Prediction of Conserved and Consensus RNA Structures

DISSERTATION
zur Erlangung des akademischen Grades
Doctor rerum naturalium

Vorgelegt der
Fakultät für Naturwissenschaften und Mathematik
der Universität Wien

von
Christina Witwer

am Institut für Theoretische Chemie und Molekulare
Strukturbiologie

im März 2003

Contents

1 Introduction

1.1 General Context

Proteins and nucleic acids, DNA and RNA, are the fundamental biopolymers in molecular genetics. In all prokaryotic and eukaryotic cells genetic information is stored in the form of DNA. In viruses the genetic material is either DNA or RNA. RNA plays an important role in the expression of genes. DNA is copied to messenger RNA (mRNA), which is subsequently decoded for protein synthesis. Translation is mediated by ribosomes, composed of ribosomal RNA (rRNA) and proteins, and transfer RNA (tRNA), which translates the codons of the mRNA into the amino acids of the protein. Proteins play a crucial role in most biological processes, the remarkable scope of their activity includes catalysis of chemical reactions, transport of small molecules and ions, control of growth and differentiation of cells, and an important role in immune protection, to mention just a few functions fulfilled by proteins.

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. In the 1980s Cech *et al.* revealed the autocatalytic splicing of the precursor of rRNA [16, 89], and 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 [51].

Since then a number of other ribozymes have been discovered. A partial list of such molecules includes small nuclear RNAs (snoRNAs) [151] that compose the pre-mRNA splicing machinery, signal recognition particle (SRP) RNA [84] necessary for protein translocation, and rRNA [116, 115]. 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.

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 [152]). 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 [17]. 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 [126, 82]. 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 [64]. Functional important RNA structures are not restricted to non-coding RNA, as the examples of the Rev response element (RRE) of HIV1, which is located within the *env* gene [101], and the cis-acting replication element located in the coding region of picornaviruses [107], show.

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. A 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 GU pairs. 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 a RNA structure which can be drawn as outerplanar graph, i.e. all base pairs can be drawn in the half-plane without intersec-

tions, 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 $tRNA^{phe}$ [4] shown in Figure 1.

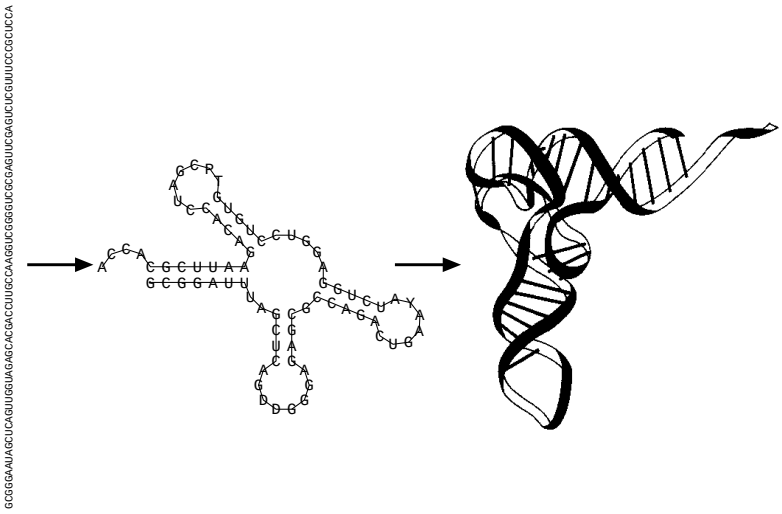


Figure 1: Sequence, secondary structure and schematic representation of the tertiary structure of tRNA

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 [38, 149] 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. Whereas the prediction of RNA structure including pseudoknots based on the same model has been proven to be NP-complete [100, 3].

1.3 Objectives of this Work

Purely phylogenetic methods can be used to derive conserved elements or a consensus structure only when a sufficiently large number of sequences is available, while the accuracy of purely thermodynamic structure prediction is often not satisfactory.

For the prediction of the consensus secondary structure for a small set of related sequences, algorithms combining thermodynamic and phylogenetic structure prediction have been developed, e.g. [97, 69]. 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.* [70, 72]. One part of this work is concerned with the identification of potential functional important structures, using `alidot`, in the genomes of picornaviruses. Picornaviruses are small RNA containing viruses, including important human and agricultural pathogens, like *Poliovirus, Hepatitis A virus* and *Foot-and-mouth disease 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 seven of the currently nine genera of the family *Picornaviridae*.

Another aspect of finding conserved RNA structures is the prediction of conserved secondary structures including pseudoknots. Thermodynamic structure predictions based on the standard energy model are very restricted in both sequence length and allowed complexity of pseudoknots. When a large number of homologous RNA sequences is available, comparative sequence analysis methods are successful in predicting the consensus structures.

Tabaska *et al.* [156] developed a method based on graph theory for RNA structure prediction including pseudoknots from an alignment of homologous RNA sequences. They use a rather simple scoring scheme which includes both thermodynamic and phylogenetic information. Their algorithm has been applied to an alignment of 33 eubacterial and archaebacterial SRP RNA sequences, and found essentially complete agreement with the phylogenetic derived structure.

It is desirable to be able to predict the consensus structure including pseudoknots based on a smaller set of sequences. Here we report an improvement of the scoring procedure that reduced the number of sequences required for a secondary structure prediction including pseudoknots.

1.4 Organization of this Thesis

In the following chapter the basic concepts of RNA secondary structure are introduced, the definitions of secondary structure and different classifications of pseudoknots are presented. Furthermore the fundamentals of the maximum weighted matching (MWM) algorithm, which is a well-known combinatorial optimization algorithm, are given. The MWM algorithm forms the

basis of the consensus structure prediction program developed during this work.

Chapter 3 provides an overview of the state of the art in computational RNA structure prediction with and without pseudoknots, followed by a detailed description of the algorithm `alidot`.

Chapter 4 starts with a brief overview of picornaviruses in general. Subsequently newly developed tools for data representation are described, followed by the prediction results of `alidot` of known and new structural elements of the genomes of picornaviruses.

Chapter 5 presents the program `hxmatch` for the consensus structure prediction of a set of homologous RNA sequences. This program is based on the MWM algorithm, like the method of Tabaska *et al.* [156], but uses an improved scoring function and an elaborated post processing. `Hxmatch` is tested on three different types of RNA known to contain pseudoknots. The presentation of the results is accompanied by a detailed discussion, including a critical evaluation of the MWM approach.

2 Basics

In this chapter essential concepts that are fundamental for later discussion will be established. These include the definition and representation of secondary and bi-secondary RNA structures, and some well known principles of graph theory.

2.1 Graphs

The classical representation of secondary structure is the drawing as a graph, and throughout this work many findings of graph theory are used. Therefore, we give several basic definitions from graph theory and some basic notation, these can be found in many textbooks, e.g. in [2].

A *graph* $G = (V, E)$ consists of a finite set V of vertices (nodes) and a finite set E of edges (arcs). 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, $(v_1, v_2, \ldots v_n)$, 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 [18, 21]. 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. [60].

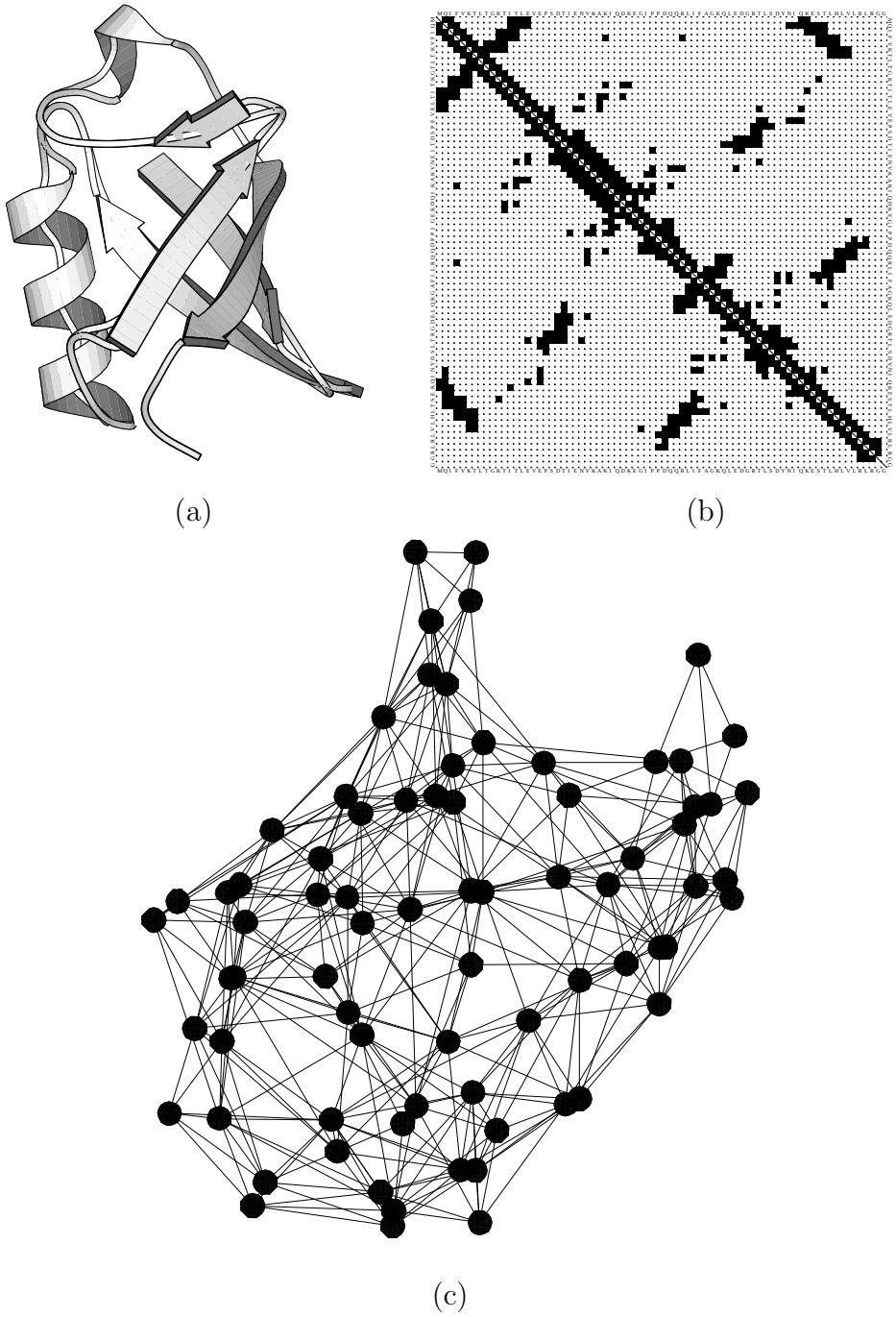
A contact structure is faithfully represented by the *contact matrix* $\mathbf{C}$ with the entries $\mathbf{C}_{ij} = 1$ if the monomers i and j are spatial neighbors without being adjacent along the backbone, and $\mathbf{C}_{ij} = 0$ otherwise. Hence $\mathbf{C}_{ij} = 0$ if $|i - j| \leq 1$. Note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

The *contact graph* has the adjacency matrix $\mathbf{A} = \mathbf{B}+\mathbf{C}$. The familiar drawing of RNA secondary structures are a much used example of bimolecular contact graphs.

Definition 2 *A diagram is called an 1-diagram if for any two arcs* $\alpha, \beta \in \Omega$ *holds* $\alpha \cap \beta = \emptyset$ *or* $\alpha = \beta$.



(a) (b)

(c)

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

2.3 RNA Contact Structures

The "classical" definition of RNA secondary structure [168] 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 [62]. Bi-secondary structures include almost all known pseudoknotted RNA structures, with the exception of the E.coli αmRNA. This chapter follows the definitions given in [62].

2.3.1 Secondary Structure

The classical definition of an RNA secondary structure [168] requires that each base interacts with at most one other nucleotide. Thus nucleic acid secondary structures are special types of 1-diagrams. The second defining condition is that arcs do not cross. 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 2:

Let $\alpha = \{i, j\}$ with $i < j$ be an arc of a diagram. We write $\bar{\alpha} \stackrel{\text{def}}{=\!=} [i, j] \subset \mathbb{R}$ 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 [148]. This leads to the following definition:

Definition 3 *A secondary structure is a 1-diagram in which any two arcs are consistent.*

As a consequence, each secondary structure can be encoded as a string s of length n in the following way: If the vertex i is unpaired, then $s_i =$ `'.'`. Each arc $\alpha = \{p, q\}$ with $p < q$ translates to $s_p =$ `'('` and $s_q =$ `')'`. Since the arcs are consistent their corresponding parentheses are either nested, `(( ))`, or next to each other, `()()`. As there are no arcs between neighboring vertices in a 1-diagram 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 [71].

Secondary structure graphs are outerplanar, 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.7.

2.3.2 Bi-Secondary Structure

A 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 4 *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 3), or in very large molecules from sufficiently complicated cross-linking patterns. Parallel-stranded RNA has not been observed (so far), see however ref. [39] on parallel-stranded DNA. Wollenzien *et al.* [15] 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.

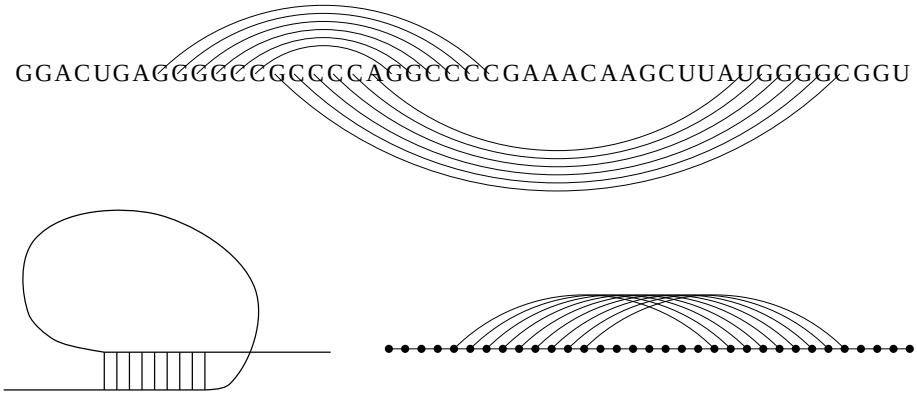


Figure 3: The contact structure of the proposed SRV-1 frame-shift signal contains a pseudo-knot, see reference [159]. Pseudo-knots such as this one belong to the class of bi-secondary structures. Knots such as the one in the lower part of the figure do not belong to the class of bi-secondary structures. Knots, in contrast to pseudo-knots, may contain parallel stranded helices which so far have not been described for RNA.

