 2000 Hervé Isambert and Eric D. Siggia [64] proposed a decomposition of generalized RNA secondary structures into so-called *nets*. Here we follow the reformulation of this work of Witwer [156] who used a more standard notation and provided proofs for the basic properties of the net-decomposition of general 1-structures.

We consider contact structures of linear or circular (bio)polymers. W.l.o.g. we assume that the vertices are labeled from 1 to n along the backbone. In the circular case the starting point of the labeling is arbitrary. The linear case is reduced to the circular case by introducing a "root vertex" 0 which is connected only to 1 and n. The Hamiltonian cycle H consisting of the vertices 0 or 1 through n and the edges $\{k-1, k\}$ and $\{0, n\}$ or $\{1, n\}$ is called the *backbone* of the structure. All other edges of the contact graph are called *bonds*.

Definition 8 *Let Γ be 1-contact structure. Consider the following edge-coloring procedure:*

1. *All bonds and all backbone edges that are contained in a stacked pair are colored in red.*

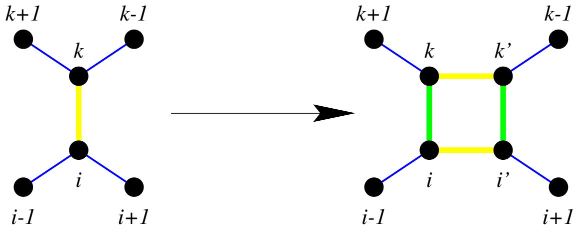


Figure 8: Modifying an IS-colored 1-contact graph in order to deal with isolated bonds.

2. *All other backbone edges are colored in blue.*
3. *Each bond that is located at the end of a stem, i.e., that has both adjacent red and adjacent blue edges is re-colored in green.*
4. *Isolated bonds, that is, red edges that have only green adjacent edges are re-colored in yellow.*

We call this edge coloring the IS-coloring of Γ*.*

It is clear that the resulting coloring is unique. An example is shown in Figure 10.

If the 1-contact structure Γ contains isolated base pairs, it will be convenient to use a modified graph Γ' with the following modified IS-coloring:

Definition 9 *The modified 1-contact structure* Γ' *is obtained from* Γ *by replacing each isolated bond* $\{i,k\}$ *by a stacked pair* $\{i,k;k',i'\}$ *such that we have the "old" backbone-edges* $\{i-1,k\}$*,* $\{i',i+1\}$*,* $\{k-1,k'\}$*,* $\{k,k+1\}$*, the two bonds* $\{i,k\}$ *and* $\{i',k'\}$*, and the two "new" backbone edges* $\{i,i'\}$ *and* $\{k,k'\}$*. The IS-coloring is modified such that the "old" backbone-edges retain their blue color, the "new" backbone-edges are colored in yellow, and the two bonds* $\{i,k\}$ *and* $\{i',k'\}$ *are colored green, see Fig 8.*

For simplicity we will refer to a modified 1-contact structure with its IS-coloring as an *IS-graph* and write $\Gamma = (V', B \cup G \cup R \cup Y)$, where B, G, R, and Y are the edges colored in blue, green, red, and yellow, respectively.

Definition 10 *Let* $\Gamma = (V', B \cup G \cup R \cup Y)$ *be an IS-graph. A* BG-subgraph *is a maximal connected subgraph of* $(V', B \cup G)$ *containing at least one edge. A stem is a maximal connected subgraph of* $(V', G \cup R \cup Y)$ *containing at least one edge.*

It is clear that each stem contains either red or yellow edges (in the latter case it represents an isolated bond). Furthermore, each stem contains exactly two green edges, its terminal base pairs.

Theorem 1 *A BG-subgraph of an IS-graph is an elementary cycle.*

Proof We show that each vertex of Γ has degree 2 in a BG-subgraph. Let $x \in V$. We have to distinguish the following cases: (i) If x is not contained in a bond, then it is incident to exactly two edges, namely either two backbone edges, a backbone edge and one of the virtual edges $\{0, 1\}$ and $\{0, n\}$, or, if $x = 0$, with both virtual edges. (ii) Suppose x is incident with a bond. If this bond is colored red, then all other edges adjacent with x are colored in red; thus x is an isolated vertex. By definition, however, a BG-subgraph does not contain isolated vertices. Hence the bond must be colored green. In this case there are exactly two other edges of the IS-graph incident with x. One of them is colored in red or yellow since green edges are obtained from re-coloring red or yellow bonds: The yellow case is clear from definition 9. In the red case, the bond is contained in a stacked pair hence has one incident red backbone edge. The other one must be blue. If it were red, then b would have been part of two stacked pairs, and hence would not have been re-colored green in the 3rd step of definition 8.

Removing the root 0 from the modified IS-graph results in the following immediate generalization of theorem 1.

Corollary 2 *The BG-subgraphs of an IS-graph Γ are elementary cycles with a single exception. The BG-subgraph containing 1 and n, which we call the exterior BG-subgraph is a connected path.*

Definition 11 *Let Γ be a IS-graph without a root. A stem Ξ is* interior *to a BG-subgraph Ψ if both green edges of Ξ are contained in Ψ. Let $\overline{\Psi}$ denote the union of a BG-subgraph and all its interior stems. A net is a two-connected component of $\overline{\Psi}$.*

With the exception of the exterior BG-subgraph, all graphs $\overline{\Psi}$ are two-connected and therefore nets of Γ.

Theorem 3 *If Γ is a secondary structure (in the classical sense) then all nets of Γ are cycle graphs, i.e., there are no stems interior to any of the BG-subgraphs of Γ.*

Proof Suppose the IS-graph Γ contains a net N with an interior stem. In this case Γ has a minor as depicted in Figure 9, which is obtained by (1) retaining only a single stem in N, (2) contracting this stem to length 2 (whether the color is red or yellow is irrelevant), and (3) retaining only a single path connecting the top and bottom cycles. Such a path must exist

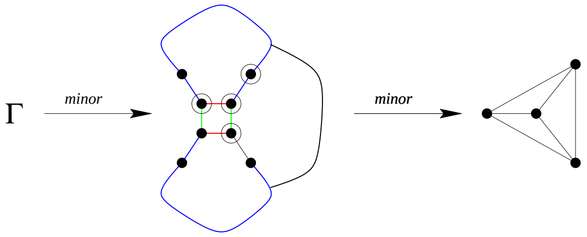


Figure 9: Proof of Theorem 3.

since the backbone must be connected, and both cycles must contain at least one blue (backbone) edge. It is clear from Figure 9 that Γ' contains K_4 as a minor. Hence Γ is not an outerplanar graph [16], and therefore not a secondary structure.

A net with exactly n interior stems will be called a n-net in the following. We call n the order of a net. A 0-net is therefore a simple cycle.

Corollary 4 *If Γ is a secondary structure graph, then the nets coincide with the "loops" of the secondary structure graph.*

Proof There are no interior stems by Theorem 3. Thus all stems connect nets. The union of the nets is therefore the union of the loops. Since the nets are edge and vertex disjoint, and so are the loops (if we replace isolated base-pairs by stems of length 2 with yellow color). Thus the nets, and equally the loops, are the exactly connected components of the union of the nets.

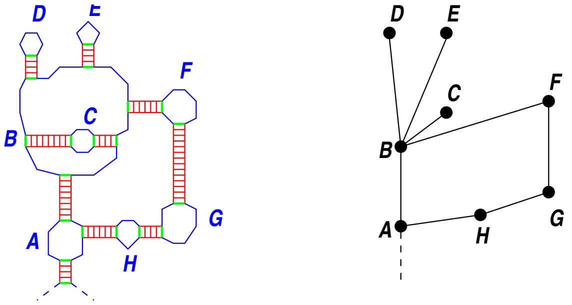


Figure 10: A counter example to the converse of Theorem 3. All nets of the 1-contact structure Γ (r.h.s.) are simple cycle, labeled A through H. Nevertheless nets B and C together with their connecting stems form a pseudoknot. The gel $\mathsf{Gel}(\Gamma)$ contains a cycle and hence is not a tree.

Remark 1 *The converse of Theorem 3 is not true as the example in Figure 10 shows.*

Definition 12 *Let* Γ *be a modified IS-graph, and let* $\mathcal{N}$ *be its set of nets. Then the Gel* $\mathsf{Gel}(\Gamma)$ *has* $\mathcal{N}$ *as its vertex set. There is an edge between two nets* N_1 *and* N_2 *if and only if there is a stem* S *that has one green edge in common with* N_1 *and the other green edge in common with* N_2 *or if* N_1 *and* N_2 *have green edges that appear one after the other on the exterior BG-path.*

Corollary 5 *The gel* $\mathsf{Gel}(\Gamma)$ *of a secondary structure (in the classical sense) is a tree.*

Proof Follows immediately from Corollary 4. It is clear from the examples in [64] that the converse cannot be true.

3 Structure prediction - State of the Art

Several methods exist for prediction of RNA secondary structure. In principle we can divide them into two broad classes: Structure prediction by *phylogenetic comparison* and *energy directed* folding.

3.1 Comparative Sequence Analysis

Given a large enough number of sequences with identical secondary structure, that structure can be deduced by examining covariances of nucleotides in these sequences. This is the principle used for structure prediction through phylogenetic comparison of homologous (common ancestry) sequences [19, 39]. Basically these methods look for compensatory mutations such as an A change to C in position i of the aligned sequences simultaneously with a change from U to G in position j, indicating a base pair (i, j). So the sequence alignment is the most complicated theoretical part (if the sequences in the set are to dissimilar).

The most common way of quantifying sequence covariation for the purpose of RNA secondary structure determination is the *mutual information* (MI) score [19, 40, 39]. The MI score of column i and j of the alignment is then given by

$$M_{ij} = \sum_{\mathsf{X,Y}} f_{ij}(\mathsf{XY}) \log \frac{f_{ij}(\mathsf{XY})}{f_i(\mathsf{X}) f_j(\mathsf{Y})} \tag{1}$$

where $f_i(\mathsf{X})$ is the frequency of base X at aligned position i, and $f_{ij}(\mathsf{XY})$ is the frequency of finding both X in i *and* Y in j.

The basic assumption is, that structure is more conserved during evolution than sequence, since it is the structure that determines function. The only experimental information needed is a large enough number of sequences. Fortunately nucleic acid sequences are nowadays one of the best accessible molecular biological informations. In fact the success of the method in the prediction of, for instance, the secondary structures of the 16S ribosomal RNAs, RNaseP RNA or the clover-leaf structure of tRNAs provides an excellent justification for this method. Since no assumptions about pairing rules are necessary, non-canonical pairs and tertiary interactions can be detected as well.

One limitation of this approach is, that a sufficiently large set of sequences which exhibit the proper amount of variation has to be provided. Another difficulty with determining the consensus structure by comparative analysis is in obtaining a good alignment of the sequences. The computer-aided recognition of strongly correlated positions in a multiple sequence alignment is followed by manual refinement of the alignment, which is an iterative, laborious process.

Nevertheless, phylogenetic comparison can generate the most reliable structure models to date and are therefore frequently used for comparison with other folding algorithms.

3.2 Thermodynamic Prediction of Secondary Structure

3.2.1 The Energy Model

The standard energy model currently used is based on the loop decomposition, introduced in the chapter 2.3.1, and assumes that the energy of a structure can be obtained as the sum over the energies of its constituent loops.

$$E(\mathcal{S}) = \sum_{l \in \mathcal{S}} E(l) \tag{2}$$

Because the energy contribution of a pair in the middle of a helix depends only on the following and previous pair, such energy rules have been termed “nearest-neighbor” rules.

To keep the number of parameters manageable, loop energies are generally split in two terms, describing the size and sequence dependency, respectively. Moreover, the sequence dependent part only considers the base pairs delimiting the loop and unpaired positions adjacent to these pairs. This still leaves a large number of parameters not all of which have been experimentally determined. The missing parameters are replaced by estimates based on physical intuition, or have been optimized to yield reasonably good predictions.

An up-to-date compilation of energy parameters for RNA was published in 1999 [93] and is available for download from the Turner group site at `http://rna.chem.rochester.edu/index.html`.

Stacking energies Energies of stacked base pairs are the most carefully measured parameters. They are particularly important since stacked base pairs provide most of the stabilizing energy for secondary structures. Values for Watson Crick pairs were among the first parameters to be measured [7], and recently modified by including a penalty for A·U and U·A pairs at the end of helices [163]. Stacking energies involving G·U pairs were added later [47] and demonstrated the shortcoming of the nearest neighbor model: The energy of the double G·U mismatch $\substack{5'\text{GU}\,3'\\3'\text{UG}\,5'}$ depends on its context. It is energetically favorable e.g. in the context $\substack{5'\text{GGUC}\,3'\\3'\text{CUGG}\,5'}$, but more often unfavorable, as in $\substack{5'\text{CGUG}\,3'\\3'\text{GUGC}\,5'}$ or $\substack{5'\text{UGUA}\,3'\\3'\text{AUGU}\,5'}$. Programs have to either look at the context beyond the nearest-neighbor model, or use some average value.

Hairpin Loops Hairpin energies are approximated as the sum of a size dependent destabilizing term plus a *mismatch* energy, which contains the favorable stacking interactions between the closing pair and the adjacent unpaired bases. Mismatch energies are not used for hairpins of size 3, which are assumed to be too tightly packed to allow stacking. The size dependent loop energy for small loops has been estimated from melting experiments, values for large loops are extrapolated logarithmically. Mismatch energies for the $6 \cdot 4 \cdot 4$ possible combinations are tabulated. Certain tetra-loops (hairpins of size four) occur much more frequently than expected in known RNA secondary structures, such as ribosomal RNA [160]. The current parameter set lists 30 such special tetra-loops and awards them bonus energies between -1.5 and -3 kcal/mol. Finally, the Turner parameters recommend a special penalty for poly-C loops [35], and a bonus of -2.2 kcal/mol for loops closed by G·U when the two bases preceding the G are also Gs [31].

Interior loops For small loops, the current energy set simply tabulates all energies instead of using the formula. This is done for 1×1 loops (a single mismatch interrupting the helix), 1×2 (size 3) interior loops, as well as symmetric 2×2 loops (two consecutive mismatches). Otherwise, interior loop energies contain a size-dependent term and mismatch energies. In addition, interior loop energy depends on the asymmetry of the loop $|n_1 - n_2|$, where n_1 and n_2 are the length of the two unpaired regions, respectively.

$$\Delta G_{\text{int.loop}} = \Delta G_{\text{size}}(n_1 + n_2) + \Delta G_{\text{asym}}|n_1 - n_2| + \Delta G_{\text{mismatch}}. \tag{3}$$

where ΔG_{size} is again tabulated for sizes up to 6 and then extrapolated. ΔG_{asym} is supposed to increase linearly up to 3kcal/mol, and mismatch energies are tabulated. Bulge loops (where all unpaired bases occur on one side) use their own tables for ΔG_{size} and a penalty for A·U or G·U pairs delimiting the loop. Also, it is assumed that bulges of size 1 do not interrupt the helix geometry, and therefore the stacking energy for the two pairs is added as well.

Multi-loops To date there are almost no thermodynamic measurements on multi-loops available. Consequently, multi-loop energies present the largest source of inaccuracy in the energy model. Furthermore, dynamic programming algorithms need an energy function that is linear in the loop size for efficient treatment of multi-loops. The usual ansatz for multi-loop energies is therefore

$$\Delta G_{\mathrm{ML}} = a + b \cdot n + c \cdot k + \Delta G_{\mathrm{dangle}}, \tag{4}$$

where n is the loop size and k the loop degree. $\Delta G_{\mathrm{dangle}}$ is an energy bonus describing the stacking interactions between a pair and one adjacent unpaired base, i.e. dangling ends work much like mismatch energies except that the mismatch energy is split into two parts stemming from the unpaired base $5'$ and $3'$ of the pair, respectively.

3.2.2 Dynamic Programming Algorithms

The additive form of the energy model allows for an elegant solution of the minimum energy problem through dynamic programming that is similar to sequence alignment. This similarity was first realized and exploited by Michael Waterman [152, 153]. His observation was the starting point for the construction of reliable energy-directed folding algorithms [56, 171].

The first dynamic programming solution was proposed by Ruth Nussinov [107, 108] originally for the "maximum matching" problem of finding the structure with the maximum number of base pairs. Michael Zuker and Patrick Stiegler [171, 172] formulated the algorithm for the minimum energy problem using the now standard energy model. Since then several variations have been developed: Michael Zuker [170] devised a modified algorithm that can generate a subset of suboptimal structures within a prescribed increment of the minimum energy. The algorithm will find any structure $\mathcal{S}$ that is optimal in the sense that there is no other structure $\mathcal{S}'$ with lower energy containing all base pairs that are present in $\mathcal{S}$. As shown by John McCaskill [96] the partition function over all secondary structures $Q = \sum_S \exp(-\Delta G(S)/kT)$ can be calculated by dynamic programming as well. In addition his algorithm can calculate the frequency with which each base pair occurs in the Boltzmann weighted ensemble of all possible structures, which can conveniently be represented in a dot-plot.

The memory and CPU requirements of these algorithms scale with sequence length n as $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$, respectively, making structure prediction feasible even for large RNAs of about 10000 nucleotides, such as the entire genomes of RNA viruses [54, 62].

`RNAfold` as part of the `Vienna RNA Package`[1] [56] reads RNA sequences from `stdin` and calculates their mfe structure, partition function and base pairing probability matrix [50, 96]. 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 4. The results of `RNAfold` are used as an input for `alidot` [54, 56].

3.3 Thermodynamic Prediction of Secondary Structure Including Pseudoknots

Folding an RNA sequence of length n into a secondary structure based on the nearest neighbor model requires $\mathcal{O}(n^2)$ time and $\mathcal{O}(n^3)$ memory. Whereas the prediction of RNA structure including all types of pseudoknots based on the same model has been proven to be NP-complete [88, 3]. However, for structure predictions including certain types of pseudoknots polynomial algorithms have been developed [37, 45, 122, 124]. Furthermore a number of algorithms which adopt heuristic search procedures exist.

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

3.3.1 Energy Models for Pseudoknots

For pseudoknots, there is not much thermodynamic information available. Experimental measurements of some model pseudoknots have shown them to be only marginally more stable than the secondary structures involved [162, 147]. Since there are no measured thermodynamic parameters for pseudoknots, we rely on approximations for the free energy of pseudoknots.

Gultyaev *et al.* [37] conceived an approximative model for H-type pseudoknots, which is described in the following. The free energy of an H-pseudoknot is mainly the sum of the free energies of stacking in the stems (stabilizing negative values), and the entropy-term of the destabilizing positive loop values. The free energy of the stems are calculated using the standard energy model for secondary structures. For the loop energies some estimate is needed. The loops are modeled as purely entropic. Using the Jacobson-Stockmayer equation [66] the free energy ΔG of formation of a loop of N nucleotides is approximated by

$$\Delta G = RT(A_{loop} + 1.75 \ln N), \tag{5}$$

where A_{loop} is a constant related to the loop type.

The model is restricted to H-pseudoknots with $|L3| = 0$. The two remaining loops are not equivalent stereo-chemically, loop $L1$ spans the deep groove of RNA stem $S2$, whereas $L2$ crosses stem $S1$ in the shallow groove [116]. Furthermore the features of the loops are dependent on the length of the corresponding stems. This is taken into account by introducing two variables $A_{deep}(S2)$ and $A_{shallow}(S1)$. Minimal loop sizes are required for bridging a stem, they are denoted by $N_{mindeep}(S2)$ and $N_{minshallow}(S1)$, respectively. Instead of just using a logarithmic increase of entropy with loop size, the dependence on the difference between the loop length and the minimally allowed length is introduced. Such an approximation can partially reflect restrictions of conformational freedom imposed by the stem end-to-end distance. Considering all these assumptions we have:

$$\Delta G_{L1} = A_{deep}(S2) + 1.75RT\ \ln(1 + N - N_{mindeep}(S2)) \quad (6)$$

$$\Delta G_{L2} = A_{shallow}(S1) + 1.75RT\ \ln(1 + N - N_{minshallow}(S1)) \quad (7)$$

Sequences of known pseudoknots that are verified by experiments and/or phylogenetic comparisons were used to estimate the parameters, assuming that the free energies of these pseudoknots are lower than those of corresponding hairpins formed by the pseudoknot stems.

Another approach for modeling free energies of secondary structures including pseudoknots has been proposed by Isambert and Siggia [64, 63]. The model is based on the Isambert-Siggia decomposition of secondary structures (see section 2.3.5), restricted to nets with a maximum order of 2. They distinguish between closed nets and open nets, see Fig. 11. Open nets are subgraphs of the exterior BG-subgraph, which are continuous sections of the path that contain a minimal number n of internal stems.

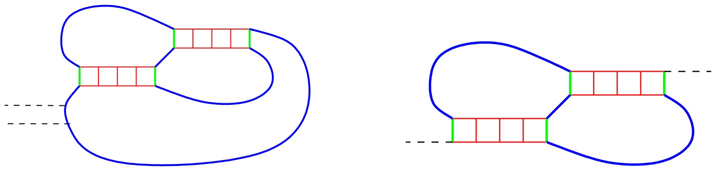


Figure 11: Closed and open nets: Example of a closed 2-net (*l.h.s.*) and an open 2-net (*r.h.s.*)

The free energy of a net is composed of the free energy of the stems, calculated using the thermodynamic parameters for base stacking [132], and the entropy of the net which is calculated using polymer theory [26]. The stems are modeled as rigid rods and the unpaired regions as Gaussian chains. The entropy of the gel is evaluated assuming that the vertices of the gel are connected by Gaussian springs. The conformational entropy of such a "Gaussian crosslinked gel" is then calculated numerically via $n - 1$ algebraic integrations, where n is the number of nets constituting the gel. The free energy of a structure is composed of the free energy of all nets, the stacking energies of stems not contained in a net, and the entropy of the gel.

3.3.2 Algorithms

Rivas and Eddy presented a dynamic programming algorithm which requires $\mathcal{O}(n^6)$ time and $\mathcal{O}(n^4)$ memory [124]. The algorithm is based on the nearest neighbor model. For the nested structures, they used the standard energy model described in section 3.2.1, for pseudoknots, they introduce a number of new parameters, which where tuned by hand, some of the pseudoknot-parameters are obtained by multiplying similar parameters for unknotted structures by a weighting parameter. The time and memory complexity of the algorithm restricts the length of sequences that can be analyzed to 130-140 bases. The program is available at `http://www.genetics.wustl.edu/eddy/software/#pk`. The type of pseudoknots included in their model is given implicitly by their recursion scheme. Furthermore, in another publication, Rivas and Eddy presented a formal grammatical representation for RNA secondary structure with pseudoknots [123], and the specific grammar that corresponds to the parsing algorithm for structure prediction by

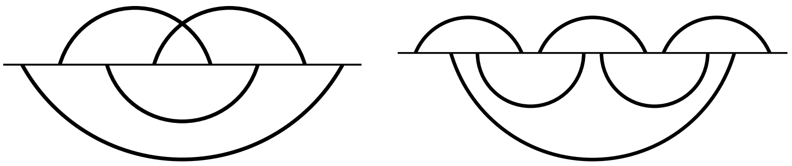


Figure 12: Structures exemplifying the class of structures the algorithm of Rivas and Eddy [124] minimizes over, helices are drawn as arcs. A non-planar structure in the class of structures minimized over (*l.h.s.*), and a planar structure not in that class (*r.h.s.*).

dynamic programming is given. The pseudoknot model allows for rather complex structures, even some non-planar structures (including the pseudoknot of α-mRNA), however, not all planar structures are included in this model as illustrated in Figure 12.

A dynamic programming algorithm, which achieves $\mathcal{O}(n^4)$ time and $\mathcal{O}(n^2)$ memory has been presented by Reeder and Giegerich [122]. The algorithm includes H-type pseudoknots, and the improvement of time and space complexity results from considering only so-called canonical pseudoknots. A pseudoknot is called canonical, if the two helices facing each other have maximal extent, i.e. $L3$ is as short as possible. For structures containing no pseudoknots the standard energy model model is used, for pseudoknots the energy is computed with a model similar to that used by Rivas and Eddy [124]. The application of the algorithm is limited to sequences of length up to 800 bases. A web interface for online RNA folding is available at `http://bibiserv.techfak.uni-bielefeld.de/pknotsrg/`.

The dynamic programming algorithm presented by Haslinger [45] requires $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ memory. It includes restricted H-pseudoknots, i.e. $|L3| < 2$ and the helices forming the pseudoknot may contain one symmetric interior loop consisting of two unpaired bases or one bulge formed by one unpaired nucleotide. Furthermore pseudoknots are not allowed to be interior to a base pair of the surrounding secondary structure. The algorithm is based on the energy model of Gultyaev [37] (Section 3.3.1).

Other methods which are capable for RNA folding including pseudoknots, adopt heuristic search procedures and sacrifice optimality. Examples of these approaches include quasi-Monte Carlo searches [1] and genetic algorithms by vanBatenburg *et al.* [150] and by Gultyaev *et al.* [38]. The method of Alexander Gultyaev *et al.* [38] allows to simulate a folding pathway of RNA, including such processes as disruption of temporarily formed structures. His findings however critically depended upon intrinsic parameters of the genetic algorithm, like population size or chain growth rate. A kinetic Monte Carlo algorithm has been presented by Isambert and Siggia [64].

3.4 Combination of Phylogenetic and Thermodynamic Structure Prediction

Comparative sequence analysis requires the knowledge of a large number of homologous RNA sequences, which is not always available. Minimum free energy structures, as predicted by dynamic programming based on a single sequence, show about 73% average accuracy (compared to a large database of known secondary structures) for sequences of less than 700 nucleotides [93, 94]. Several algorithms have been developed to combine phylogenetic and thermodynamic structure prediction to predict the consensus structure for a small set of related RNA sequences. Those methods fall into two broad groups: algorithms starting from a multiple sequence alignment and algorithms that attempt to solve the alignment problem and the folding problem simultaneously.

3.4.1 Algorithms Based on a Set of Unaligned Sequences

Sankoff [127] proposed that a dynamic programming algorithm could solve the alignment and folding problem simultaneously for a set of N sequences of length n. The algorithm requires time n^{3N} and storage n^{2N}, i.e. n^6 for the prediction of the consensus structure of two sequences.

Gorodkin *et al.* [33] reduced the time complexity to $\mathcal{O}(n^4)$ for predicting the structures of two sequences by optimizing the number of base pairs instead of the free energy and by forbidding multibranch loops in their algorithm FOLDALIGN.

Another algorithm based on the recursion of Sankoff has been given by Mathews and Turner: DYNALIGN [95]. They introduce an upper bound, M, for the maximum distance between aligned nucleotides, and restrict themselves to two sequence alignments. This reduces the complexity to $\mathcal{O}(M^3n^3)$.

Perriquet *et al.* [114] recently presented an algorithm called CARNAC for pairwise folding of unaligned sequences which has an empirically observed complexity of about $\mathcal{O}(n^4)$. The first step is generating a list of possible stems, using stacking energies and tetra-loop bonuses [93], only stems with lower energy then a (distance dependent) threshold are taken into account. Then so-called anchor points are detected in the primary sequence alignment, using classical recursions for sequence alignment, except that indels are not allowed. Pairs of matchable stems are created, depending on consistency with anchor points and covariation. The subset of the matchable stems forming a secondary structure with lowest energy is found by dynamic programming. The recursion formula is similar to the formula of Sankoff, the improved time complexity is achieved by restriction of the search space and by considering potential stems, not single base pairs.

Notredame *et al.* [106] developed a genetic algorithm (RAGA) that finds the structure of a sequence given a second, related sequence with known structure. Chen *et al.* [17] apply a genetic algorithm to a set of related RNA sequences to find common RNA secondary structures. The fitness function is based on free energy of a structure and a measure of structure conservation among the sequences.

Hofacker *et al.* [52] use a variant of Sankoff's algorithm [127], called `pmmach`. Instead of attempting to solve the folding and the alignment problem simultaneously, they align base pairing probability matrices, which were obtained by McCaskill's algorithm, and which efficiently incorporate the information on the energetics of each sequence. A novel, simplified variant of Sankoff's algorithms can then be employed to extract the maximum weight common secondary structure and an associated alignment. In [53] Hofacker *et al.* describe an algorithm (`pmz`), which uses stochastic backtracking to compute e.g. the probability that a prescribed sequence-structure pattern is conserved between two RNA sequences. Their matrix of matching probabilities can be computed $O(n^4)$ memory and $O(n^6)$ CPU.

3.4.2 Algorithms Based on a Multiple Sequence Alignment

Most of the alignment based methods start from thermodynamics-based folding for each sequence and use the analysis of sequence covariations or mutual information for post processing.

Le and Zuker [83] presented an algorithm that generates a number of suboptimal structures (whose energy is close to the minimum free energy) for each sequence, and helices, identical in position and occurring in most of the sequences, are combined into a consensus structure.

The program `alidot` developed by Hofacker and Stadler [58] is based on the base pairing probabilities calculated by RNAfold [50] for each sequence. The sequence covariation is taken into account by assigning a bonus to base pairs where different pairing combinations occur (refer to Section 3.5).

Lück *et al.* [86] also take the base pairing probabilities of each sequence as starting point, sequence covariation is taken into account by means of the MI score.

Hofacker *et al.* [51] have developed an algorithm (`alifold`) which integrates the thermodynamic and phylogenetic information into a modified energy model to predict the consensus secondary structure of a set of aligned RNA sequences.

Juan and Wilson [68] assign an energy score to each potential pairing region that does not in any way account for the entropic cost of closing the loop between the two unpaired regions. For this reason a term that penalizes large loop formation is added, and including a covariation score this gives the overall score for a helix. A secondary structure, or a structure including pseudoknots, respectively, is then progressively built depending on the scores.

Tabaska et al. [140] described a method based on the Maximum Weighted Matching algorithm for RNA structure prediction including pseudoknots from an alignment of homologous RNA sequences. To each possible base pair, that can be formed, a weight is assigned. This gives a weighted graph, where the nucleotides form the vertex set, and the edge set is built from all base pairs with positive weight. With the help of the MWM algorithm the matching which has the maximum total weight is extracted. Helices with a length shorter then 3 base pairs are removed from the outcome. An additional way of output filtration is the removal of base pairs that have been rematched during the run of the MWM algorithm. They present different methods for assigning edge weights, which may be combined into hybrid sets. Helix plots combine phylogenetic and thermodynamic information to yield base pair scores. For each sequence of the alignment of length N a $N \times N$ scoring matrix is generated. A 'good pair' score is assigned for Watson-Crick and G-U pairs, a larger negative 'bad pair' score for every other type of base pair, and an even larger 'paired gap' score for base-gap. Then the entries with positive score are scanned for potential helices, all base pairs in helices with length smaller then 3 receive the 'bad pair' score, and to all base pairs in helices with length greater then 3 a bonus score proportional to the length of the helix is added. Then the individual scoring matrices are summed, this gives the scores for MWM. Other scoring methods used include the use of MI scores, and a thermodynamic score based on calculating the minimum free energy of the structure containing a given base pair by means of `mfold` [170].

A related approach `ilm` by Ruan *et al.* [125] uses the same weight matrix as Tabaska's program but replaces the solution of the MWM Problem by an iterated loop matching algorithm. One first computes the Maximum Circular Matching [108] to obtain a pseudoknot-free secondary structure and then repeats the computation on the remaining un-paired bases in order to insert pseudoknots, iterating the procedure until no further base pairs can be found. This approach, which is implemented in the program `ilm`, appears to reduce the number of spurious base pairs and works well on alignments of smaller sets of sequences.

The algorithm `hxmatch` developed by Witwer *et. al.* [159] uses MWM but differs from Tabaska's approach in two respects: (i) it uses a significantly improved scoring scheme for assigning edge weights (described in the following), (ii) in the post-process of the results it is restricted to bi-secondary structures as described in 2.3.3. The scoring matrix is generated from the combination of a thermodynamic score, derived from the stacking energies of helices, and a covariance score, which is based on the number of mutations for a given alignment position. The thermodynamic score is derived from the energy of each helix which is calculated using the experimentally determined standard energy model for thermodynamic RNA folding [93]. The weight of a base pair in each sequence is the energy of the longest helix the base pair is part of. The entry in the combined scoring matrix for a specific base pair in the alignment is then given by the sum of weights for this base pair over all sequences. In order to give more weight to conserved helices Witwer applies a covariance score, which was originally introduced in [55]. Thus positive score is attributed to compensatory and consistent mutations, a score of 0 is given to not "legal" base pairs. Additionally the fraction of inconsistent sequences for a specific base pair incurs a penalty. Thus thermodynamic helix score, covariance score and non-consistent penalty score are combined to a combined weight. In a next step all optimal helices of length at least 3 are collected and a weight of the helix is determined by the sum over all combined weights of its base pairs. Finally to each base pair Witwer assigns the weight of the helix with the largest weight that passes through it. The `hxmatch` algorithm uses $O(n^3)$ time and $O(n^2)$ memory, where n is the length of the alignment.

3.5 The Algorithm alidot

alidot (ALIgned DOT-plots) [54, 58, 50], has been used for the prediction of conserved secondary structure of the genomes of the virus family *flaviviridae* (see Section 4), therefore this section gives a detailed description of the algorithm.

The method requires an independent thermodynamic prediction of the secondary structure for each of the sequences and a multiple sequence alignment that is obtained without any reference to the predicted secondary structures. In this respect alidot is similar to programs such as construct [86, 85] and x2s [68], see also [83]. In contrast to efforts to simultaneously compute alignment and secondary structures e.g. [127, 33, 114, 95] this approach emphasizes that the sequences may have common structural motifs but no global common structure. In this sense alidot combines structure prediction and motif search [22].

The algorithm implements a combination of thermodynamic structure prediction and phylogenetic comparison. In the first step a set of thermodynamically plausible candidate base-pairs is obtained by computing the matrix of base pairing probabilities using McCaskill's partition function algorithm [96] for each sequence and retaining all pairs with a thermodynamic equilibrium probability greater than 3×10^{-3}. The computations were performed using the Vienna RNA Package [56], based on the energy parameters published in [93].

The multiple sequence alignments can be obtained, for example, using ClustalW [148], dialign [99], or code2aln [139]. The quality of the alignment has a strong effect on the results, as small errors in the alignment can easily hide a conserved feature. While false positives remain rare, the number of conserved structures that are found decreases with the diversity of the sequences analyzed, when using an automated alignment. Best results are obtained when the sequence diversity is large enough to provide many compensatory mutations, but low enough to allow accurate alignments, typically at pairwise identity of, say, 80%.

The gaps in the alignment are then inserted into the corresponding probability matrices. Now it is possible to superimpose the probability matrices of the individual sequences to produce a *combined dot plot*. 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 a color coding is used 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 Fig. 14.

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 $(i.j)$, this is called a *consistent* mutation. If there are combinations such as **GC** and **CG** or **GU** and **UA**, where both positions are mutated at once it is called a *compensatory* mutation. The occurrence of consistent and, in particular, compensatory mutations strongly supports a predicted base pair, at least in the absence of non-consistent mutations.