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 4. However it is possible to define the *normal form* of

a bi-secondary structure by means of the following rule: The leftmost arc of two arcs that are not consistent belongs to Ω_U.

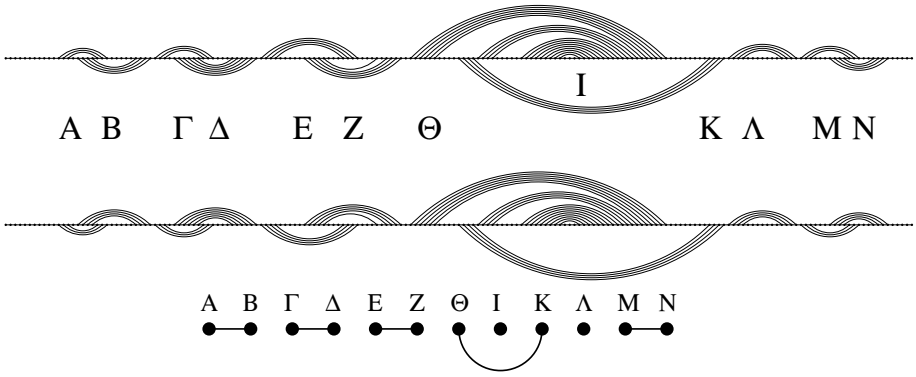


Figure 4: Two diagrams encoding the 3' non-coding region of tobacco mosaic virus RNA [1]. The upper diagram corresponds to the normal form, the lower diagram maximizes the number of upper arcs. Stems are labeled by uppercase Greek letters. The third line shows the inconsistency graph of the tmvRNA structure.

2.3.3 The Inconsistency Graph of a Diagram

Definition 5 *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.*

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 $p = 0, \ldots, h-1$. 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 4 and 5.

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 [158]. Compensatory mutations have defined an unusual pseudo-knotted structure [157], the thermodynamics of which were subsequently investigated in detail [46]. The diagram of its contact structure cannot be drawn without intersections, see Figure 5.

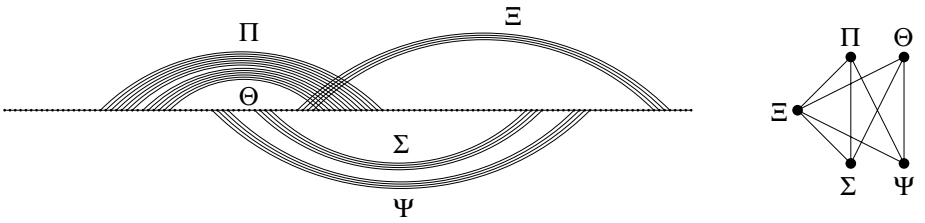


Figure 5: Diagram of the contact structure of E. coli α-mRNA. The structure contains 5 stems, labeled by uppercase Greek letters. We may choose the color partition if $\Theta(\Delta)$ such that all arcs in a stem have the same color. It therefore suffices to draw the inconsistency graph for stems (r.h.s. of the figure).

2.3.4 Color Partition of a Graph

Definition 6 *A* color partition *of a graph* Γ *is the partition* $V = V_1 \cup V_2 \cup \cdots \cup V_c$ *of its vertex set into* c *subsets* V_i *such that no two vertices in* V_i *are adjacent. The* chromatic number $\chi(\Gamma)$ *is the smallest number* c *of colors for which a color partition of* Γ *can be found.*

An arbitrary diagram Δ can be decomposed into substructures by means of the following obvious result: Let $\Delta = ([n], \Omega)$ be a diagram and let $\mathcal{V} : \Omega = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_c$ be a partition of the set of arcs. Then the sub-diagram

$([n], \Omega_i)$, $i = 1, \ldots, c$, can be drawn without intersection if and only if $\mathcal{V}$ is a color partition of the inconsistency graph $\Theta(\Delta)$. Noticing that $\chi(\Gamma) = 1$ if Γ contains no edges and $\chi(\Gamma) = 2$ if Γ is bipartite with non-empty edge set the following characterization follows immediately:

- (i)Δ is a secondary structure iff $\chi(\Theta(\Delta)) = 1$;
- (ii)Δ is a bi-secondary structure iff $\chi(\Theta(\Delta)) \leq 2$.

The chromatic number $\chi(\Theta(\Delta))$ may therefore serve as a measure for the structural complexity of a contact structure.

2.3.5 Loop Decomposition of Secondary Structures

A vertex 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 7 *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 vertices which are immediately interior to some base pair (p, q)*, the "closing" pair of the loop. The number of these vertices is called the size of the loop.*

(iii) An external vertex is an unpaired vertex which does not belong to a loop. A collection of adjacent external vertices is called an external element. If it contains the vertex 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 8 *A stem* $[(p, q), ..., (p+k, q-k)]$ *is called terminal if* $p - 1 = 0$ *or* $q + 1 = n + 1$ *or if the two vertices* $p - 1$ *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 9 *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.*

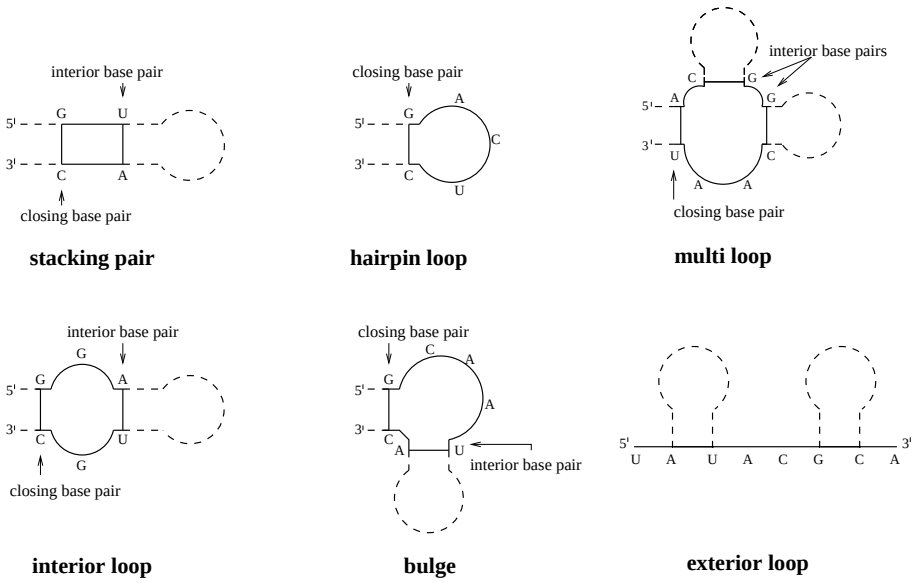


Figure 6: Basic loop types

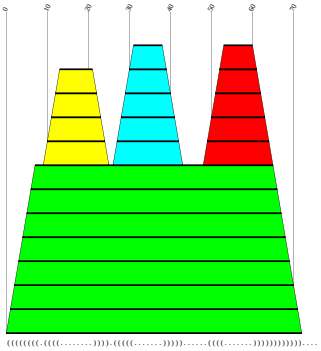
2.3.6 RNA Secondary Structure Representation

Figure 7 shows a variety of different representation forms for RNA secondary structure. Beside the conventional drawing as a planar graph and the string notation, which is defined in section 2.3.1 a secondary structure can be represent 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 McCakill'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.

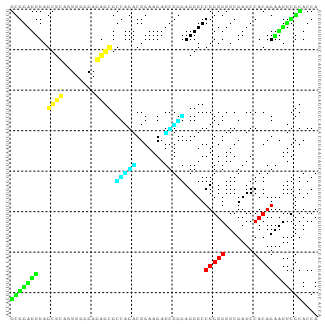
Especially useful to compare even large structures is the *mountain*-representation (or *mountain plot*) [73]. The three symbols of the string representation '.', '(' and ')' are assigned to three directions "horizontal', 'up' and 'down' in the plot. The structural elements match certain secondary structure 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.

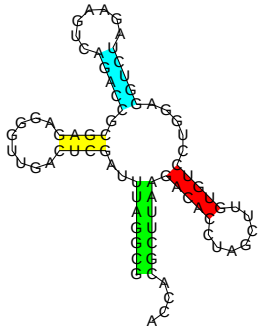
In the *linked diagram* representation the sequence is arranged along the x-axis and the base pairs are drawn as arcs confined to the upper half-plane. The *circle* representation places the sequence along a circle and the base pairs are represented by the arcs.



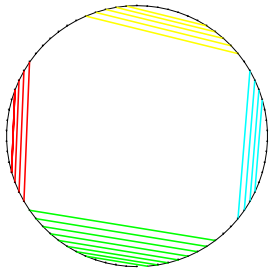
Mountain plot



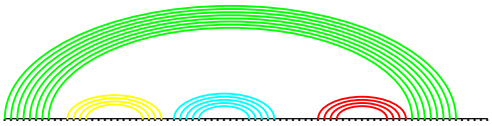
Dot plot



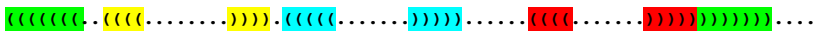
Conventional



Circle



Linked diagram



String

Figure 7: Different representations of RNA secondary structure. All drawings show the same structure and use the same colors to mark the different stems.

2.4 Classifications of Pseudoknots

The consideration of overlapping base pair interactions yields an enormous enlargement of the structure space. The number of possible structures grows faster than exponentially. We are not able to deal with the resulting complexity by computational means. Additionally, many of them can never be realized by an RNA sequence (e.g. parallel β-sheets). Thus the term "pseudoknot" gives not at all a sufficient definition to what extend the structure space should be enlarged. Simply calling all structures containing an overlapping base pair a pseudoknot is not of great help. A number of different classifications of pseudoknots have been proposed. One possibility is the consideration of bi-secondary structures (Section 2.3.2), which includes almost all known pseudoknotted RNA structures, with the exception of *E.coli* αmRNA. Other definitions of special types of pseudoknots include the definition of H-type pseudoknots [132], and the decomposition of generalized RNA structures into so-called nets proposed by H. Isambert and E.D. Siggia [77].

2.4.1 H-type Pseudoknots

The usual definition for an H-type pseudoknot is, that nucleotides from a hairpin-loop are basepaired with a single-stranded region outside of the hairpin [132]. This section follows a more restrictive definition of H-type pseudoknots given by C. Haslinger [61].

Definition 10 *A building block $B_{i,k,l,j}$, $i \leq k \leq l \leq j$ is a secondary structure on the interval $[i,j]$ with a gap at $[k+1, l-1]$, where all base pairs fulfill the condition:*

$$B_{i,k,l,j} = \{(p,q) | i \leq p \leq k \text{ and } l \leq q \leq j\} \tag{1}$$

Note that, because of its gap, a building block alone is not a valid secondary structure.

Definition 11 *Two building blocks $B_{i,k,l,j}$ and $B_{i',k',l',j'}$ are H-type generating if their associated intervals are disjoint:*

$$\{[i,k] \cup [l,j]\} \cap \{[i',k'] \cup [l',j']\} = \emptyset \tag{2}$$

and arranged in an alternating way:

$$i < i' : [i,k].[i',k'].[l,j].[l',j'] \tag{3}$$

$$i' < i : [i',k'].[i,k].[l',j'].[l,j] \tag{4}$$

Definition 12 *An H-pseudoknot $Pk_{i,j'}$ is obtained when we merge two H-type generating building blocks. We can distinguish an upstream $B^u_{i,k,l,j}$ and a downstream $B^d_{i',k',l',j'}$ building block:*

$$PK_{i,j'} = B^u_{i,k,l,j} \cup B^d_{i',k',l',j'} \tag{5}$$

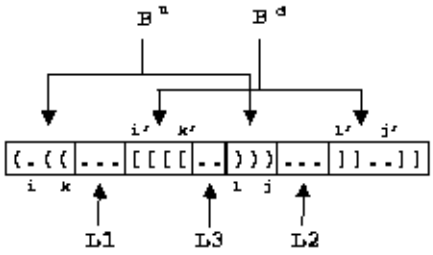


Figure 8: Two H-type generating building blocks merged resulting in a H-pseudoknot.

When (k,l), (k',l') and (i,j), (i',j') are base pairs, we produce three unpaired regions:

$$L1 = \{n | k < n < i'\} \tag{6}$$

$$L3 = \{n | k' < n < l\} \tag{7}$$

$$L2 = \{n | j < n < l'\} \tag{8}$$

All vertices $L1$ are immediately interior to (k, l), vertices $L2$ are immediately interior to (k', l'). All vertices $L3$ are immediately interior to both (k, l) and (k', l'). These three regions are in the literature referred as loops although they do not meet our definition of loops (definition 7), because it is not possible to uniquely assign the vertices $L3$ to just one loop.

Definition 13 *A simple H-pseudoknot is an H-pseudoknot where* $B^u_{i,k,l,j}$ *and* $B^d_{i',k',l',j'}$ *are stems.*

Figure 9 shows an example of a simple H-pseudoknot.

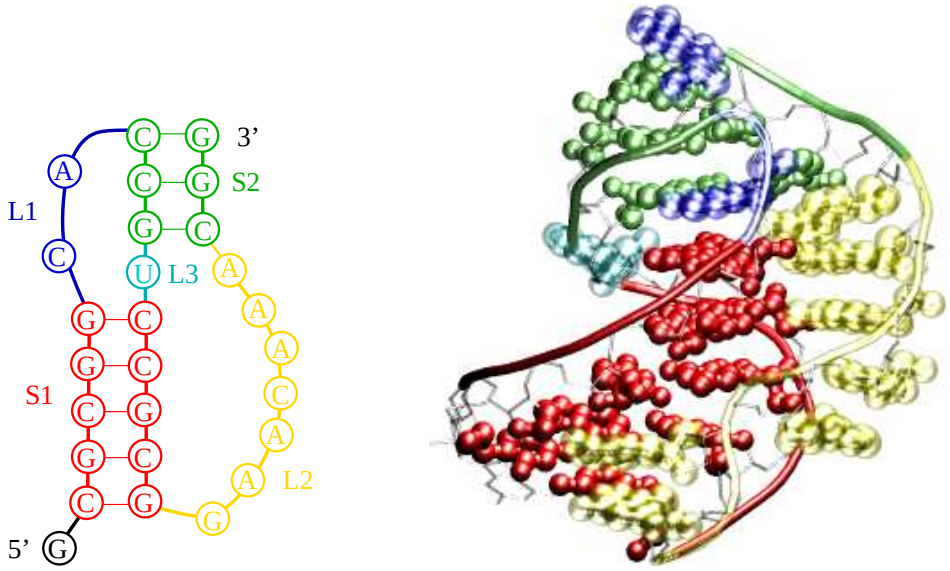


Figure 9: A simple H-pseudoknot: Two-dimensional representation (*l.h.s.*) and crystal structure (*r.h.s.*) of the beet western yellow virus ribosomal frame shifting pseudoknot, as solved by X-ray diffraction [155].

The common use of the term *H-type* pseudoknots usually matches definition 12 of H-pseudoknots, in that additional structure elements within the loops $L1$ and $L2$ are not allowed. Furthermore it is assumed that the two stems $S1$ and $S2$ stack coaxially, therefore the size of $L3$ is restricted to be 0 or 1.

2.4.2 Isambert-Siggia Decomposition of Secondary Structures

In a recent study Hervé Isambert and Eric D. Siggia [77] proposed a decomposition of generalized RNA secondary structures into so-called *nets*. Here we reformulate their work in more standard notation and provide 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 14 *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.*
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.

Is is clear that the resulting coloring is unique. An example is shown in Figure 12.

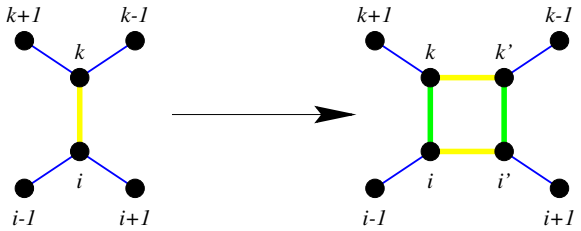


Figure 10: Modifying an IS-colored 1-contact graph in order to deal with isolated bonds.

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 15 *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 10.*

Remark 1 *The IS-coloring of a 1-contact graph can be computed in $\mathcal{O}(n)$ steps.*

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 16 *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.*

Is 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 15. 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 14.

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 17 *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

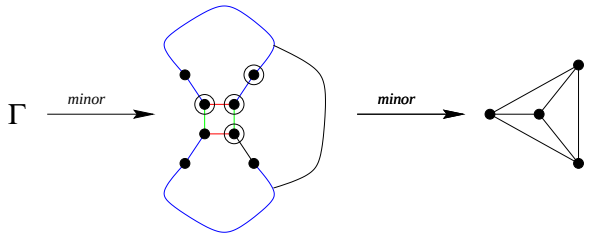


Figure 11: Proof of Theorem 3.

this case Γ has a minor as depicted in Figure 11, 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 since the backbone must be connected, and both cycles must contain at least one blue (backbone) edge. It is clear from Figure 11 that Γ' contains K_4 as a minor. Hence Γ is not an outerplanar graph [19], 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.

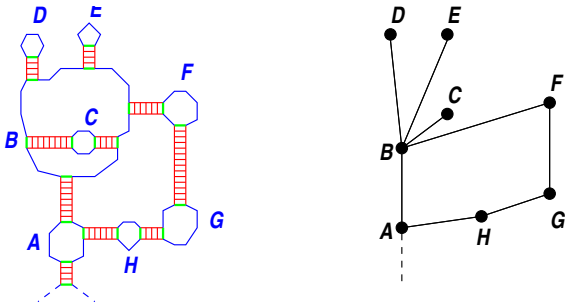


Figure 12: A counter example to the converse of Theorem 3. All nets of the 1-contact structure Γ (r.h.s.) are simple cycle, labeled A through H. Nevertheless nets B and C together with their connecting stems form a pseudoknot. The gel $\mathsf{Gel}(\Gamma)$ contains a cycle and hence is not a tree.

Remark 2 *The converse of Theorem 3 is not true as the example in Figure 12 shows.*

Definition 18 *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 [77] that the converse cannot be true.

As part of this thesis an algorithm which finds the IS-coloring of a general graph and constructs its gel has been implemented. The man-page of this program, called `iscolor`, is given in A.4.

2.5 Matching Theory

In chapter 5 we present our algorithm for predicting conserved RNA structures including pseudoknots based on maximum weighted matching (MWM). Since the MWM algorithm is an essential part of our method, this section will describe the fundamentals of the algorithm, which are also important for the interpretation of the results.

A *matching* on a graph is a set of edges no two of which share an endpoint. A maximum *cardinality* matching includes the maximum number of edges which form a matching. Let w be a weight function on the edges of G. The *maximum-weight matching* problem is then to find a matching M having maximum total weight.

The basis of nearly all matching algorithms is the concept of *augmenting paths* [7]. The first efficient algorithm for weighted matching was presented by Edmonds in 1965 [33], it is based on the *blossom-shrinking algorithm* and is outlined in detail in the following section. The algorithm of Edmonds requires run-time $O(n^4)$, where n is the number of vertices. Gabow presented an improved implementation of Edmonds' algorithm with an asymptotic running-time to $O(n^3)$ [41]. A detailed description of the maximum weighted matching algorithm and its variants can be found for example in [147].

2.5.1 Definitions and Matching Concepts

A *weighted graph* $G = (V, E, w)$ has given a weight function $w : E \to \mathbb{R}$ on the edges of G. A subset $M \subseteq E$ of edges is called a *matching* on graph G, if no two edges of M share a common vertex. All edges uv in M are said to be *matched*, and all edges in the difference $E \setminus M$ are *unmatched*. A vertex v is called *matched* if there exists a matched edge incident to that vertex, otherwise v is *unmatched* or *free*.

A *k-matching*, for $k \geq 0$ contains exactly k edges.
A *maximum cardinality matching* contains the maximum number of edges of a graph that form a matching.
In a *maximum weighted matching* the sum of the weights of the edges forming the matching is maximal.

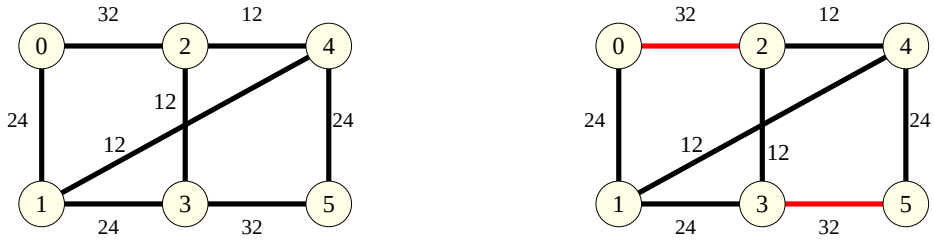


Figure 13: Left: A weighted graph. Right: 2-Matching on a weighted graph, the matched edges are colored red.

An *alternating path* with respect to a matching M is a path where the edges connecting the vertices are alternately in M and not in M. An *alternating cycle* is an alternating path that starts and ends at the same vertex.

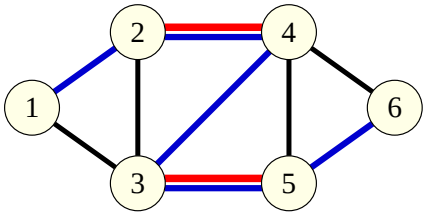
An *augmenting path* with respect to a matching M is an alternating path that starts and ends with an unmatched edge.

The *symmetric difference* for two sets S_1 and S_2 is defined as

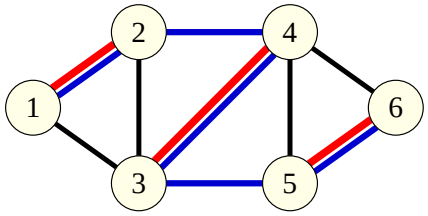
$$S_1 \oplus S_2 = (S_1 \cup S_2) - (S_1 \cap S_2)$$

An augmenting path p can be used to augment the current matching M by forming the symmetric difference of M and p: $M' = M \oplus p$. It can be seen easily that M' is a matching and contains one more edge then M, compare figure 14. An augmenting thus changes the pairing partners of previously matched vertices, these vertices are said to have been *rematched*.

Central to matching theory is *Berge's lemma* [7].



graph containing a 2-matching
M = {(2, 4), (3, 5)}
and an augmenting path
P = {(1, 2), (2, 4), (4, 3), (3, 5), (5, 6)}



graph containing a 3-matching
$M' = M \oplus P = \{(1,2),(4,3),(5,6)\}$

Figure 14: Left: Augmenting path in a graph containing a 2-matching; Right: 3-Matching on the graph after augmenting. Matched edges are colored red, the path is colored blue, black edges indicate unmatched edges not contained in the path.

Berge's Lemma (Theorem 12.8 in [2]) A matched graph (G, M) has an augmenting path if and only if M is not a maximum cardinality matching.

An equivalent formulation to Berge's lemma is the
Augmenting Path Theorem: If a node r is unmatched in a matching M, and this matching contains no augmenting path that starts at vertex r, then node r is unmatched in some maximum matching.

This suggests the following approach for solving the matching problem. Start with some k-matching (which might be a 0-matching) and try to identify an augmenting path starting at some unmatched vertex r. If such a path is found, augment the matching; otherwise delete the vertex and all edges incident to it from the graph. This step is repeated for every free vertex of the graph.

Therefore the matching problem is reduced to finding whether or not the graph contains an augmenting path starting at a free vertex r.

2.5.2 Maximum Cardinality Matching

For the search for an augmenting path, initially all unmatched vertices of the graph are labeled *even* (+), all matched vertices are *unlabeled* (*). Starting with an unmatched vertex r^+ an *alternating tree* T is grown, such that each path from a vertex $u \in T$ to the root-vertex r is alternating with respect to the matching.

Let $v^* \notin T$ be adjacent to any vertex $u^+ \in T$. T is extended by taking the unmatched edge uv and the matched edge vw to T. v is labeled *odd* (-) and w is labeled *even* (+). This procedure is called a *grow step*.

Figure 15: Grow step: A matched vertex v adjacent to an even vertex r^+ gets an odd label, its mate an even label.

When an even vertex $u^+ \in T$ is adjacent to any vertex $v^+ \notin T$, an augmenting path $p = (v, u, \ldots, r)$ with respect to M has been found.

Figure 16: Finding an even labeled vertex $v^t \notin T$ adjacent to an even vertex $u^+ \in T$ results in an augmenting path

In a general graph the labeling of the vertices is not unique. Whenever there is an alternating cycle with respect to a matching, the vertices in this cycle can be labeled either even or odd, see Figure 18.

Let $G = \{V, E\}$ be a general graph. For any subset $S \in V$ we denote $\delta(S)$

the set of edges having exactly one endpoint in S, and $\gamma(S)$ the set of edges having both endpoints in S:

$\delta(S) = \{uv \in E : u \in S \text{ and } v \notin S\}$
$\gamma(S) = \{uv \in E : u \in S \text{ and } v \in S\}$

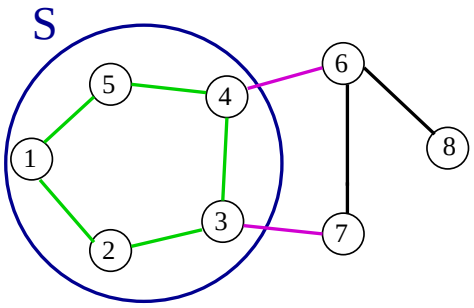


Figure 17: The subset $S \in V$ is indicated by the circle.
$\delta(S) = \{(3,7),(4,6)\}$
$\gamma(S) = \{(1,2),(2,3),(3,4),(4,5),(1,5)\}$

Blossoms A *blossom* is an odd length alternating cycle C. There exists exactly one vertex in a blossom $\mathcal{B}$, which is either free, or whose matching edge is not in $\gamma(\mathcal{B})$. This vertex is called the *base* b of $\mathcal{B}$.

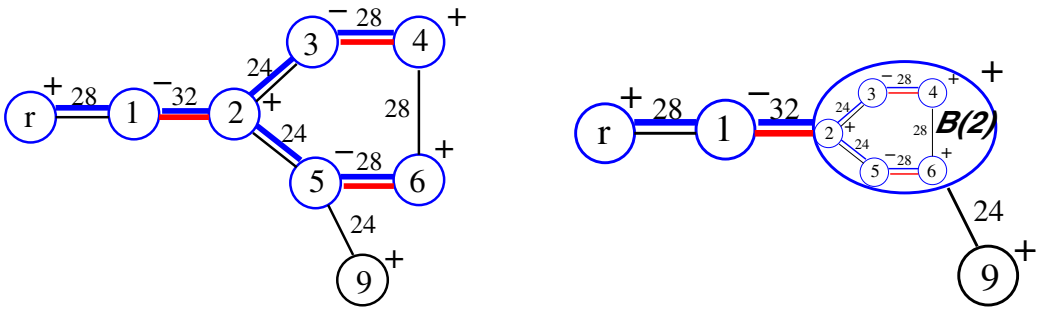


Figure 18: Shrink step: Two even vertices of the tree, which are adjacent to each other identify a blossom. The base b of the blossom in our example is vertex 2. The vertices and edges contained in the alternating cycle are *'shrinked'*. Left: original graph G. Right: Contracted graph G'.

If a vertex $u^+ \in T$ adjacent to another vertex $v^+ \in T$ is encountered during a search, a blossom is found. This leads to a so-called *shrink step*, i.e. a contracted graph G' is built according to the following rules:

$$G' = (V', E')$$

$$V' = (V \setminus \mathcal{B}) \cup \{b\}$$

$$E' = \gamma(V \setminus \mathcal{B}) \cup \{ub : uv \in \delta(\mathcal{B}) \text{ and } u \notin \mathcal{B}\}$$

The matching $M' = M \setminus \gamma(\mathcal{B})$ is then the matching on G' corresponding to M. The *blossom-shrinking approach* is due to Edmonds [34].