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 two conditions for secondary structure defined in section 2.3.1 The remainder of this section describes how to extract credible secondary structures from the list of base pairs. The individual base pairs are ranked 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

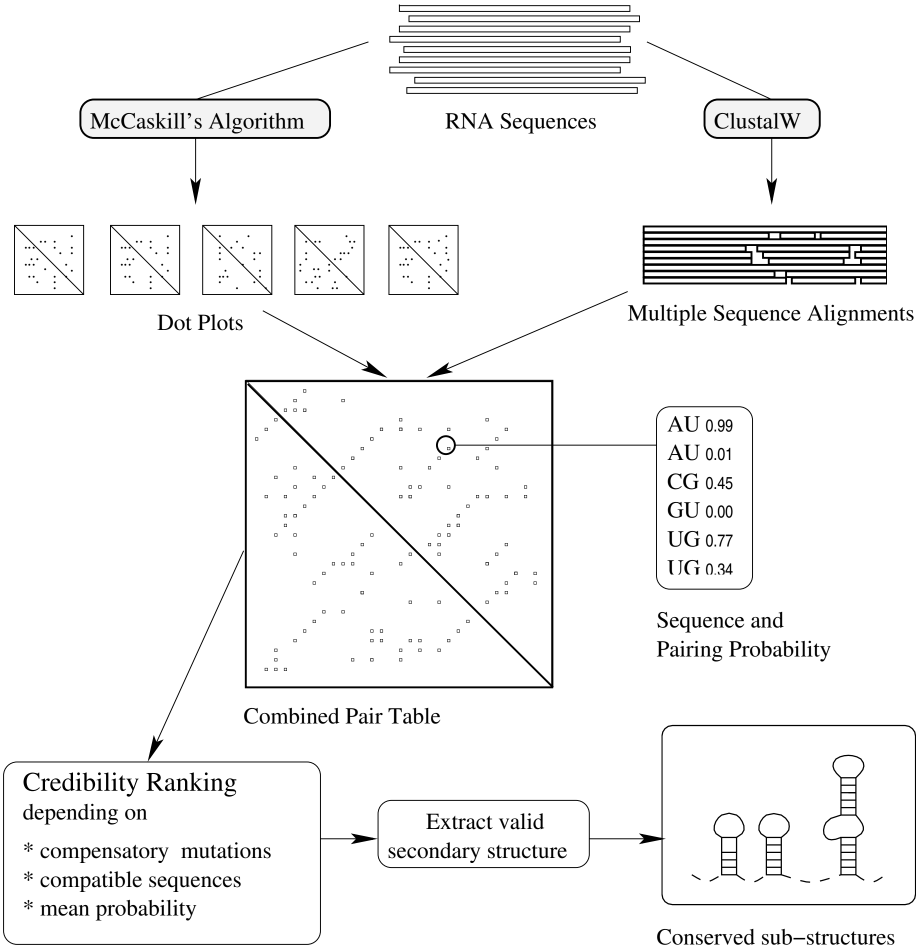


Figure 13: Flow diagram of the algorithm. A multiple sequence alignment is calculated using, for instance, `ClustalW`. RNA genomes are folded using McCaskill's partition function algorithm as implemented in `RNAfold`. The alignment and the structure predictions are joined together to the combined pair table. The sequence information and the mean pairing probability of the base pairs provide the basis of the credibility ranking. In the final step a valid secondary is extracted of the ranked list of possible base pairs.

number of different pairing combinations.

Then the sorted list is scanned and all base pairs that conflict with a higher ranked pair by violating conditions (i) or (ii) are removed.

The list now represents a valid secondary structure, albeit still containing ill-supported base pairs. A series of additional “filtering” steps is used to minimize the number of false positives: First, all pairs with more than two non-compatible sequences are removed, as well as pairs with two non-compatible sequences adjacent to a pair that also has non-compatible sequences. Next, all isolated base pairs are omitted. 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 `alidot` program. A flow diagram of the algorithm is given in Fig. 13.

Results are presented as conventional secondary structure drawing, as *colored* mountain plots, see Fig. 14, or dot plots. Colored mountain plots and dot plots contain information about both sequence variation (color code) and thermodynamic likeliness of a base pair (indicated by the height of the slab and the size of the dot, respectively). Colors in the order red, ocher, green, cyan, blue, violet indicate 1 through 6 different types of base pairs. Pairs with one or two inconsistent mutation are shown in (two types of) pale colors.

In the conventional graphs paired positions with consistent mutations are indicated by circles around the varying position, Fig. 14 shows an example of an annotated structure drawing. Compensatory mutations thus are shown by circles around both pairing partners. Inconsistent mutants are indicated

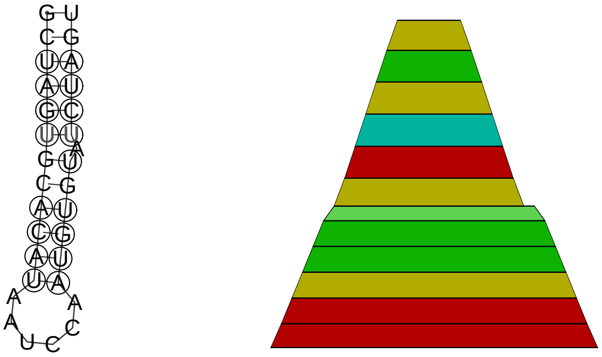


Figure 14: Annotated structure drawing (left) and colored mountain plot (right). The example shows a conserved secondary structure element located in the 5'-non-coding region of the rhinovirus genome. In the conventional drawings, consistent and compensatory mutations are indicated by circles around bases that have mutations. Gray letters indicate inconsistent mutations. Colors indicate the number of consistent mutations 🟥 1, 🟨 2, 🟩 3, 🟦 4 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 non-compatible sequences.

by gray instead of black lettering.

Vienna RNA Viewer. 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` [24]. 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 presents the colored dot plot, allows zooming in, and, for example, the drawing of a colored `Mountain Plot` and an annotated conventional secondary structure representation for a region enclosed by a selected base pair.

4 Conserved Structures in Flavivirus Genomes

The family *Flaviviridae* is subdivided into the three genera *Flavivirus*, *Pestivirus*, *Hepacivirus* and the group of GB virus C/hepatitis G viruses (GBV-C) with a currently uncertain taxonomic classification [151]. They are small enveloped particles which possess a single stranded positive sense RNA genome (ss+RNA). Although the different genera have diverse biological properties and do not show serological cross-reactivity, they appear to be similar in terms of virion morphology and genome organization (Fig. 15).

The RNA genome has a size of $9.6 - 12.3$kb and in all genera is the only mRNA found in infected cells. It consists of one single long open reading frame containing the information first for structural proteins, necessary for the construction of the virus capsid and membrane, followed by several non-structural proteins (NS) including a protease, a helicase and an RNA-dependant RNA-polymerase. Viral proteins are synthesized as one single polyprotein, which is co- and posttranslationally cleaved by viral and cellular proteinases.

The coding region is flanked by a 5' and 3' untranslated region (UTR). These are known to form into specific secondary structures required for genome replication and translation. Based on the control of these two processes it is useful to consider two subgroups of *Flaviviridae*: The first group is formed by the genus *Flavivirus* and is characterized by a type I cap structure at the 5'UTR [8] and a highly structured 3'UTR. In this group there is evidence that the 5' and 3' ends stack together to cause a cyclization of the genome (sometimes referred to as “panhandle structure”) which might be an important feature for RNA-replication [42, 72]. These regions are called “cyclization domains”.

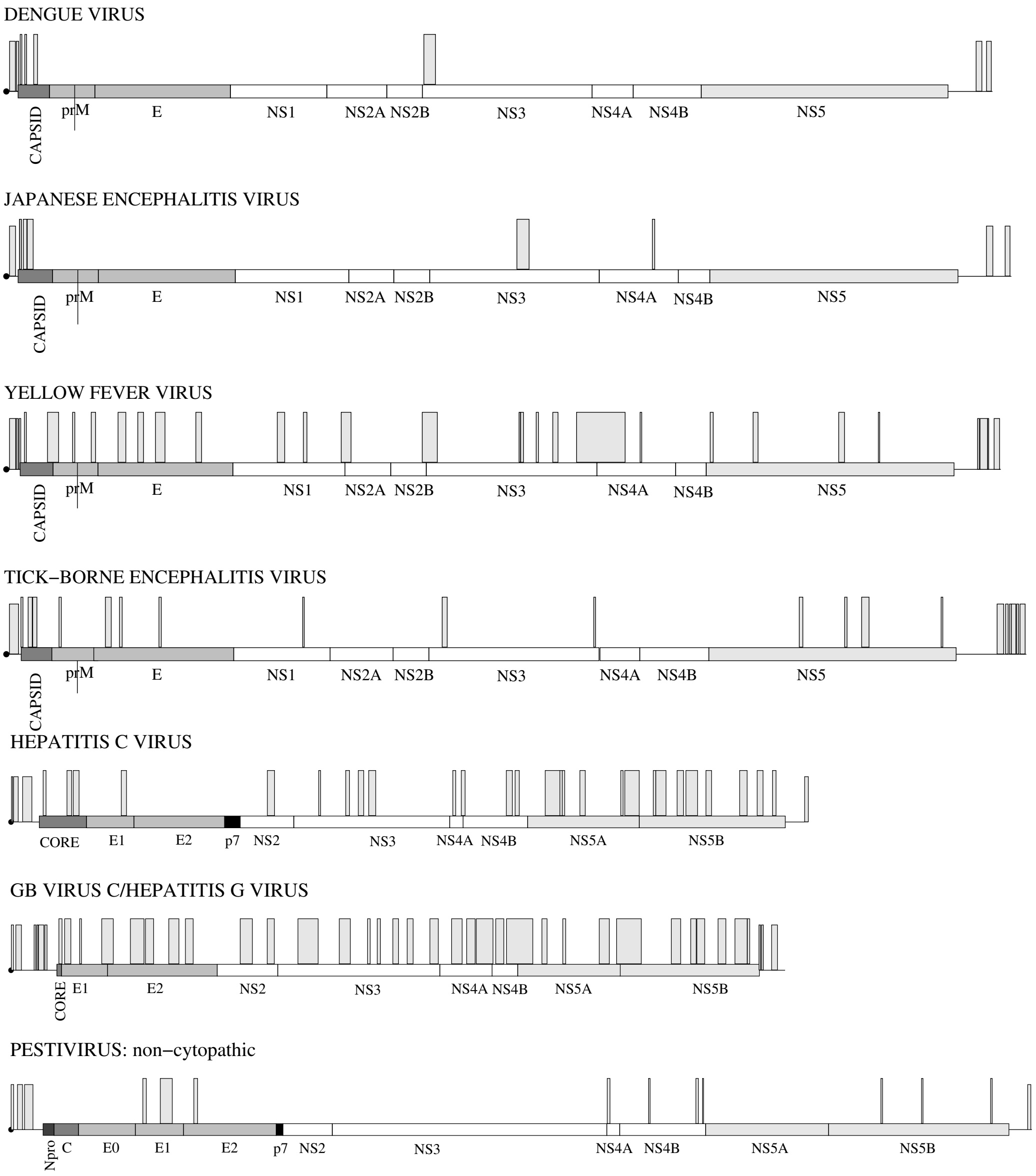


Figure 15: Genome map for the Flavivirus groups dengue virus, Japanese encephalitis virus, yellow fever virus and tick-borne encephalitis virus and hepatitis C virus and GB virus C/hepatitis G virus and the genus *Pestivirus*. Putative conserved secondary structures are indicated by the boxes above the RNA sequence.

Table 1: Number of analyzed sequences N, length of our alignments, length of 5'UTR, IRES (if present), coding region and 3'UTR with the respective mean pairwise sequence identities σ.

Group	N	length	σ	5'UTR	σ	IRES	3'UTR	σ
Group 1: *Flavivirus*								
DEN	16	10775	80.4	1..98	87.4		10289..10775	85.4
JEV	17	10979	95.5	1..95	98.9		10395..10979	95.6
YFV	7	10863	96.4	1..121	99.8		10352..10863	91.7
TBE	6	11143	86.5	1..132	91.2		10375..11143	69.8
Group 2: *Pestivirus*, *Hepacivirus*, and GB virus C/hepatitis G Virus								
GBV-C	10	9397	89.8	1..556	94.2	45-556	9086..9397	96.7
HCV	9	9679	87.1	1..342	98.3	43-354	9400..9679	85.1
PESTI	11	12393	74.9	1..388	80.7	65-388	12115..12393	62.0

The second group consisting of hepatitis C virus (HCV), GBV-C, and the genus *Pestivirus* (PESTI), controls translation by means of an IRES in the 5'UTR and has a short, less structured 3'UTR. PESTI and HCV have very similar IRES regions [115]; the IRES of GBV-C is 50% longer and structurally quite different [134]. Here we will treat these two groups separately.

While the 5' and 3'UTRs of the members of the virus family *Flaviviridae* have been the object of several studies, very little is known, however, about the secondary structures of the coding regions despite some evidence that the coding region might also contain functional RNA motifs [133, 149].

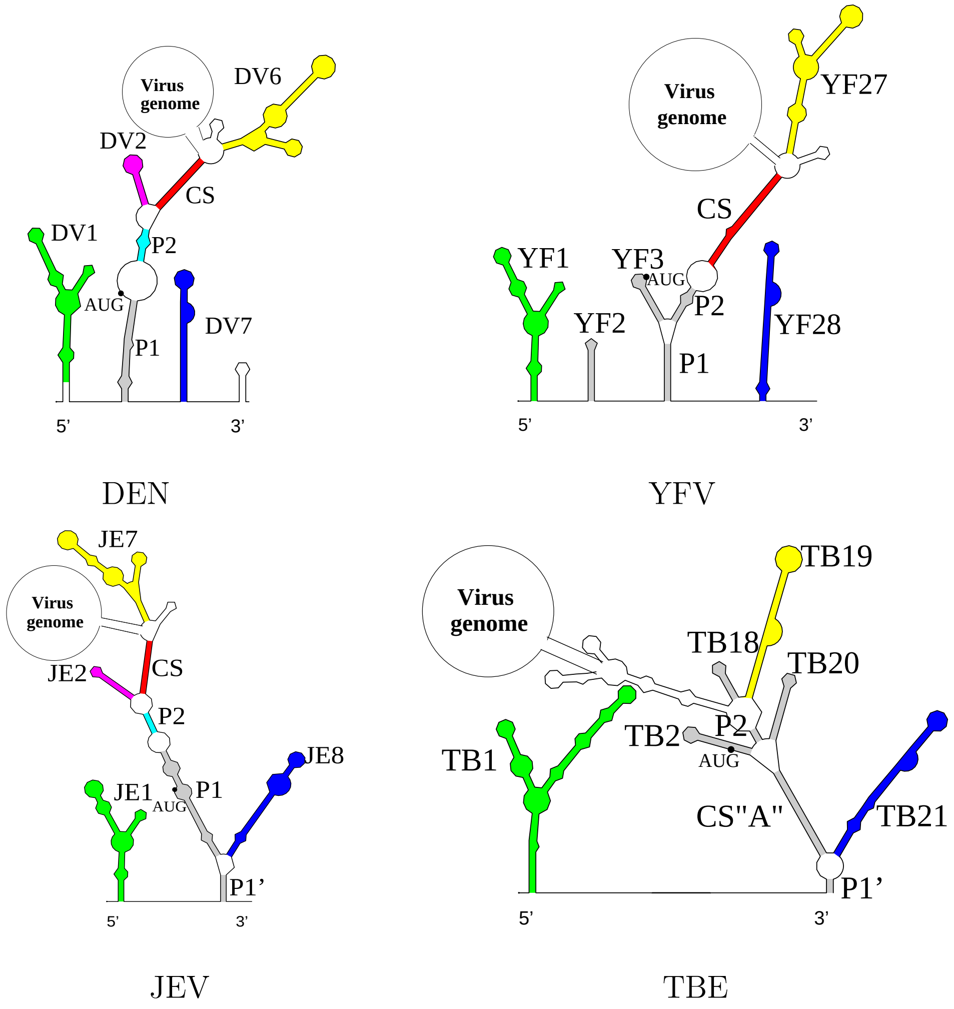
4.1 Flavivirus

The genus *Flavivirus* comprises almost 70, mostly arthropod-borne viruses including a number of human pathogens of global medical importance, such as yellow fever virus (JFV), Japanese encephalitis virus (JEV), dengue virus (DEN) and tick-borne encephalitis virus (TBE) and many others. Most members of this genus are transmitted to vertebrates by chronically infected tick- or mosquito-vectors. The spectrum of diseases caused ranges from a mild fever to hepatitis, hemorrhagic disease and encephalitis. Since at least 6 sufficiently diverged genomic sequences of each group are necessary for our analysis method, we focused on the groups DEN, JEV, YFV and TBE. In App. A.1 we define the sequences used for the respective group.

In Fig. 15 we show a genome map of the different virus groups and indicate the localization of conserved secondary structures we found relative to the genome regions. Tab. 1 lists the length of the alignments, the number of the respective aligned sequences for all investigated virus groups and the mean pairwise sequence identities of the single genome regions.

Table 2: Sequence positions of putative motives which take part in genome cyclization for the Flavivirus groups DEN, JEV, YFV and TBE. P1', P1, P2 and CS/CS"A" are shown in Fig. 16.

	DEN	JEV	YFV	TBE
P1'		73-10970		109-11143
		78-10965		111-11141
P1	83-10710	108-10883	134-10779	
	98-10694	112-10879	140-10773	
P2	112-10686	79-10910	100-10789	166-10958
	117-10681	102-10887	109-10780	168-10956
CS/CS"A"	141-10680	136-10876	147-10767	115-11073
	151-10670	146-10866	165-10750	129-11059



DEN

YFV

JEV

TBE

Figure 16: The minimum free energy structure of one sequence of the respective virus group is represented. Colored backgrounds mark regions which the folding algorithm and selection criteria allowed for all sequences of the respective group. The same color is used for equivalent structures in different groups, gray motifs are conserved only within a single group. The nomenclature of the structures corresponds to the website atlas of structures, see `rna.tbi.univie.ac.at`. Conserved secondary structures where 5' and 3'UTRs are involved in genome cyclization are called P1', P1, and P2; CS and CS"A" are taken from [42] and [72] respectively. Distances along the x-axis are not to scale; the exact positions of the structure elements are given in Tab. 2

Fig. 16 shows an overview of the conserved secondary structure elements of the 5' and 3' ends and the cyclization domains called P1', P1, P2, and CS or CS"A", respectively, found by applying the alidot algorithm. In Tab. 2 the exact positions of the respective cyclization regions are listed.

4.1.1 Genome Cyclization

Hahn *et al.* found complementary sequences (cyclization sequences "CS") close to the 5' end and the 3' end of the genome and concluded that the two ends of the genome of the genus *Flavivirus* stick together in a panhandle-like structure [42]. Recently, it has been shown that RNA synthesis *in vitro* requires both 5' and 3' ends present, either connected in the same RNA sequence, or added in trans [167]. Another piece of evidence for the cyclization of the genomic RNA is the finding that the 1st stem in 5'UTR and the last stem in 3'UTR together with the cyclization sequences (CS) are necessary and sufficient for virus translation and replication [72].

The mean pairwise sequence identity of all four groups of the genus *Flavivirus* (less than 50%) was too small to yield good alignments. The species TBE differs most from the other species in both, sequence and structure. From an alignment of the remaining species, DEN, JEV and YFV, a common structure for the CS as shown in Fig. 16 is obtained, which supports the prediction of Hahn [42]. Then only DEN and JEV were compared. In the result data the CS contained no sequence variation but was predicted with pair probabilities close to 1. Adjacent to the CS there is a further stem which participates in genome cyclization and which contains several sites of sequence variation (see P2" in Fig. 17). Between CS and P2" there is a well conserved hairpin structure supported by numerous compensatory mutations (see DV2/JE2 in Fig. 17).

Tick-borne Encephalitis Virus. The conserved cyclization motifs first reported by Hahn in [42] for mosquito-borne viruses are absent in the tick-

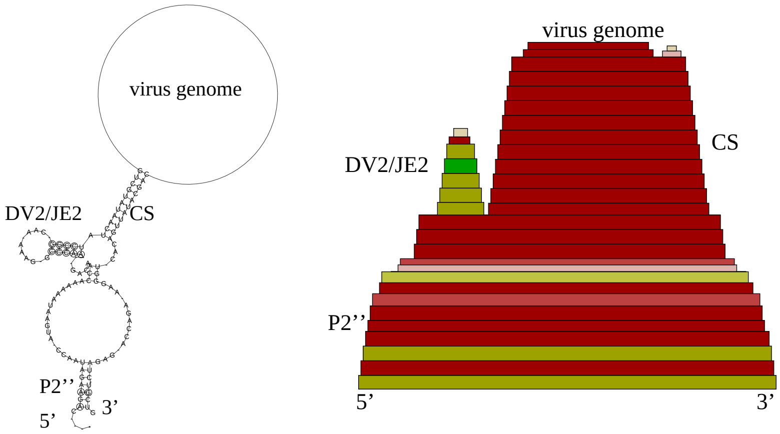


Figure 17: Conserved structures close to the CS of DEN and JEV. P2" is a common stem of DEN P2 and JEV P2. The scheme of the annotations in the conventional drawings as well as the color code are explained in the caption of Fig. 14.

borne encephalitis virus group. Putative CSs were proposed for Powassan virus RNA [91] and for TBE virus [72].

In all proposed motifs for genome cyclization again no mutations in sequences were found, thus the method could not be used to confirm the predicted structure by means of sequence covariation. Thermodynamic folding, however, provided strong evidence for the CS"A"-motif in Fig. 16 [72, 91] because these base pairs appeared with probabilities close to 1 in the folds of the complete genome. Khromykh's region CS"B" was folded only by one single sequence (TEU27491) and thus could not be considered as a common motif for all members of TBE.

4.1.2 5'UTR

The 5'UTRs of DEN, JEV and YFV formed into a very similar secondary structure, while the structure for TBE turned out significantly different, see

Fig. 16 and Fig. 18, DV1, JE1, YF1 and TB1 respectively. For DEN there was structural conservation, while the sequences of JEV and YFV were very conserved. A manually improved alignment of JEV and YFV to DEN of this region showed that there was a significant structure conservation among all three groups. Furthermore all structures contained an interior loop of three Us (one on the 5' and two on the 3' strand), see DV1, JE1 and YF1 in Fig. 18.

For DEN a structure similar to DV1 is proposed by Leitmeyer in [84] and Khromykh in [72]. A stem-loop structure from positions 80-105 reported in [84] is predicted thermodynamically, but has a conserved sequence, and thus is not supported by sequence covariation.

A stem carrying the initiator AUG proposed by Hahn *et al.* [42] for YFV was found to fold in all sequences of this group but is not supported by sequence covariation.

The 5'UTR structure proposed by Khromykh in [72] for TBE was inconsistent with the available sequence data. We found a quite different structure that was confirmed by several mutations, both consistent and compensatory, see TB1 in Fig. 18.

4.1.3 Coding Region

Several conserved secondary structures were found in the coding regions of DEN, JEV, YFV and TBE. The structures are shown in App. A.2. So far, no functions have been proposed for these regions. Stem-loop DV2 was proposed already for Den-2 virus [42].

4.1.4 3'UTR

Conserved structures in the 3'UTR are shown in Fig. 18, the structures show strong similarity between groups. Sequence variation in the stem DV6a was

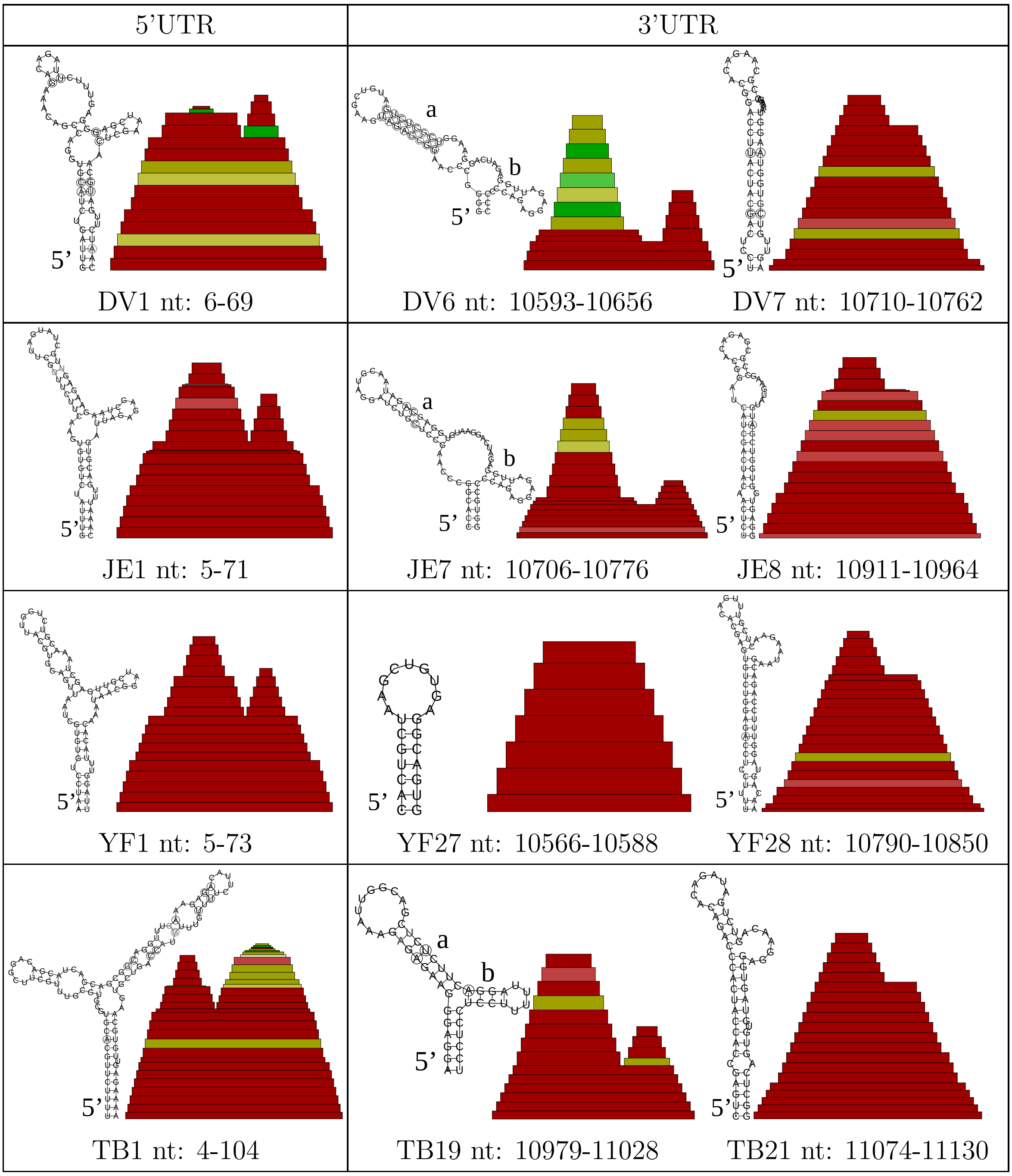


Figure 18: Conserved secondary structures of the *Flavivirus* species DEN, JEV, YFV and TBE in the 5'UTR (first column) and 3'UTR (second and third column).
The scheme of the annotations in the conventional drawings as well as the color code are explained in the caption of Fig. 14.

high in DEN, and present in JEV. For YFV we found a stem corresponding to the DV6a and JE7a.

Structures similar to DV6, JE7, YF27 or TB19 were also proposed by Hahn in [42] for DEN 2 and YFV, by Khromykh in [72] for DEN, YFV, JEV and TBE, by Rauscher in [120] (B for DEN, YFV, and JEV and I, II and III for TBE), by Proutski in [118] (TL1/RCS2 or TL2/CS2 for DEN and JEV, and "stem loop 1 in subregion I" for YFV), and by Leitmeyer in [84] for DEN.

Dengue virus. For the DEN 3'UTR the same structures were found as in [120], where the analysis was restricted to the isolated 3'UTR. None of the long-range interactions interfered with any of these structural motifs. Leitmeyer proposes additional base pairings that we could not find, because they conflicted with the cyclization domains [84].

Only parts of the secondary structures proposed by Proutski in [117] for DEN2 as conserved for all DEN species were found. In particular his structures I2 and I3, II1 and III except region 3'LSH (our DV7) were absent in our data. DV6 and DV7 are also discussed [118] for DEN4. All other structures reported in that study are disrupted by the cyclization of the viral genome. Assuming that cyclization of the genome is vital the deletion studies reported by Men in [98] can be re-interpreted in the following way: deletion of DV6a (TL2) yields a delayed and reduced growth in simian and mosquito cells. When the deletions extended more to the 3' end of the sequence the CS region is destroyed [mutant 3'172-83 in [98]], hence no viable viruses are found. A non-viable mutant 3'172-107 may be explained by the importance of the sequence motif CAAAAA for virus propagation [98]. Our data indicate that in this case the sequence motif rather than any structure associated with it is important. For the mutants 3'd333-183 and 3'384-183 Men *et al.* measure a greatly delayed and reduced growth in living cells. We would argue that these deletions destroy a possible prolongation of the cyclization region that we found for dengue viruses (data not shown). Our data indicate that each single sequence allows additional stems for cyclization in this region even though their exact positions vary slightly among the different sequences. It is plausible that such an extended cyclization region adds to the efficiency of viral replication but is not necessarily essential for its viability.

Yellow Fever Virus and **Japanese Encephalitis Virus.** The sequences in this data set had about $\sigma = 91.7\%$ pairwise identity in the 3'UTR. Only a small number of compensatory mutations was observed to verify structural features predicted based on the thermodynamic algorithm. The results contained essentially the same structures as proposed by Rauscher *et al.* in [120]; again none of the structures reported by Rauscher conflicted with CS regions. YF28 was shorter by 9 base pairs than reported in [42]. YF28 and YF27 corresponded to 3'LSH and I1 respectively, JE7 and JE8 to 3'LSH and II2 respectively as proposed in [117]. More structures could not be found for similar reasons as explained for DEN above.

Tick-borne Encephalitis Virus. The structures recovered here are very similar to those reported by Mandl *et al.* in [92] and Rauscher *et al.* in [120]. In particular TB17 and TB18 corresponds to their IV and VI respectively, TB19 contains stem III, and TB16 corresponds to VII, VIII and IX. Structure A1 reported by Mandl was shorter because of conflicts with cyclization sequences P1' and CS"A". Structure A2 did not seem to be conserved. For structure MS and V there was evidence from thermodynamic folding. However, these two structures conflict with P2. TB16 to TB21 go conform with structures proposed in [117].

4.2 Pestivirus, Hepacivirus, and GB Virus C

4.2.1 The 5'UTR

The 5'UTRs of these virus groups contain an IRES. For parts of HCV's IRES even studies about tertiary structure are available [73, 87]. The sequences of 5'UTRs of GBV-C and HCV are significantly higher conserved than the

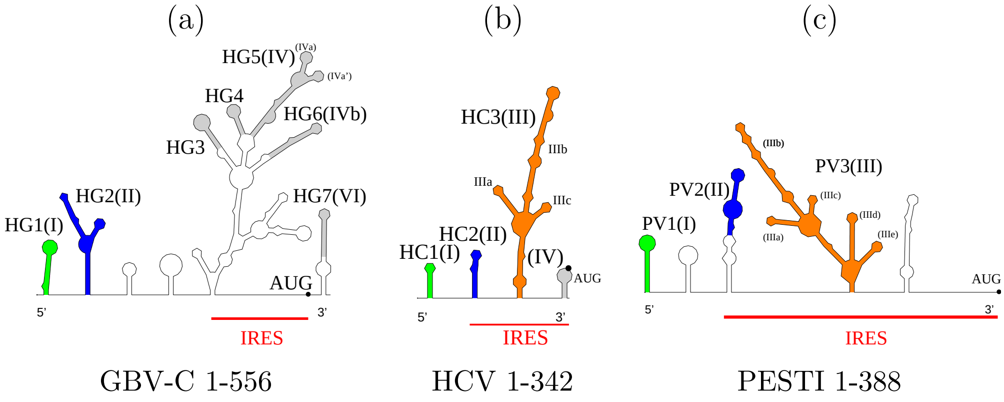


Figure 19: Schematic illustration of 5'UTRs of GBV-C, HCV and PESTI. Conserved structures are discussed in the text. Notations in brackets correspond in (a) to [134], (b) to [60] and (c) to [9].

rest of their respective genomes, see Tab. 1. For these two virus groups we found that the secondary structure of the 5'UTR is less conserved than we expected from literature (due to the few sequence covariations); an overview is given in Fig. 19. This was consistent with the data reported by Witwer for *Picornaviridae* [158]. Contrarily, the IRES of PESTI turned out to be highly conserved.

The IRES structures of HCV and PESTI shared a common overall structure despite the fact that they were not comparable at the sequence level. Nevertheless, they shared a few significant details: the IIIa stem carried a completely conserved loop sequence and stem IIIc was conserved in its sequence.

GB Virus C. The 5'UTR sequences of GBV-C were much higher conserved than the rest of the genome (Tab. 1). Most of the sequence variation occurred around nucleotide (nt) positions 410 to 437, which comprised the structure element HG6 (IVb), Fig. 20(a). This motif was predicted also in previous studies [134, 136].

Stem HG2 (FIG. 19(a)) was shorter and more shifted to the 5' end of the IRES than stem loop II reported by Simons [134]. The prediction was supported by compensatory mutations (data not shown). The reason for the discrepancy was the formation of a panhandle-like structure by means of a base pairing interaction from nt 163-175 with nt 9213-9201 (see section 4.2.4).

The sequences were too conserved in the remainder of the 5'UTR to support predicted structures by means of sequence covariation. The thermodynamic prediction, however, found structures similar to those proposed previously [70, 134, 136].

Hepatitis C Virus. The 5'UTR of HCV comprises 341-342 nucleotides. The `RNAfold` algorithm recovered structures similar to those reported in previous studies [20, 61, 69, 74, 78, 109, 119, 137, 144], see Fig. 19(b). The algorithm was not designed to predict pseudoknots. However nucleic acids which are known to be involved in pseudoknots [115] do not pair to other parts of the sequence.

Due to high sequence conservation (Tab. 1) we found only two sites with compensatory mutations in HC3 (called IIIa, b, and c in [61]) in this data set of 9 complete genomic sequences. When additional sequences of the IRES region were included in the analysis, the structure was well supported by compensatory mutations (data not shown). This structure HC3 has received considerable attention since it appears to act as binding site for the eIF3-40S complex. It has an internal loop which is twisted in itself [20]. Even though we had a mean identity of 98.3% in this region, there were two compensatory mutations just before and after this highly structured part of the HCV IRES. This confirms Collier's interpretation that here the shape of the backbone rather than the sequence composition is important for translation initiation.