Lemma (Lemma 12.11 in [2]) Let G' be a contracted graph obtained by shrinking a blossom as described above. If an augmenting path p' with respect to the matching M' exists, the original graph G contains an augmenting path p with respect to M.

Let blossom $\mathcal{B}$ with base b be part of the augmenting path p'. p' can be split into p_1,b and p_2: $p' = (p_1, b, p_2)$. Let p_2 be the path which starts with the unmatched edge bv. If b is free, then $p' = p_1$ and p_2 is empty, hence p' does not traverse the blossom and is therefore identical to an augmenting path p in the original graph G. If b is matched, there is a vertex $u \in \mathcal{B}$ with the unmatched edge $uw \in G$ ($w \notin \mathcal{B}$). Since a blossom is an alternating cycle of odd length, it is always possible to find an even length alternating path $p_\mathcal{B}$ between the base b and an arbitrary vertex of $\mathcal{B}$. Let $p_\mathcal{B} = b, \ldots, u$ denote this path, then the augmenting path in the original graph G is given by $(p_1, p_\mathcal{B}, p_2)$. This procedure, restoring the original graph G and finding the augmenting path $p \in G$, is called an *expand step* for blossom $\mathcal{B}$.

We distinguish between *trivial* and *non-trivial* blossoms. Each vertex $v \in V$ corresponds to a trivial blossom $\mathcal{B} = \{v\}$. Shrink steps can be applied iterative, i.e. a non-trivial blossom $\mathcal{B}_1$ may be part of another blossom $\mathcal{B}_2$ (blossoms can be *nested*). $\mathcal{B}_1$ is then called a *subblossom* of $\mathcal{B}_2$. If a blossom

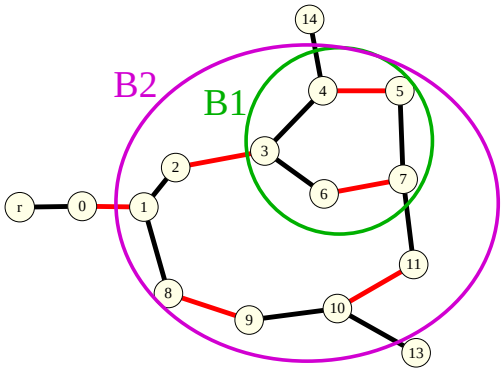


Figure 19: Nested blossoms: $B1$ is a subblossom of $B2$. $B2$ is a (non-trivial) surface blossom. The vertices r, 1 and 9 are trivial surface blossoms.

is not part of any other blossom it is called a *surface blossom*, compare figure 19.

When an augmenting path contains nested blossoms, the blossoms are expanded one by one, until the augmenting path p in the original graph is obtained.

We are now in the position to describe the algorithm for the search of an augmenting path. Let M be a matching on graph G, all free vertices are labeled even and all matched vertices are unlabeled. Let r^+ be the only vertex in the alternating tree $T = \{r^+\}$.

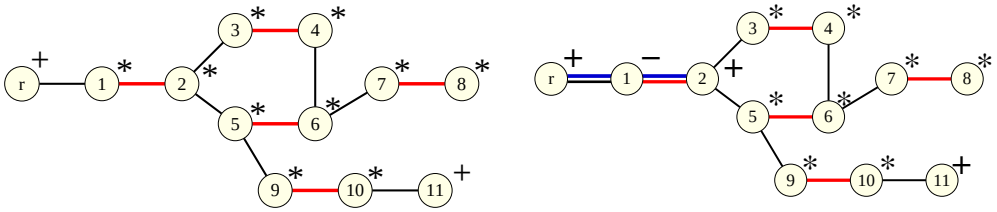
For each even labeled vertex $u^+ \in T$, the incident edges not in T are checked. Whenever an edge uv with $u^+ \in T$ and $v^+ \notin T$ exists, an augmenting path is detected. If the path contains blossoms, they are expanded one by one, until the augmenting path in the original graph is obtained. While no augmenting path has been found, the tree T is extended, if an edge uv with $u^+ \in T$ and $v^* \notin T$ exists (*grow step*). Otherwise, if an edge uv with $u^+ \in T$ and $v^+ \in T$ exists, a blossom is identified, and the contracted graph G' is built by means of a *shrink step*. When non of the above cases applies, no edges

Table 1: Algorithm for finding an augmenting path

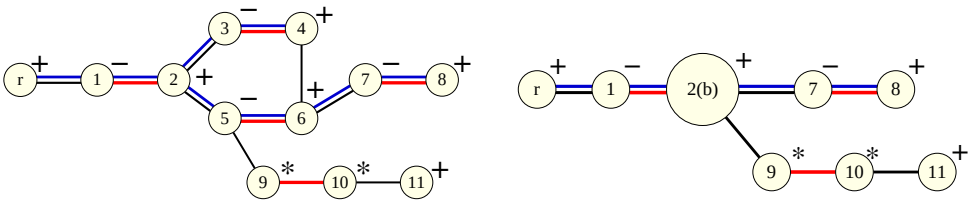
Search for an Augmenting Path
let M be a matching on a graph G
let r be the only vertex of T
while no edge uv with $u^+ \in T$ and $v^+ \notin T$ exists {
if an edge uv with $u^+ \in T$ and $v^* \notin T$ exists: *grow step*
else if an edge uv with $u^+ \in T$ and $v^+ \in T$ exists: *shrink step*
else terminate,
T is abandoned since no augmenting path
for r exists
}
an edge uv with $u^+ \in T$ and $v^+ \notin T$ exists
expand step $\Rightarrow$ *augmenting path in* G

from any even labeled vertex in T to a vertex not in T exists and r^+ is the only free vertex in T. Therefore no augmenting path starting in r exists and T is *abandoned*, i.e. all vertices contained in T will never be looked again. The search for an augmenting path is summarized in Table 1. An example is shown in Figure 20.

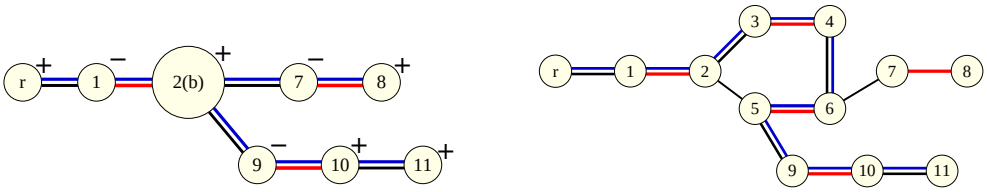
The algorithm for finding the maximum cardinality matching on a general graph is summarized in Table 2. The search for an augmenting path is repeated for each free vertex r of the graph. If no augmenting path starting in r is found, r is not matched in a maximum cardinality matching, and therefore r is deleted from the graph. Otherwise the matching is augmented by forming the symmetric difference $M' = M \oplus p$. and the search is repeated for another free vertex, until all remaining vertices are matched.



Starting from vertex r^+ an alternating tree is grown, which is indicated by the blue line.



Two even labeled vertices of the tree are adjacent (4^+ and 6^+), a blossom is found, and subsequently *shrinked*.



The alternating tree is grown further and an augmenting path is detected (r^+ to 11^+), and the blossom contained in the path is *expanded* to obtain an augmenting path in the original graph G.

Figure 20: Example of searching an augmenting path. Matched edges are colored red, unmatched edges black, the alternating tree is colored blue.

Table 2: Algorithm for finding a Maximum Cardinality Matching

Maximum Cardinality Matching
let M be an arbitrary matching in G

label all free vertices even, unlabel all matched vertices
for each vertex r in G {
if r is matched continue with an other vertex
search for an augmenting path
if an augmenting path has been found {
augment $M' = M \oplus P$
unlabel all vertices contained in T
delete all blossoms of T
destroy T
}
else T has been abandoned
continue with another vertex
}

M is a maximum cardinality matching

2.5.3 Maximum Weighted Matching

Linear programing and duality theory is the basis for solving the maximum-weight matching problem. For a detailed description of the theory see e.g. Bertsimas and Tsitsiklis [6], Chvatal [23] and Papadimitriou and Steiglitz [125], an overview is given in Ahuja [2].

Let $G = (V, E, w)$ be a weighted graph. An incidence vector x is associated to the edges of the graph, it represents the matching on the graph. The components x_e of the vector are 1 or 0, depending on whether edge e is contained in the matching or not:

$$x_e = \begin{cases} 0 & \text{if} \quad e \notin M \\ 1 & \text{if} \quad e \in M \end{cases}$$

A linear program is an optimization problem with a linear objective function, a set of linear constraints, and a set of non-negativity restrictions imposed upon the underlying restriction variables. The maximum-weight matching can be formulated as zero-one *integer linear program*.

$$\text{Maximize} \sum_{e \in E} w_e x_e$$

$$\text{subject to} \sum_{e \in \delta(u)} x_e \leq 1 \qquad \text{for all } u \in V$$

$$x_e \in \{0, 1\} \qquad \text{for all } e \in E$$

To obtain a non-integer linear program, the second restriction is relaxed, which yields following linear programing formulation:

linear programing relaxation:

$$\text{Maximize} \sum_{e \in E} w_e x_e$$

$$\text{subject to} \sum_{e \in \delta(u)} x_e \leq 1 \qquad \text{for all } u \in V$$

$$x_e \geq 0 \qquad \text{for all } e \in E$$

The linear program as stated above does not have zero-one solutions only. Consider, for example, the graph shown in figure 21. An optimal solution of the integer linear program is, for example, $x_{1-2} = 1$, $x_{3-4} = 1$ and $x_e = 0$ for the other edges, which yields the value $\sum_{e \in E} w_e x_e = 5$ for the objective function. In contrast the optimal solution for the linear programing relaxation is $x_e = 0.5$ for all edges of the given graph, yielding $\sum_{e \in E} w_e x_e = 6$ for the objective function.

Figure 21: Example graph whose optimal solution of the linear programing relaxation has non-integer values for the incidence vector x_e.

The problem is connected to odd cycles (blossoms). Therefore constraints are added that prevent non-integer solutions from being feasible.

Let $\mathcal{O}$ denote the set of all non-singleton odd cardinality subsets of V:

$O = \{B \subseteq V : |B| \text{ is odd and } |B| \geq 3\}$

linear program:
Maximize $\sum_{e \in E} w_e x_e$

$$\text{subject to} \quad \sum_{e \in \delta(u)} x_e \leq 1 \qquad \text{for all } u \in V$$

$$\sum_{e \in \gamma(B)} x_e \leq \lfloor |B|/2 \rfloor \qquad \text{for all } B \in O$$

$$x_e \geq 0 \qquad \text{for all } e \in E$$

Lemma [33] The general maximum-weight matching problem is equivalent to the linear program stated above.

The next step is the development of a primal-dual problem that computes an optimal solution to the linear programming formulation.

Duality theory defines for each linear program (the so-called *primal problem*) a closely related associated linear programming problem , called the *dual problem*. Furthermore a set of *complementary slackness conditions* can be defined.

Primal:

Maximize $\sum_{uv \in E} w_{uv} x_{uv}$

$$\text{subject to} \quad \sum_{uv \in \delta(u)} x_{uv} \leq 1 \qquad \text{for all } u \in V \qquad (P1)$$

$$\sum_{uv \in \gamma(B)} x_{uv} \leq \lfloor |B|/2 \rfloor \qquad \text{for all } B \in O \qquad (P2)$$

$$x_{uv} \geq 0 \qquad \text{for all } uv \in E \qquad (P3)$$

Dual:

Minimize $\sum_{u \in V} y_u + \sum_{B \in O} \lfloor |B|/2 \rfloor z_B$

$$\text{subject to} \quad y_u \geq 0 \qquad \text{for all } u \in V \qquad (D1)$$

$$z_B \geq 0 \qquad \text{for all } B \in O \qquad (D2)$$

$$y_u + y_v + \sum_{\substack{B \in O \\ uv \in \gamma(B)}} z_B \geq w_{uv} \qquad \text{for all } uv \in E \qquad (D3)$$

Complementary Slackness Conditions:

$$x_{uv} > 0 \Longrightarrow \quad \pi_{uv} = 0 \qquad \text{for all } uv \in E \qquad (CS1)$$

$$y_u > 0 \Longrightarrow \quad \sum_{uv \in \delta(u)} x_{uv} = 1 \qquad \text{for all } u \in V \qquad (CS2)$$

$$z_B > 0 \Longrightarrow \quad \sum_{uv \in \gamma(B)} x_{uv} = \lfloor |B|/2 \rfloor \quad \text{for all } B \in O \qquad (CS3)$$

Reduced Cost: $\pi_{uv} = y_u + y_v - w_{uv} + \sum_{\substack{B \in O \\ uv \in \gamma(B)}} z_B$

Lemma (Theorem C.5 in [2]) Given a feasible solution x of the primal problem and a feasible solution (y, z) of the dual problem, these solutions are optimal if the complementary slackness conditions hold.

Finding Initial Solutions We can start with an empty matching, which clearly is a feasible solution to the primal problem, since $x_e = 0$ for each edge $e \in E$.
The potential of each vertex u is set to $y_u = max\{w_e/2 : e \in \delta(u)\}$, and $z_B = 0$ for each $B \in O$. This constitutes a feasible solution to the dual problem. Furthermore this initial solution fulfills the complementary slackness conditions CS(1) and CS(3). So the only condition which is violated is CS(2). The algorithm will alter the solutions x and (y, z) such that the violations of CS(2) will be reduced while maintaining the other constraints.

Reducing Violations of (CS2)

Let r be a vertex that violates (CS2), i.e. r is unmatched and $y_r > 0$. To fulfill the complementary slackness condition (CS2), either r has to be matched or the dual solution (y, z) has to be adjusted such that $y = 0$.

Matching a free vertex r As in the unweighted case, matching a vertex can be obtained by searching for an augmenting path. An alternating tree is grown as described in the proceeding section. The search algorithm guarantees that the augmentation of an augmenting path found is still a feasible

solution to the primal problem. And the dual solution is not altered by this procedure. However, only tight edges may be used in order to maintain (CS1). All details of the search algorithm apply, except in the case when no further tight edges are incident to any vertex $u^+ \in T$, a dual adjustment (as described below) is initiated instead of abandon tree T . Blossoms with $z > 0$ retain their identity.