Stem HC2 corresponds to IIa proposed by Honda [60]. For the nucleotides following stem IIa, the prediction favored long range interactions with nu-

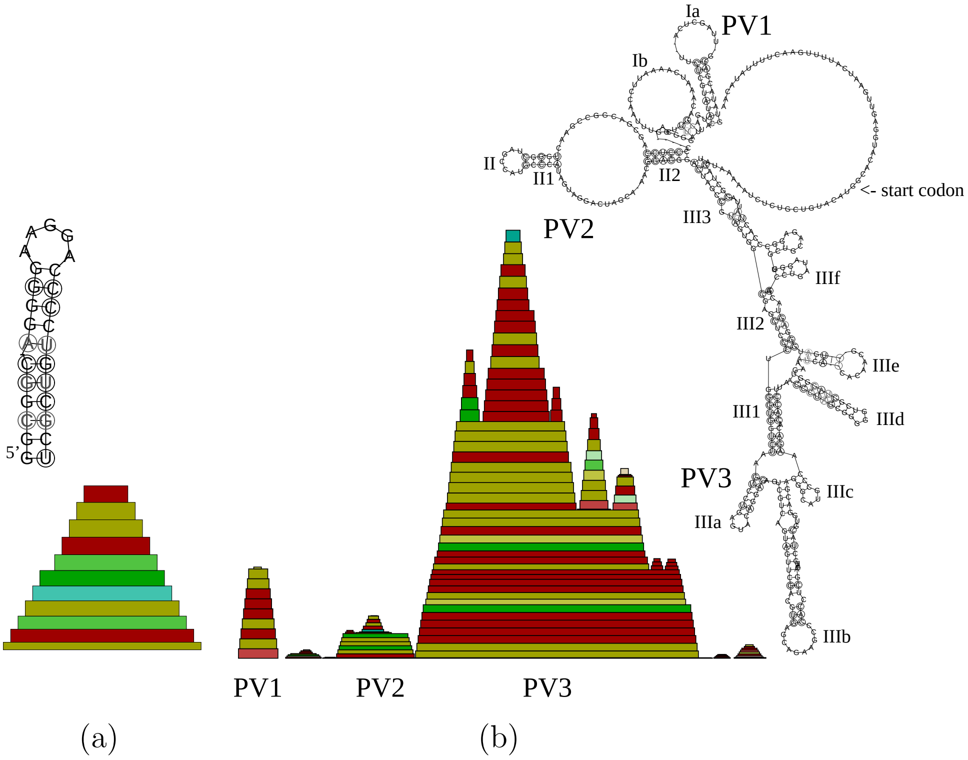


Figure 20: (a) GBV-C: 5'UTR nt: 410-437, IRES conserved element HG6(IVb) (b) *Pestivirus* 5'UTR nt:1-420; the IRES is supposed to begin with stem PV2(II).

cleotides 8571 to 8552 (NS5B), see HCVCS2 in section 4.2.4. When the isolated IRES region (i.e. nt 44-357) was folded separately, stem IIa and IIb were recovered as proposed in [60].

Pestivirus. As with HCV and GBV-C the sequence of the 5'UTR region was more conserved than the rest of the genome (see Tab. 1), but still we found a considerable amount of consistent and compensatory mutations.

Stem PV1 was proposed as Ia by Brown in [9] and as domain A by Deng in

[23], Fig. 20(b).

Fletcher observed that a deletion of nucleotides comprising stem PV2 (II in [9], domain C in [23]) produces a decrease of IRES activity to 19% [25]. Though the pair probabilities in stem PV2 were small, see the mountain-plot of Fig. 20(b), we found no inconsistencies and a considerable amount of compensatory mutations. This might point out the importance of the structure rather than the sequence to IRES function in this region.

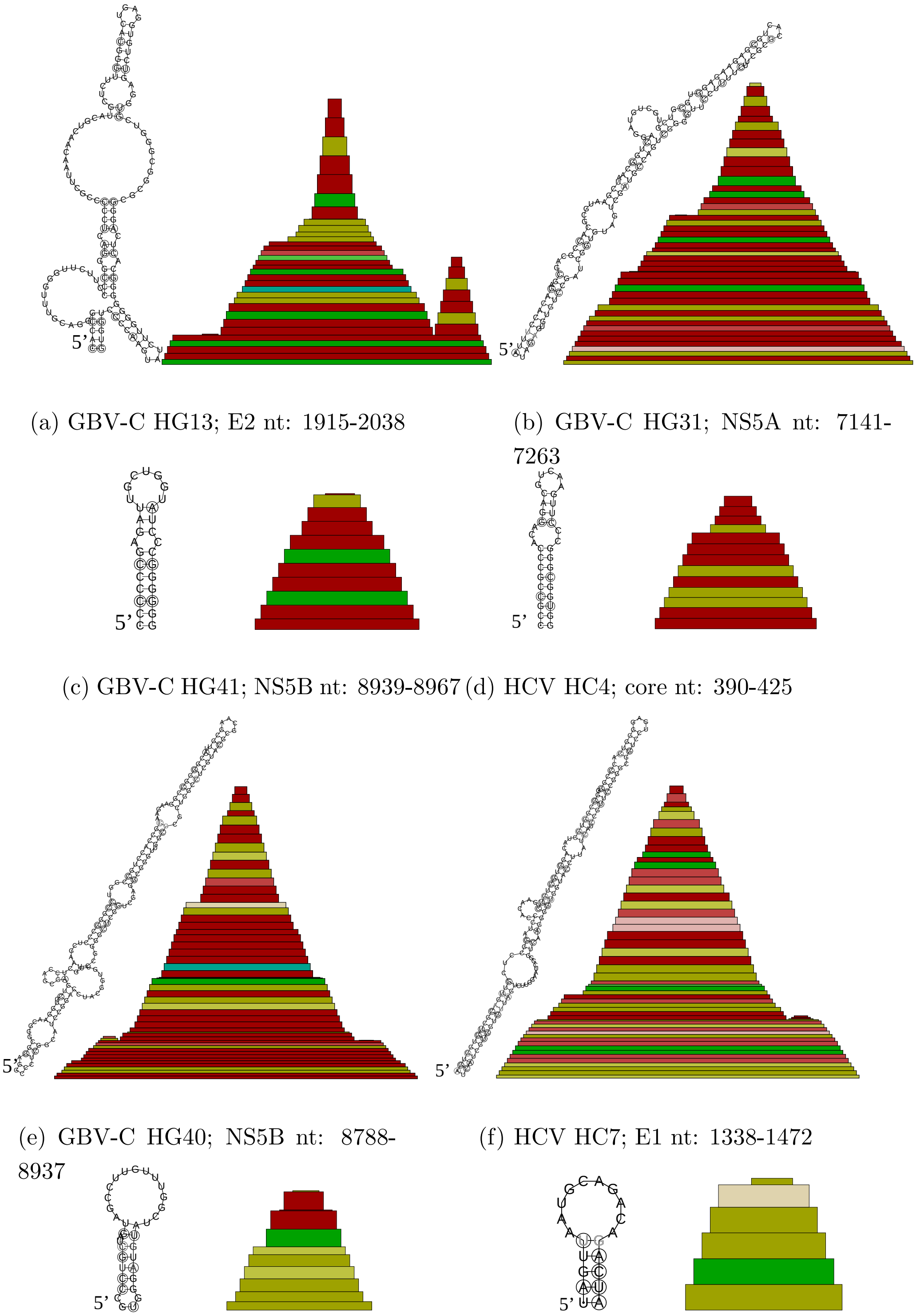
As in previous studies [23, 25, 79, 100] stem PV3 is detected as an important feature of *Pestivirus* IRES structure. Even though our algorithm does not allow pseudoknots, in the case of PESTI thermodynamic prediction showed slight probability for the existence of the pseudoknot reported in [115].

4.2.2 Coding region

GB Virus C. We found two significantly conserved stems (HG9 and HG10) in the E1 region, which were proposed previously by Simmonds based on a different algorithm [133] (data presented in the supplemental material).

Conserved secondary structures seemed to be concentrated in the NS5A and NS5B region of the GBV-C genome, Fig. 21. Some of these were proposed already by Cuceanu in [21], Fig. 21(c) and (e). Furthermore HG38 corresponds to SL_{NS5B}V and HG39 to SL_{NS5B}IV. The SL_{NS5B}I motif is completely conserved in the sequence, in SL_{NS5B}VI more inconsistent than compensatory mutations are found (data not shown), and the proposed SL_{NS5B}VII structure could not be found with our method.

Hepatitis C Virus. Again we found most of the conserved structures in NS5A and NS5B regions. Some of these have been reported previously as important for the efficiency of the IRES function [149, 169]. One of the motifs Tuplin detected in [149] is HC4, shown in Fig. 21(d). Tuplin further finds HC6 as SL443, HC27 as SL8828 and HC28 as SL9011.



(a) GBV-C HG13; E2 nt: 1915-2038

(b) GBV-C HG31; NS5A nt: 7141-7263

(c) GBV-C HG41; NS5B nt: 8939-8967

(d) HCV HC4; core nt: 390-425

(e) GBV-C HG40; NS5B nt: 8788-8937

(f) HCV HC7; E1 nt: 1338-1472

(g) PESTI PV8; NS4A nt: 7238-7271

(h) PESTI PV14; NS5B nt: 11892-11911

Figure 21: Examples of conserved secondary structures in the coding region of GBV-C, HCV, and PESTI. (a) to (c) and (e) conserved structures in GBV-C coding region; (c) and (e) were already proposed by Cuceanu in [21] (SL_{NS5B}II and SL_{NS5B}III, respectively). (d) and (f) examples from HCV, (d) was first proposed by Tuplin [149]. (g) and (h) proposed conserved structures in PESTI coding region.

According to the data of this Thesis there was no evidence for the existence of SL7730 and SL9118. SL8926 showed too many inconsistencies, and SL8376 was not folded because of interactions of this region with the 3'UTR (see section 4.2.4).

Ray argues that the HCV persistence is associated with sequence variability in putative envelope genes E1 and E2 [121]. We found a conserved RNA structure, HC7, in the E1 region, Fig. 21(f).

Pestivirus All putative conserved secondary structure elements in the coding region of PESTI were very short. A stem loop downstream of the initiator AUG appears in our data to have too many inconsistencies and thus cannot be considered as a conserved feature of PESTI, in agreement with the analysis of Myers [102]. The most prominent stems found in the coding region are shown in Fig. 21(g) and (h).

4.2.3 3'UTR

GB Virus C. The 3'UTR sequences of GBV-C are highly conserved ($\sigma = 96.7\%$). Not surprisingly, we predicted structures similar to those reported previously [70, 112, 164] but not all of them were supported by sequence covariation (data not shown). Some of the previously proposed structures conflict with long range interactions to the 5'UTR predicted by our method (see section 4.2.4). One example well supported by sequence covariation is structure HG43 that was also proposed by Cuceanu [21] and Xiang [164].

Hepatitis C Virus. The 3'UTR consists of a short sequence of variable length and composition (variable region), an U rich stretch (poly-U-UC region) variable in its length and a highly conserved sequence of approximately 100 nucleotides at the 3' end (conserved region, X-tail) [80, 141, 165]. Within this X-tail we found only a single mutation (which is compatible with the predicted structure). Our stem HC29 corresponds to SL1 as reported previously [6, 65, 165]. Stems SL2 and SL3, as proposed in [6] and [65], compete in our data with the formation of two long range interactions LR1 and LR2, see section 4.2.4 The probability of base pairs in LR1 was around $p = 0.54$, significantly higher than HC29 (SL1). The elements SL2 and SL3 were thermodynamically unfavorable in the genomic context and could only be detected when a sequence window was used that was too small to contain the long-range interactions. Most recently Yi described several point mutations in the X-tail of HCV's 3'UTR [166]. Their results could not provide a proof for the existence of SL2 or SL3 but indicated that there are stringent requirements for the sequence in this region.

Pestivirus. *Pestiviruses* are very heterogeneous in their 3'UTR region, due to extended AU rich insertions in some strains. The only RNA feature that was shared among all available sequences is the terminal stem PV15 that was originally described in [23], see also [5, 168].

4.2.4 Genome Cyclization

Surprisingly, we discovered strong evidence for genome cyclization not only in the genus *Flavivirus*, where this effect has already been described in the literature, but also within HCV and GBV-C. The most prominent cyclization domains of HCV and GBV-C are shown in Fig. 23.

In GBV-C genome cyclization is localized to base pairings between nt 33-48 with nt 9367-9353 (HGCS1: pair probabilities ≈ 0.6), nt 128-140 with nt 9224-9214 (HGVCS2) and nt 163-175 with nt 9213-9201 (HGVCS3) (both with pair probabilities ≈ 0.7), see the last one in Fig. 23. These domains are very conserved in sequence. We found only one consistent mutation in the base pair (42,9357). On the other hand there was one sequence carrying an inconsistent mutation at base pair (130,9222).

In HCV putative cyclization domains comprised base pairs of nt 1-3 with nt 8627-8625 (HCVCS1), 88-92 with 8602-8606 (HCVCS2) and 95-110 with 8556-8571 (HCVCS3). Within HCVCS3 we found two sites of compensatory mutations, see Fig. 23. In HCV nucleotides from the IRES region (nt 1-3, 88-92 and 95-110) paired with nucleotides within the coding region for the protein NS5B. At the same time we observed two regions of the 3'UTR to fold forward to the NS5B region as well: (i) LR1: nt 8628-8661 (NS5B) paired with nt 9599-9633 (3'UTR) and (ii) LR2: nt 8978-8995 (NS5B) with nt 9583-9598 (3'UTR). This brought 5' and 3' regions into very close proximity, as is illustrated in Fig. 22. Sequence position 8627 is involved in interaction with the IRES; the adjacent nt 8627 pairs the 3'UTR.

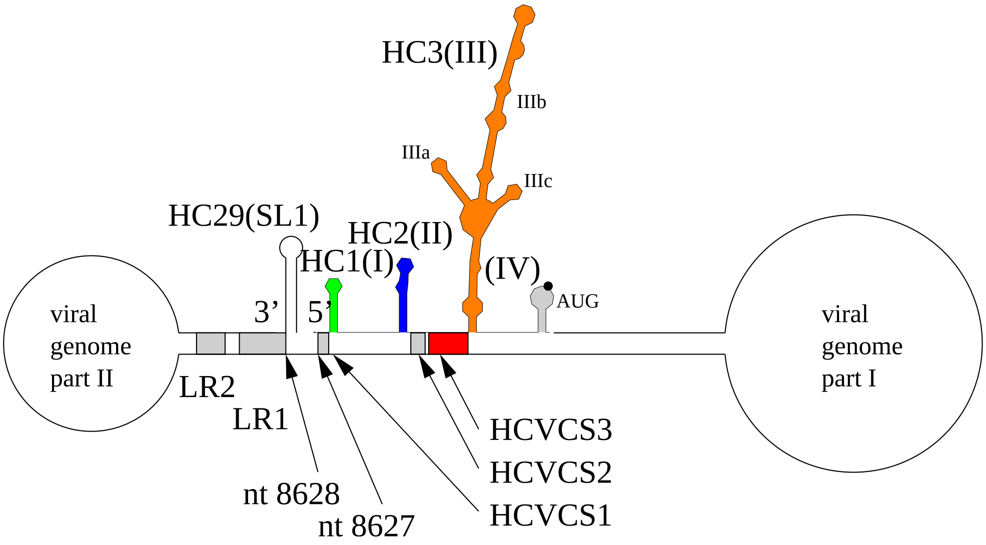


Figure 22: HCV 5' and 3'UTRs brought together very closely by interactions with the NS5B region of the genome.

All of the mutations (15 point mutations and six double mutations) studied in [166] exhibit reduced or no replication activity. Most of them would disrupt base pairs in either LR1 or LR2, supporting our proposed interactions. However, five of the point mutations are in predicted loop regions and would be expected to cause only minor secondary structure changes. This could indicate that there are sequence constraints beyond conservation of secondary structure. However, to prove or disprove the existence of LR1 and LR2 more mutation experiments would be needed.

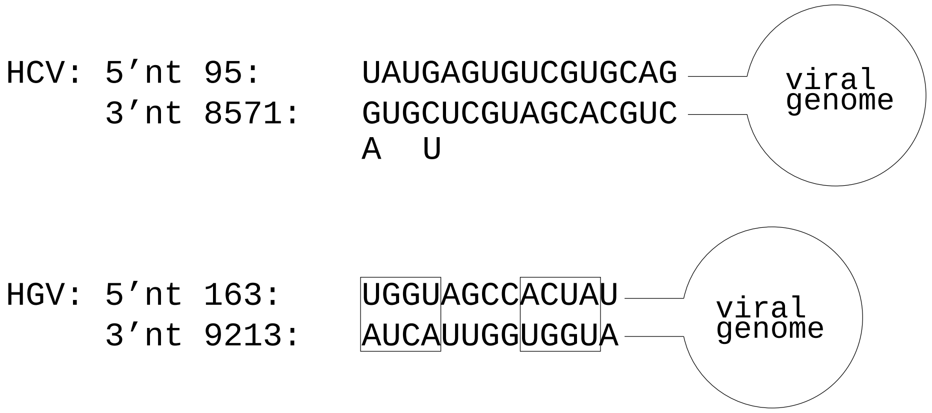


Figure 23: Putative cyclization regions HCVCS3 and HGVCS3 in the genomes of HCV and GBV-C respectively. The boxed areas point out sequences that might be read as palindrome sequences and maybe play a functional role in replication processes.

4.3 Discussion

In the genus *Flavivirus* cyclization of the genome was already described in the literature and localized to very conserved cyclization sequences. Apart from recovering these known cyclization sequences, we detected further sequences which took part in cyclization for all species in this study (P1’, P1 and P2). These sequences varied considerably in sequence, length and position. Men *et al.* [98] showed, that deletion of these sequences led to a greatly delayed and reduced growth in simian and mosquito cells. It is possible that these additional cyclization domains are not strictly necessary for virus viability, but only support and stabilize viral genome cyclization.

We found viral genome cyclization most surprisingly also in GBV-C and HCV, which had not been reported before. Yi *et al.* [166] suppose a cyclization of HCV genome by the assistance of some cellular protein. Our algorithm made out base pair probabilities for both, previously reported secondary structures in 5’ and 3’UTRs as well as for genome cyclization. For both cases, our data revealed no inconsistencies. Thus previously proposed structures compete with genome cyclization. Our evaluation conditions favored genome cyclization based on both, thermodynamic prediction and, in the case of HCV, even sequence covariation. This result can be interpreted either as a relict of ancient ancestors between these genera and the genus *Flavivirus* or, more speculatively, as a switch providing different functions in different states of the viral life cycle (e.g. a switch between replication and translation states of the virus).

While in *Flavivirus* and GBV-C the 5' and 3' end of the genome pair by forming a "pan handle" like structure, we found base pairing in HCV between the 5' end and the 3' end to a region some 1000 nt upstream of the 3' end (i.e. a region within the NS5B protein). Most interestingly we observed that in this way 5' and 3' ends were brought closely together. This could be a reason for the particular importance of the NS5B region as assumed in the literature [110, 111]. It may also explain the results of Friebe *et al.* [30] and Kim *et al.* [75] who observed that domain HC1(I) and HC2(II) in the 5'UTR are essential for replication, while domain HC3(III) helps to facilitate replication, but is not absolutely required.

In Appendix A.2 we present a large number of secondary structure elements that have not been described before, most importantly within the coding region. This information could be used as a basis for experimental research for additional regions which might be important for virus viability and propagation.

5 Designs for Experiments

Based on the results obtained by this work, a collaboration with two different groups working experimentally on viruses came into life.

5.1 Tick-Borne Encephalitis Virus

The Group of Prof. Christian Mandl at the Institut für Virologie, Medizinische Universität Wien, works on TBE virus strain Neudoerfl (wt) (Acc. no. U27495) and an infectious cDNA clone, pTNd-C, which differs from wt in several mutations, see Fig.24. From both genomes they had derivatives, bearing idifferent lengths of deletions in the 5'ends of the coding regions.

Replicons are subgenomic RNA molecules that are competent for autonomous RNA translation and replication but due to a delition of stuctural protein genes are incapable of forming infectious virions. Such replicons are useful vehicles for studying replication without having to handle with the infectious virus. dCME represents such a replicon of TBE virus wt and dCME[pTNd] is the respective c-DNA clone. dCME lost its infectivity through complete deletion of pre-membrane protein prM and substantial parts of proteins C and E. There exist two further replicons, C15 and C17 Tab. 5.1 gives an overview of the derivatives and their clones and the extend of their deletions.

We were asked, whether we could predict from structural inspection of the genomes and comparison with the infectious (wt) and not infectious but known to be vital genomes (dCME) that either of the replicons C15 and C17 could have kept its vital functions. When constucting a replicon, a c-DNA colone had to be produced. During this procedure several point mutations occured. The effect of these mutations to structural changes was an other feature of interest.

Fig. 24 summarizes the results. The coding region of the genome of each virus

Table 3: Overview of the replicons of the genome of TBE wt.

derivative	ν	Δ
dCME	+	213-2391
C15	?	177-2386
c17	?	183-2386

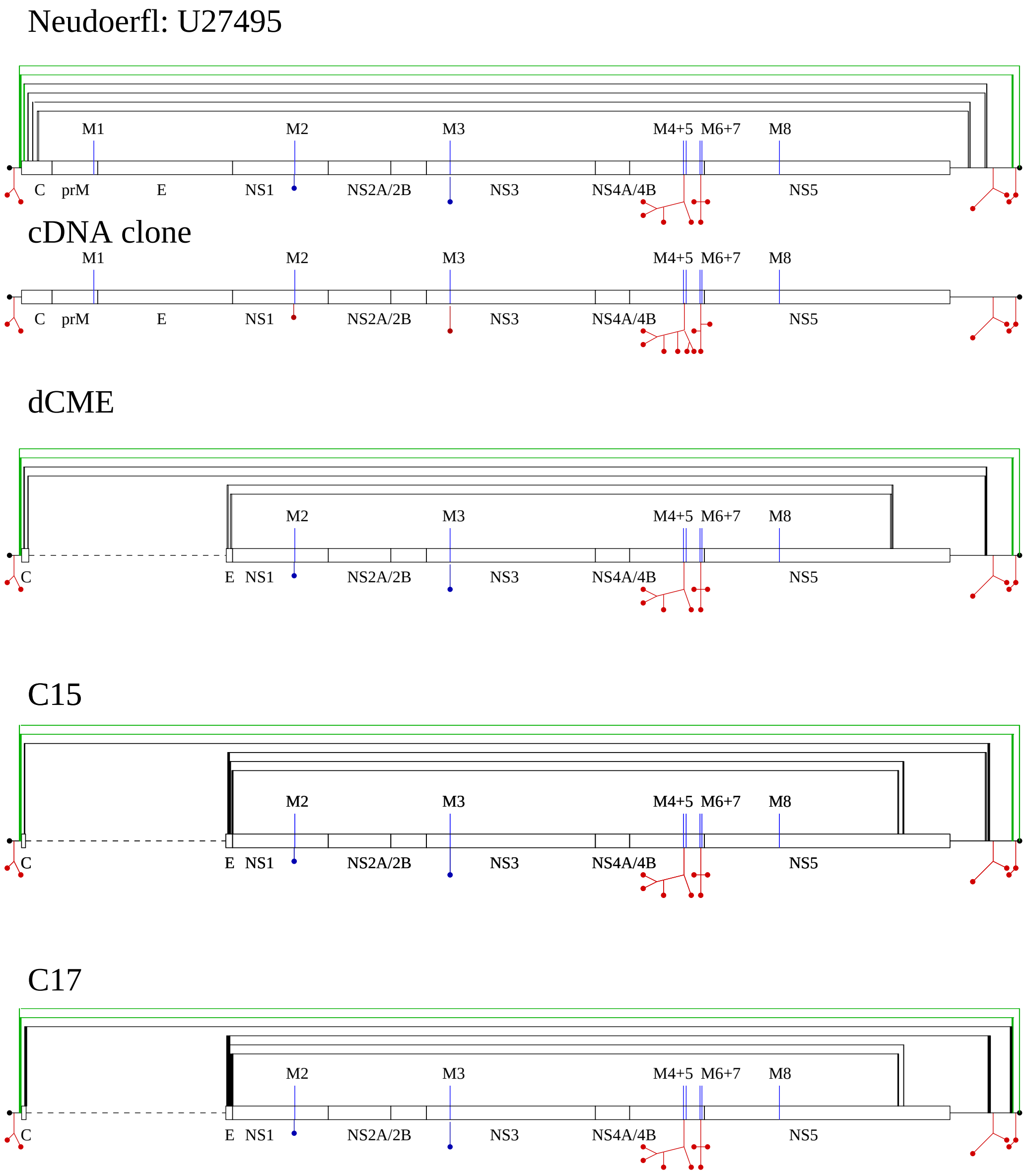
+ indicates that the genome is viable, ? viability (ν) is unknown, Δ gives the nt region of deletions of the derivative with respect to the sequence of wt.

is represented by a horizontal bar, the names of the encoded proteins are indicated below. The coding region is flanked on both sides by untranslated regions, represented by a line. Dashed lines substitute the parts of the coding regions which were deleted in the respective replicon. Arcs above the genome indicate long range interactions. Green arcs are interactions that can be found in all TBE virus strains, black arcs are formed only in the respective genome. The exact positions of long range interactions are listed in Tab. 5.1. Blue vertical lines indicate mutations that differ wt from its c-DNA clone, and are marked M1 - M8. Beneath the genome, blue structure representations indicate structures, that do not change between wt and its clone. Where the structure is colored red, there is a slight structural change. While mutations M1 and M2 lie in a non-structured region, all other mutations coincide with structural motifs. Mutations M2 and M3 do not cause any change of the structure. Mutations M4 and M5 lie within the same structural motif and cause slight change in secondary structure. So do mutations M6 and M7.

In all genomes cyclization domains P1', CS“A”, and P1 are detected, see the green arcs in Fig. 24. (For definition of cyclization domains P1', CS“A”, and P1 in genomes of *Flaviviruses* see section 4.1.1.) Black arcs indicate regions which prolongate cyclization, but which vary considerable in nt positions among different strains of the genus *Flaviviruses*, but are present in every strain. As discussed in section 4.3 we suppose them to aid stabilizing cyclization, but not to be essential. Such longrange interactions are predicted in all genomes of wt and replicons. When large parts of the coding region are deleted and original longrange interactions are not possible anymore, new regions more upstream of the genome build longrange interactions to the 3'end of the genome.

We find longrange interactions similar to those predicted for the genome of the wt and the replicon dCME also in the replicons C15 and C17. Given that dCME contains a viable but not infectious genome, we would estimate, that the replicons C15 and C17 should be viable as well.

Introducing mutations by producing cDNA clones seems not to affect viability. Since structural effects are the same also in the genomes of C15 and C17, we do not expect the mutations to have relevant influence on their viability.



Neudoerfl: U27495

cDNA clone

dCME

C15

C17

Figure 24: Genome-maps of TBE strain Neudoerfl, its cDNA clone, and the replicons dCME, C15 and C17, as used for experimental structure-analysis by Mandl and Kofler [90]. The green arcs above the genome indicate longrange interactions which are found in all TBE strains, the black ones are formed only in the specific genome. The exact positions are listed in Tab. 5.1.

Table 4: Long-range interactions from 5'end of TBE to its 3'UTR.

	Neudoerfl	dCME		C15		C17	
		pos.	rel. Nd.	pos.	rel. Nd.	pos.	rel. Nd.
P1'	109-11141	109-8978	109-11141	109-8996	109-11141	109-9002	109-11141
	111-11139	111-8976	111-11139	111-8994	111-11139	111-9000	111-11139
CS"A"	115-11071	115-8908	115-11071	115-8926	115-11071	115-8932	115-11071
	129-11057	129-8894	129-11057	129-8912	129-11057	129-8918	129-11057
P1	157-10781	157-8618	157-10781	163-8676	163-10821	168-8919	168-11058
	166-10772	166-8609	166-10772	169-8670	169-10815	191-8899	296-11038
D	202-10765	202-8602	202-10765	199-8666	728-10811	193-8682	717-10821
	207-10760	207-8597	207-10760	219-8648	1053-10793	225-8654	1053-10793
E	255-10597	228-7583	2172-9746	220-7657	1052-9802	226-7663	1054-9802
	258-10594	240-7572	2402-9735	231-7647	1264-9792	229-7660	1057-9799
F	307-10591	265-7568	2427-9731	246-7600	2390-9745	252-7606	2396-9745
	322-10578	279-7554	2441-9717	258-7590	2402-9735	264-7596	2406-9735

P1', CS"A" and P1 are conserved among all TBEs, while D, E and F form specifically to the strain or the clone respectively. The interactions are indicated by their starting and their closing base pair. For the replicons also the nucleotide positions relative to Neudoerfl are given, in order to render them easier to compare.

5.2 Hepatitis C

The group of Prof. Ralf Bartenschlager at the Abteilung Molekulare Virologie, Universität Heidelberg was interested in the idea that the genome of HCV might engage in genome cyclization as shown in Fig. 23. In fact they could show that the first 125 nt of the HCV genome were sufficient for RNA replication, although replication was significantly better when the entire 5'UTR was present [30].

The objective was to design mutations within the cyclization regions in order to destroy and then repair the predicted structure, see Fig. 22.

In Fig. 25 we show the cyclization motif HCVCS3 in detail. Since the 3'part of the structure lies within the coding region, we had to respect the genetic code in that mutations should not affect the information for encoded amino acids. Possible mutations at appropriate sites are indicated above and below the respective strand.

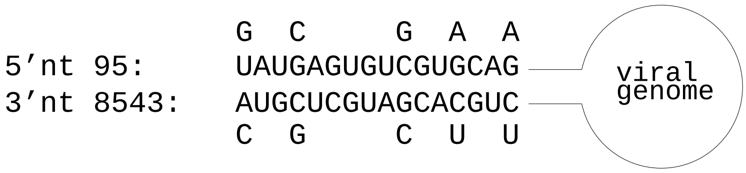


Figure 25: Long range interaction HCVCS3 and introduced mutations

We propose three different mutants, see Tab. 5. M1 has mutations only in the 5'UTR, destroying HCVCS3. M2 additionally has silent mutations in the NS5B region, restoring the predicted helix HCVCS3. M3 contains only the mutations in the NS5B region, which are also expected to destroy HCVCS3. The mutations are listed in Tab. 5.

The calculation of the mfe structure of the 3 mutants shows: As expected, HCVCS3 is no more contained in M1 and M3, while it is restored in M2. For

M1 RNAfold predicts other long range interactions of the nucleotides 94-110 (which form the 5' part of HCVCS3 in the original sequence), the closing base pair for the new formed stem of M1 is 94-8965. For M3 the dot plot contains a stem with closing base pair 94-8991 with a high probability.

The predicted structure of the IRES region for M2 is identical to the original sequence, while for M1 only stems I and IIIa are contained in the mfe structure, see Fig. 19. The predicted mfe structure for M3 contains stems I and IIIa-c. Therefore M1 probably does not form the IRES, while M3 is more likely to form the IRES.

Table 5: List of mutation sites proposed for HCV (Accession no. AJ238799) and the corresponding nucleotides. For comparison, HCV indicates the original nucleotides. The proposed mutated clones are M1, M2 and M3.

pos	95	98	104	107	110	8528	8531	8534	8540	8543
HCV	U	G	C	G	G	C	C	G	C	A
M1	G	C	G	A	A					
M2	G	C	G	A	A	U	U	C	G	C
M3						U	U	C	G	C

6 alidot Goes Pseudoknot

Alidot uses for the search for secondary structures an input-set of thermodynamic base pairing probabilities. These are calculated by a program called RNAfold [57] which relies on McCaskill's algorithm [96].

We pointed out in 2.3.3 that most pseudoknots are comprised in the class of bi-secondary structures. In definition 6 we defined bi-secondary structures as superposition of two individual stems a and b. As discussed in section 3.3.1 the energy of a pseudoknot is approximated to be mainly the sum of its two constituent stems. Putting these ideas together, base pair probabilities of both stems of a pseudoknot ought to be seen in the form of competing stems in a set of thermodynamically possible base pairs, although one of them with eventually very low probabilities.

The limits of the prediction are obvious:

1 Only base pairs could be found that were predicted with a minimal probability of 3×10^{-3} which is the usual prediction threshold of RNAfold.

2 The inclusion of one more base pair to a stack contributes approximately 2kcal/mol to the stack. In other words, a difference of one base pair in length between two stems might result in a difference in probability of a factor of 10^{-1}. If the difference in stem-length is sufficiently large, the shorter stem will not reach the threshold of prediction of RNAfold, and thus become invisible.

The idea is, first to scan through the input dataset of base pair probabilities and search for a valid and most probable secondary structure. These accepted base pairs are taken out of the input set. In a next search through the remaining base pairs, the algorithm looks for further secondary structures. Finally, the algorithm tests, which of the new stems can be overlayed to the already accepted secondary structure and accepts these as pseudoknots.

To adopt the algorithm to this task, we applied two different ways that are outlined in the following. In the layer decomposition of base pairs we treat the single base pairs independent from each other while determining secondary structure and pseudoknots. In the stack-based layer decomposition base pairs are first combined into stacks, before secondary structure and pseudoknots are assigned. Results of applying both versions of `alidot` are presented and discussed in section 6.3

6.1 Layer Decomposition

The layer decomposition of `alidot` is based on the ranking of base pairs. In a first step base pairs are ranked by the base pair credibility criteria lined out in section 3.5. The most believable base pair is put into the first layer. Than, following the rank list of base pairs, the next base pair in the ranking list is compared to the first one. If it conflicts, it is put into a higher layer, otherwise it is accepted into the first layer. Following, the ranking list of base pairs, the next base pairs are compared to the accepted base pairs in the layers. Base pairs, that violate the non crossing condition in a layer, are separated into the next higher layers, until they reach a layer where they do not cross other base pairs. Thus in the first layer we receive the usual secondary structure as before. Every following layer contains a valid secondary structure by its own. In a last step, we try to move as much stems as possible from every higher layer into the first layer, and accept them as pseudoknots.

The algorithm is used by applying `alidot -p -L`. When we refer in the following to the layer decomposed version of `alidot` we will call it `alidotLD`.

6.2 Stack-Based Layer Decomposition

We observed that many pseudoknots could not be detected by pure layer decomposition, although pair probabilities were present in the input data set.

The reason was that base pairs of one stem were split into different layers. To overcome this problem, we combined all base pairs first into stacks.

Because of limiting criterion 2 in section 6, usually the shorter stem appeared with very low probability in the input dataset of base pair probabilities. Therefore we decided not to emphasize too much on stem probabilities but to lay more weight on the effects of compensatory, compatible, and incompatible mutations.

Thus the combined stacks are evaluated and ranked by the following criteria called the stack-credibility:

- number of incompatible sequences in the stack
- number of compensatory and consistent mutations in the stack
- number of incompatible positions
- number of compatible positions
- probability of the stack

Stacks whose terminal base pairs do not conflict are arranged in the same layer. Wherever base-pairs of different stacks within a layer intersect, those base-pairs which are less credible (in terms of base pair credibility, as described in section 3.5) are canceled from the list, following the criteria of secondary structure. Only stacks longer or equal 3 are accepted in the layer. Thus, again, in every layer we have a valid secondary structure by its own, the first layer contains a secondary structure composed by the most probable stacks.

Finally we try to fit stems from higher layers into the first layer, in order to combine them to pseudoknots. When intersections between base pairs of stacks of different layers occur, the bases pairs of the less credible stack (in terms of stack-credibility) are discarded. An example is given in fig. 26.

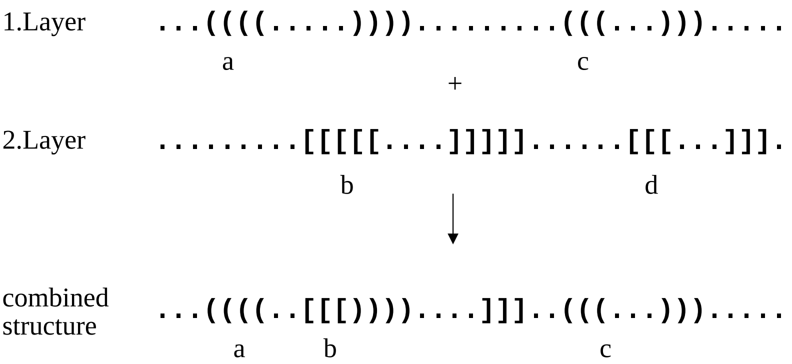


Figure 26: Example for fitting stacks from higher layers (here from the 2nd layer) into the secondary structure contained in the first layer.

Stem *a* of layer 1 intersects with stem *b* of layer 2. In the combined structure, those pairs are discarded, which belong to the less credible stem. Only stacks longer or equal 3 are accepted to the final structure prediction. If stem *d* of layer 2 in fig. 26 is combined with stem *c* of layer 1 to a pseudoknot, in the combined structure stem *d* would have al length shorter than 3. Therefore it is not accepted to the final structure.

The stack-based layer decomposed version of `alidot` will be referred to in the following as `alidotSD`. The algorithm is used by: `alidot -p -Ls`.

6.3 Results and Discussion

The algorithm was tested on three different types of RNA known to contain pseudoknots: Signal recognition Particle RNA (SRP RNA), Ribonuclease P RNA (RNase P RNA), and tmRNA. SRP RNA (Fig. 27 and 28) has one long double helical stem (stem 5 in Fig. 27) and one pseudoknot structure close to the 5'end [82], which can be viewed as 'kissing hairpins' (stems 3 and 4 in Fig. 27). The overall structure of RNase P RNA (Fig. 35) is more globular, with rather short double helical domains, and it contains two long-range pseudoknots [43]. The structure of tmRNA (Fig. 32) contains four H-type pseudoknots and is roughly globular [173].

As observed in preliminary calculations, the quality of the results depends strongly on the quality of the applied alignment. Therefore we used alignments which were taken from the following sources: SRP RNA: SRPDB [34]; tmRNA: tmRNA Database [77]; RNase P RNA: RNase P Database [10]; In table 6.3 we list the reference organisms to which our results were compared.

Table 6: Sequences and reference organisms used for prediction

	Reference organism	*len*	*RP*	*PK*
SRP RNA	*Methanocaldococcus jannaschii*	305	86	1
	Bacillus subtilis	305	86	1
	Halobacterium halobium	305	86	1
tmRNA	*Escherichia coli*	362	106	4
RNase P RNA	*Agrobacterium tumefaciens*	404	124	2

We list the name of the reference organism, its sequence length, the number of base pairs *RP* of the reference structure and the number of pseudoknots *PK* of the reference structure.

To compare the quality of the results with data achieved by applying `hxmatch` [157] and `ilm` [125] in section 6.4, we introduce the terms base pair sensitivity S_{bp} and base pair specificity P_{bp}. $S_{bp} = (TP/RP) \times 100$ and $P_{bp} = (TP/(TP + FP)) \times 100$, where RP is the number of base pairs in the reference structure (reference pairs), TP is the number of correctly predicted base pairs (true positives) and FP is the number of predicted base pairs that are not contained in the reference structure (false positives). We are not particularly interested in the detection of single base pairs. The intended application of alidot was more to recover possible stems of pseudoknots, which might be longer or shorter for different organisms. We therefore we extend the terms of S_{bp} and P_{bp} to stem-sensitivity S_s and stem-specificity P_s. Thus $S_s = (TS/RS) \times 100$ and $P_s = (TS/(TS + FS)) \times 100$, where TS is the number of correctly predicted stems, RS is the number of stems in the reference structure and FS is the number of stems not contained in the reference structure.

6.3.1 Signal Recognition Particle SRP

The Signal Recognition Particle (SRP) is a phylogenetically highly conserved ribonucleoprotein, which associates with ribosomes, recognizes target sequences of nascent secretory and membrane proteins and binds to receptors in membranes of the endoplasmic reticulum. Thus SRP contributes crucially to translocation of secretory proteins across biological membranes. For a review see e.g. [71].

The reference structures for the bacterial and archaeal SRP RNAs shown in Fig. 27 (a) and (b) and Fig. 28 were obtained by comparative sequence analysis [82]. The structures are based on an alignment of 39 sequences. Closely related sequences were aligned first. Then a profile-alignment of the groups was performed. In regions with high sequence variability secondary structure elements were used as additional markers. Positive evidence is given by compensating base changes (Watson-Crick and GU base pairs), negative evidence by a mismatch. A base pair is considered as 'true' if there is at least twice as much positive evidence than negative evidence.

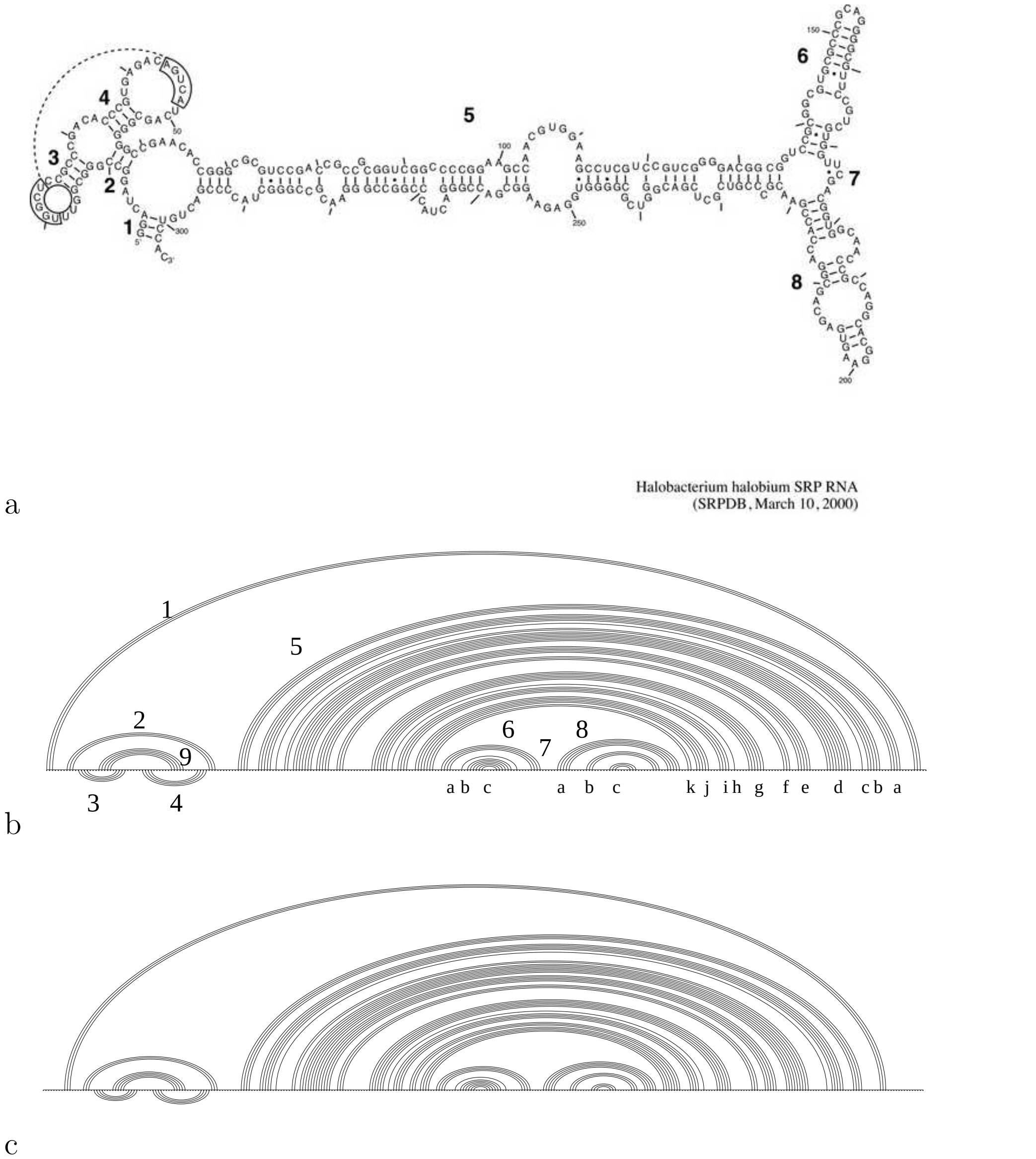


Figure 27: SRP reference structure of archaeal bacteria: (a) and (b) structure representation an representation in two dimensional diagram of *Halobacterium halobium*, image (a) is adapted from [82]. (c) Representation in two dimensional diagram of *Methanocaldococcus jannaschii*.

In Fig. 27(b) we annotated the stem nomenclature of Fig. 27(a), which was originally introduced by [82], to the two dimensional representation of SRP RNA of *Halobacterium halobium*. In Fig. 27(c) we apply the same nomenclature to *Methanocaldococcus jannaschii*. In contrast to *Halobacterium halobium* and *Methanocaldococcus jannaschii*, stem 6 is missing in *Bacillus subtilis*, and thus stem 7 results in a prolongation of stem 5, see Fig. 28.

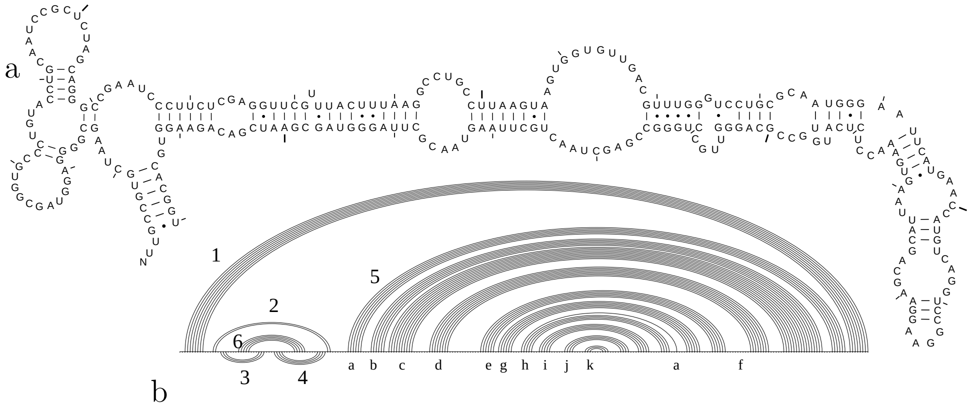


Figure 28: Reference Structure of SRP of *Bacillus subtilis*: (a) secondary structure representation, image adapted from [82], (b) representation in two dimensional diagram.

The consensus structure for a set of aligned sequences was computed using both variants of the program `alidot`. Our dataset comprises 14 archaeal and 15 bacterial SRP RNA sequences, and `alidot` was tested on different subsets of the alignment taken from the Signal Recognition Particle Database [34]. The sequences contained in each subset are listed in App. A.3. All used datasets comprise both, archaeal and bacterial sequences, although datasets S1 and S2 contain mere sequences of archaeal bacteria. Datasets S3 and S4 contain a more balanced set of bacterial sequences. In the following all sets are compared to both archaeal and bacterial reference structures.

Tab. 7 shows the results for all four datasets. Values for base pair sensitivity are always higher than for base pair specificity. This reflects our intention to reduce the number of falsely predicted base pairs on cost of the number of eventually correctly detected base pairs, in order to optimize specificity. All stems predicted in all datasets (except for S1 calculated with `alidotSD`) calculated by both variants of `alidot` are true, which results in a stem specificity of 100. In dataset S2 all stems of the reference structure were predicted correctly. For dataset S2 the kissing hairpin is found by both variants of `alidot`, see Fig. 6.3.1. The color code is explained in the caption of Fig. 6.3.1. For dataset S3 only the upstream pseudoknot (composed by stems 4 and 6) can be identified, and for dataset S4 only stem 6 of the pseudoknot was predicted.

Table 7: SRP RNA compared to *Methanocaldococcus jannaschii*

aln	N	μ	vers.	RP	TP	FP	S_{bp}	P_{bp}	RS	TS	FS	S_S	P_S
S1	6	51.2	`LD`	86	70	33	81.4	68.0	8	6	0	75	100
			`SD`		67	35	77.9	50.8		7	1	87.5	87.1
S2	8	58.7	`LD`	86	78	26	90.7	75	8	8	0	100	100
			`SD`		71	29	82.6	71		8	0	100	100
S3	13	52.8	`LD`	86	52	23	60.5	69.3	8	5	0	62.5	100
			`SD`		61	20	70.9	75.3		5	0	62.5	100
S4	29	54.5	`LD`	86	38	6	44.2	86.4	8	4	0	50	100
			`SD`		45	16	52.3	73.8		4	0	50	100

N number of sequences in the alignment, μ mean pairwise identity of the alignment, vers. version of `alidot`: LD for `alidotLD` and SD for `alidotSD`, RP number of base pairs in the reference structure, TP truly predicted pairs, FP falsely predicted pairs, S_{bp} sensitivity, P_{bp} specificity, TS true stems, RS reference stems, FS false stems, S_S stem sensitivity, P_S stem specificity.

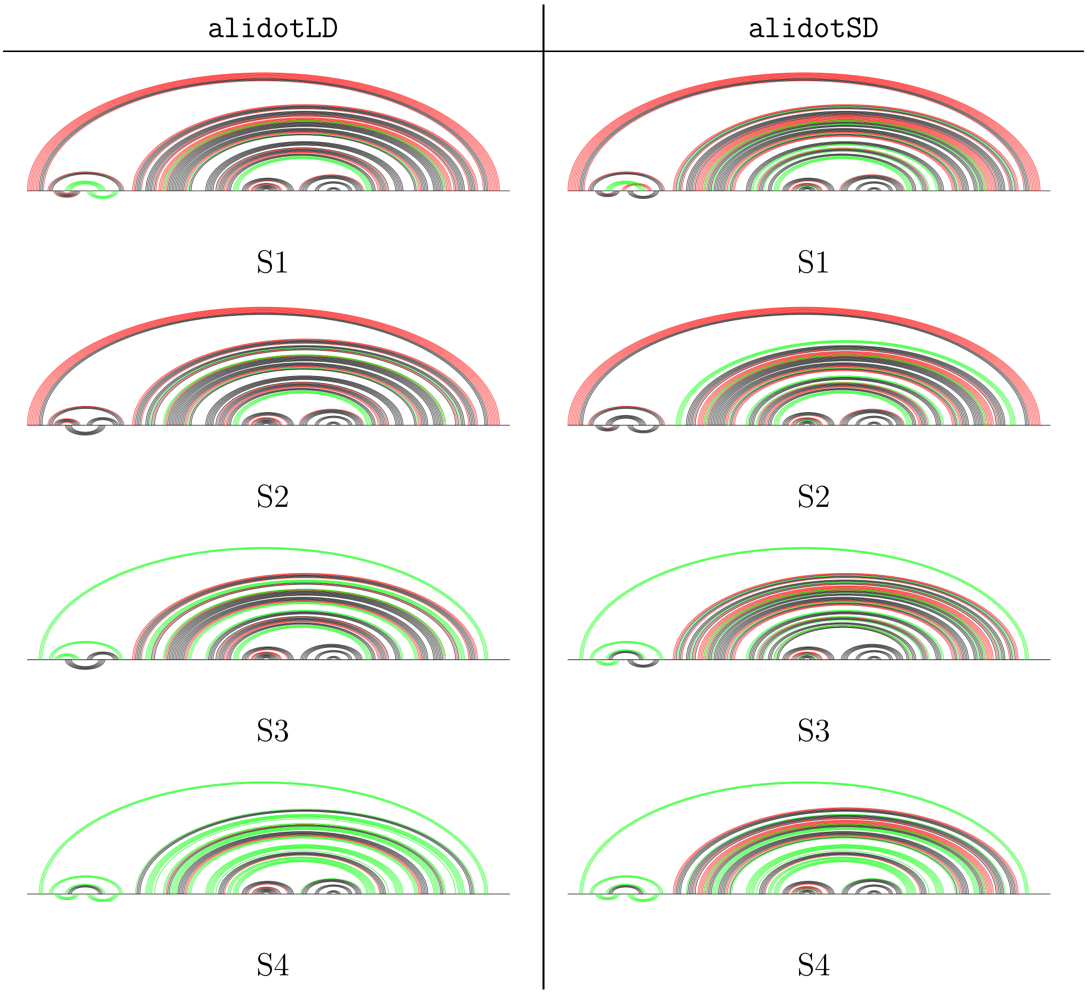


Figure 29: Prediction of consensus structure for the used datasets S1, S2, S3, and S4, calculated with both variants, `alidotLD` and `alidotSD`. The predicted structures are compared to the phylogenetically derived structure of *Methanocaldococcus jannaschii* SRP RNA [82]. We use this example to explain the representation of the results: base pairs which are predicted by `alidot` and which are part of the reference structure are shown in black, base pairs which are predicted by `alidot` but which are not part of the reference structure are shown red, and base pairs which are not predicted but which can be found in the reference structure are colored green.

In the following SRP RNA is compared to *Halobacterium halobium.* The same datasets are used as for comparison with *Methanocaldococcus jannaschii* with two exceptions: datasets S1b and S3b contain all sequences as are contained in datasets S1 and S3, respectively, but additionally have the sequence of *Halobacterium halobium* included.

Base pair specificity is always higher than 90%, and higher than sensitivity, see Tab. 8. In no case we predict stems that do not exist in the structure of *Halobacterium halobium.* In dataset S2 we detected all existing stems (stem sensitivity = 100%). As for observed for *Methanocaldococcus jannaschii*, we find the 'kissing hairpin' motif, which is composed by stems 3, 4, and 9, in dataset S2. In dataset S1b this motive is only found partly (pseudoknot stems 3 and 4), as well as in dataset S3b, where the pseudoknot formed by stems 4 and 9 is correctly predicted.

Table 8: SRP RNA compared to *Halobacterium halobium*

aln	N	μ	vers.	RP	TP	FP	S_{bp}	P_{bp}	RS	TS	FS	S_S	P_S
S1b	7	54.7	LD	86	66	7	76.7	90.4	8	7	0	87.5	100
			SD		66	7	76.7	90.4		7	0	87.5	100
S2	9	54.5	LD	86	79	8	91.9	90.8	8	8	0	100	100
			SD		72	8	83.7	90		8	0	100	100
S3b	14	54.7	LD	86	53	5	61.6	91.4	8	5	0	62.5	100
			SD		53	4	61.6	93		5	0	62.5	100
S5	29	54.5	LD	86	38	2	44.2	95	8	4	0	50	100
			SD		46	4	53.5	92		4	0	50	100

N number of sequences in the alignment, μ mean pairwise identity of the alignment, vers. version of `alidot`: LD for `alidotLD` and SD for `alidotSD`, RP number of base pairs in the reference structure, TP truly predicted pairs, FP falsely predicted pairs, S_{bp} sensitivity, P_{bp} specificity, TS true stems, RS reference stems, FS false stems, S_S stem sensitivity, P_S stem specificity.

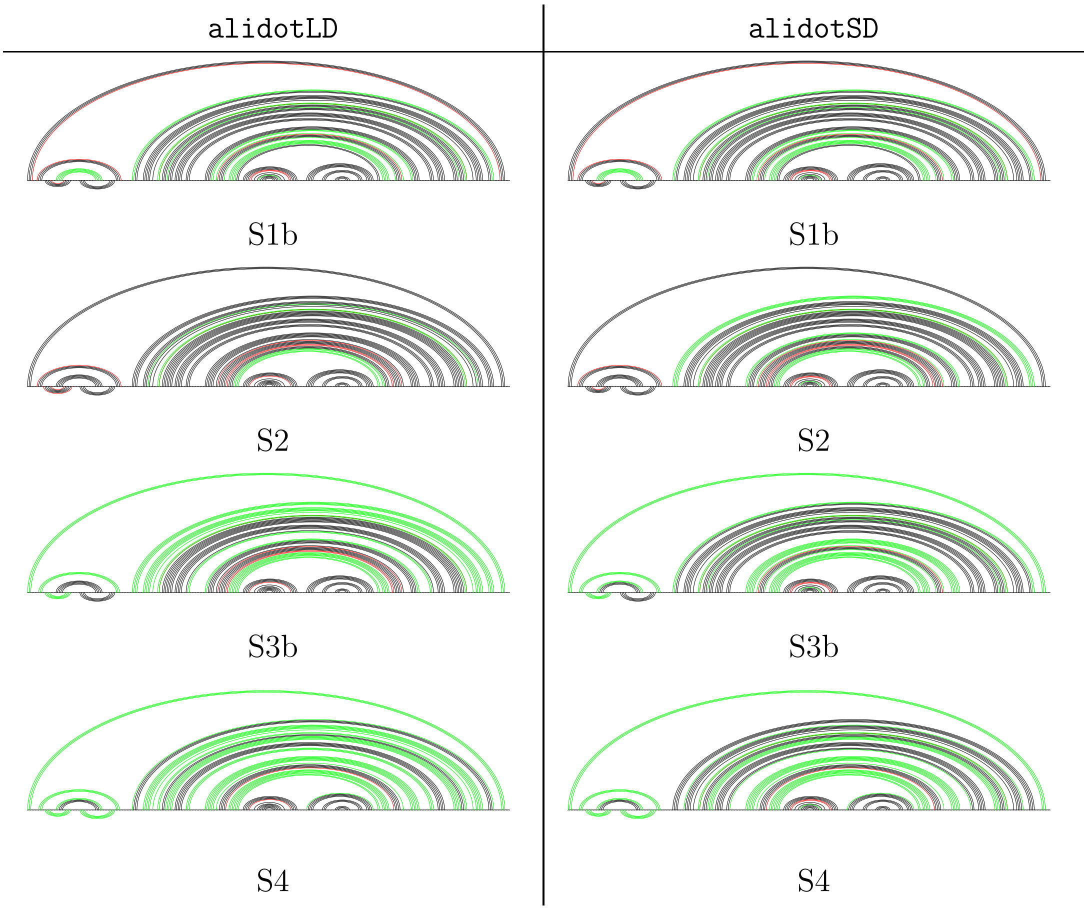


Figure 30: Prediction of consensus structure for the used datasets S1b, S2 S3b and S4, calculated with both variants, `alidotLD` and `alidotSD`. The predicted structures are compared to the phylogenetically derived structure of *Halobacterium halobium* SRP RNA [82].

Finally the datasets are evaluated against the bacterial stem *Bacillus subtilis* as reference structure, see Tab. 9. Datasets S1a and S2a again are the same as S1 and S2 but do additionally contain the sequence of *Bacillus subtilis*, respectively. Values for base pair specificity in all datasets are higher than 95%. Again, they are higher than values for base pair sensitivity. Especially in datasets S1a and S2a, but also in datasets S3a and S4, we find a notably low amount of falsely predicted base pairs in respect to data obtained when referring to the archaeal stems *Methanocaldococcus jannaschii* and *Halobacterium halobium*. In all datasets we received stem specificity of 100%, and in datasets S1a and S2a even stem sensitivity is high.

Fig. 6.3.1 shows that in datasets S1a and S2a we retrieve the 'kissing hairpin' motif very well. In dataset S3a we predict only pseudoknot stems 4 and 6 of *Bacillus subtilis* the 'kissing hairpin' motif, and in dataset S4 we find only stem 6.

Table 9: SRP RNA compared to *Bacillus subtilis*

aln	N	μ	vers.	RP	TP	FP	S_{bp}	P_{bp}	RS	TS	FS	S_S	P_S
S1a	7	49.9	LD	81	73	2	90.1	97.3	6	6	0	100	100
			SD		66	3	81.5	95.7		6	0	100	100
S2a	9	56.5	LD	81	70	2	86.4	97.2	6	6	0	100	100
			SD		68	4	84	94.4		6	0	100	100
S3a	14	52.5	LD	81	54	1	66.7	98.2	6	3	0	50	100
			SD		56	2	69.1	96.6		3	0	50	100
S4	29	54.5	LD	81	30	1	37.0	96.8	6	2	0	33.3	100
			SD		43	1	53.1	97.7		2	0	33.3	100

N number of sequences in the alignment, μ mean pairwise identity of the alignment, vers. version of `alidot`: LD for `alidotLD` and SD for `alidotSD`, RP number of base pairs in the reference structure, TP truly predicted pairs, FP falsely predicted pairs, S_{bp} sensitivity, P_{bp} specificity, TS true stems, RS reference stems, FS false stems, S_S stem sensitivity, P_S stem specificity.

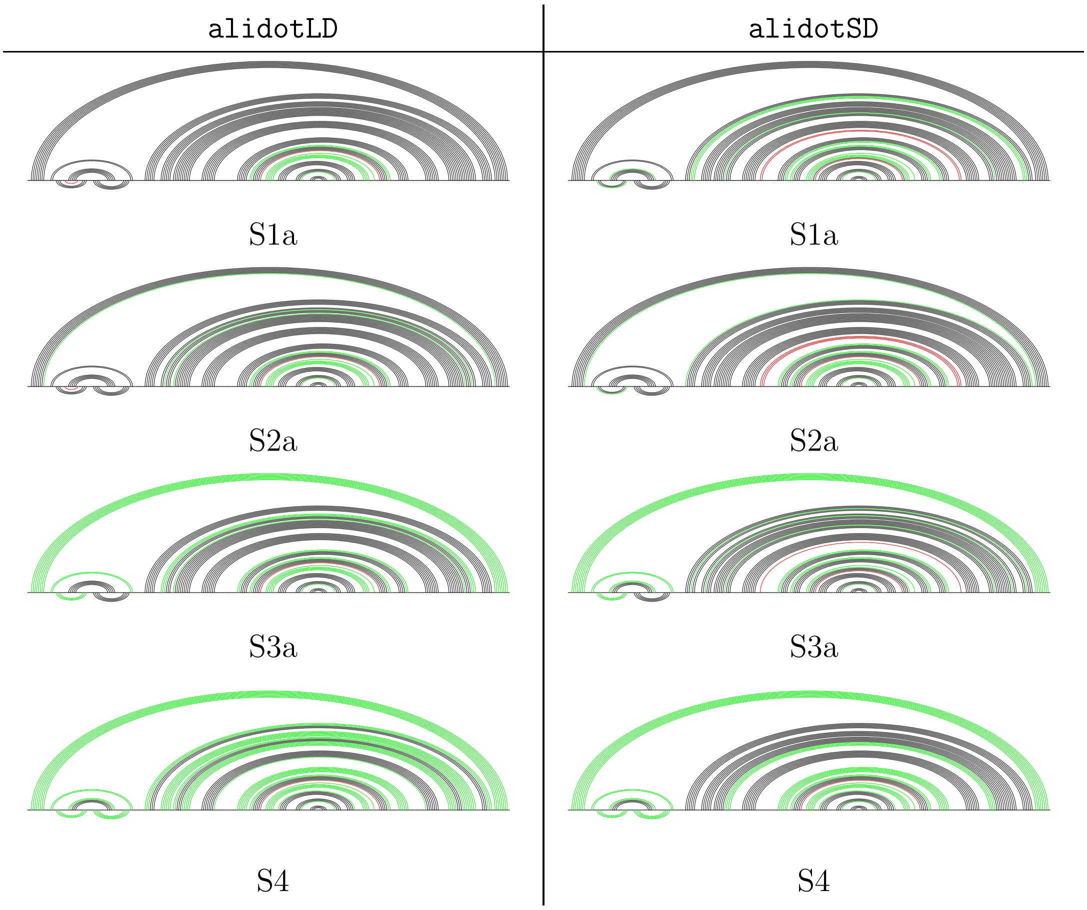


Figure 31: Prediction of consensus structure for the used datasets S1a, S2a, S3a and S4, calculated with both variants, `alidotLD` and `alidotSD`. The predicted structures are compared to the phylogenetically derived structure of *Bacillus subtilis* SRP RNA [82].

6.3.2 tmRNA

tmRNA (transfer-messenger RNA) is a cytoplasmic RNA found in bacteria. It is also known as 10SRNA or SsRNA, and combines both properties, tRNA and mRNA, in one molecule. For reviews see [101, 161, 154]. In archaea and eucaryota so far no homologous RNA has been found. tmRNA is believed to rescue ribosomes stalled on a truncated mRNA lacking a stop codon, and to attach a tag-protein to the truncated protein, by being the template by its own. This tag-protein serves as a signal for the proteolytic destruction of the defective protein.

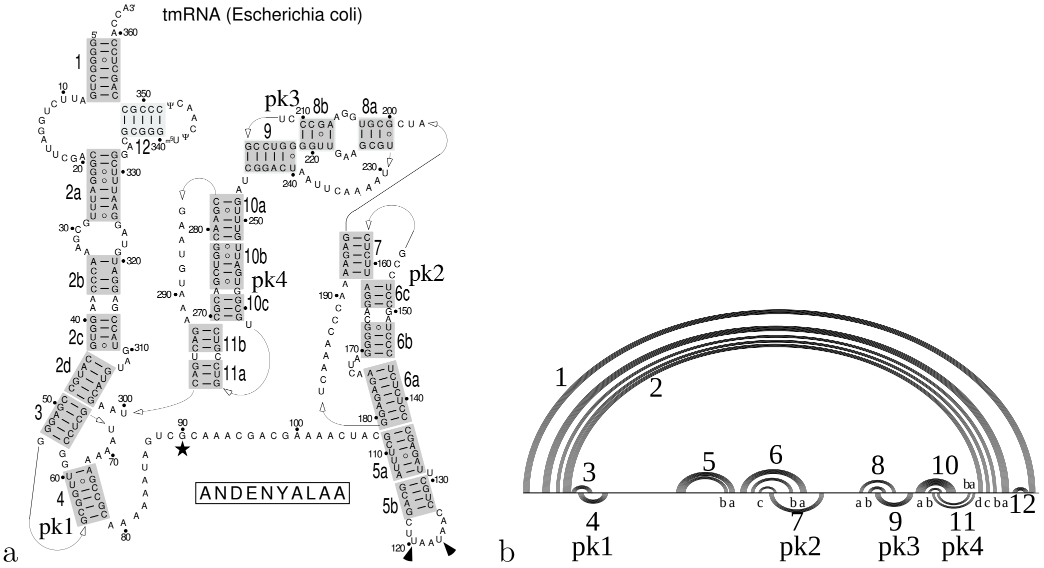


Figure 32: Reference Structure: tmRNA of *Escherichia coli*: (a) secondary structure representation (image adapted from [173]) and (b) representation in two dimensional diagram.

The consensus structures predicted by `alidot` are based on three different subsets of the alignment taken from the tmRNA Database [77], containing 5, 8, 22 bacterial tmRNA sequences, respectively. The sequences used in the respective datasets are listed in appendix A.3. Tab. 10 gives the mean pairwise sequence identity of the alignments and the number of correctly predicted base pairs and helices.

In dataset T2 we have a stem sensitivity of 100, i.e. we predict all stems correctly that exist in the reference structure. All pseudoknots are found by `alidotLD`, see Fig. 6.3.2, but there are two false stems predicted. `alidotSD` predicts only one false stem but does not find pseudoknot 4 (see also Fig. 32).

In dataset T3 we predict a considerable amount of false base pairs, but all detected stems do exist in the tmRNA of *Escherichia coli.* There are no predicted stems that do not exist in tmRNA of *Escherichia coli*, but the amount of predicted structures lies about 50%, due to considarable contribution of incompatible sequences.

Table 10: tmRNA compared to *Escherichia coli*

aln	N	μ	vers.	RP	TP	FP	S_{bp}	P_{bp}	RS	TS	FS	S_S	P_S
T1	5	73.7	LD	106	66	30	62.3	68.7	12	9	8	75	52.9
			SD		56	17	52.8	52.8		10	5	83.3	66.7
T2	8	60.2	LD	106	78	8	73.6	90.7	12	12	2	100	85.7
			SD		68	4	64.2	64.2		11	1	91.7	91.7
T3	22	66.1	LD	106	41	2	38.7	95.3	12	7	0	58.3	100
			SD		39	1	36.8	36.8		6	0	50	100

N number of sequences in the alignment, μ mean pairwise identity of the alignment, vers. version of `alidot`: LD for `alidotLD` and SD for `alidotSD`, RP number of base pairs in the reference structure, TP truly predicted pairs, FP falsely predicted pairs, S_{bp} sensitivity, P_{bp} specificity, TS true stems, RS reference stems, FS false stems, S_S stem sensitivity, P_S stem specificity.

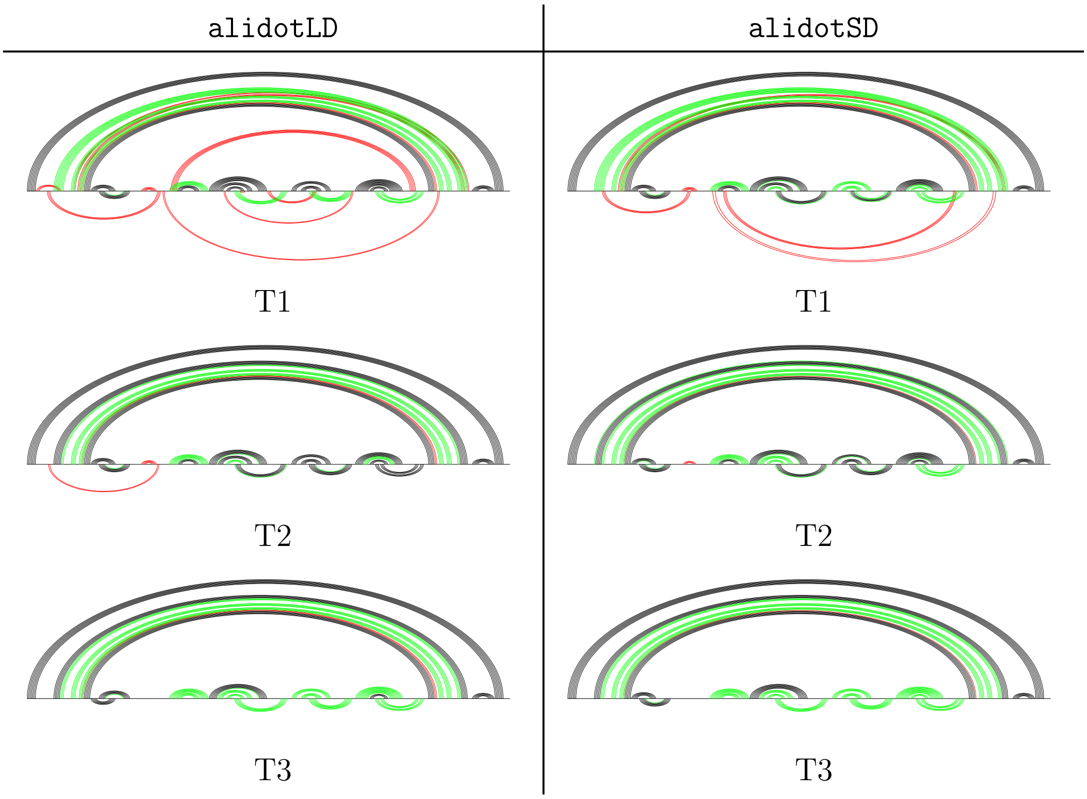


Figure 33: Prediction of consensus structure for the used datasets T1, T2, T3, calculated with both variants, `alidotLD` and `alidotSD`. The predicted structures are compared to the phylogenetically derived structure of *Escherichia coli* tmRNA.