Requirements to the Dual Adjustment The dual solution (y, z) gets adjusted to (y', z') such that

R1: the objective value $(\sum_{u \in V} y_u + \sum_{B \in O} \lfloor |B|/2 \rfloor z_B)$ of the dual problem strictly decreases

R2: the solution x is still feasible to the Primal

R3: the solutions (y, z) are still feasible to the Dual

R4: (CS1) and (CS3) remain true for (y', z')

R5: y_r strictly decreases

(R1) ensures that the dual solution converges with its optimum. After a series of dual adjustments either enough tight edges exist, so that r can be matched, or y_r will go down to zero (due to (R8)).

Performing a Dual Adjustment Whenever the search for an augmenting path fails, because no further tight edges exist incident to a vertex $u^+ \in T$, the dual solution (y, z) is altered. The new dual solution (y', z') can be obtained by the following rules for $\delta > 0$. The value of δ arises from the requirements stated above.

$$
\begin{aligned}
y'_v &= y_v - \delta && \text{for all } v^+ \in T,\\
y'_v &= y_v + \delta && \text{for all } v^- \in T,\\
y'_v &= y_v && \text{for all } v^{\{*|+\}} \notin T,\\
z'_B &= z_B + 2\delta && \text{for all } B^+ \in T,\\
z'_B &= z_B - 2\delta && \text{for all } B^- \in T,\\
z'_B &= z_B && \text{for all } B^{\{*|+\}} \notin T,
\end{aligned}
$$

z_B is only adjusted when B is a non-trivial surface blossom. The adjustment of y_u applies for vertices contained in blossoms as well. All vertices of a blossom have the same label (even or odd) as the blossom itself. Note that only blossoms (vertices) contained in the alternating tree T are affected.

Let us proof that requirement (R1) holds. Let $f = \sum_{u \in V} y_u + \sum_{B \in O} \lfloor |B|/2 \rfloor z_B$ denote the objective value before, and f' the objective value after the dual adjustment, and $\Delta f = f' - f$. The contribution of a trivial surface blossom of T is $\Delta f_{v^+} = -\delta$ for $v^+ \in T$ and $\Delta f_{v^-} = \delta$ for $v^- \in T$. For an even labeled non-trivial surface blossom of T, Δf is given by

$$\Delta f_{B^+} = |B|(-\delta) + \lfloor |B|/2 \rfloor (2\delta) = -|B|\delta + (|B| - 1)\delta = -\delta$$

and for $B^- \in T$,

$$\Delta f_{B^-} = |B|\delta + \lfloor |B|/2 \rfloor (-2\delta) = |B|\delta - (|B| - 1)\delta = \delta$$

Since each (trivial or non-trivial) surface blossom of the tree, except the root r, is matched with an odd surface blossom, the number of even blossoms n^+ of an alternating tree exceeds the number of odd blossoms n^- by one, $n^+ = n^- + 1$. Therefore $\Delta f = n^+(-\delta) + n^-(\delta) = -\delta$, and since δ is positive, the objective value of the dual problem decreases.

Since x is not altered, the new solution is still feasible to the primal problem and (R2) holds.

Maintaining (R3) (the solution is still feasible to the dual problem) brings about restrictions on the value of δ. Since y_u and $z_{\mathcal{B}}$ have to be positive due to the non-negativity restrictions (D1) and (D2) of the dual, y_u of even vertices and $z_{\mathcal{B}}/2$ of odd blossoms are upper bounds on the value of δ:

$$\delta \leq y_u \text{ for all } u^+ \in T$$
$$\delta \leq z_B/2 \text{ for all } B^- \in T$$

The linear constraint $(D3)$ on the variables (y, z) stated in the dual problem (the reduced cost of an edge must be non-negative) requires closer inspection.

$$\pi_{uv} = y_u + y_v - w_{uv} + \sum_{\substack{B \in O \\ uv \in \gamma(B)}} z_B \geq 0$$

The reduced cost of an edge uv is affected by a dual adjustment only when at least one endpoint of the edge lies in T. So we have to take a closer look at all possible combinations of vertices forming such an edge.
Let denote π, y, z and π', y', z' the variables before and after the dual adjustment, respectively.

Case 1: $uv \notin \gamma(B)$, therefore the reduced cost of edge uv is given by
$\pi_{uv} = y_u + y_v - w_{uv} \geq 0 \; e \notin \gamma(B)$

Case 1a: $u^+ \in T$ and $v^- \in T$:
$y'_u = y_u - \delta, \quad y'_v = y_v + \delta$
$\pi'_{uv} = \pi_{uv}$

Case 1b: $u^+ \in T$ and $v^+ \in T$:
$y'_u = y_u - \delta, \quad y'_v = y_v - \delta$
$\pi'_{uv} = \pi_{uv} - 2\delta$, this restricts δ to $\delta \leq \pi_{uv}/2$

Case 1c: $u^+ \in T$ and $v^{\{*|+\}} \notin T$:
$y'_u = y_u - \delta, \quad y'_v = y_v$
$\pi'_{uv} = \pi_{uv} - \delta$, this restricts δ to $\delta \leq \pi_{uv}$

Case 1d: $u^- \in T$ and $v^- \in T$:

$y'_u = y_u + \delta, \quad y'_v = y_v + \delta$

$\pi'_{uv} = \pi_{uv} + 2\delta$

Case 1e: $u^- \in T$ and $v^{\{*|+\}} \notin T$:

$y'_u = y_u + \delta, \quad y'_v = y_v$

$\pi'_{uv} = \pi_{uv} + \delta$

Case 2: $e = uv \in \gamma(B)$

Case 2a: $B^+ \in T$:

$y'_u = y_u - \delta, \quad y'_v = y_v - \delta, \quad z'_B = z_B + 2\delta$

$\pi'_{uv} = \pi_{uv}$

Case 2b: $B^- \in T$:

$y'_u = y_u + \delta, \quad y'_v = y_v + \delta, \quad z'_B = z_B - 2\delta$

$\pi'_{uv} = \pi_{uv}$

(R4) demands that the complementary slackness conditions (CS1) and (CS3) remain true for (y', z'). Since x_{uv} is not altered and $\pi'_{uv} = \pi_{uv}$ for all matched edges, (CS1) ($x_{uv} > 0 \Rightarrow \pi_{uv} = 0$) stays true during a dual adjustment. The validity of (CS3) remains, since x_{uv} is not altered and z_B is changed only for non-trivial surface blossoms, which are by definition full, i.e. $\sum_{uv \in \gamma(B)} x_{uv} = \lfloor |B|/2 \rfloor$.

The root r is an even labeled vertex and in T, therefore y_r strictly decreases, and (R5) is fulfilled.

All together we obtain the value of δ as
$\delta = min\{\delta_1, \delta_2, \delta_3, \delta_4, \}$, where
$\delta_1 = \min_{u \in V} \{y_u : u^+ \in T\}$

$$\delta_2 = \min_{uv \in E} \{\pi_{uv} : u^+ \in T,\ v^{\{*|+\}} \notin T\}$$

$$\delta_3 = \min_{uv \in E} \{\pi_{uv}/2 : u^+ \in T,\ v^+ \in T\}$$

$$\delta_4 = \min_{B \in O} \{z_B/2 : B^- \in T\}$$

Whichever of the bounds $\delta_1, \ldots, \delta_4$ becomes the effective bound on δ has different consequences on the choice of the step that follows a dual adjustment. The vertex, edge or blossom that is responsible for one of the bounds is called the *responsible* vertex, edge or blossom.

In the case $\delta = \delta_1$, the potential of the responsible vertex u^+ becomes zero. Therefore this vertex may be unmatched without violating the complementary slackness condition (CS2). Since the responsible vertex is even, an even length alternating path from u to r exists and r can be matched by replacing M by $M \oplus p$.

If $\delta = \delta_2$ the responsible edge uv becomes tight and can be used to extend T.

When $\delta = \delta_3$ the responsible edge uv becomes tight and can be used to shrink a new blossom.

Finally, in case $\delta = \delta_4$, $z_{\mathcal{B}}$ of $\mathcal{B}$ will drop to zero, $\mathcal{B}$ cannot participate in another dual adjustment. Then blossom $\mathcal{B}$ is expanded, which in turn leads to another dual adjustment.

The expand step for an odd Blossom is similar to the expand step in the unweighted matching algorithm. $\mathcal{B}$ lies in the alternating path, therefor a matched edge and an unmatched edge are incident to $\mathcal{B}$. Let v denote the vertex of $\mathcal{B}$ that is incident to the matched edge, and w the vertex of $\mathcal{B}$ incident to the unmatched edge. $\mathcal{B}$ is expanded by incorporating the even length alternating path from v to w to the tree and deleting the vertices of the blossom from the tree, that are not part of the even length path from

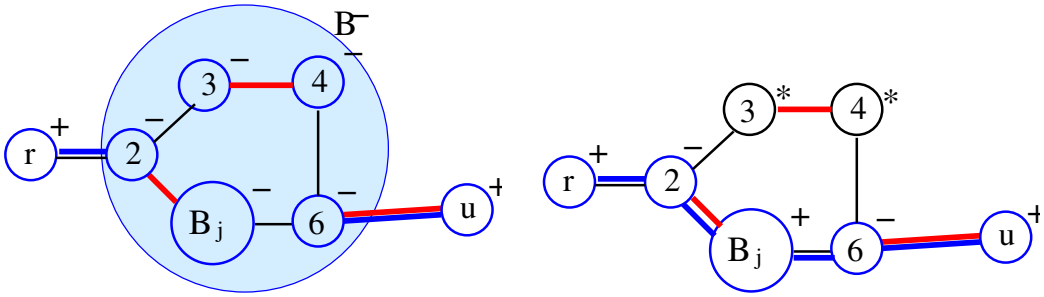


Figure 22: Expand step for an odd blossom: The even length alternating path from $v = 6$ to $w = 2$ is added to the tree, vertices 3 and 4 are deleted from the tree.

v to w (compare Fig 22). The vertices of the path from v to w are labeled accordingly.

The algorithm for finding the maximum-weight matching in a general graph is summarized in table 3, and Figure 23 shows an example.

Table 3: Algorithm for the maximum weighted matching

Maximum Weighted Matching

Initial Solution:

let M be the empty matching

$y_u = max\{w_e/2 : e \in E\}$ for each vertex $u \in G$

label each vertex $u \in G$ even

for each vertex $r \in G$ {

if r is matched or $y_r = 0$ continue

let B_r be the only blossom of T

repeat {

if an vertex $u^+ \in T$ with $y_u = 0$ exists {

let p be the alternating path from u to r

replace M by $M \oplus p$

}

else if an edge uv with $u^+ \in T$ and $\pi_{uv} = 0$ exists {

case $v^* \notin T$: grow step

case $v^+ \in T$: shrink step

case $v^+ \notin T$: augment step

}

else if there exists an odd blossom $B^- \in T$ with $z_B = 0$ {

expand step for B

}

else {

determine δ

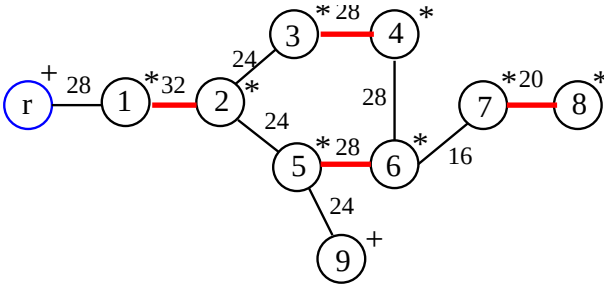
perform dual adjustment

}

} until r is matched or $y_r = 0$

}

node u	1	2	3	4	5	6	7	8	9	r
y_u	16	16	14	14	14	14	10	10	12	14



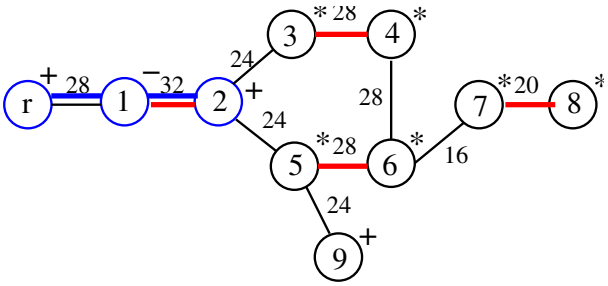
starting the search with $r^+ \in T$
$\pi(r^+, 1^*) = 2$
no tight edges are incident to $r^+ \Rightarrow$ *dual adjustment*

$\delta_1 = 14$, $\delta_2 = 2$, $\delta_3 = \infty$, $\delta_4 = \infty$
$\delta = 2$

Dual adjustment:

node u	1	2	3	4	5	6	7	8	9	r
y_u	16	16	14	14	14	14	10	10	12	14
y'_u	16	16	14	14	14	14	10	10	12	12

edge $(r, 1)$ is tight $\Rightarrow$ *grow step*



$\pi(2^+, 3^*) = 6$, $\pi(2^+, 5^*) = 6$
no tight edges are incident to $2^+ \Rightarrow$ *dual adjustment*

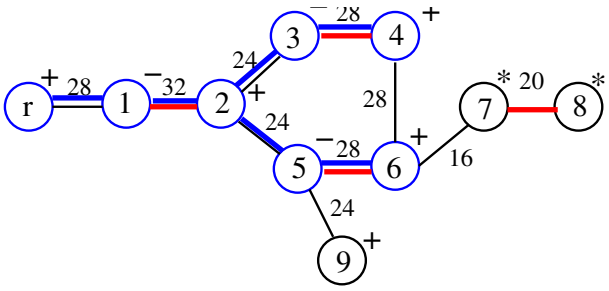
$\delta_1 = 12$, $\delta_2 = 6$, $\delta_3 = \infty$, $\delta_4 = \infty$
$\delta = 6$

Dual adjustment:

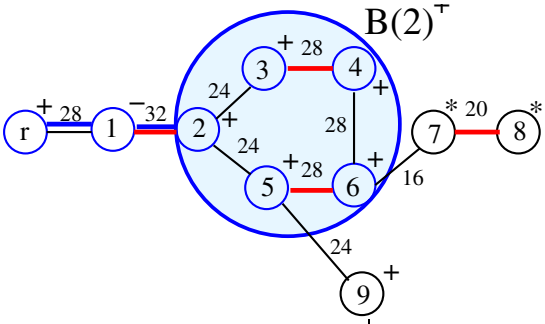
node u	1	2	3	4	5	6	7	8	9	r
y_u	16	16	14	14	14	14	10	10	12	12
y'_u	22	10	14	14	14	14	10	10	12	6

edge $(2^+, 3^*)$ is tight $\Rightarrow$ *grow step*, edge $(2^+, 5^*)$ is tight $\Rightarrow$ *grow step*

Figure 23: Search for an augmenting path on a weighted graph.The search starts with a 4-matching on the graph. Matched edges are colored red, unmatched edges black, and the alternating tree is colored blue.



edge $(4^+, 6^+)$ is tight $\Rightarrow$ *shrink step*



$\pi(5^+, 9^+) = 2$, $\pi(6^+, 7^*) = 8$

no tight edges are incident to $(B(2))^+$ $\Rightarrow$ *dual adjustment*

$\delta_1 = 6$, $\delta_2 = 2$, $\delta_3 = \infty$, $\delta_4 = \infty$

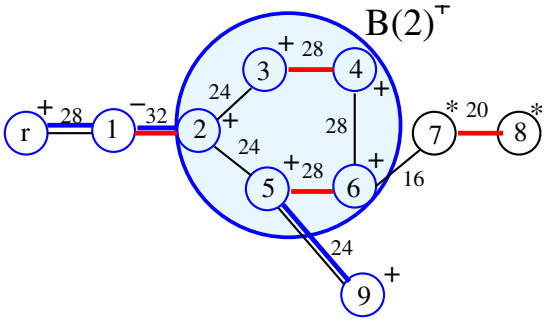
$\delta = 2$

Dual adjustment:

node u	1	2	3	4	5	6	7	8	9	r
y_u	22	10	14	14	14	14	10	10	12	6
y'_u	24	8	12	12	12	12	10	10	12	4

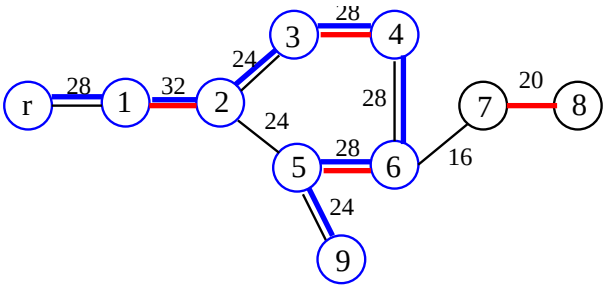
$z_B(B(2)) = 4$

edge $(5^+, 9^+)$ is tight, $9^+ \notin T$ $\Rightarrow$ *augmenting path found*

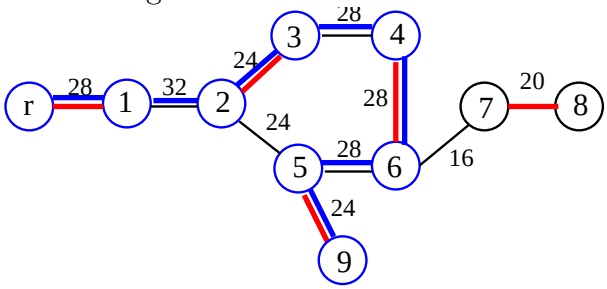


$\Rightarrow$ *expand* the blossom

Figure 23: Search for an augmenting path on a weighted graph (continued).



$\Rightarrow$ *augment* the matching



Maximum weighted matching on the graph.

Figure 23: Search for an augmenting path on a weighted graph (continued).

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: Folding 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. 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 determination is the *mutual information* (MI) score [22, 55, 54]. The MI score of column i and j of the alignment is then given by

$$M_{ij} = \sum_{\mathsf{X},\mathsf{Y}} f_{ij}(\mathsf{XY}) \log \frac{f_{ij}(\mathsf{XY})}{f_i(\mathsf{X}) f_j(\mathsf{Y})} \tag{9}$$

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 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 2.3.5, and assumes that the energy of a structure can be obtained as the sum over the energies of its constituent loops.

$$E(\mathcal{S}) = \sum_{l \in \mathcal{S}} E(l) \tag{10}$$

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 recently published [104] 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 [8], and recently modified by including a penalty for A·U and U·A pairs at the

end of helices [177]. Stacking energies involving G·U pairs were added later [63] and demonstrated the shortcoming of the nearest neighbor model: The energy of the double G·U mismatch $^{5'\mathsf{GU}\,3'}_{3'\mathsf{UG}\,5'}$ depends on its context. It is energetically favorable e.g. in the context $^{5'\mathsf{GGUC}\,3'}_{3'\mathsf{CUGG}\,5'}$, but more often unfavorable, as in $^{5'\mathsf{CGUG}\,3'}_{3'\mathsf{GUGC}\,5'}$ or $^{5'\mathsf{UGUA}\,3'}_{3'\mathsf{AUGU}\,5'}$. 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 tetraloops (hairpins of size four) occur much more frequently than expected in known RNA secondary structures, such as ribosomal RNA [173]. The current parameter set lists 30 such special tetraloops and awards them bonus energies between -1.5 and -3 kcal/mol. Finally, the Turner parameters recommend a special penalty for poly-C loops [50], and a bonus of -2.2 kcal/mol for loops closed by G·U when the two bases preceding the G are also Gs [45].

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.

$$\Delta G_{\text{int.loop}} = \Delta G_{\text{size}}(n_1 + n_2) + \Delta G_{\text{asym}}|n1 - n2| + \Delta G_{\text{mismatch}}. \qquad (11)$$

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

$$\Delta G_{\text{ML}} = a + b \cdot n + c \cdot k + \Delta G_{\text{dangle}}, \tag{12}$$

where n is the loop size and k the loop degree. ΔG_{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 The Algorithm

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 [168, 169]. His observation was the starting point for the construction of reliable energy-directed folding algorithms [71, 183].

The first dynamic programming solution was proposed by Ruth Nussinov [119, 120] originally for the "maximum matching" problem of finding the struc-

ture with the maximum number of base pairs. Michael Zuker and Patrick Stiegler [183, 184] formulated the algorithm for the minimum energy problem using the now standard energy model. Since then several variations have been developed: Michael Zuker [182] 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 [106] 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 [70, 75].

`RNAfold` as part of the `Vienna RNA Package`[1] [71] reads RNA sequences from `stdin` and calculates their mfe structure, partition function and base pairing probability matrix [68, 106]. It returns the mfe structure in bracket notation, its energy, the free energy of the thermodynamic ensemble and the frequency of the mfe structure in the ensemble to `stdout`. It also produces `PostScript` output files with plots of the resulting secondary structure graph and a dot plot of the base pairing matrix. The dot plot shows a matrix of squares with area proportional to the pairing probability in the upper half, and one square for each pair in the mfe structure in the lower half, see Figure 7. The results of `RNAfold` are used as an input for `alidot` [70, 71].

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

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 pseudoknots based on the same model has been proven to be NP-complete [100, 3]. However, for structure predictions including certain types of pseudoknots polynomial algorithms have been developed. Furthermore a number of algorithms which adopt heuristic search procedures exist.

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 [176, 160]. Since there are no measured thermodynamic parameters for pseudoknots, we rely on approximations for the free energy of pseudoknots.

Gultyaev *et al.* [52] 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 [81] the free energy ΔG of formation of a loop of N nucleotides is approximated by

$$\Delta G = RT(A_{loop} + 1.75\ lnN), \tag{13}$$

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 stereochemically, loop $L1$ spans the deep groove of RNA stem $S2$, whereas $L2$ crosses stem $S1$ in the shallow groove [133], see Figure 9. 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 loopsizes 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:

$$\Delta G_{L1} = A_{deep}(S2) + 1.75RT\ ln(1 + N - N_{mindeep}(S2)) \quad (14)$$

$$\Delta G_{L2} = A_{shallow}(S1) + 1.75RT\ ln(1 + N - N_{minshallow}(S1)) \quad (15)$$

Sequences of known pseudoknots that are evidenced 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 [77, 76]. The model is based on the Isambert-Siggia decomposition of secondary structures (Section 2.4.2), restricted to nets with a maximum order of 2. They distinguish between closed nets, which match our definition of nets (definition 17) and open nets. 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.

The free energy of a net is composed of the free energy of the stems, calculated using the thermodynamic parameters for base stacking [150], and the entropy of the net which is calculated using polymer theory [37]. The stems

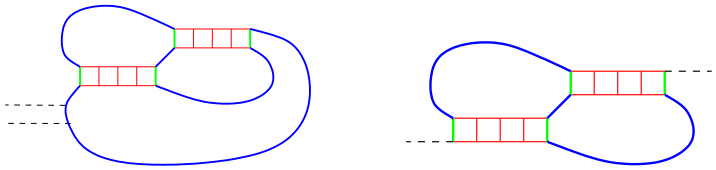


Figure 23: Closed and open nets: Example of a closed 2-net (*l.h.s.*) and an open 2-net (*r.h.s.*)

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 [139]. 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 [138], and the specific gram-

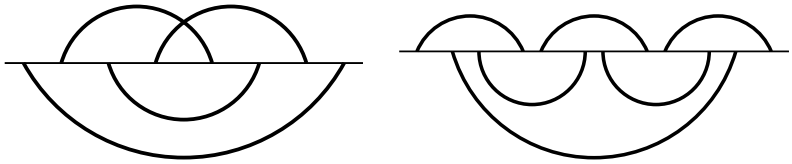


Figure 24: Structures exemplifying the class of structures the algorithm of Rivas and Eddy [139] minimizes over, helices are drawn as arcs. A non-planar structure in the class of structures minimized over (*l.h.s.*), and a planar structure not in that class (*r.h.s.*).

mar that corresponds to the parsing algorithm for structure prediction by 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 24.

A dynamic programming algorithm, which achieves $\mathcal{O}(n^4)$ time and $\mathcal{O}(n^2)$ memory has been presented by Reeder and Giegerich [137]. 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 [139]. 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 [61] 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 [52] (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 [53, 163]. A kinetic Monte Carlo algorithm has been presented by Isambert and Siggia [77].

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 [104, 103]. 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 [146] 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.* [48] 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.

Another algorithm based on the recursion of Sankoff has been given by Mathews and Turner [105], 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.* [128] recently presented an algorithm for pairwise folding of

unaligned sequences which has an empirically observed complexity of about $\mathcal{O}(n^2)$. The first step is generating a list of possible stems, using stacking energies and tetraloop bonuses [104], 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.* [118] developed a genetic algorithm that finds the structure of a sequence given a second, related sequence with known structure. Chen *et al.* [20] 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.

3.4.2 Algorithms Based on a Multiple Sequence Alignment

Most of the alignment base 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 [93] 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 [72] is based on the base pairing probabilities calculated by RNAfold [68] 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.* [98] 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.* [69] have developed an algorithm 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 [83] 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. [156] 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 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` [182].

3.5 The Algorithm `Alidot`

`Alidot` (ALIgned DOT-plots) [70, 72, 68], has been used for the prediction of conserved secondary structure of the genomes of picornaviruses (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` [98, 97] and `x2s` [83], see also [93]. In contrast to efforts to simultaneously compute alignment and secondary structures e.g. [146, 48, 128, 105] this approach emphasizes that the sequences may have common structural motifs but no single common structure. In this sense `alidot` combines structure prediction and motif search [27].

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 [106] 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` [71], based on the energy parameters published in [104].

The multiple sequence alignments can be obtained, for example, using `ClustalW` [161] or `code2aln` [154]. 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 proba-

bility 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 Figure 26.

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 following two conditions: (i) No nucleotide takes part in more than one base pair. (ii) Base pairs never cross, that is, there may not be two base pairs $(i.j)$ and $(k.l)$ such that $i < k < j < l$. 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 number of different pairing combinations.

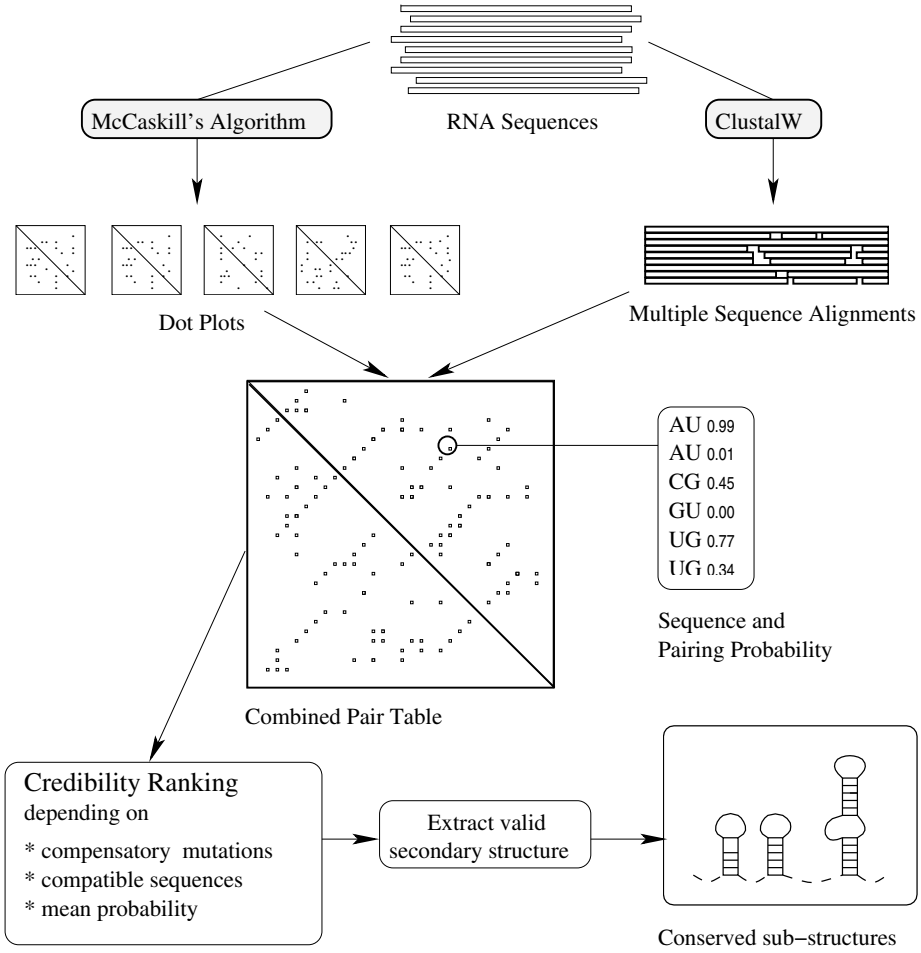


Figure 25: Flow diagram of the algorithm. A multiple sequence alignment is calculated using, for instance, `ClustalW`. RNA genomes are folded using McCaskill's partition function algorithm as implemented in `RNAfold`. The alignment and the structure predictions are joined together to the combined pair 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.