6.3.3 RNase P RNA

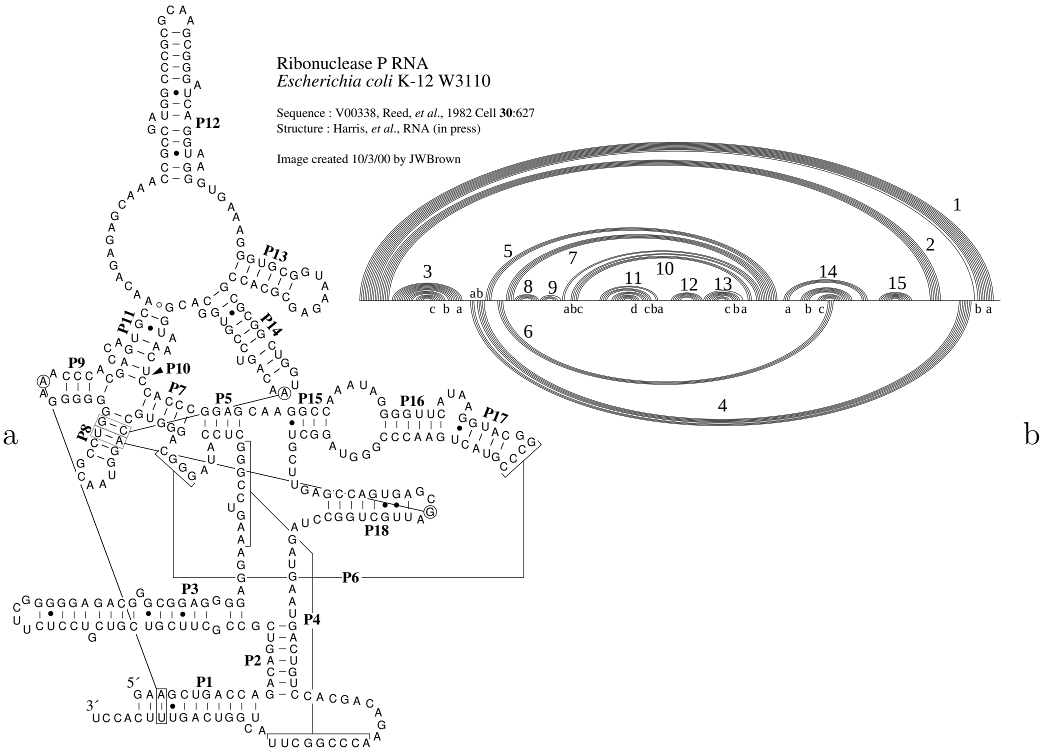

Ribonuclease P RNA
Escherichia coli K-12 W3110
Sequence : V00338, Reed, *et al.*, 1982 Cell **30**:627
Structure : Harris, *et al.*, RNA (in press)
Image created 10/3/00 by JWBrown
P1
P2
P3
P4
P5
P6
P7
P8
P9
P10
P11
P12
P13
P14
P15
P16
P17
P18
5´
3´
a
b
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
c b a
ab
abc
d cba
cba
a b c
b a

Figure 34: Reference Structure of RNase P of *Escherichia coli*: (a) secondary structure representation adopted from [10, 43], and (b) representation in two dimensional diagram. Note that the name of stems used by Brown (a) does not correspond to the stem names we used to identify similarities in (b). We aimed to follow as far as possible the names introduced in Fig. 35

Ribonuclease P (RNase P) is a ribonucleincomplex that catalyzes the removal of leader sequences from precursor tRNA (for review, see e.g. [29]). This ribozyme is present in all cells and organelles that carry out tRNA synthesis. Bacterial RNase P is composed of two subunits, an RNA (350-400 nucleotides) and a protein (about 120 amino acids). The RNA subunit of bacteria is catalytically active *in vitro* in the absence of the protein [36].

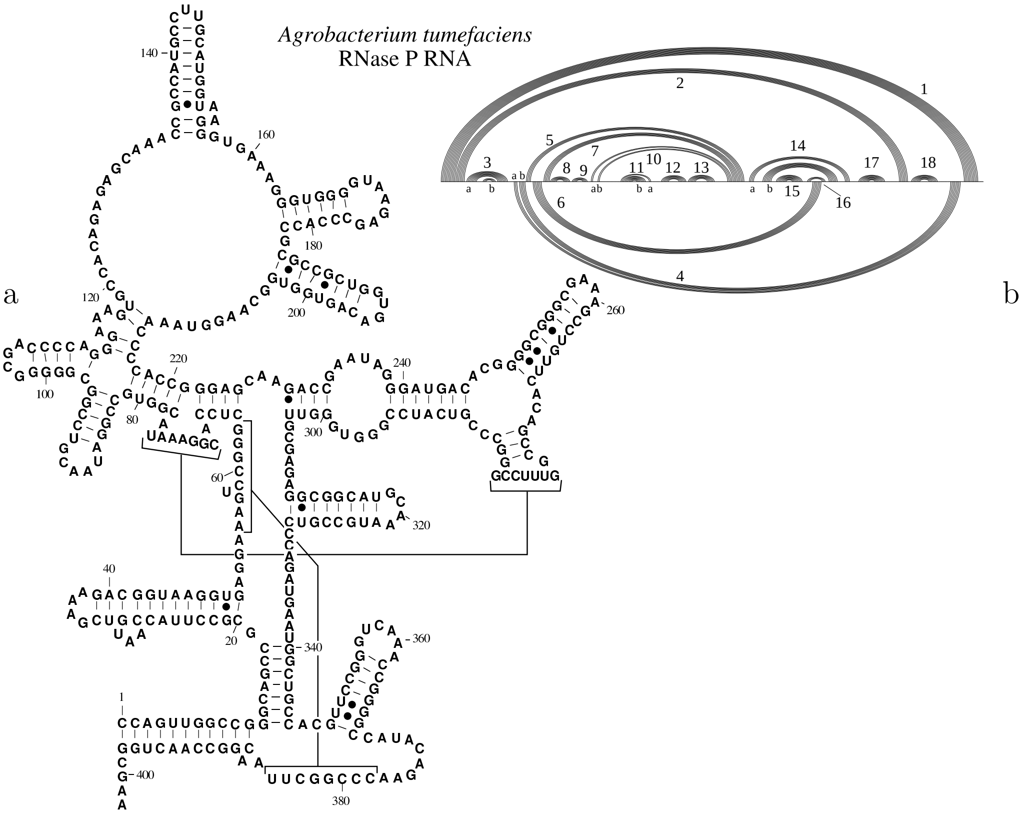


Figure 35: Reference Structure of RNase P of *Agrobacterium tumefaciens*: (a) secondary structure representation adopted from [10], and (b) representation in two dimensional diagram.

The reference structures of *Escherichia coli*, shown in Fig. 34, and *Agrobacterium tumefaciens* (Fig. 35) were obtained by comparative sequence analysis. The sequences were aligned manually, and sequence covariation was analyzed using the mutual information score combined with manual inspection [11, 43].

Bacterial RNase P RNAs fall into two broad classes, type A is the main form of RNase P RNA in bacteria, whereas type B is found only in the low G+C gram-positive bacteria. There is structural variation not only between the two types, but also between the instances of each structure type [41]. *Escherichia coli* and *Agrobacterium tumefaciens* are both representatives of type A RNase P RNA. Their structure differs mainly in that stems 15 and 18 of *Agrobacterium tumefaciens* are missing in *Escherichia coli.*

When comparing our dataset with the reference structure of *Escherichia coli* (see Tab. 11 and Fig. 6.3.3) `alidotSD` detects clearly less false base pairs and no false stems (except in dataset R1, where one false stem is predicted). Best results are obtained however for dataset R6, where both pseudoknots are recovered. In datasets R3 and R4 we find one of the two pseudoknots, respectively.

Table 11: RNase P RNA compared to *Escherichia coli*

aln	N	μ	vers.	RP	TP	FP	S_{bp}	P_{bp}	RS	TS	FS	S_S	P_S
R1	5	54.2	`LD`	124	94	13	75.8	87.9	15	13	2	86.7	86.7
			`SD`		75	6	60.5	92.6		12	1	80	92.3
R2	8	64.1	`LD`	124	86	13	69.4	86.7	15	11	3	73.3	78.6
			`SD`		85	3	68.5	96.6		11	0	73.3	100
R3	8	58.4	`LD`	124	92	9	74.2	91.1	15	14	3	93.3	82.4
			`SD`		80	0	64.5	100		13	0	86.7	100
R4	10	59.7	`LD`	124	91	7	73.4	92.9	15	14	2	93.3	87.5
			`SD`		84	1	67.7	98.8		14	0	93.3	100
R5	20	63.2	`LD`	124	49	8	60.5	86.0	15	14	2	93.3	87.5
			`SD`		56	11	45.2	83.6		14	0	93.3	100

N number of sequences in the alignment, μ mean pairwise identity of the alignment, vers. version of `alidot`: LD for `alidotLD` and SD for `alidotSD`, RP number of base pairs in the reference structure, TP truly predicted pairs, FP falsely predicted pairs, S_{bp} sensitivity, P_{bp} specificity, TS true stems, RS reference stems, FS false stems, S_S stem sensitivity, P_S stem specificity.

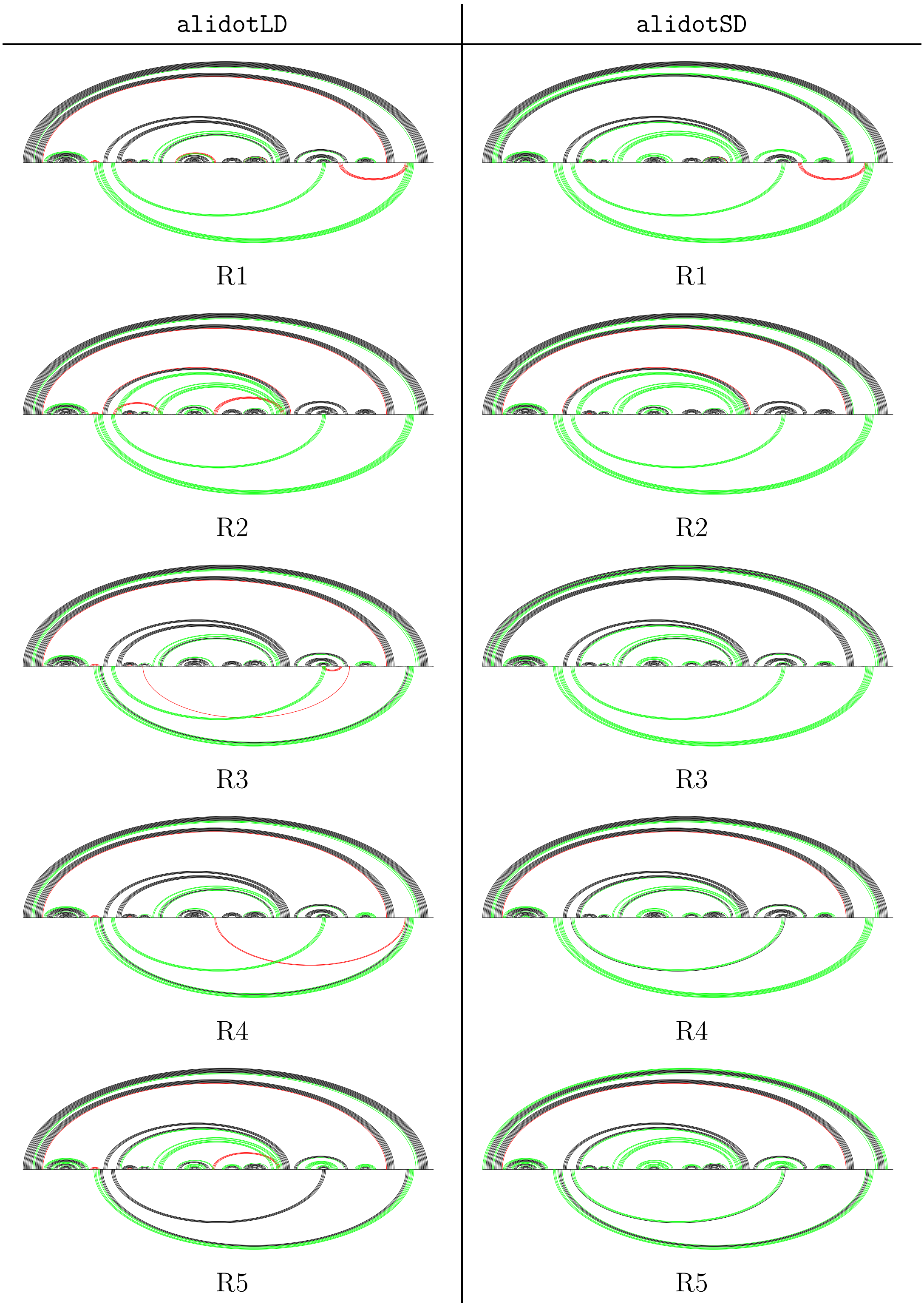


Figure 36: Prediction of consensus structure for the used datasets R1, R2, R3, R4 and R5 calculated with both variants, `alidotLD` and `alidotSD`. The predicted structures are compared to the phylogenetically derived structure of *Escherichia coli* RNase P RNA.

In the following the same data are compared to *Agrobacterium tumefaciens*. In analogy to previous results, also here in dataset R5 the two pseudoknots are predicted correctly. For dataset R4 each of the two alternatives of `alidot` recovered one of the two pseudoknots, respectively. While `alidotSD` does not predict false stems, `alidotLD` retrives one false stem. In dataset R3 stem 4 of the pseudoknot is correctly predicted by `alidotLD`, and the amount of falsely predicted stems is very high. For dataset R1 both variants of `alidot` find one false pseudoknot. In dataset R2 `alidotLD` predicts one pseudoknot correctly but retrieves 3 false stems.

Table 12: RNase P RNA compared to *Agrobacterium tumefaciens*

aln	N	μ	vers.	RP	TP	FP	S_{bp}	P_{bp}	RS	TS	FS	S_S	P_S
R1	5	54.2	`LD`	124	94	20	75.8	82.5	18	15	3	83.3	83.3
			`SD`		79	10	63.7	88.8		15	1	83.3	93.8
R2	8	64.1	`LD`	124	94	20	75.8	82.5	18	15	3	83.3	83.3
			`SD`		88	5	71	94.6		14	0	77.8	100
R3	8	58.4	`LD`	124	93	16	75	85.3	18	17	3	94.4	85
			`SD`		83	6	66.9	93.3		15	0	83.3	100
R4	10	59.7	`LD`	124	92	13	74.2	87.6	18	16	3	88.9	84.2
			`SD`		89	7	71.8	92.7		16	0	88.9	100
R5	20	63.2	`LD`	124	84	10	67.7	89.4	18	17	2	94.4	89.5
			`SD`		75	3	60.5	96.2		16	0	88.9	100

N number of sequences in the alignment, μ mean pairwise identity of the alignment, vers. version of `alidot`: LD for `alidotLD` and SD for `alidotSD`, RP number of base pairs in the reference structure, TP truly predicted pairs, FP falsely predicted pairs, S_{bp} sensitivity, P_{bp} specificity, TS true stems, RS reference stems, FS false stems, S_S stem sensitivity, P_S stem specificity.

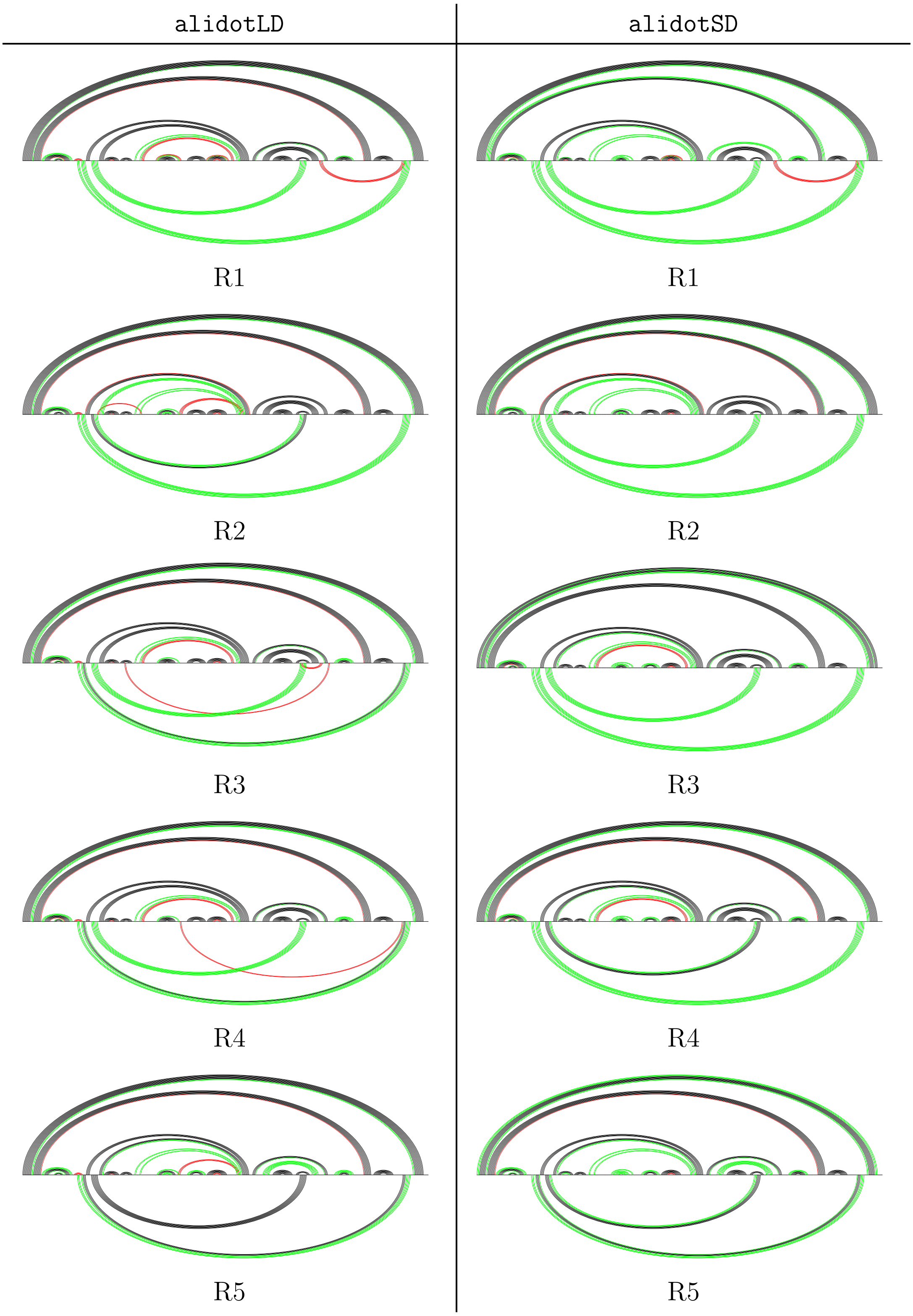


Figure 37: Prediction of consensus structure for the used datasets R1, R2, R3, R4 and R5 calculated with both variants, `alidotLD` and `alidotSD`. The predicted structures are compared to the phylogenetically derived structure of *Agrobacterium tumefaciens* RNase P RNA.

We observed that given a specific alignment the choice of the reference structure does not influence the quality of the prediction of secondary and bisecondary structures. Nor does the number of aligned sequences have a substantial effect on the predicted structures. The crucial variable in reaching good predictions with `alidot` is, as it is also for other algorithms, the quality of the alignment. Here we used alignments which were optimized by manual inspection, mutual information scores and other elaborate techniques. From these alignments we choose sequence sets with a mean pairwise identity between 50% and 60%, as this is the range where other algorithms work best. Still we observe a notable influence of the sequences used in the alignment. For instance, R2 and R3 both contain 8 sequences, both alignments are in a similar range of mean pairwise identity, but the predicted secondary and bi-secondary structures differ notably for `alidotLD` and `alidotSD` (Fig. 6.3.3 and Fig. 6.3.3).

6.4 Critical Comparison of our Predictions with Other Algorithms

We compared the predicted secondary structure of `alidot` to the predictions of two other algorithms. These are `ilm` of Ruan *et al.* [125] and `hxmatch` of Witwer *et al.* [157]. Both algorithms are introduced in section 3.4.2.

In Tab. 13 and Tab. 14 we present the results of our calculations. All calculations were carried out on the same datasets. We used the datasets Witwer *et al.* applied in [157]. For SRP RNA we used the alignment S2 with the reference structure *Bacillus subtilis*. For tmRNA we calculated with the alignment T2 and compared with *Escherichia coli*, and for RNase P RNA we used dataset R3 with *Agrobacterium tumefaciens* as reference structure.

Compared to `ilm` and `hxmatch` the quality of the prediction of secondary and bi-secondary structures of `alidotLD` as well as `alidotSD` range among the other methods, being a little better than `ilm` and a little worse than

Table 13: Quality of predictions of S_{bp} and P_{bp} compared to ilm and hxmatch

	ilm			hxmatch			alidotLD		
	S_{bp}	P_{bp}	PK	S_{bp}	P_{bp}	PK	S_{bp}	P_{bp}	PK
SRP RNA	86.0	66.6	0/1	91.9	84.9	1/1	90.9	90.8	1/1
tmRNA	89.6	71.4	4/4	84.0	90.8	4/4	73.6	90.7	4/4
RNase P RNA	75.8	76.4	1/2	77.4	88.9	2/2	75.0	85.3	1/2

For comparison of base pair sensitivity S_{bp} and specificity P_{bp}, both are defined in section 6.3, with the algorithms ilm and hxmatch we used alidotLD. PK gives the number of pseudoknots found in relation to the number of pseudoknots in the reference structure.

hxmatch. While specificity (Tab. 13) of alidotLD is comparable to hxmatch, it is notably better than specificity of ilm. Sensitivity (Tab. 13) of hxmatch is always best, while ilm and alidotLD are approximately equal. hxmatch finds all pseudoknots in all three tested datasets. ilm and alidot find all pseudoknots in tmRNA and one of two pseudoknots of RNase P RNA. In SRP RNA alidot finds the only pseudoknot.

Table 14: Quality of predictions of S_S and P_S compared to ilm and hxmatch

	ilm			hxmatch			alidotSD		
	S_S	P_S	FS	S_S	P_S	FS	S_S	P_S	FS
SRP RNA	87.5	87.5	1	100	100	0	100	100	0
tmRNA	91.7	73.3	4	100	85.7	2	91.7	91.7	1
RNase P RNA	88.9	80.0	4	94.4	100	0	83.3	100	0

For comparison of stem sensitivity S_s and specificity P_s, both are defined in section 6.3, with the algorithms ilm and hxmatch we used alidotSD. PK gives the number of pseudoknots found in relation to the number of pseudoknots in the reference structure.

In Tab. 14 we compare stem sensitivity and stem specificity obtained by alidotSD with hxmatch and ilm. For stem sensitivity hxmatch obviously obtains the best results, by finding all predictable stems for SRP RNA and tmRNA and reaching 94% of stems in the case of RNase P RNA. For SRP RNA `alidot` predicts all stems. For tmRNA `alidot` recovers 91.7% of stems, i.e. the same as `ilm`. In two of the three tested sets `hxmatch` and `alidot` reach a stem specificity of 100%, i.e. no false stems were predicted. In the case of tmRNA only `alidot` detected one single false stem. In this case `alidot` obtains an even higher stem specificity than `hxmatch`.

7 Automated Text Categorization

Bibliographic search for experimental evidence for secondary structures and further information thereon in a given group of viruses turned out to be more tedious than the work on the actual sequence and structure data.

There are several reasons for this difficulty: (i) RNA secondary structure is usually referred to only as " secondary structure" or simply as "structure" since the context "RNA" is clear. The term "secondary structure", however, appears much more frequently in the context of protein structures for the same virus group because proteins are usually discussed more frequently and in much more detail. (ii) RNA secondary structures are rarely the main topic of research papers on viruses. Rather, only one or a few paragraphs are devoted to them. (iii) With few exceptions there is no well-established nomenclature of RNA features in viruses so that keyword searches for specific structural motifs are not very effective. (iv) Relevant articles are written by authors from rather diverse scientific communities, from clinical virologists to structural biologists.

We therefore set out to develop an automated text categorization tool for bibliographic search. Lukas Faulstich, from the Insitut für Informatik, Universität Leipzig, Germany, was engaged in programming, combining algorithms, and training of the tool. As input for training he used elaborate data sets of hundrets of manually labeled and classified documents. These were provided for the virus family *Picornaviridae* by Christina Witwer, from the Universität Wien, and for the particularily feasible virus family *Flaviviridae* as part of this thesis.

Our target topic of "conserved RNA secondary structure in viral genomes" consists of several subtopics, each dedicated to a specific group of RNA viruses (e.g., *Picornaviridae*, *Flaviviridae*, *Coronaviridae*, or *Hepadnaviridae*). For some of these subtopics, we supplied manually labeled document corpora. The question addressed in this exploratory study is whether classifiers trained for one subtopic can be applied successfully to other subtopics. This would be in particular attractive for subtopics with a large amount of available literature, e.g., on the HIV virus in the case of *Retroviridae*. In this context, successful means a high recall (e.g., 80%) with a not too low precision (e.g., 30%) because the emphasis is on finding most of the relevant literature with a tolerable overhead caused by false positives.

The results presented here indicate that a classifier trained on one virus group can be applied successfully to search the literature on other virus groups. Therefore, a system for supporting bibliographic search based on automated text categorization seems feasible for our target topic.

Training data has been obtained from searching the Pubmed collection via the entrez interface[2] and then downloading the referenced articles as PDF documents (as far as available). The search queries (see Tab. 15) have been specified for *Picornaviridae* by Christina Witwer and for *Flaviviriae* as part of this thesis. The resulting corpora are referred to as picorna and flavi.

Since corpus picorna is quite small and corpus flavi contains only few positive examples, we decided to add more documents from our bibliographical collections. The resulting corpora are referred to as picorna2 and flavi2. A document is considered a positive example within its corpus if it contains information on the secondary structure of the RNA of viruses belonging to the virus group (*Picornaviridae*, *Flaviviridae*) the corpus is dedicated to.

7.1 Methods

7.2 Data Preparation

The PDF documents where converted into text using the Unix tools pdftotext and ps2ascii. The ConceptComposer text analysis suite [49] was used to build

[2] http://www.ncbi.nlm.nih.gov/Entrez/

Table 15: The training corpora.

Corpus	Source	Size	Positive
picorna	Pubmed query: picornavirus RNA secondary structure	40	68%
picorna2	picorna + 24 extra documents	64	58%
flavi	Pubmed query: RNA AND (IRES OR "secondary structure" OR "conserved structure" OR "5'utr" OR "3'utr" OR "coding region") AND ("hepatitis C virus" OR "hepatitis G virus" OR pestivirus OR dengue OR "japanese encephalitis virus" OR "yellow fever virus"OR "tick-borne encephalitis virus")	153	8%
flavi2	flavi + 34 extra documents	187	12%

a full text index of the resulting text documents in a relational database (mysql).

Based on this index, the documents were transformed into vector representation using a SQL script. We computed term weights according to the standard tfidf method (see e.g. [126]). Each corpus is stored in a separate mysql database.

For feature selection we implemented the term relevance measures *Odds Ratio* and *Mutual Information* (see [131]). In addition we implemented derived term relevance measures where the original relevance value for a term is weighted with its frequency in the test database that is used for evaluation.

7.3 Text Categorization

We built the Java application litsift on top of the Weka 3 machine learning software [155] to classify the document corpora. This enabled us to experiment with the variety of classifiers provided by Weka. Further parameters that can be varied are

- the term relevance measure to use for feature selection
- the number of features to be taken into account
- the target recall when evaluating a classifier on the test corpus
- classifier specific parameters

The application reads class labels for documents and their term weights for the selected features from the training database and creates a set of Weka instances from it. This instance set is either used for cross evaluation on the training corpus or it is used to train a classifier that is evaluated on a separate test corpus. In the latter case, only those documents are classified as positive whose predicted class-membership probability exceeds a certain threshold. This threshold is adjusted automatically to achieve at least the chosen target recall (if possible at all) in a trade-off with the achieved precision. The threshold is found by computing histograms on the number of positives and true positives over the predicted probabilities.

7.4 Results

Before we assess the applicability of classifiers trained on one corpus to another corpus, we present cross-evaluation results on each corpus as a base line for comparison.

Table 16: Filtering results for different corpora, and relevance measures, with target recall 100%. Column “avg. pre.” shows the average precision over all feature counts where the target recall is exceeded. Column “max. pre.” shows the maximum precision. The minimum feature count at which the maximum precision is reached is shown in “no. f.”. The recall achieved with this number of featured is shown in “rec.”.

corpus	rel. measure	avg. pre.	max. pre.	rec.	no. f.
flavi	mutual information	11.2%	23.1%	100.0%	20
flavi	odds ratio	7.8%	7.9%	100.0%	10
flavi2	mutual information	20.2%	40.7%	100.0%	20
flavi2	odds ratio	11.8%	11.8%	100.0%	10
picorna	mutual information	76.7%	100.0%	100.0%	20
picorna	odds ratio	67.6%	69.2%	100.0%	10
picorna2	mutual information	69.3%	100.0%	100.0%	30
picorna2	odds ratio	58.0%	59.7%	100.0%	10

7.4.1 Feature Selection

To assess the performance of different term relevance measures, we varied the number N of features. From the corpus we filtered those documents that contained at least one of the N best terms of the chosen measure. Then we computed precision and recall of this filter by counting the selected documents as positives and the rest of the corpus as negatives. The results are shown in Table 16. It shows that 10–30 features are always sufficient to retrieve all positive examples. Moreover it shows that the corpora picorna and picorna2 are quite trivial since they can be classified completely and correctly by using just the first 20 (picorna) or 30 (picorna2) features selected by Mutual Information.

7.4.2 Cross Evaluation on Each Corpus

As a base line for comparison we cross-evaluated several classifiers from the Weka toolkit in combination with the available term relevance measures on each corpus. It shows that

1. on flavi and flavi2 the target recall of 80% can be reached only by the NaiveBayes classifier
2. with few exceptions, less than 50 features are needed to achieve maximum recall
3. corpora picorna and picorna2 can be almost perfectly classified in most cases
4. J48 seems sensitive with respect to the relevance measure: on flavi2, Odds Ratio performs much better, on picorna2, Mutual Information performs much better.

7.4.3 Validation on Separate Test Corpus

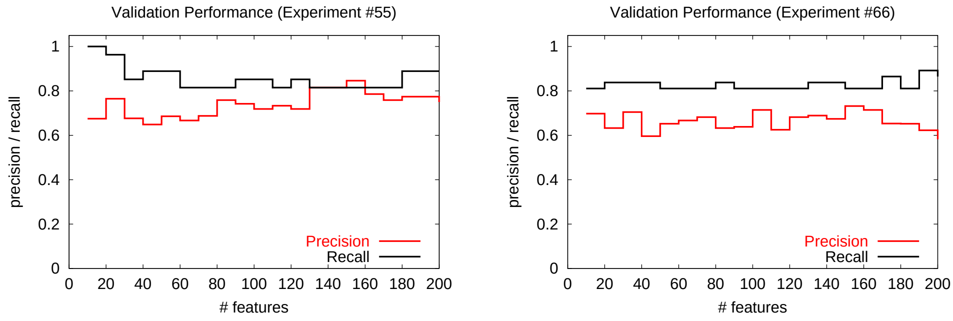
We first present some exemplary experiments with SMO and then give in an overview of all experiments in form of a table.

Training on flavi, Validation on picorna

A SMO classifier trained on flavi with Odds Ratio measure evaluated on picorna2 reaches the target recall of 80% beginning with 30 features. The precision reaches a maximum of 80% at about 150 features (see Fig. 38a). Using Mutual Information yields similar results.

Training on flavi2, Validation on picorna2

Compared to Sec. 7.4.3, the average precision of SMO drops slightly from 74% to 66% (see Fig. 38b) which is still quite acceptable for bibliographic



(a) (b)

Figure 38: Performance for Weka SMO with Odds Ratio, target recall 80%: (a) on corpus picorna after training on corpus flavi, (b) on corpus picorna2 after training on corpus flavi2.

search.

Training on picorna, Validation on flavi

While SMO trained on corpus flavi can be successfully applied to the corpora picorna and picorna2, the inverse setting is not as successful. At 80% recall, SMO achieves a maximum precision of 23% precision at 60 features (see Fig. 39a).

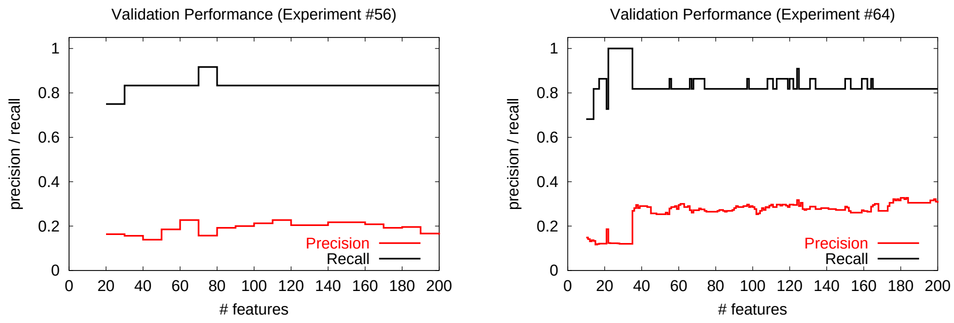
With Mutual Information, the precision is even lower (about 10%).

Using a derived term relevance measure (Odds Ratio, weighted with term frequencies from flavi) did not yield any improvement, either.

23% precision may not seem high, but in our application to bibliographic search it is still more tolerable than in other fields of text classification.

Training on picorna2, Validation on flavi2

Compared to Sec. *Training on picorna, Validation on flavi*, precision of SMO increases to 30% starting from 40 features (see Fig. 39b).



(a) (b)

Figure 39: Performance for Weka SMO with Odds Ratio, target recall 80%: (a) on corpus flavi after training on corpus picorna, (b) on corpus flavi2 after training on corpus picorna2.

7.5 Discussion

We may summarize the results as follows:

1. Corpora picorna and picorna2 can quite successfully be classified after training on flavi and flavi2, respectively.
   (a) With J48 or NaiveBayes, 100% recall can be achieved with maximum precisions above 70%, using only few features (10–30).
   (b) Mutual Information seems to perform better than Odds Ratio.
2. Corpora flavi and flavi2 can not as easily classified after training on picorna and picorna2, respectively.
   (a) The best maximum precision is achieved by SMO with Odds Ratio
   (b) Corpus flavi2 is easier to classify than flavi
3. In most cases the difference between average and maximum precision is quite small. This supports the observation from Figs. 38 and 39 that precision does not depend too much on the number of features.

The asymmetry between the picorna* and flavi* corpora can to some extent be explained by the fact that the *Flaviviridae* virus group is more heterogenous than the *Picornaviridae* group. For instance, while all *Picornaviridae* genomes have so-called IRES (Internal Ribosomal Entry Site) regions, this does not hold for all *Flaviviridae*. This means that a classifier trained on a picorna* corpus only finds those positive examples in flavi* that are similar to those in the training corpus. In the other direction this partition within a flavi* corpus seems to be sufficient to learn the characteristics of the positive examples in the picorna* corpora. The additional positives in corpus flavi2 might be more "picorna"-like which would explain the better performance when testing on flavi2 instead of flavi.

8 Conclusion and Outlook

The knowledge about the spatial conformation of functional RNA molecules is a crucial prerequisite to understand how they work. In order to obtain a model that is theoretically and computationally easier to manage, the RNA secondary structure, an interesting class of contact structures, is introduced. Secondary structures provide a coarse graining of the 3D structure by regarding base pair pattern only.

We have employed a combination of structure prediction based on thermodynamic rules and the evaluation of consistent and compensatory mutations to search the genomes of the virus family *Flaviviridae* for functional RNA structure motifs. While the UTRs of some of these viruses have been studied previously, this thesis reports a comprehensive survey of structural features across the full genomes of the whole family *Flaviviridae*.

The 5' untranslated region (UTR) of a number of these viruses has been studied previously because of the particular interest in the IRES region. Our automatic approach confirms many of the patterns identified previously based on smaller data sets. However, we find that in many cases the parts of these features that are conserved base-pair by base-pair are significantly smaller than reported. This conclusion is mainly based on the fact that some sequences that are now contained in the database simply cannot form parts of the structures that have previously been reported as conserved. The same conclusion can be drawn for the 3'UTR.

Furthermore, instead of using a "sliding window" technique, all predictions were carried out for the complete genomic RNA sequences. This enables our algorithm to find long range interactions, in particular we found significant probability for genome cyclization in all genera except *Pestivirus*.

For the genus *Flavivirus* cyclization domains of the genome were proposed already previousy. Our data suggest that cyclization regions are larger as supposed. This finding is going to be addressed by the group of Prof. Christian Mandl at the institute of Virology, Vienna, For the strain Neudoerfl, a member of the species tick-borne encephalitis virus.

Most surprisingly we found evidence for viral genome cyclization also in GBV-C and HCV, which had not been reported before, although Yi *et al.* [166] suppose a cyclization of HCV genome by the assistance of some cellular protein. Our algorithm detected base pair probabilities for both, previously reported secondary structures in 5' and 3'UTRs as well as for genome cyclization. For both cases, our data revealed no inconsistencies. Thus we suppose that known structures compete with genome cyclization. This finding is at present experimentally studied on the genome of HCV by the group of Prof. Bartenschlager in Heidelberg.

Furthermore we present a large number of secondary structure elements that have not been described before, most importantly within the coding region. This information could be used to identify additional regions which might be important for virus viability and propagation, and thus to gain more insight into the life-cycle of the members of the family *Flaviviridae*.

Our algorithm combines successfully covariational and thermodynamic information to predict consensus secondary structure from a small set of homologous sequences. RNA pseudoknots mediate several biological functions, like translational and replicational control, others are necessary to form the reaction center in ribozymes. Therefore it is desirable to allow our algorithm also to predict pseudoknots.

Bi-secondary structures are often contained in the thermodynamic data in the form of competing stems. In the second part of this thesis we therefore extended the algorithm to search through the data for possible alternative stems and combine them to pseudoknots. We followed two different ways to achieve this task and compared them with each other. While one method considers base pairs as independent and only in the last step combines them to stems, the other method first combines all possible base pairs to stems, then sorts, ranks and filters them.

The program was tested on three different types of RNA known to contain pseudoknots: Signal Recognition Particle RNA, Ribonuclease P RNA, and tmRNA. The alignments used had a mean pairwise identity about 55%. While the first method is more sensitive to the prediction of single base pairs, the second method reduces significantly the number of false positive stems and pseudoknots.

A substantial improvement of the algorithm could probably be achieved by a more elaborate hierarchy of patterns of stem credibility. For instance, we give too much negative weight on incompatibility. Apart from simply reducing the contribution of incompatible mutations to stem ranking, it should also be evaluated, whether incompatibilities are caused by one single sequence, in that this sequence does not express the respective structure element, or different sequences contribute with single incompatible mutations to stem rating. An improvement could also be to completely neglect probabilities, when stems of higher layers are ranked. Maybe it would be worth while to combine both methods we developed. In a first step predict secondary structures based on classical base pair ranking and then combine the remaining base pairs to stems and proceed as used in stack-based layer decomposed `alidot`.

The results obtained by automated text categorization in bibliographic search are rather heterogeneous. Nevertheless, they indicate that classifiers trained on one subtopic can be applied to other subtopics and achieve precisions (here 20% – 100%) that will result in cost savings when searching for relevant literature while not too many (here 20%) relevant documents are lost.

The complications of bibliographic search that plague the case of RNA secondary structure features in viral RNAs are not restricted to this particular topic. Whenever the available literature has to be searched for information that is rarely the main focus of the publication keyword-based searches tend to either have low recall or low precision. Regulatory sequences associated with certain classes of genes may serve as another example.

We thus plan to extend the litsift application into a bibliographic search tool that sends a user query to a bibliographic database such as Pubmed, retrieves the search results and the articles cited therein, and ranks the results according to the predictions of a classifier previously trained using the same tool. The user may choose to re-label some of the results manually and to retrain the classifier in order to enhance the classifier.

A Appendix

A.1 List of Flaviviridae Sequences

Table 17: List of *Flavivirus* sequences partI

Flavivirus				
ID	Acc.No.	Length	Organism	Strain/(Isolate/Clone)
Serotype: *Dengue virus* (DEN)				
AF350498	AF350498	10735	DEN type 1	GZ/80
NC_001477	NC_001477	10735	DEN type 1	(clone=”45AZ5”)
AF226686	AF226686	10735	DEN type 1	FGA/NA d1d
AF317645	AF317645	10696	DEN type 3	80-2
AF208496	AF208496	10722	DEN type 2	DEN2/H/IMTSSA-MART/98-703
DEN2JAMCG	M20558	10723	DEN type 2	Jamaica/N.1409
AF119661	AF119661	10723	DEN type 2	(China 04)
AF022434	AF022434	10724	DEN type 2	ThNH-7/93
AF022437	AF022437	10723	DEN type 2	ThNH-p11/93
AF022435	AF022435	10723	DEN type 2	ThNH-28/93
AF169678	AF169678	10723	DEN type 2	ThNH29/93
AF204177	AF204177	10723	DEN type 2	44
AF276619	AF276619	10723	DEN type 2	FJ-10
DENRCG	M19197	10703	DEN type 2	S1 vaccine strain
NC_001474	NC_001474	10703	DEN type 2	
AF326573	AF326573	10694	DEN type 4	814669

Table 18: List of *Flavivirus* sequences partII

Flavivirus				
ID	Acc.No.	Length	Organism	Strain/(Isolate/Clone)
Species: *Japanese encephalitis virus* (JEV)				
AF014160	AF014160	10976	JEV	RP-2ms
AF014161	AF014161	10976	JEV	RP-9
AF069076	AF069076	10977	JEV	JaGAr 01
AF098735	AF098735	10976	JEV	HVI
AF098736	AF098736	10976	JEV	TC
AF098737	AF098737	10976	JEV	TL
AF315119	AF315119	10976	JEV	SA14-14-2
JEU15763	U15763	10976	JEV	
D90194	D90194	10976	JEV	SA14
AF221499	AF221499	10976	JEV	CH2195LA
AF221500	AF221500	10976	JEV	CH2195SA
JEVCG	M18370	10976	JEV	JaOArS982
NC_001437	NC_001437	10976	JEV	JaOArS982
AF075723	AF075723	10976	JEV	GP78
JEU47032	U47032	10976	JEV	p3
L48961	L48961	10976	JEV	Beijing-1
AF080251	AF080251	10977	JEV	(Vellore P20778)
AF045551	AF045551	10963	JEV	K94P05
AF217620	AF217620	10964	JEV	FU

Table 19: List of *Flavivirus* sequences partIII

Flavivirus				
ID	Acc.No.	Length	Organism	Strain/(Isolate/Clone)
Species: *Yellow fever virus* (YFV)				
AF094612	AF094612	10760	YFV	Trinidad 79A (788379)
U21055	U21055	10862	YFV	French neurotropic virus
NC_002031	NC_002031	10862	YFV	
U17067	YFU17067	10862	YFV	vaccine strain 17D-213
U17066	YFU17066	10862	YFV	vaccine strain 17DD
U21056	YFU21056	10862	YFV	French viscerotropic virus
U54798	YFU54798	10862	YFV	85-82H Ivory Coast
Species: *Tick-borne encephalitis virus* (TBE)				
TEU27491	U27491	11141	TBE	263
TEU27495	U27495	11141	TBE	Neudoerfl
TEU39292	U39292	10835	TBE	Hypr
AB062063	AB062063	11100	TBE	Oshima 5-10
AB062064	AB062064	10894	TBE	Sofjin-HO
L40361	L40361	10927	TBE	Vasilchenko

Table 20: List of *GB Virus C/Hepatitis G virus* sequences

GB Virus C/Hepatitis G virus				
ID	Acc.No.	Length	Group	Strain/(Isolate/Clone)
D87709	D87709	9391	GBV-C/HGV	K1737
D87711	D87711	9391	GBV-C/HGV	K1789
AB008342	AB008342	9387	GBV-C/HGV	(HGV-IM71)
D87714	D87714	9391	GBV-C/HGV	K1668
D87263	D87263	9391	GBV-C/HGV	(GSI93)
D87712	D87712	9391	GBV-C/HGV	K1916
D87710	D87710	9391	GBV-C/HGV	K1741
D87713	D87713	9391	GBV-C/HGV	K2141
AF121950	AF121950	9395	GBV-C/HGV	Iowan
NC_001710	NC_001710	9392	GBV-C/HGV	(45255)

Table 21: List of *Hepatitis C virus* sequences

Hepatitis C virus				
ID	Acc.No.	Length	Group	Strain/(Isolate/Clone)
AF054247	AF054247	9595	HCV	HC-J4K1737(pCV-J4L6S)
AF054248	AF054248	9595	HCV	HC-J4(pCV-J4L6S)
AF046866	AF046866	9425	HCV	type 3a (CB)
AF054249	AF054249	9596	HCV	HC-J4(pCV-J4L4S)
AF139594	AF139594	9616	HCV	HCV-N
D89815	D89815	9548	HCV	
HCJ238799	AJ238799	9605	HCV	RB(Con1)
AF009606	AF009606	9646	HCV	(H77)
AF177037	AF177037	9611	HCV	pH77CV-J6S

Table 22: List of *Pestivirus* sequences

Pestivirus				
ID	Acc.No.	Length	Group	Strain/(Isolate/Clone)
type 1				
BVI133739	AJ133739	12308	Bovine viral diarrhea virus	non-cytopathic NADL
BVDPOLYPRO	M96751	12308	Bovine viral diarrhea virus	
PTU86600	PTU86600	12267		ILLNC
type 2				
AY072924	AY072924	12229	Classical swine fever virus	Paderborn
HCVCOMSEQ	X96550	12297	Classical swine fever virus	CAP
NC_002657	NC_002657	12301	Classical swine fever virus	Eystrup
AF091661	AF091661	12297	Classical swine fever virus	Brescia
NC_003100	removed from submission	12297	Classical swine fever virus	39
HCVCGSA	J04358	12297	Hog cholera virus	Alfort
type 3				
AF037405	AF037405	12333	Border disease virus	X818
unclassified				
AF144618	AF144618	12318	reindeer	reindeer-1 V60-Krefeld

A.2 Conserved Secondary Structures in Flaviviridae

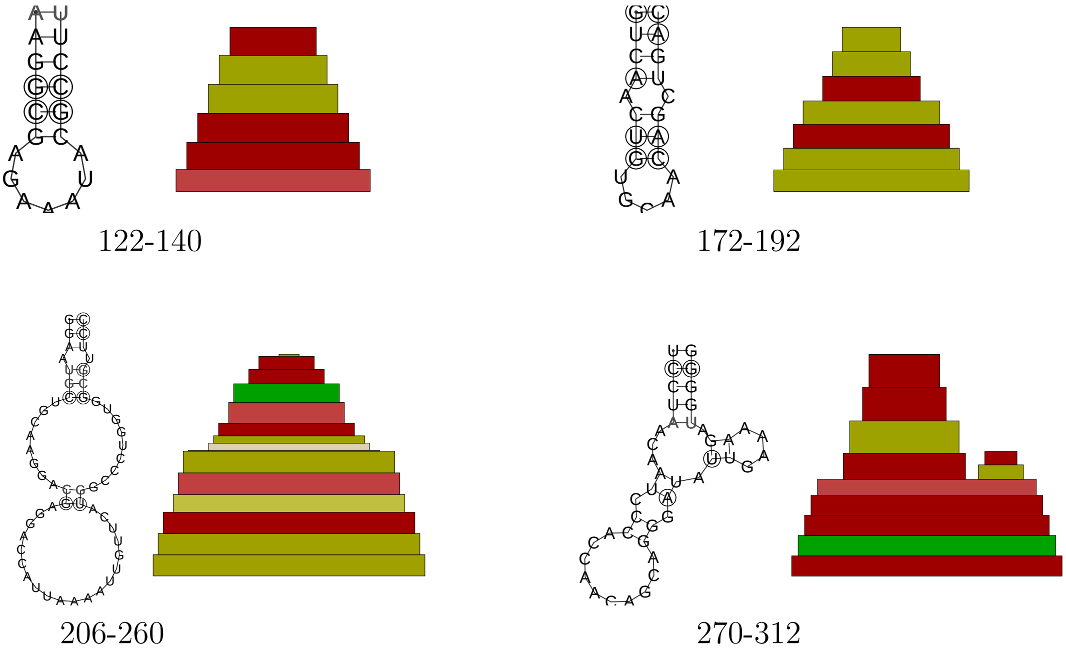
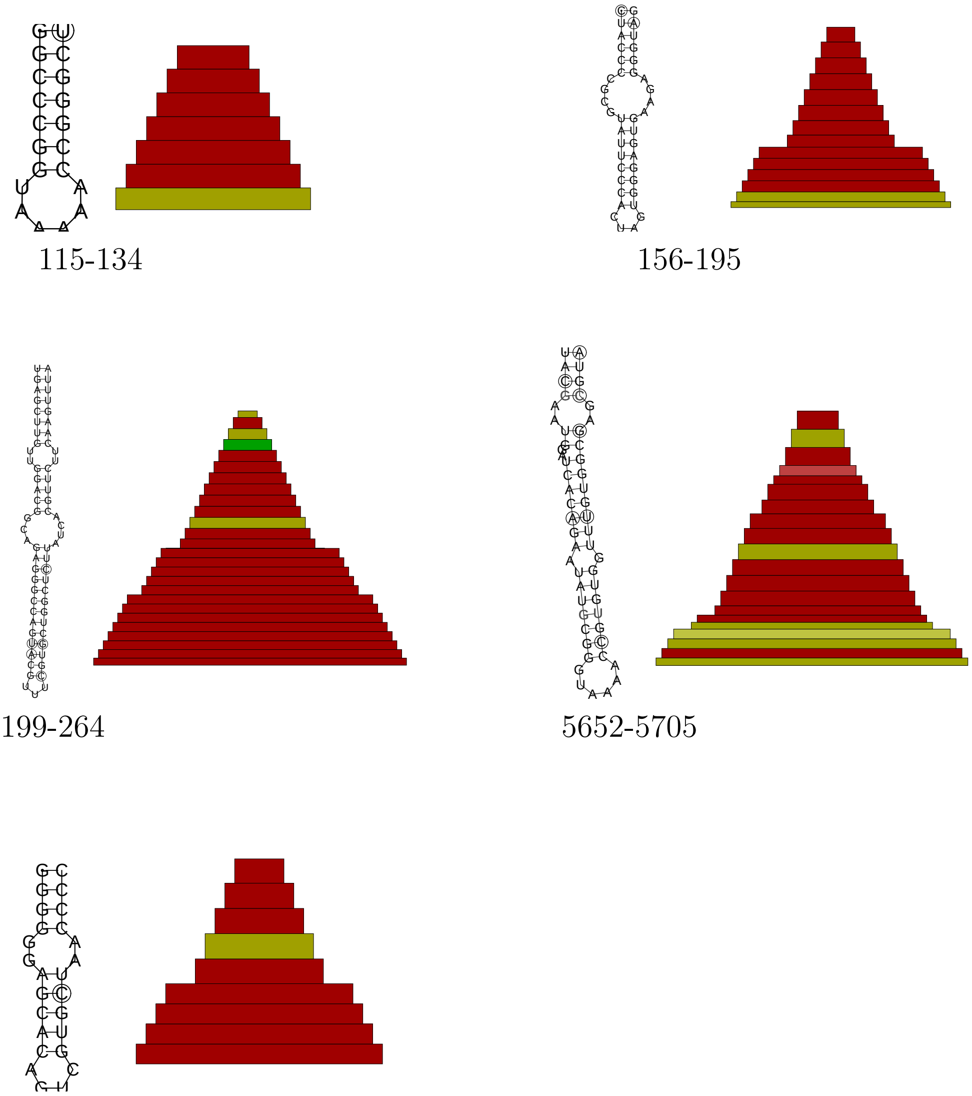


Figure 40: Conserved secondary structure elements in the coding region of DEN



115-134

156-195

199-264

5652-5705

7047-7072

Figure 41: Conserved secondary structure elements in the coding region of JEV

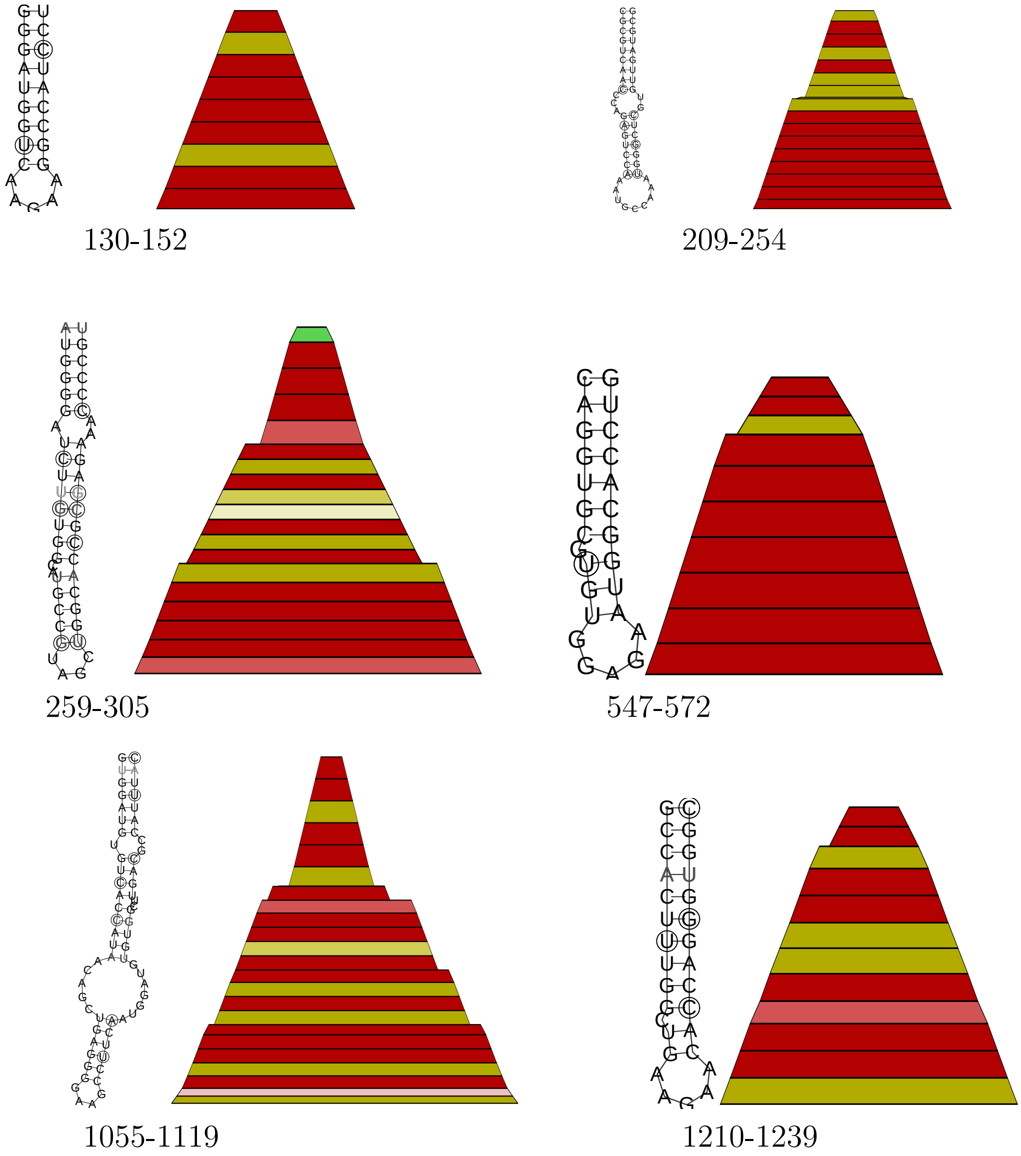
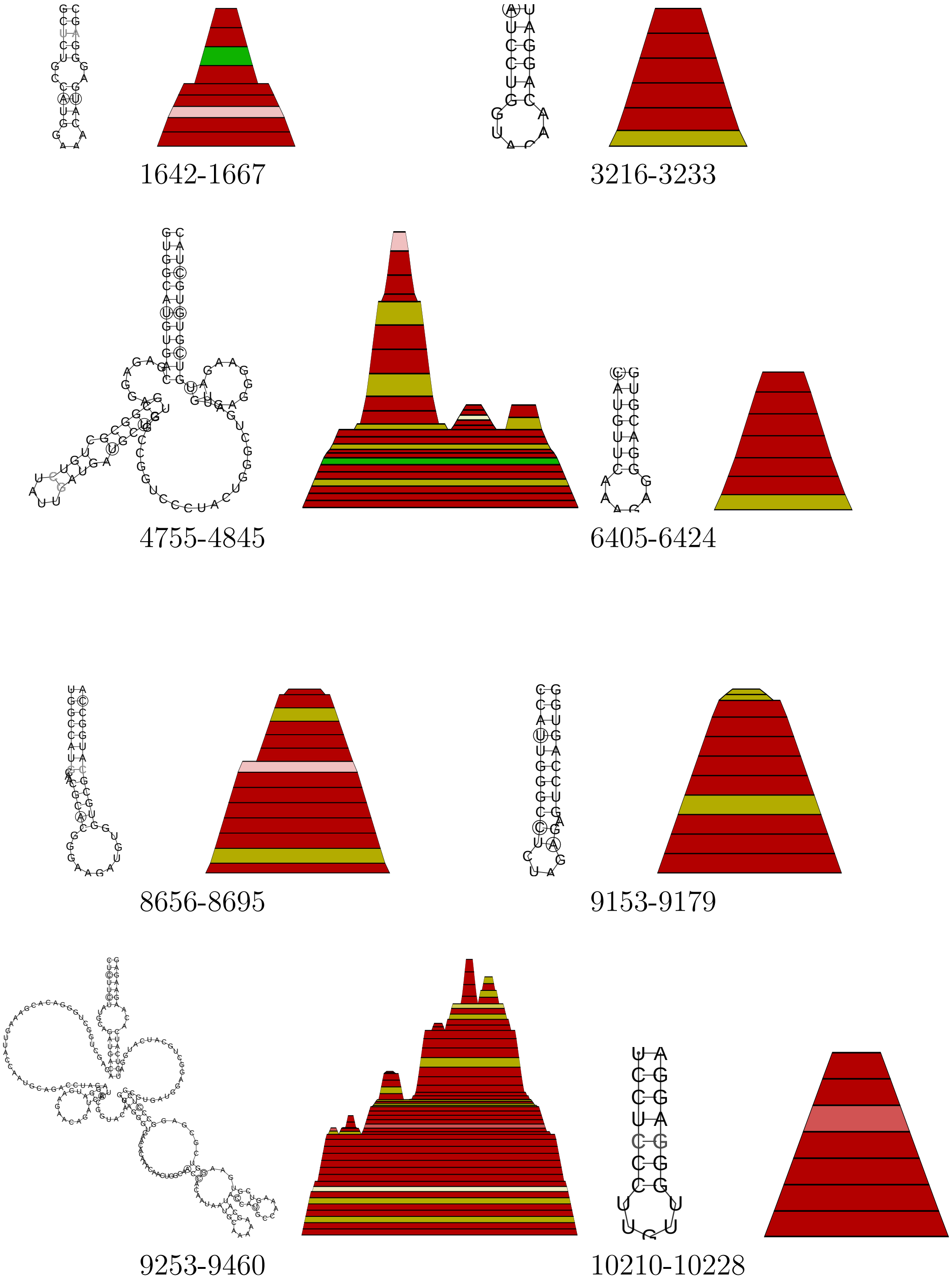