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 Figure 25.

Results are presented as conventional secondary structure drawing, as *colored* mountain plots, see Fig. 26, 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, Figure 26 shows an example of an annotated structure drawing. Compensatory mutations thus are shown by circles around both pairing partners. Inconsistent mutants are indicated by gray instead of black lettering.

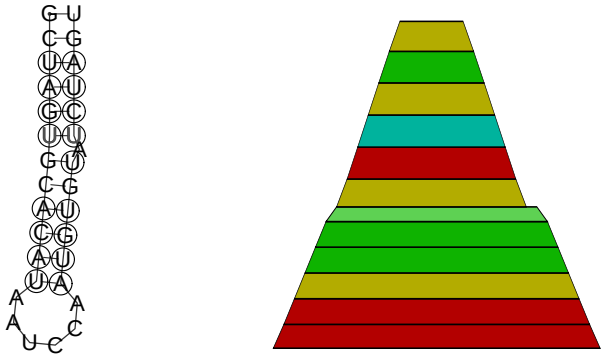


Figure 26: Annotated structure drawing (left) and colored mountain plot(right). The example shows a conserved secondary structure element located in the 5'-non-coding region of the rhinovirus genome. Colors indicate the number of consistent mutations ■ 1, ■ 2, ■ 3, ■ 4 different types of base pairs. 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.

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` [35]. 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 Structural Elements of Picornaviruses

4.1 About Picornaviruses

The members of the family *Picornaviridae* are among the smallest mammalian ribonucleic acid-containing viruses known. The virus particles are non-enveloped and the capsid, which has an icosahedral symmetry, is made up of four structural proteins. The protein shell encloses a single copy of the single-stranded positive-sense genomic RNA (Fig. 28). The family is currently divided into nine genera [85].

The genus *Aphthovirus* contains two species, *Foot-and-mouth disease virus* (FMDV) and *Equine rhinitis A virus* (ERAV). The first animal virus described was FMDV virus by F.Löffler and P.Frosch in 1898 [96]. Foot-and-mouth disease is a highly contagious disease of cloven hoofed animals, ERAV causes a mild respiratory infection in horses [143].

Members of the genus *Cardiovirus* are isolated mostly from rodents, they cause encephalitis and myocarditis (encephalomyocarditis - inflammation affecting the heart and the brain). The genus currently comprises of two virus species: *Encephalomyocarditis virus* (EMCV) and *Theilovirus* (ThV) [124].

The genus *Enterovirus* is divided into eight species: *Poliovirus (PV)*, *Human enterovirus A* (HEV-A), *Human enterovirus B* (HEV-B) , *Human enterovirus C* (HEV-C), *Human enterovirus D* (HEV-D), *Bovine enterovirus* (BEV), *Porcine enterovirus A* (PEV-A) and *Porcine enterovirus B* (PEV-B). Enteroviruses are so called because the most inhabit the alimentary (enteric) tract. Enterovirus infection can cause a wide spectrum of clinical symptoms, the most common forms of infection are asymptomatic or mild. Among the serious outcomes of enterovirus infection is paralytic poliomyelitis caused by polioviruses. HEV-A are the cause of hand, foot and mouth disease, HEV-B

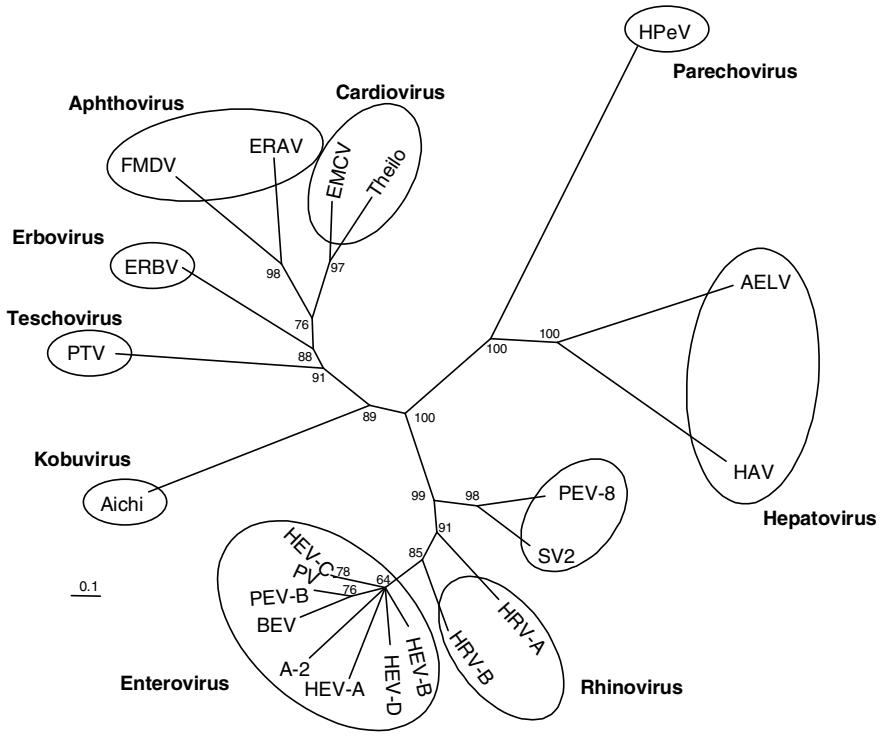


Figure 27: Phylogenetic relationship between picornaviruses based on the sequence of protein 3D [90].

are associated with aseptic meningitis. Other clinical symptoms effected by enterovirus infection are encephalitis, pneumonia, paralysis and carditis [108].

The genus *Erbovirus* contains only one species, the *Equine rhinitis B virus* (ERBV), it causes mild respiratory disease in horses which resembles the common cold in men [175].

The genus *Hepatovirus* contains the species *Hepatitis A virus* (HAV) and *Avian encephalomyelitis-like viruses* (AEV). HAV is the most common cause

of acute viral hepatitis - probably something like half of all cases are due to this virus. Avian encephalomyelitis is a disease of young chickens, pheasants, quail and turkeys. AEV can cause slight reduction in egg production and can be transmitted in embryos, which results in reduced hatching resp. tremors and/or ataxia of the chicks [102].

The genus *Kobuvirus* contains one species, the *Aichi virus* (AIV). AIV was first isolated in 1989 from a stool specimen of a patient with oyster associated non-bacterial gastroenteritis. This and other findings of the virus strongly suggest that AIV is one of the causative agents of acute gastroenteritis in humans [178].

The genus *Parechovirus* comprehends the species *Human parechovirus* (HPEV) and *Ljungan virus* (LV). HPEV disease is similar to that caused by enteroviruses and may include aseptic meningitis, gastroenteritis, encephalitis and neonatal sepsis-like disease [121]. LV is a suspected human pathogen recently isolated of bank voles [114].

The genus *Rhinovirus* is divided into *Human rhinovirus A* (HRV-A) and *Human rhinovirus B* (HRV-B). Rhinoviruses inhabit the respiratory tract, they are a cause of the common cold [25].

The genus *Teschovirus* comprises of one species, the *Porcine teschovirus* (PTV). The PTV are described as causative agents of severe and mild neurological disorders, fertility disorders and dermal lesions of swine [179].

4.1.1 The Virion

The virions of picornaviruses are roughly spherical, with no lipid envelope, their diameters ranging from 24 to 30 nm. The virion contains an RNA core, tightly packed in the central cavity of a thin protein shell. The capsid consists of four structural proteins in an icosahedral arrangement of 60 protomers. Each protomer is made up of the four proteins (see Fig. 28) [136, 141, 5, 99].

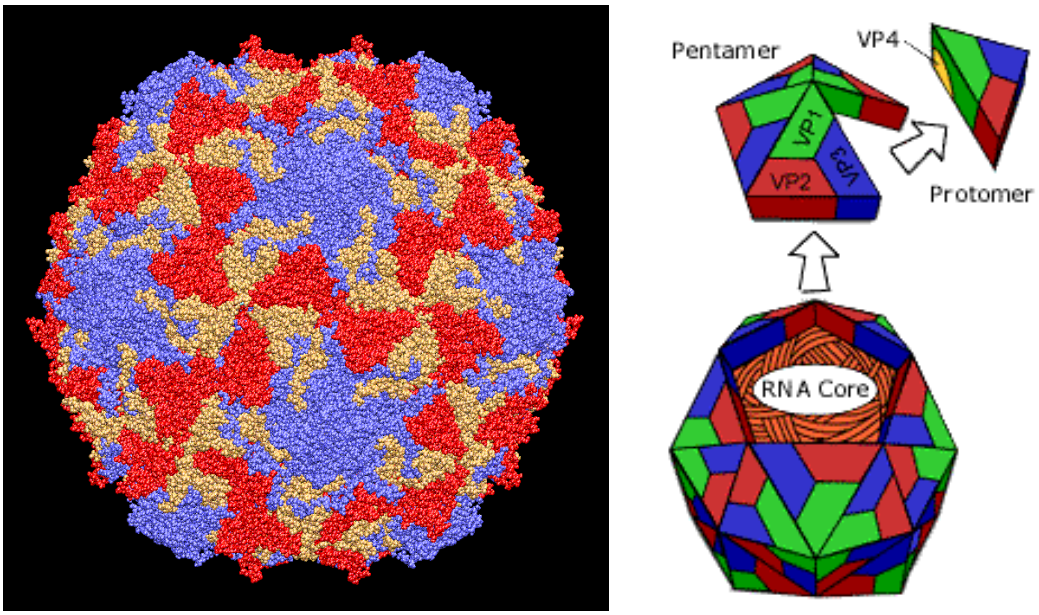


Figure 28: Left: Molecular surface of Poliovirus, as solved by X-ray crystallography [134]. Right: Schematic presentation of the virus capsid [135].

4.1.2 The Genome

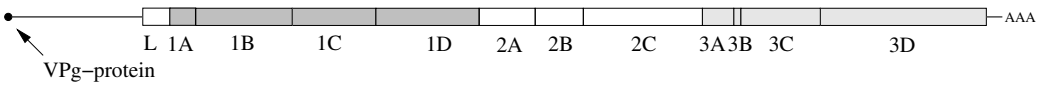


Figure 29: Genomic structure of the picornavirus genome: The VPg-protein is covalently attached to the 5' end of the RNA. The protein-coding region is indicated by the rectangle. Proteins: leader protein L (only present in aphtho-, cardio- and teschovirus), capsid proteins 1A-1D, viral protease 2A, proteins involved in RNA synthesis 2B, 2C, unknown function 3A, VPg 3B, major viral protease 3C, RNA-dependent RNA-polymerase 3D; [144, 164, 143]

Although the members of the different genera share little sequence identity, they all have similar genomic structure and gene organization (Fig. 29). The genome consists of a single strand messenger-active (+) RNA of 7,200 to 8,500 nts that is polyadenylated at the 3' terminus and carries a small protein (virion protein, genome; VPg) covalently attached to its 5' end. The major part of the RNA consists of a large open reading frame (ORF) encoding a polyprotein. The 5' non-translated region (5'NTR) is unusually long and contains multiple AUG triplets prior to the initiator of the viral translation. Aphthoviruses, the species EMCV of the genus *Cardiovirus* and teschoviruses carry a poly-(C)-tract on the 5'NTR. All Picornaviruses have an internal ribosomal entry site (IRES) instead of a 5'cap structure [65, 117, 126, 82].

4.1.3 The Viral Life-Cycle

The first stage in picornavirus infection is attachment of the virion to specific receptors embedded in the cell membrane (Fig. 30, step 1), after conformational alteration of the protein shell the RNA is released to the cytoplasm (step 2). The positive-sense genomic RNA serves three functions: (i) It acts as messenger RNA from which the polyprotein is translated (step 4), which is cleaved co- and post-translationally. (ii) The genomic RNA serves as a template for minus-strand synthesis (step 3) and (iii) the newly synthesized

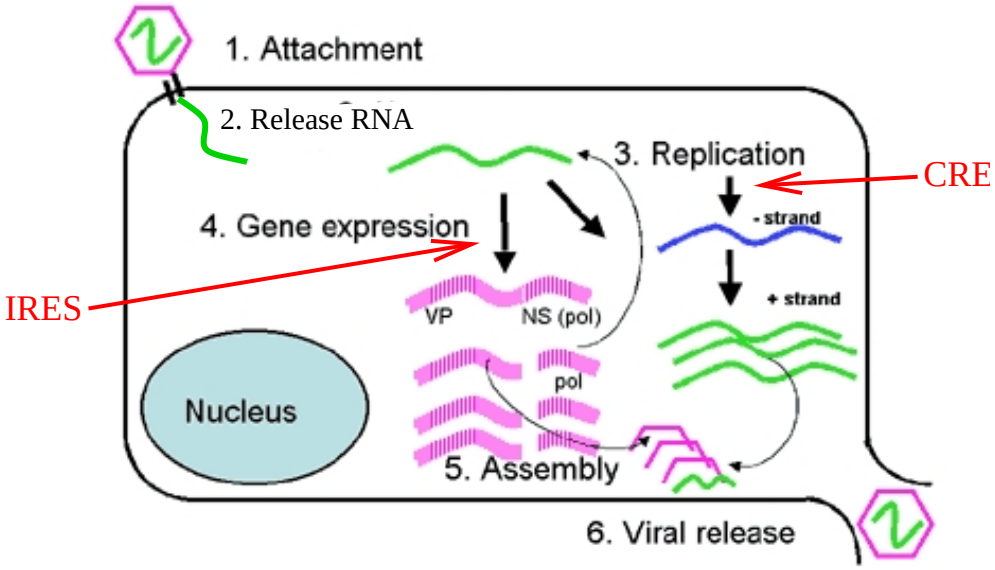


Figure 30: Overview of viral life-cycle. IRES: Internal Ribosome Entry Site, CRE: Cis-acting Replication Element (Image adapted from Y. Hahn [57])

plus-strand RNA, which is copied from the minus-strand, is packed into the capsids to form new virus particles (step 5).