Figure 42: Conserved secondary structure elements in the coding region of TBE



1642-1667

3216-3233

4755-4845

6405-6424

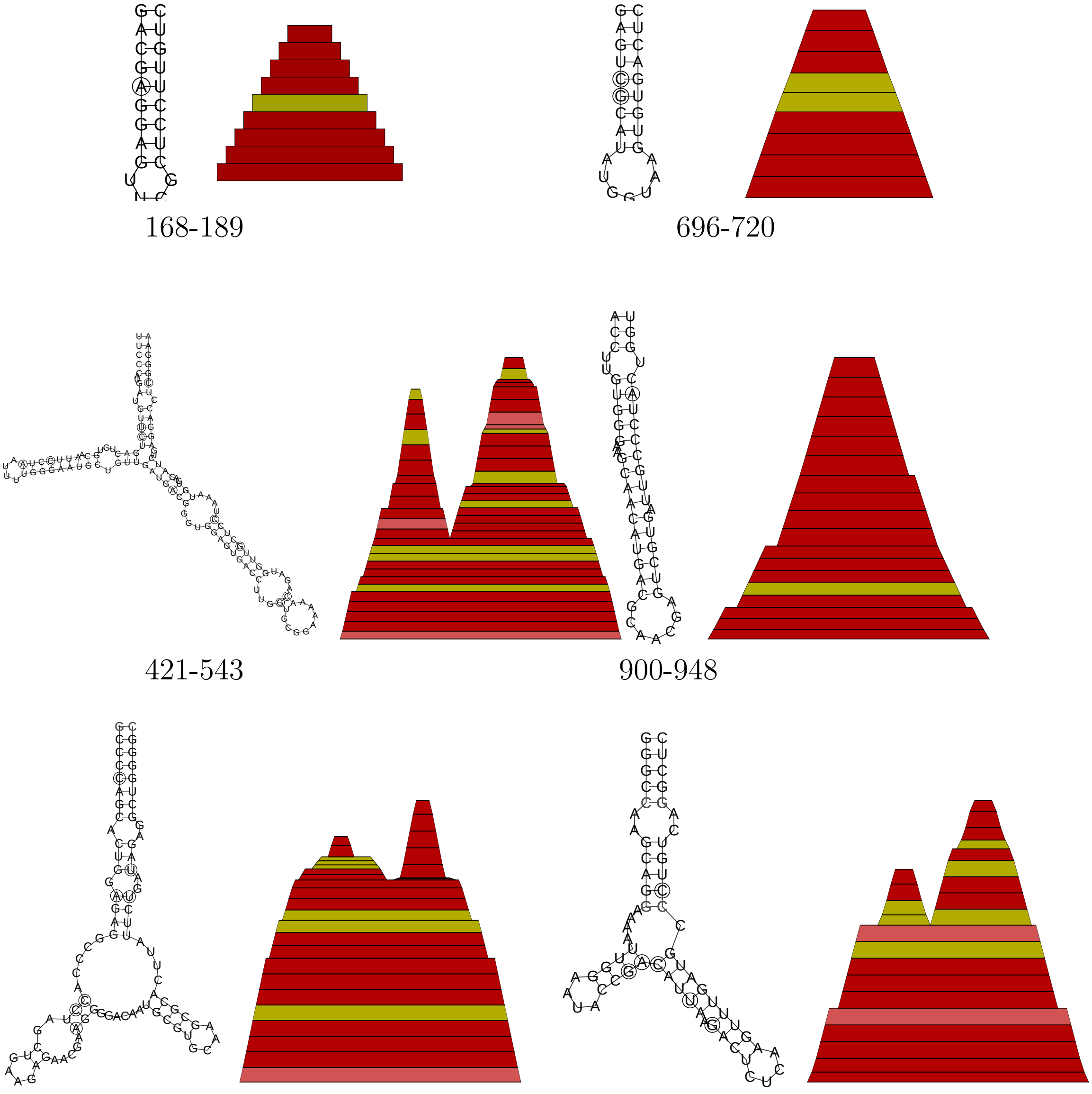
8656-8695

9153-9179

9253-9460

10210-10228

Figure 43: Conserved secondary structure elements in the coding region of TBE



168-189

696-720

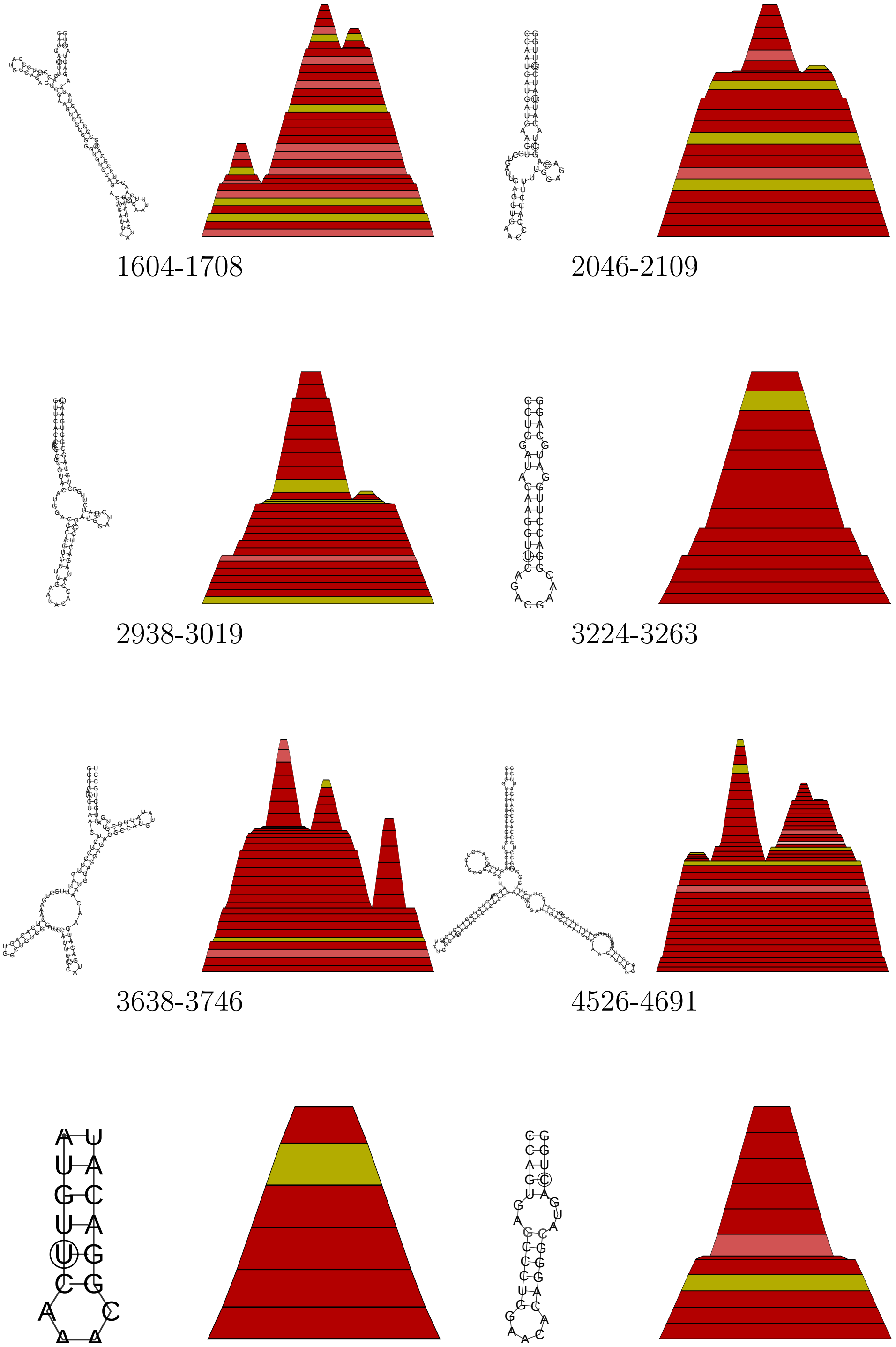
421-543

900-948

1194-1279

1410-1473

Figure 44: Conserved secondary structure elements in the coding region of YFV



1604-1708

2046-2109

2938-3019

3224-3263

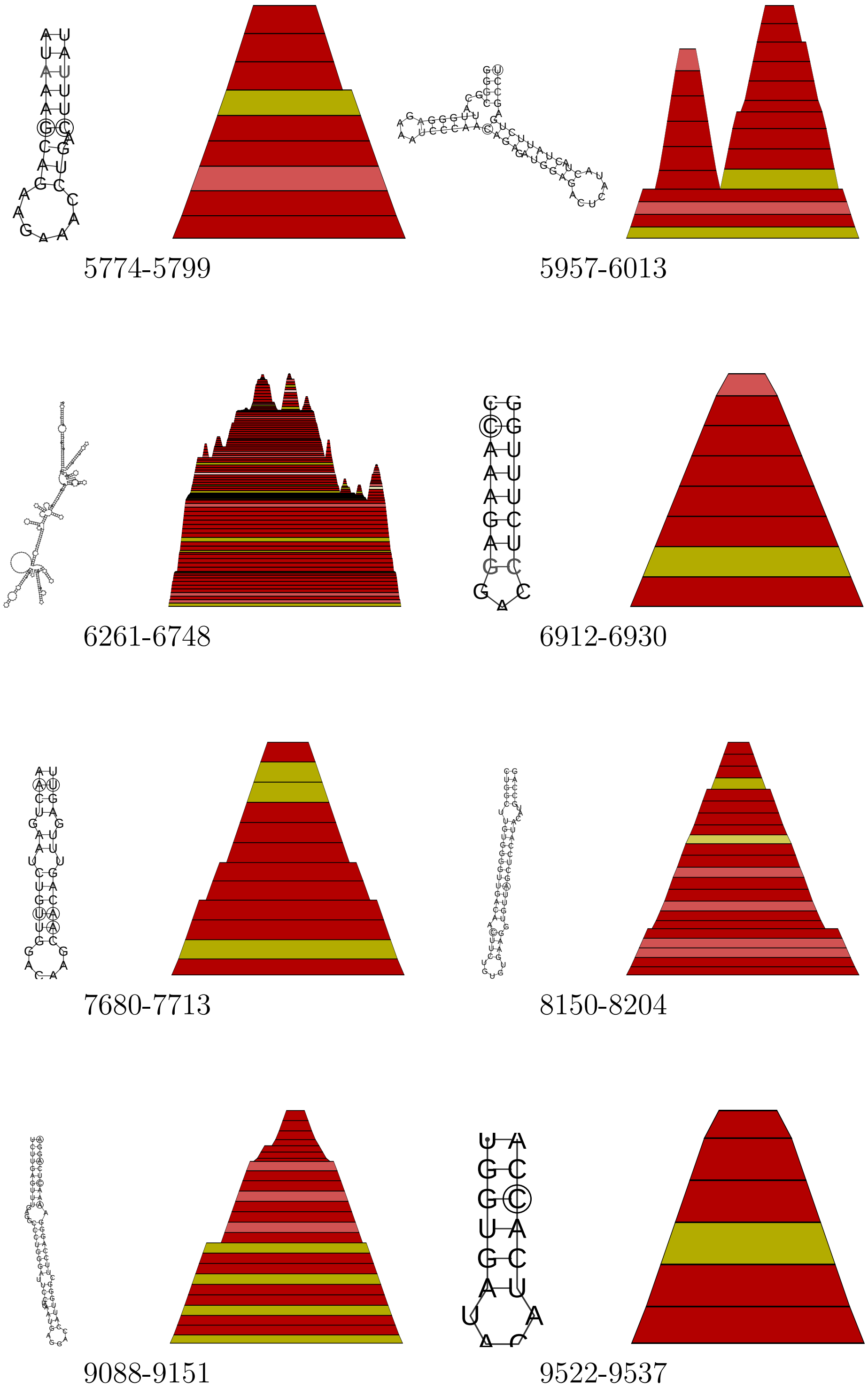
3638-3746

4526-4691

5586-5601

5604-5635

Figure 45: Conserved secondary structure elements in the coding region of YFV



5774-5799

5957-6013

6261-6748

6912-6930

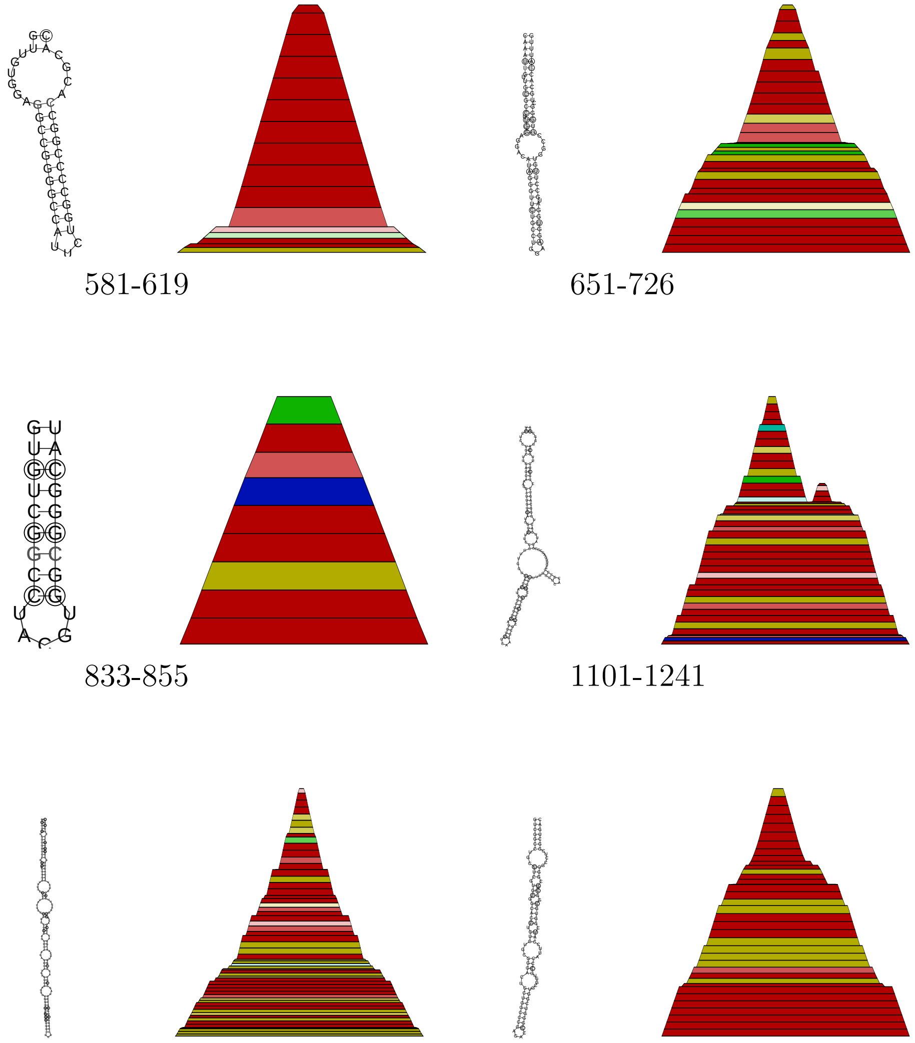
7680-7713

8150-8204

9088-9151

9522-9537

Figure 46: Conserved secondary structure elements in the coding region of YFV



581-619

651-726

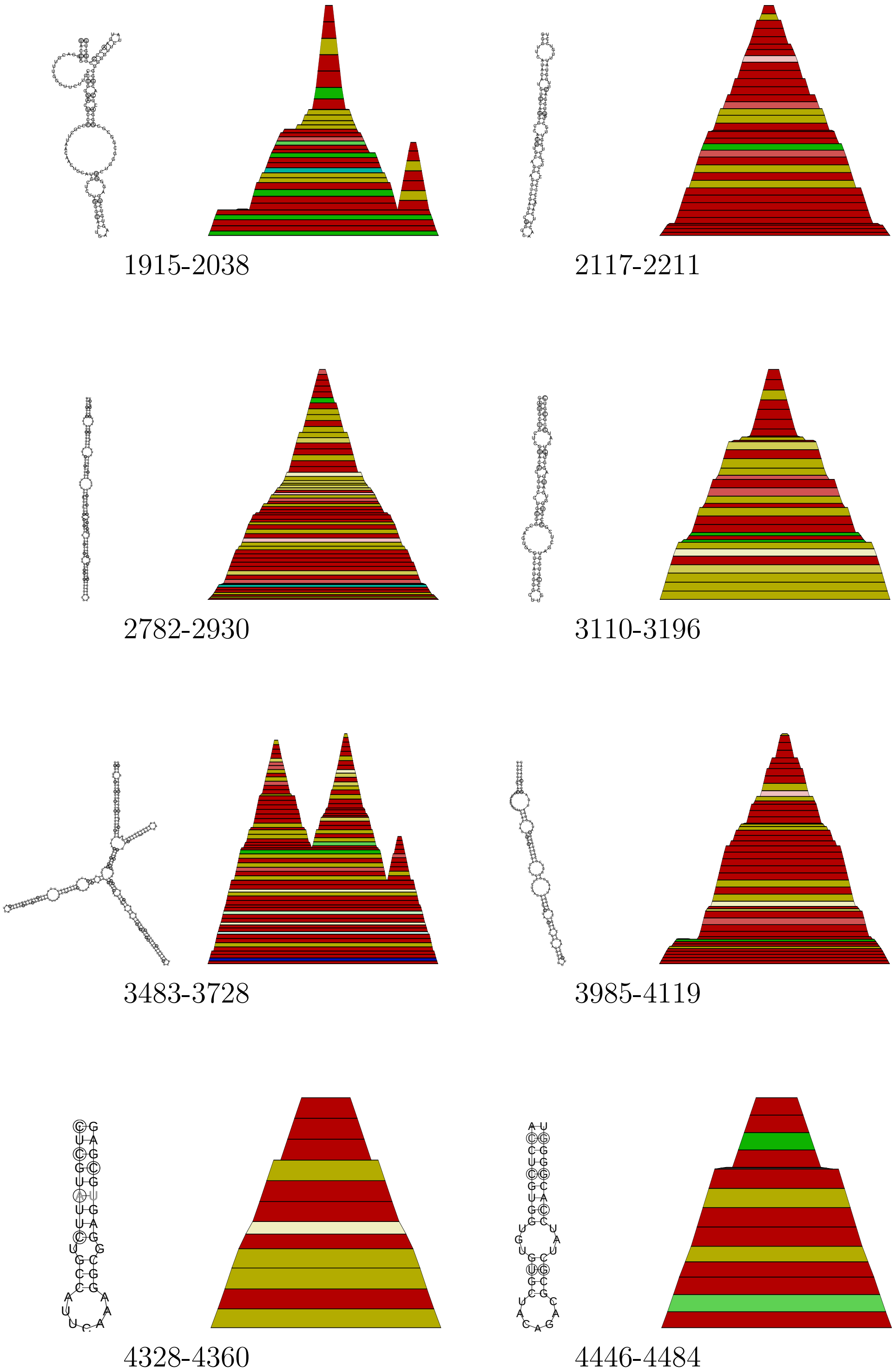
833-855

1101-1241

1449-1614

1632-1731

Figure 47: Conserved secondary structure elements in the coding region of GBV-C



1915-2038

2117-2211

2782-2930

3110-3196

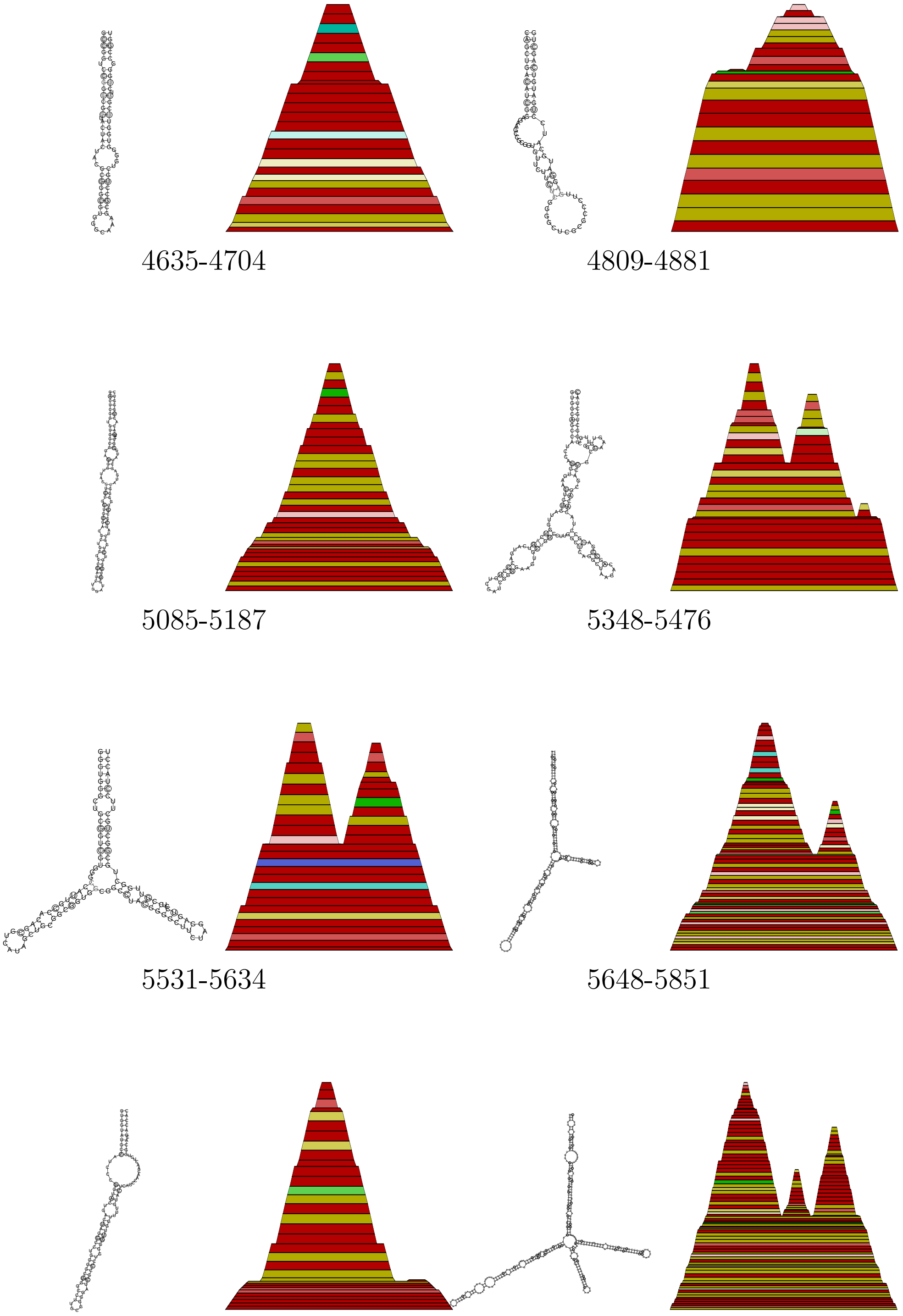
3483-3728

3985-4119

4328-4360

4446-4484

Figure 48: Conserved secondary structure elements in the coding region of GBV-C



4635-4704

4809-4881

5085-5187

5348-5476

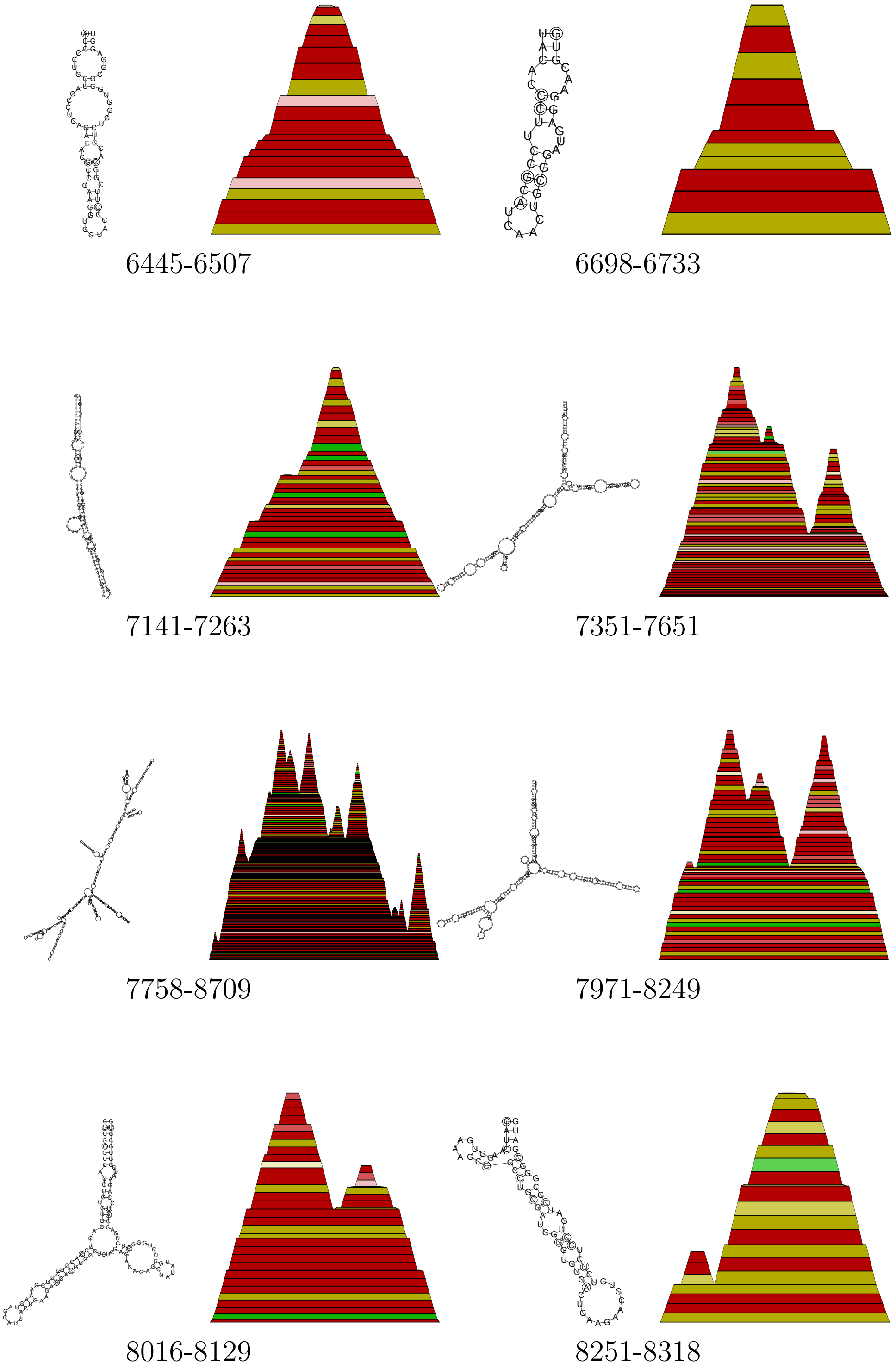
5531-5634

5648-5851

5884-5982

6015-6334

Figure 49: Conserved secondary structure elements in the coding region of GBV-C



6445-6507

6698-6733

7141-7263

7351-7651

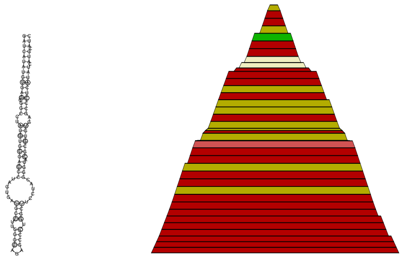
7758-8709

7971-8249

8016-8129

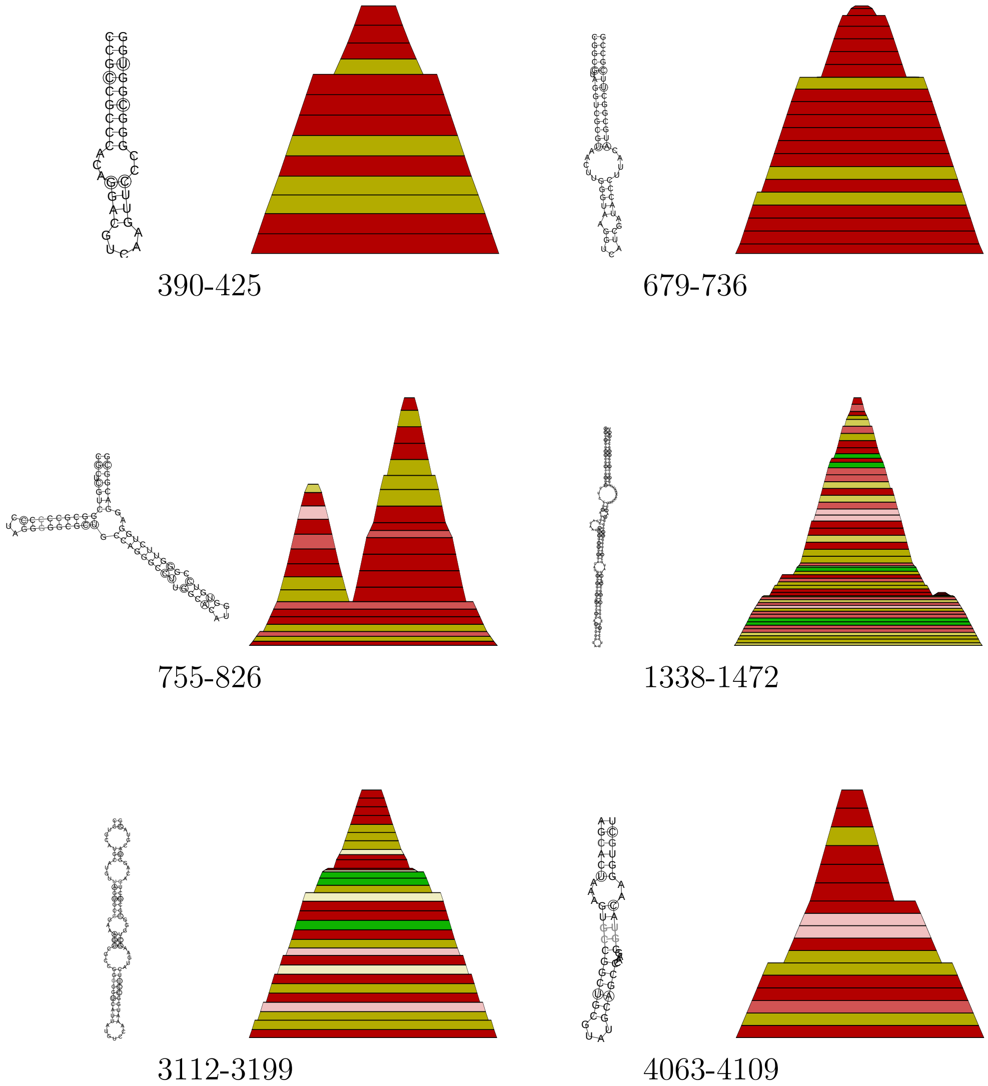
8251-8318

Figure 50: Conserved secondary structure elements in the coding region of GBV-C



8326-8421

Figure 51: Conserved secondary structure elements in the coding region of GBV-C



390-425

679-736

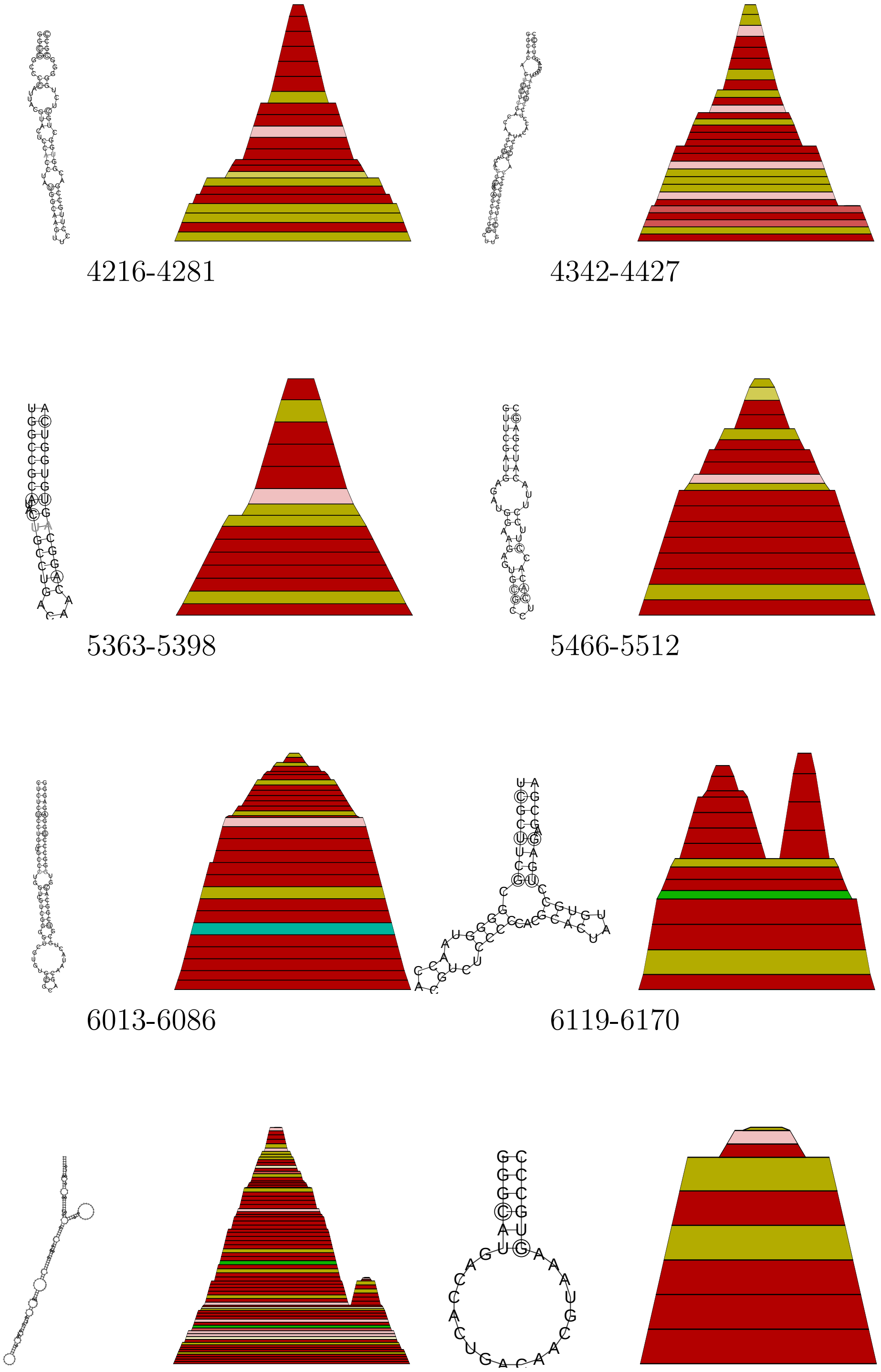
755-826

1338-1472

3112-3199

4063-4109

Figure 52: Conserved secondary structure elements in the coding region of HCV



4216-4281

4342-4427

5363-5398

5466-5512

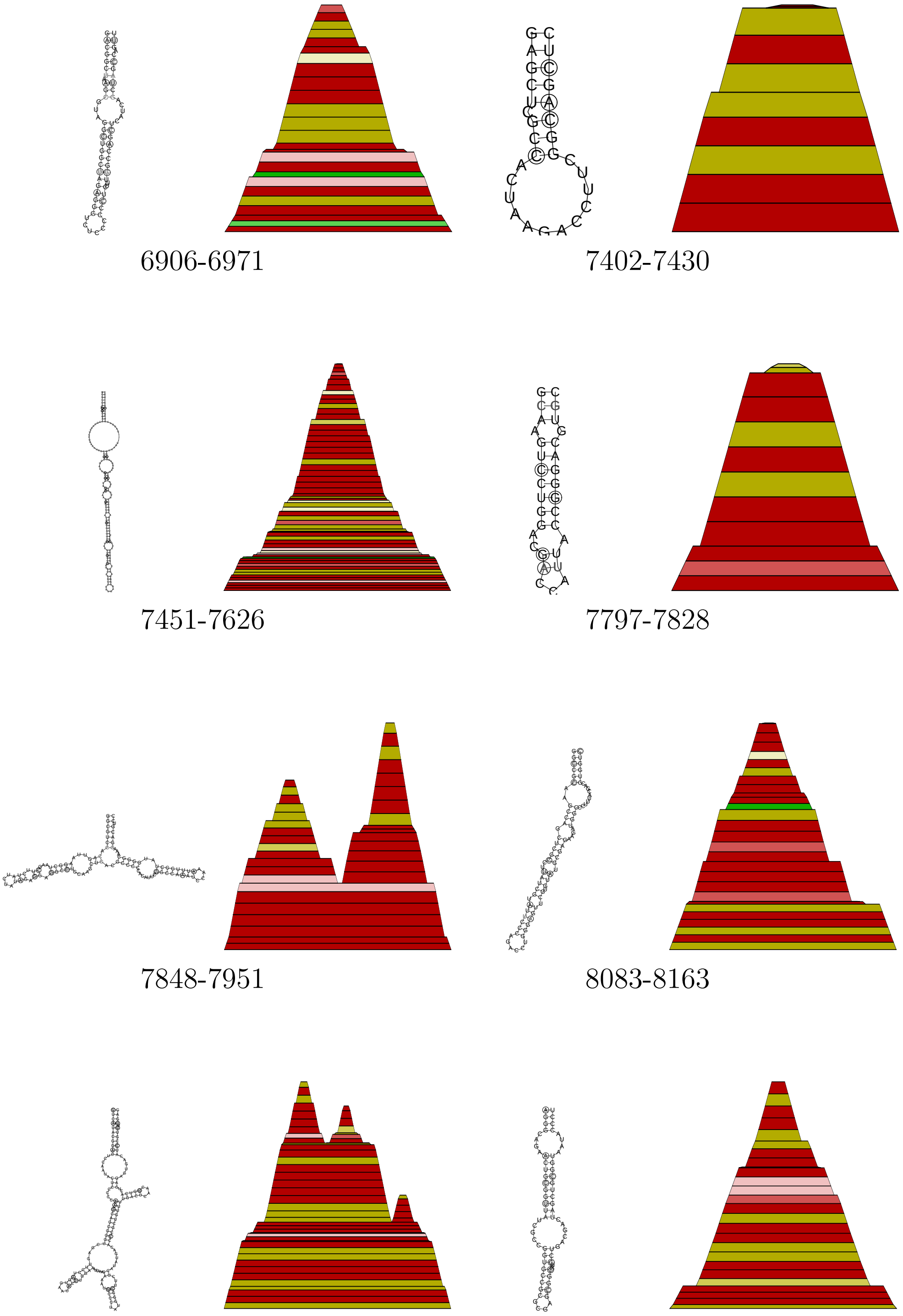
6013-6086

6119-6170

6485-6721

6663-6692

Figure 53: Conserved secondary structure elements in the coding region of HCV



6906-6971

7402-7430

7451-7626

7797-7828

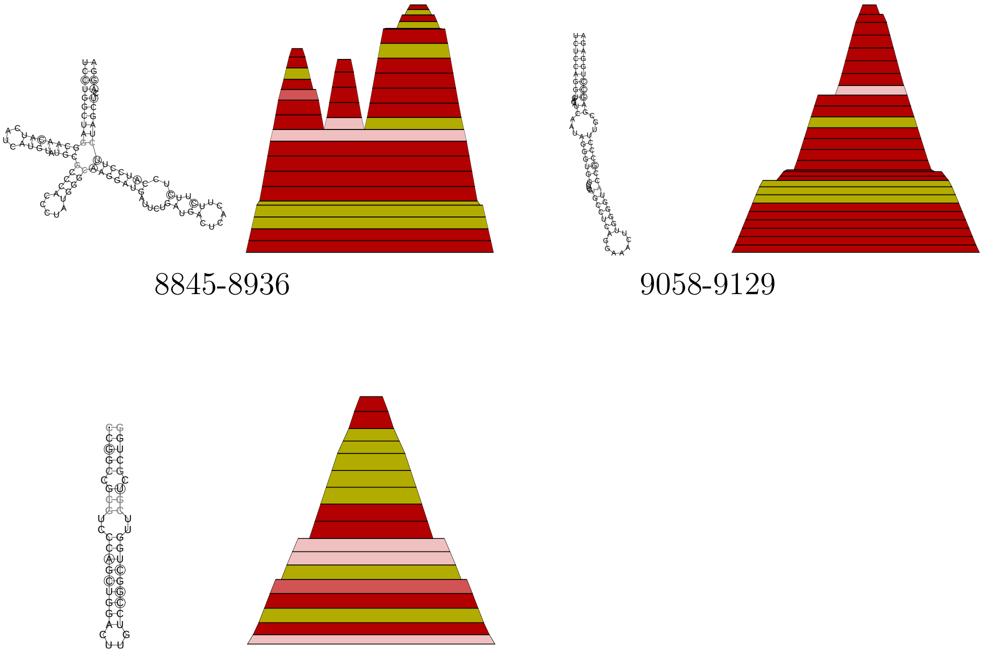
7848-7951

8083-8163

8193-8334

8436-8504

Figure 54: Conserved secondary structure elements in the coding region of HCV

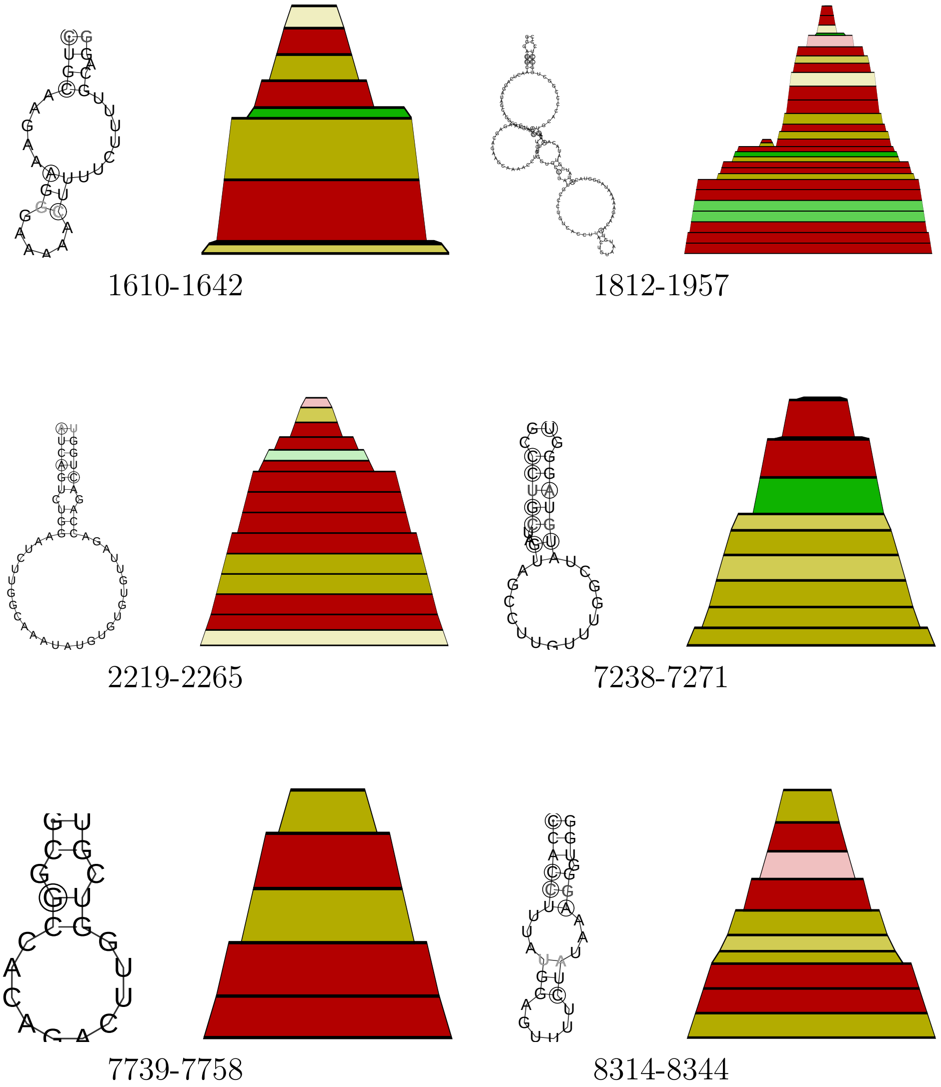


8845-8936

9058-9129

9244-9287

Figure 55: Conserved secondary structure elements in the coding region of HCV



1610-1642

1812-1957

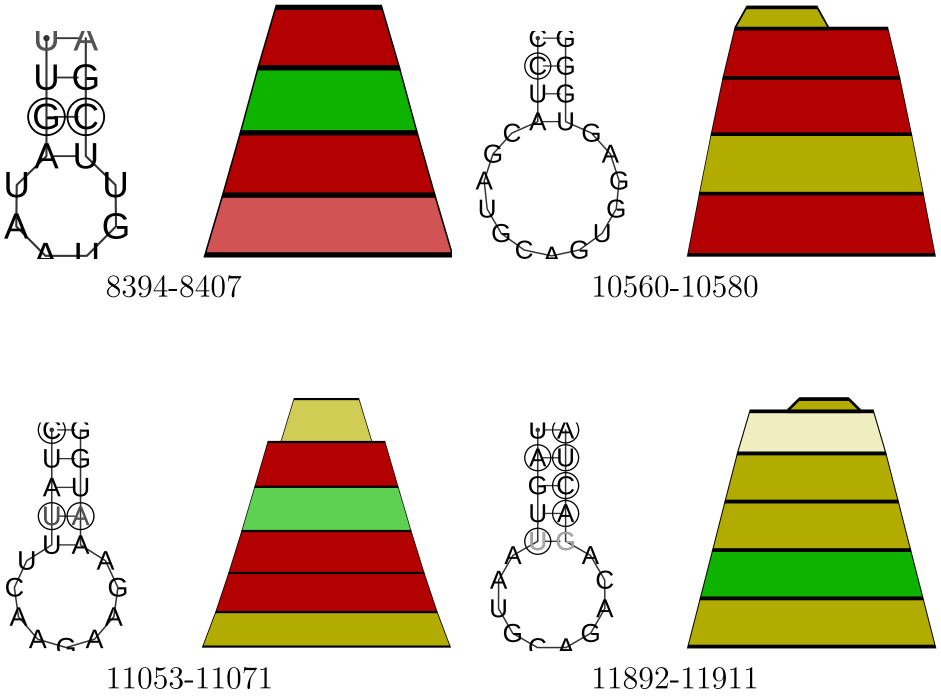
2219-2265

7238-7271

7739-7758

8314-8344

Figure 56: Conserved secondary structure elements in the coding region of PESTI



8394-8407

10560-10580

11053-11071

11892-11911

Figure 57: Conserved secondary structure elements in the coding region of PESTI

A.3 Sequences Used for Prediction of Pseudoknots

SRP RNA:
S1 = { *Aeropyrum pernix, Archaeoglobus fulgidus, Methanosarcina acetivorans, Methanothermus fervidus, Methanococcus jannaschii, Methanobacterium thermoautotrophicum* }
S1a = S1 + { *Bacillus subtilis* }
S1b = S1 + { *Halobacterium halobium* }
S2 = { *Archaeoglobus fulgidus, Halobacterium halobium, Methanothermus fervidus, Methanococcus jannaschii, Methanobacterium thermoautotrophicum, Pyrococcus horikoshii, Pyrodictium occultum, Thermococcus celer* }
S3 = S1 + { *Methanococcus voltae, Pyrococcus abyssi, Pyrococcus horikoshii, Pyrodictium occultum, Sulfolobus solfataricus, Sulfolobus solfataricus, Thermococcus celer* }
S3a = S3 + { *Bacillus subtilis* }
S3b = S3 + { *Halobacterium halobium* }
S4 = S3 + { *Bacillus alcalophilus, Bacillus amyloliquefaciens, Bacillus brevis, Bacillus cereus, Bacillus circulans, Bacillus macerans, Bacillus megaterium, Bacillus polymyxa, Bacillus pumilus, Bacillus sphaericus, Bacillus stearothermophilus, Bacillus subtilis, Bacillus thuringiensis, Brevibacillus brevis, Clostridium perfringens, Bacillus subtilis* }

tmRNA:
T1 = { *Actinobacillus actinomycetemcomitans, Aeromonas salmonicida, Escherichia coli, Haemophilus influenzae, Vibrio cholerae* }
T2 = T1 + { *Marinobacter hydrocarbonoclasticus, Pseudomonas aeruginosa, Pseudoalteromonas haloplanktis,* }
T3 = T2 + { *Acidithiobacillus ferrooxidans, Alteromonas haloplanktis, Dichelobacter nodosus, Francisella tularensis, Haemophilus ducreyi, Klebsiella pneumoniae, Legionella pneumophila, Pasteurella multocida, Pseudomonas putida, Salmonella paratyphi, Salmonella typhimurium, Shewanella putrefaciens, Xylella fastidiosa, Yersinia pestis* }

RNaseP RNA:
R1 = { *Alcaligenes eutrophus, Agrobacterium tumefaciens, Bacteroides thetaiotaomicron, Corynebacterium diphtheriae, Escherichia coli* }
R2 = { *Acidithiobacillus ferrooxidans, Aspergilluns nidulans, Agrobacterium tumefaciens, Carboxydothermus hydrogenoformans, Chromatium vinosum, Desulfovibrio desulfuricans, Erwinia agglomerulans, Escherichia coli* }
R3 = *Alcaligenes eutrophus, Agrobacterium tumefaciens, Bacteroides thetaiotaomicron, Clostridium acetobutylicum, Clostridium difficile, Corynebacterium diphtheriae, Escherichia coli, Mycobacterium avium* }
R4 = { *Alcaligenes eutrophus, Agrobacterium tumefaciens, Bacteroides thetaiotaomicron, Clostridium acetobutylicum, Caulobacter crescentus, Clostridium difficile, Corynebacterium diphtheriae, Carboxydothermus hydrogenoformans, Escherichia coli, Mycobacterium avium* }
R5 = R2 + { *Alcaligenes eutrophus, Bacteroides thetaiotaomicron, Clostridium acetobutylicum, Caulobacter crescentus, Clostridium difficile, Corynebacterium diphtheriae, Campylobacter jejuni, Mycobacterium avium, Mycobacterium bovis, Mycobacterium leprae, Prochlorococcus marinus, Vibrio cholerae* }

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] R.K. Ahuja, T.L. Magnati, and J.B. Orlin. *Network Flows: Theory, Algorithms, and Applications.* Prentice-Hall, Inc., New Jersey, 1993.

[3] T. Akutsu. Dynamic programming algorithms for RNA secondary structure with pseudoknots. *Discrete Applied Mathematics*, 104:45–62, 2000.

[4] J.L. Sussman ans S. Kim. Three-dimensional structure of a transfer rna in two crystal forms. *Science*, 192:853–858, 1976.

[5] P. Becher, M. Orlich, and H. J. Thiel. Complete genomic sequence of border disease virus, a pestivirus from sheep. *J. Virol.*, 72(6):5165–5173, 1998.

[6] K. J. Blight and C. M. Rice. Secondary structure determination of the conserved 98-base sequence at the 3' terminus of hepatitis C virus genome RNA. *J. Virol.*, 71(10):7345–7352, 1997.

[7] P. N. Borer, B. Dengler, Ignatio Tinoco Jr, and O. C. Uhlenbeck. Stability of ribonucleic acid doublestranded helices. *J. Mol. Bio.*, 86:843–853, 1974.

[8] M. A. Brinton and J. H. Dispoto. Sequence and secondary structure analysis of the 5'-terminal region of flavivirus genome RNA. *Virology*, 162:290–299, 1988.

[9] 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(19):5041–5, 1992.

[10] J.W. Brown. The ribonuclease P database. *Nucl. Acids Res.*, 27(1):314, 1999. http://www.mbio.ncsu.edu/RNaseP/home.html.

[11] J.W. Brown, J.M. Nolan, E.S. Haas, M.A.T. Rubio, F. Major, and N.R. Pace. Comparative analysis of ribonuclease P RNA using gene sequences from natural microbial populations reveals tertiary structural elements. *Proc. Natl. Acad. Sci. USA*, 93:3001–3006, 1996.

[12] C. R. Cantor, P. L. Wollenzien, and J. E. Hearst. Structure and topology of 16S ribosomal RNA. an analysis of the pattern of psoralen crosslinking. *Nucl. Acids Res.*, 8:1855–1872, 1980.

[13] T.R. Cech, A.J. Zaug, and P.J. Grabowski. In vitro splicing of the ribosomal RNA precursor of tetrahymena: involvement of a guanosine nucleotide in the excision of the intervening sequence. *Cell*, 27:487–496, 1981.

[14] M. Chamorro, N. Parkin, and H. E. Varmus. An RNA pseudoknot and an optimal heptameric shift site are required for highly efficient ribosomal frameshifting on a retroviral messenger RNA. *Proc. Natl. Acad. Sci. USA*, 89:713–717, 1992.

[15] Hue Sun Chan and Ken A. Dill. Interchain loops in polymers: Effects of excluded volume. *J. Chem. Phys.*, 90:492–508, 1988.

[16] G. Chartrand and F. Harary. Planar permutation graphs. *Ann. Inst. Henri Poincarè B*, 3:433–438, 1967.

[17] J. Chen, S. Le, and J.V. Maizel. Prediction of common secondary structures of RNAs: a genetic algorithm. *Nucl. Acids Res*, 28:991–999, 2000.

[18] Shi-Jie Chen and Ken A. Dill. Statistical thermodynamics of double-stranded polymer molecules. *J. Chem. Phys.*, 103:5802–5808, 1995.

[19] D. K. Chiu and T. Kolodziejczak. Inferring consensus structure from nucleic acid sequences. *CABIOS*, 7:347–352, 1991.

[20] A. J. Collier, J. Gallego, R. Klinck, P. T. Cole, S. J. Harris, G. P. Harrison, F. Aboul-Ela, G. Varani, and S. Walker. A conserved RNA structure within the HCV IRES eIF3-binding site. *Nat. Struct. Biol.*, 9(5):375–380, 2002.

[21] N. M. Cuceanu, A. Tuplin, and P. Simmonds. Evolutionarily conserved RNA secondary structures in coding and non-coding sequences at the 3' end of the hepatitis G virus/GB-virus C genome. *J. Gen. Virol.*, 82(4):713–722, 2001.

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

[23] R. Deng and K. V. Brock. 5' and 3' untranslated regions of pestivirus genome: primary and secondary structure analyses. *Nuc. Acids Res.*, 21(8):1949–1957, 1993.

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

[25] S. P. Fletcher and R. J. Jackson. Pestivirus internal ribosome entry site (IRES) structure and function: elements in the 5' untranslated region important for IRES function. *J. Virol.*, 76(10):5024–5033, 2002.

[26] P.J. Flory. *Principles of Polymer Chemistry.* Cornell Univ. Press, Ithaca, 1953.

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

[28] I. Fortsch, H. Fritzsche, E. Birch-Hirschfeld, E. Evertsz, R. Klement, T. M. Jovin, and C. Zimmer. Parallel-stranded duplex dna containing da.du base pairs. *Biopolymers*, 38:209–220, 1996.

[29] D.N. Frank and N.R. Pace. Ribonuclease P: unity and diversity in a tRNA processing ribozyme. *Annu. Rev. Biochem.*, 67:153–180, 1998.

[30] P. Friebe, V. Lohmann, N. Krieger, and R. Bartenschlager. Sequences in the 5' nontranslated region of hepatitis C virus required for RNA replication. *J. Virol.*, 75(24):12047–12057, 2001.

[31] M. R. Giese, K. Betschart, T. Dale, C.K. Riley, C. Rowan, K.J. Sprouse, and M.J. Serra. Stability of RNA hairpins closed by wobble base pairs. *Biochemistry*, 37:1094–1100, 1998.

[32] T. C. Gluick and D. E. Draper. Thermodynamics of folding a pseudoknotted mRNA fragment. *J. Mol. Biol.*, 241:246–262, 1994.

[33] J. Gorodkin, 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.

[34] J. Gorodkin, B. Knudsen, C. Zwieb, and T. Samuelsson. SRPDB signal recognition particle database. *Nucl. Acids Res.*, 29(1):169–170, 2001. http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html.

[35] D. R. Groebe and O. C. Uhlenbeck. Characterization of RNA hairpin loop stability. *Nucl. Acids Res.*, 16:11725–11735, 1988.

[36] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35:849–857, 1983.

[37] Alexander P. Gultyaev, F. H. D. van Batenburg, and Cornelis W. A. Pleij. An approximation of loop free energy values of RNA H-pseudoknots. *RNA*, 5:609–617, 1999.

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

[39] R. R. Gutell, A. Power, G. Z. Hertz, E. J. Putz, and G. D. Stormo. Identifying constraints on the higher-order structure of RNA: continued development and application of comparative sequence analysis methods. *Nucl. Acids Res.*, 20:5785–5795, 1992.

[40] R. R. Gutell and C. R. Woese. Higher order structural elements in ribosomal RNAs: Pseudo-knots and the use of noncanonical pairs. *Proc. Natl. Acad. Sci. USA*, 87:663–667, 1990.

[41] E.S. Haas and J.W. Brown. Evolutionary variation in bacterial Rnase P RNAs. *Nucl. Acids Res.*, 26(18):4093–4099, 1998.

[42] C. S. Hahn, Y. S. Hahn, C. M. Rice, E. Lee, L. Dalgarno, E. G. Strauss, and J. H. Strauss. Conserved elements in the 3'untranslated region of flavivirus RNAs and potential cyclization sequences. *J. Mol. Biol.*, 198(1):33–41, 1987.

[43] J.K. Harris, E.S. Haas, D. Williams, and D.N. Frank. New insight into RNase P RNA structure from comparative analysis of the archaeal RNA. *RNA*, 7:220–232, 2001.

[44] W. E. Hart and S. Istrail. Lattice and off-lattice side chain models of protein folding: linear time structure prediction better than 86% of optimal. *J. Comput. Biol.*, 4:241–259, 1997.

[45] C. Haslinger. *Prediction algorithms for restricted RNA pseudoknots.* PhD thesis, Universität Wien, 2001.

[46] C. Haslinger and P.F. Stadler. RNA structures with pseudo-knots. *Bull. Math. Biol.*, 61:437–467, 1999.

[47] L. He, Ryszard Kierzek, John SantaLucia Jr., Amy E. Walter, and Douglas H. Turner. Nearest-neighbor parameters for GU mismatches. *Biochemistry*, 30, 1991.

[48] C.U.T. Hellen and P. Sarnow. Internal ribosome entry sites in eukaryotic mRNA molecules. *Genes Dev.*, 15:1593–1612, 2001.

[49] Gerhard Heyer, Uwe Quasthoff, and Christian Wolff. Automatic analysis of large text corpora — A contribution to structuring WEB communities. In H. Unger, T. Böhme, and A. Mikler, editors, *Innovative Internet Computing Systems*, volume 2346 of *Lecture Notes in Computer Science*, pages 15–26. Springer-Verlag, Heidelberg, 2002.

[50] 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.

[51] I.L. Hofacker, M. Fekete, and P.F. Stadler. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 319:1059–1066, 2002.

[52] Ivo L. Hofacker, Stephan H. F. Bernhart, and Peter F. Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 2004. in press.

[53] Ivo L. Hofacker and Stadler Peter F. The partition function variant of Sankoff's algorithm. In *ICCS 2004 Proceedings*, 2004. in press.

[54] Ivo L. Hofacker, Martin Fekete, Christoph Flamm, Martijn A. Huynen, Susanne Rauscher, Paul E. Stolorz, and Peter F. Stadler. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucl. Acids Res.*, 26:3825–3836, 1998.

[55] Ivo L. Hofacker, Martin Fekete, and Peter F. Stadler. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 319:1059–1066, 2002. SFI Preprint 01-11-067.

[56] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[57] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.

[58] Ivo L. Hofacker and Peter F. Stadler. Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comp. & Chem.*, 23:401–414, 1999.

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

[60] M. Honda, E. A. Brown, and S. M. Lemon. Stability of a stem-loop involving the initiator AUG controls the efficiency of internal initiation of translation on hepatitis C virus RNA. *RNA*, 2(10):955–68, 1996.

[61] M. Honda, L. H. Ping, R. C. Rijnbrand, E. Amphlett, B. Clarke, D. Rowlands, and S. M. Lemon. Structural requirements for initiation of translation by internal ribosome entry within genome-length hepatitis C virus RNA. *Virology*, 222(1):31–42, 1996.

[62] 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.

[63] H. Isambert and E.D. Siggia. Technical appendix for the RNA folding manuscript. http://uqbar.rockefeller.edu/˜siggia/RNA_folding/technical.app.ps, 1999.

[64] H. Isambert and E.D. Siggia. Modeling RNA folding paths with pseudoknots: Application to hepatitis delta virus ribozyme. *PNAS*, 97(12):6515–6520, 2000.

[65] T. Ito and M. M. C. Lai. Determination of the secondary structure of and cellular protein binding to the 3'-untranslated region of the hepatitis C virus RNA genome. *J. Virol.*, 71(11):8698–8706, 1997.

[66] H. Jacobson and W.H. Stockmayer. Intramolecular reaction in polycondensations. I. the theory of linear systems. *J. Chem. Phys.*, 18:1600–1606, 1950.

[67] S.K. Jang, H.G. Krausslich, M.J. Nicklin, G.M. Duke, A.C. Palmenberg, and E. Wimmer. A segment of the 5' nontranslated region of encephalomyocarditis virus RNA directs internal entry of ribosomes during in vitro translation. *J. Virol.*, 62:2636–2643, 1988.

[68] Veronica Juan and Charles Wilson. RNA secondary structure prediction based on free energy and phylogenetic analysis. *J. Mol. Biol.*, 289(4):935–947, 1999.

[69] K. I. Kalliampakou, L. Psaridi-Linardaki, and P. Mavromara. Mutational analysis of the apical region of domain II of the HCV IRES. *FEBS Lett.*, 511(1-3):79–84, 2002.

[70] K. Katayama, S. Fukushi, C. Kurihara, N. Ishiyama, H. Okamura, and A. Oya. Full-length GBV-C/HGV genomes from nine japanese isolates: characterization by comparative analysis. *Arch. Virol.*, 143:1–13, 1998.

[71] R.J. Keenan, D.M. Freymann, R.M. Stroud, and P. Walter. The signal recognition particle. *Annu. Rev. Biochem.*, 70:755–775, 2001.

[72] A. A. Khromykh, H. Meka, K. J. Guyatt, and E. G. Westaway. Essential role of cyclization sequences in flavivirus RNA replication. *J. Virol.*, 75(14):6719–6728, 2001.

[73] J. S. Kieft, K. Zhou, A. Grech, R. Jubin, and A. Doudna. Crystal structure of an RNA tertiary domain essential to HCV IRES-mediated translation initiation. *Nat. Srtuct. Biol.*, 9(5):370–374, 2002.

[74] J. S. Kieft, K. Zhou, R. Jubin, and J. A. Doudna. Mechanism of ribosome recruitment by hepatitis C IRES RNA. *RNA*, 7(2):194–206, 2001.

[75] Y. K. Kim, C. S. Kim, S. H. Lee, and S. K. Jang. Domains I and II in the 5' nontranslated region of the HCV genome are required for RNA replication. *Biochem. Biophys. Res. Commun.*, 290(1):105–112, 2002.

[76] Z. Kiss-Laszlo, Y. Henry, J. P. Bachellerie, M. Caizergues-Ferrer, and T. Kiss. Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. *Cell*, 85:1077–1088, 1996.

[77] B. Knudsen, J. Wower, C. Zwieb, and J. Gorodkin. tmRDB (tmRNA database). *Nucl. Acids Res.*, 29(1):171–172, 2001. http://psyche.uthct.edu/dbs/tmRDB/tmRDB.html.

[78] V. G. Kolupaeva, T. V. Pestova, and C. U. Hellen. An enzymatic footprinting analyses of the interaction of 40S ribosomal subunits with the internal ribosomal entry site of hepatitis C virus. *J. Virol.*, 74(14):6242–6250, 2000.

[79] V. G. Kolupaeva, T. V. Pestova, and C. U. Hellen. Ribosomal binding to the internal ribosomal entry site of classical swine fever virus. *RNA*, 6(12):1791–807, 2000.

[80] A. A. Kolykhalov, S.M. Feinstone, and C. M. Rice. Identification of a highly conserved sequence element at the 3' terminus of hepatitis C virus genome RNA. *J. Virol.*, 70(6):3363–71, 1996.

[81] K. Kruger, P.J. Grabowski, A.J. Zaug, J. Sands, D.E. Gottschling, and T.R. Cech. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell*, 31(1):147–157, 1982.

[82] N. Larsen and C. Zwieb. SRP-RNA sequence alignment and secondary structure. *Nucl. Acids Res.*, 19(2):209–215, 1991.

[83] S. Y. Le and M. Zuker. Predicting common foldings of homologous RNAs. *J. Biomol. Struct. Dyn.*, 8:1027–1044, 1991.

[84] K. C. Leitmeyer, D. W. Vaughn, D. M. Watts, R. Salas, I. Villalobos de Chacon, C. Ramos, and R. Rico-Hesse. Dengue virus structural differences that correlate with pathogenesis. *J. Virol.*, 73:4738–4747, 1999.

[85] R. Lück, S. Gräf, and G. Steger. ConStruct: A tool for thermodynamic controlled prediction of conserved secondary structure. *Nucl. Acids Res.*, 27:4208–4217, 1999.

[86] 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.

[87] P. J. Lukavsky, I. Kim, G. A. Otto, and J. D. Puglisi. Structure of HCV IRES domain II determined by NMR. *RNA*, 7(2):194–206, 2001.

[88] R.B. Lyngsø and C.N.S. Pedersen. RNA pseudoknot prediction in energy based models. *J. Comp. Biol.*, 7(3/4):409–428, 2000.

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

[90] C. Mand and R. Kofler. personal communication.

[91] 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. *Virol.*, 194(1):173–184, 1993.

[92] C. W. Mandl, H. Holzmann, T. Meixner, S. Rauscher, P. F. Stadler, S. 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. Virol.*, 72(3):2132–2140, 1998.

[93] D. H. Mathews, J. Sabina, M. Zucker, and H. Turner. Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J. Mol. Biol.*, 288:911–940, 1999.

[94] D.H. Mathews, J.M. Diamond, and D.H. Turner. The apllication of thermodynamics to the modeling of RNA structure. In E. Di Cera, editor, *Thermodynamics in biology*, pages 177–201. Oxford Univeraity Press, Oxford, 2000.

[95] D.H. Mathews and D.H. Turner. Dynalign: An algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.*, 317:191–203, 2002.

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

[97] K. 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.

[98] R Men, M Bray, D Clark, R M Chanock, and C J. Lai. Dengue type 4 virus mutants containing deletions in the 3' noncoding region of the RNA genome: analysis of growth restriction in cell culture and altered viremia pattern and immunogenicity in rhesus monkeys. *J. Virol.*, 70:3930–3937, 1996.

[99] B. Morgenstern. Multiple DNA and protein sequence alignment at bibiserv. *Nucleic. Acids Res.*, page in press, 2004.

[100] C. Moser, A. Bosshart, J. D. Tratschin, and M. A. Hofmann. A recombinant classical swine fever virus with a marker insertion in the internal ribosome entry site. *Virus Genes*, 23(1):63–68, 2001.

[101] A. Muto, C. Ushida, and H. Himeno. A bacterial RNA that functions as both tRNA and an mRNA. *Trends Biochem. Sci.*, 23(1):25–29, 1998.

[102] T. M. Myers, V. G. Kolupaeva, E. Mendez, S. G. Baginski, I. Frolov, C. U. Hellen, and C. M. Rice. Efficient translation initiation is required for replication of bovine viral diarrhea virus subgenomic replicons. *J. Virol.*, 75(9):4226–4238, 2001.

[103] P. Nelson, M. Kiriakidou, A. Sharma, E. Maniataki, and Z. Mourelatos. The microRNA world: small is mighty. *Trends Biochem. sci.*, 28:534–540, 2004.

[104] P. Nissen, J. Hansen, N. Ban, P.B. Moore, and T.A. Steitz. The structural basis of ribosome activity in peptide bond synthesis. *Science*, 289:920–930, 2000.

[105] H.F. Noller, V. Hoffarth, and L. Zimniak. Unusual resistance of peptidyl transferase to protein extraction procedures. *Science.*, 256:1416–1419, 1992.

[106] C. Notredame, E.A. O'Brien, and D.G. Higgins. Raga: RNA sequence alignment by genetic algorithm. *Nucl. Acids Res.*, 25(22):4570–4580, 1997.

[107] 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.

[108] Ruth Nussinov, George Piecznik, Jerrold R. Griggs, and Daniel J. Kleitman. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35(1):68–82, 1978.

[109] F. E. Odreman-Macchioli, S. G. Tisminetzky, M. Zotti, F. E. Baralle, and E. Buratti. Influence of correct secondary and tertiary RNA folding on the binding of cellular factors to the HCV IRES. *Nucl. Acids Res.*, 28(4):875–885, 2000.

[110] J. W. Oh, T. Ito, and M. M. Lai. A recombinant hepatitis C virus RNA-dependent RNA polymerase capable of copying the full-length viral RNA. *J. Virol.*, 73(9):7694–7702, 1999.

[111] J. W. Oh, G. T. Sheu, and M. M. Lai. Template requirement and initiation site selection by hepatitis C virus polymerase on a minimal viral RNA template. *J. Biol. Chem.*, 275(23):17710–17717, 2000.

[112] H. Okamoto, H. Nakao, T. Inoue, M. Fukuda, J. Kishimoto, H. Iizuka, F. Tsuda, Y. Miyakawa, and M. Mayumi. The entire nucleotide sequences of two GB virus C/hepatitis G virus isolates of distinct genotypes from japan. *J. Gen. Virol.*, 78(4):737–745, 1997.

[113] J. Pelletier and N. Sonenberg. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature*, 334:320–325, 1988.

[114] O. Perriquet, Touzet H, and M. Dauchet. Finding the common structure shared by two homologous RNAs. *Bioinformatics*, 19(1):108–116, 2003.

[115] T. V. Pestova, I. N. Shatsky, S. P. Fletcher, R. J. Jackson, and C. U. Hellen. A prokaryotic-like mode of cytoplasmic eukaryotic ribosome binding to the initiation codon during internal translation initiation of hepatitis C and classical swine fever virus RNAs. *Genes Dev.*, 12(1):67–83, 1998.

[116] C.W. Pleij, K. Rietveld, and L. Bosch. A new principle of RNA folding based on pseudoknotting. *Nucl. Acids Res.*, 13(5):1717–1731, 1985.

[117] V. Proutski, E. A. Gould, and E. C. Holmes. Secondary structure of the 3' untranslated region of flaviviruses: similarities and differences. *Nucl. Acids Res.*, 25(6):1194–1202, 1997.

[118] V. Proutski, T. S. Gritsun, E. A. Gould, and E. C. Holmes. Biological consequences of delitions within the 3'-untranslated region of flaviviruses may be due to rearrangements of RNA secondary structure. *Virus Research*, 64:107–123, 1999.

[119] L. Psaridi, U. Georgopoulou, A. Varaklioti, and P. Mavromara. Mutational analysis of a conserved tetraloop in the 5' untranslated region of hepatitis C virus identifies a novel RNA element essential for the internal ribosome entry site function. *FEBS Lett.*, 453(1-2):49–53, 1999.

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

[121] S. C. Ray, Y. M. Wang, O. Laeyendecker, J. R. Ticehurst, S. A. Villano, and D. L. Thomas. Acute hepatitis C virus structural gene sequences as predictors of persistent viremia: hypervariable region 1 as a decoy. *J. Virol.*, 73(4):2938–2946, 1999.

[122] J. Reeder and R. Giegerich. Improved efficiency of RNA secondary structure prediction including pseudoknots. *unpublished*, ECCB 2002 poster; http://bibiserv.techfak.uni-bielefeld.DE/pknotsrg/, 2002.

[123] E. Rivas and S.R. Eddy. The language of RNA: a formal grammar that includes pseudoknots. *Bioinformatics*, 16(4):334–340, 2000.

[124] Elena Rivas and Sean R. Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of Molecular Biology*, 285:2053–2068, 1999.

[125] J. Ruan, G. D. Stormo, and W. Zhang. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics*, 20:58–66, 2004.

[126] Gerard Salton and Chris Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513–523, 1988.

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

[128] W. R. Schmitt and M. S. Waterman. Linear trees and RNA secondary structure. *Discr. Appl. Math.*, 12:412–427, 1994.

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

[130] Peter Schuster and Peter F. Stadler. Discrete models of biopolymers. In M. Drew J. Crabbe, A. Konopka, editor, *Handbook of Computational Chemistry and Biology*. Marcel Dekker, New York, 2002. in press.

[131] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[132] M.J. Serra and D.H. Turner. Predicting thermodynamic properties of RNA. *Methods Enzymol.*, 259:242–61, 1995.

[133] P. Simmonds and D. B. Smith. Structural constraints on RNA virus evolution. *J. Virol.*, 73(7):5787–5794, 1999.

[134] J. N. Simons, S. M. Desai, D. E. Schultz, S. M. Lemon, and I. K. Mushahwar. Translation initiation in GB viruses A and C: evidence for internal ribosome entry and implication for genome organization. *J. Virol.*, 70:6126–6135, 1996.

[135] C.M. Smith and J.A. Steitz. Sno storm in the nucleolus: new roles for myriad small RNPs. *Cell*, 89(5):669–672, 1997.

[136] D. B. Smith, N. Cuceanu, F. Davidson, L. M. Jarvis, J. L. Mokili, S. Hamid, C. A. Ludlam, and P. Simmonds. Discrimination of hepatitis G virus/GBV-C geographical variants by analysis of the 5' non-coding region. *J. Gen. Virol.*, 78(7):1533–1542, 1997.

[137] C. M. Spahn, J. S. Kieft, R. A. Grassucci, P. A. Penczek, K. Zhou, J. A. Doudna, and J. Frank. Hepatitis C virus IRES RNA-induced changes in the conformation of the 40s ribosomal subunit. *Science*, 291(5510):1959–1962, 2001.

[138] A. Spicher, O.M. Guicherit, L. Duret, A. Aslanian, E.M. Sanjines, N.C. Denko, A.J. Giaccia, and H.M. Blau. Highly conserved RNA sequences that are sensors of environmental stress. *Mol. Cell Biol.*, 18(12):7371–7382, 1998.

[139] Roman Stocsits. *Nucleic Acid Sequence Alignments of Partly Coding Regions*. PhD thesis, Universität Wien, 2003.

[140] J.E. Tabaska, R.B. Cary, H.N. Gabow, and G.D. Stormo. An RNA folding method capable of identifying pseudoknots and base triples. *Bioinformatics*, 14(8):691–699, 1998.

[141] 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(5):3307–12, 1996.

[142] C. K. Tang and D. E. Draper. An unusual mRNA pseudoknot structure is recognized by a protein translation repressor. *Cell*, 57:531–536, 1989.

[143] C. K. Tang and D. E. Draper. Evidence for allosteric coupling between the ribosome and repressor binding sites of a translationally regulated mRNA. *Biochemistry*, 29:4434–4439, 1990.

[144] S. Tang, A. J. Collier, and R. M. Elliott. Alterations to both the primary and predicted secondary structure of stem-loop IIIc of the hepatitis C virus 1b 5' untranslated region (5'UTR) lead to mutants severely defective in translation which cannot be complemented in trans by the wild-type 5'UTR sequence. *J. Virol.*, 73(3):2359–2364, 1999.

[145] Andrea Tanzer and Peter F. Stadler. Molecular evolution of a microRNA cluster. *J. Mol. Biol.*, 2004. in press.

[146] E. ten Dam, I. Brierly, S. Inglis, and C. Pleij. Identification and analysis od the pseudoknot-containing *gag-pro* ribosomal frameshift signal of simian retrovirus-1. *Nucl. Acids Res.*, 22:2304–2310, 1994.

[147] C.A. Theimer, Y. Wang, D.W. Hoffman, H.M. Krisch, and D.P. Giedroc. Non-nearest neighbor effects on the thermodynamics of unfolding of a model mrna pseudoknot. *J. Mol. Biol.*, 279:545–564, 1998.

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

[149] A. Tuplin, J. Wood, D. J. Evans, A. H. Patel, and P. Simmonds. Thermodynamic and phylogenetic prediction of RNA secondary structures in the coding region of hepatitis C virus. *RNA*, 8(6):824–841, 2002.

[150] F.H. van Batenburg, A.P. Gultyaev, and C.W.A. Pleij. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *J. Theor. Biol.*, 174:269–280, 1995.

[151] M. H. V. van Regenmortel, C.M. Fauquet, D.H.L. Bishop, E.B. Carstens, M.K. Estes, S.M. Lemon, J. Maniloff, M.A. Mayo, D.J. McGeoch, C.R. Pringle, and R.B. Wickner. *Virus Taxonomy: The Classification and Nomenclature of Viruses. The Seventh Report of the*

International Committee on Taxonomy of Viruses. Academic Press, SanDiego, 2000. `http://www.ncbi.nlm.nih.gov/ICTVdb/`.

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

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

[154] J.H. Withey and D.I. Friedman. The biological roles of trans-translation. *Curr. Opin. Microbiol.*, 5(2):154–159, 2002.

[155] I. H. Witten and E. Frank. Nuts and bolts: Machine learning algorithms in java,. In *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations.*, pages 265–320. Morgan Kaufmann, 1999.

[156] C. Witwer. *Prediction of Conserved and Consensus RNA Structures.* PhD thesis, Universität Wien, 2001.

[157] C. Witwer, I. L. Hofacker, and P. F. Stadler. Prediction of consensus RNA secondary structures including pseudoknots. *submitted*, 2004.

[158] C. Witwer, S. Rauscher, I. L. Hofacker, and P. F. Stadler. Conserved RNA secondary structures in picornaviridae genomes. *Nucl. Acids Res.*, 29(24):5079–5089, 2001.

[159] Christina Witwer, Susanne Rauscher, Ivo L. Hofacker, and Peter F. Stadler. Conserved RNA secondary structures in Picornaviridae genomes. *Nucl. Acids Res.*, 29:5079–5089, 2001.

[160] C. R. Woese, Winker S., and R. R. Gutell. Architecture of ribosomal RNA: Constraints on the sequence of tetra-loops. *Proc. Natl. Acad. Sci., USA*, 87:8467–8471, 1990.

[161] J. Wower, I.K. Wower, B. Kraal, and C.W. Zwieb. Quality control of the elongation step of protein synthesis by tmRNP. *J. Nutr.*, 131(11):2978S–2982S, 2001.

[162] J.R. Wyatt, J.D. Puglisi, and I. Tinoco. RNA pseudoknots: Stability and loop size requirements. *J. Mol. Biol.*, 214:455–470, 1990.

[163] T. Xia, J. SantaLucia Jr., M. E. Burkard, R. Kierzek, S. J. Schroeder, X. Jiao, C. Cox, and Douglas H. Turner. Parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick pairs. *Biochemistry*, 37:14719–14735, 1998.

[164] J. Xiang, S. Wunschmann, W. Schmidt, J. Shao, and J. T. Stapleton. Full-length GB virus C (hepatitis G virus) RNA transcripts are infectious in primary CD4-positive T cells. *J. Virol.*, 74:9125–9133, 2000.

[165] N. Yamada, K. Tanihara, Takada T. Yorihuzi, M. Tsutsumi, H. Shimomura, T. Tsuji, and T. Date. Genetic organization and diversity of the 3' noncoding region of the hepatitis C virus genome. *Virology*, 223(1):255–261, 1996.

[166] M. K. Yi and S. M. Lemon. 3' nontranslated RNA signals required for replication of hepatitis C virus rna. *J. Virol.*, 77(6):3557–3568, 2003.

[167] S. You and R. Padmanabhan. A novel in vitro replication system for dengue virus. initiation of RNA synthesis at the 3'-end of exogenous viral RNA templates requires 5'- and 3'-terminal complementary sequence motifs of the viral RNA. *J. Biol. Chem.*, 274(47):3714–3722, 1999.

[168] H. Yu, C. W. Grassmann, and S. E. Behrens. Sequence and structural elements at the 3' terminus of bovine viral diarrhea virus genomic RNA: functional role during RNA replication. *J. Virol.*, 73(5):3638–3648, 1999.

[169] W. D. Zhao and E. Wimmer. Genetic analysis of a poliovirus/hepatitis C virus chimera: new structure for domain II of the internal ribosomal entry site of hepatitis C virus. *J. Virol.*, 75(8):3719–3730, 2001.

[170] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52, 1989.

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

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

[173] C. Zwieb, I. Wower, and J. Wower. Comparative sequence analysis of tmRNA. *Nucl. Acids Res.*, 27(10):2063–2071, 1999.