Two conserved secondary structure elements of the viral RNA are known to be of functional importance. The IRES is essential for translation and a Cis-acting Replication Element (CRE) located in the coding region of the genome is required for synthesis of the minus strand RNA.

Translation Unlike most eukaryotic mRNAs the initiation of translation in picornaviruses is cap-independent. Instead of the m^7G cap structure picornavirus-RNA has the virion protein (VPg) covalently attached to the 5'end. The 5′-non-translated region shows highly conserved RNA secondary structure domains which form the IRES. The first IRES elements described were those present in picornavirus genomes [126, 82]. In addition, a number of IRES-containing eukaryotic mRNAs have been detected recently [64].

The picornavirus IRES elements are classified into three groups (Fig 32): type I IRES of enteroviruses and rhinoviruses, type II of aphthoviruses, cardioviruses and parechoviruses and type III of hepatoviruses, see e.g. [79, 175].

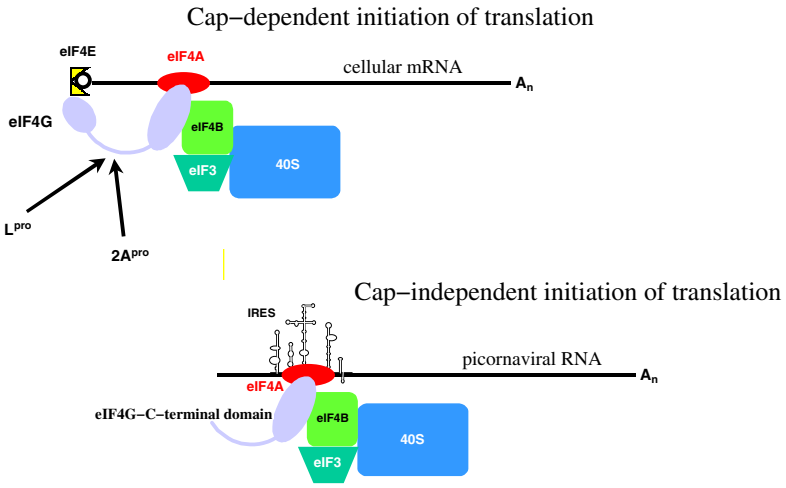


Figure 31: Comparison of cap-dependent and cap-independent initiation of translation. The eukaryotic initiation factor eIF4G is cleaved by the viral protease $2A^{pro}$ of entero- and rhinoviruses and by the leader protein L^{pro} of aphthoviruses (Image adapted from R. Zell [181])

The cap-dependent initiation of translation involves binding of the eukaryotic initiation factor (eIF) 4E to the cap structure. The N-terminal domain of protein eIF4G binds to eIF4E, while the C-terminal domain of eIF4G, together with eIF4A, eIF4B and eIF3, binds to the ribosomal 40S subunit

(see Fig. 31). This complex then scans along the RNA until the first AUG initiation codon is reached and the translation begins [86, 181].

By contrast, the cap-independent initiation of translation is mediated by the IRES. The eukaryotic translation initiation factors are required for the internal initiation of translation, with the exception of the actual cap-binding protein eIF4E [78]. The 5'end of the picornavirus RNA contains multiple AUG triplets prior the initiating AUG codon. The IRES enables binding of the ribosome downstream the 5'end, that way overriding the non-initiating AUG triplets. The synthesis of host proteins is shut down during entero-, rhino- and aphthovirus infection through cleavage of the eukaryotic initiation factor eIF4G. Cardioviruses do not cleave eIF4G, they have evolved an unusually efficient IRES and appear to simply outcompete host mRNAs for utilization of the translation machinery [143, 144].

The structural elements important for IRES function in entero- and rhinovirus sequences are elements II, IV, V and VI [144]. The binding sites for the initiation factors are not well known yet, there is one study showing that eukaryotic initiation factor eIF4B binds to domain V in poliovirus [123]. Experimental studies revealed that the structure elements H to M are essential for type II IRES function [144] and the Y-shaped J, K element is the binding site for eIF4B [110, 122] as well as for eIF4G and eIF4A [88, 145]. Structural elements IV and V of hepatoviruses have been shown to be necessary for cap-independent translation of HAV[10]. In contrast to type I and type II IRES the uncleaved eukaryotic initiation factor 4G, as well as the cap-binding protein eIF4E is required for HAV IRES activity [9].

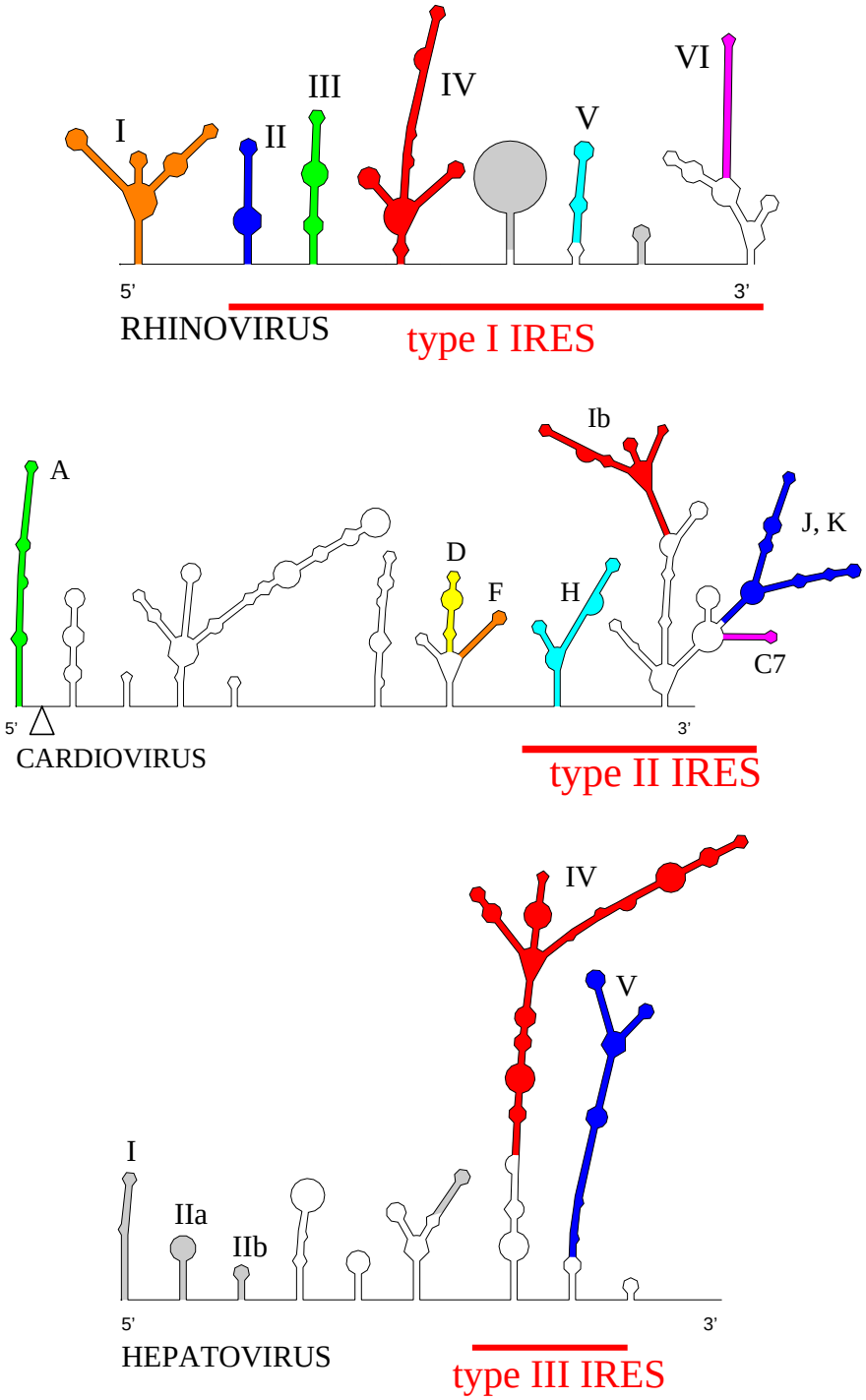


Figure 32: Conserved secondary structure of the 5′-NTR of different picornaviruses. Colored regions mark the conserved domains, the IRES region is underlined red. Top: type I IRES; Middle: type II IRES; Bottom: type III IRES

4.2 Methods and Tools

Conserved structural elements have been identified by using the algorithm `alidot` [70, 72] (see Section 3.5). Multiple sequence alignments were generated by `ClustalW` [161], the alignments are used without further modifications (except where stated explicitly). The thermodynamic structure predictions for each sequence were calculated using `RNAfold` [71] (see Section 3.2.2).

To obtain publication-quality secondary structure drawings, which facilitate structural homology comparisons, a program for schematic drawing of secondary structures has been developed. The program, termed `splot`, is described in the following section.

The distribution of biological information has become critically dependent on the Internet, enabling access to information in a platform independent manner. Therefore we implemented a `CGI` script, which allows access to our prediction results of conserved structural elements of RNA virus genomes. The resulting database, called Vienna Atlas of Viral RNA Structures, is presented in Section 4.2.2.

4.2.1 Schematic Drawing

Simple structure representations such as mountain plots, or dot plots are particularly useful for comparing structures. Nevertheless, classical drawings are often required. The automatic generation of RNA structure drawings is therefore an important task in large surveys of structural RNAs, because interactive layouting using editing tools such as `SStructview` [36] and `XRNA` [170] is very time-consuming. Programs such as `Naview` [14], `RNAplot` (part of the `Vienna RNA Package`), `RnaViz` [28], `RNA-d2` [127], `rnasearch` [112], or `VizQFolder` [58] are designed for drawing complete RNA secondary structures, where most of the nucleotides are paired. When drawing only the

conserved parts of a secondary structure, long unpaired regions may emerge, since there may be long range interactions spanning long unconserved regions. In this case standard drawing tools often produce layouts that do not clearly display the relevant structural information. Therefore we developed a drawing program, called `splot`, which generates clear and readable displays, and conserves structural similarities in different structures, thus greatly aiding structural homology comparisons.

The first step of our algorithm is the simplification of the secondary structure graph Γ to a graph Γ', which contains only the information which is essential for calculating the layout. The reduced graph Γ' consists of the gel $\mathsf{Gel}(\Gamma)$ (see Section 2.4.2) and the reduced exterior BG-subgraph. The reduced exterior BG-subgraph $\mathcal{R}$ consists of the first and the last vertex of the sequence, vertices derived by substitution of each green edge, edges which connect those vertices to the corresponding component of the gel, and edges replacing the connected path of blue edges between two green edges of the exterior BG-subgraph.

The layout is calculated for the reduced graph Γ'. The reduced exterior BG-subgraph $\mathcal{R}$ is drawn as a line, (at the beginning) with fixed distances between two vertices of $\mathcal{R}$. Since the loops of a secondary structure coincide with the nets (corollary 4), the terms are used interchangeably in the remainder. Starting from each vertex in $\mathcal{R}$ the corresponding components are drawn successively according to the following rules: All edges incident to a vertex in $\mathcal{R}$ are drawn at an angle of 90°. The length of each edge is proportional to the length of the corresponding stem. The angles between the edge corresponding to the closing of the loop and the other edges incident to this loop have a fixed value depending on the degree of the loop. The radius of all multiloops depends on the degree of the loop, the radius of loops of degree 1 or 2 is proportional to their size. Having calculated the coordinates of all components, those are shifted along the x-axis until there is no more overlap between the individual components. Finally the coordinates

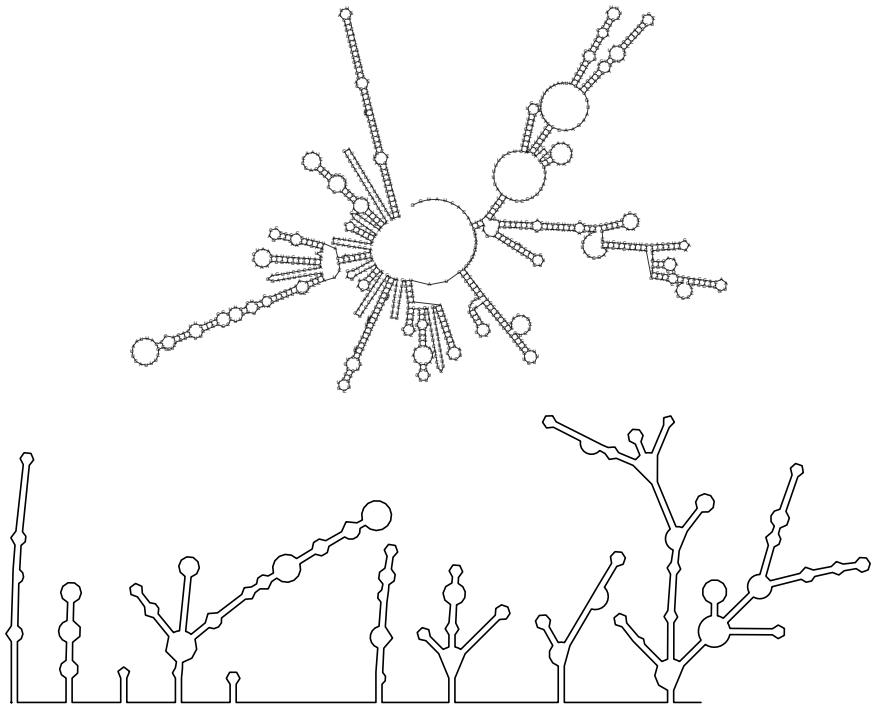


Figure 33: Conventional drawing of the 5'NTR of cardiovirus: Layout produced by `RNAplot` (*top*) and by `splot` (*bottom*).

of the original graph are calculated. The algorithm has been implemented in `ANSI C`, it takes a secondary structure graph in `GML` format [66, 67] as input and produces a `PostScript` file. A `Perl` script transforms the secondary structure given in bracket notation to `GML` format. The man-page of `splot` is given in A.4.

4.2.2 Vienna Atlas of Viral RNA Structures

The *Vienna Atlas of Viral RNA Structures* is a database containing conserved secondary structures of a growing number of RNA virus families. At present data of the family Flaviviridae [162], Hepadnaviridae [153] and Picornaviridae (this work, [172]) are included.

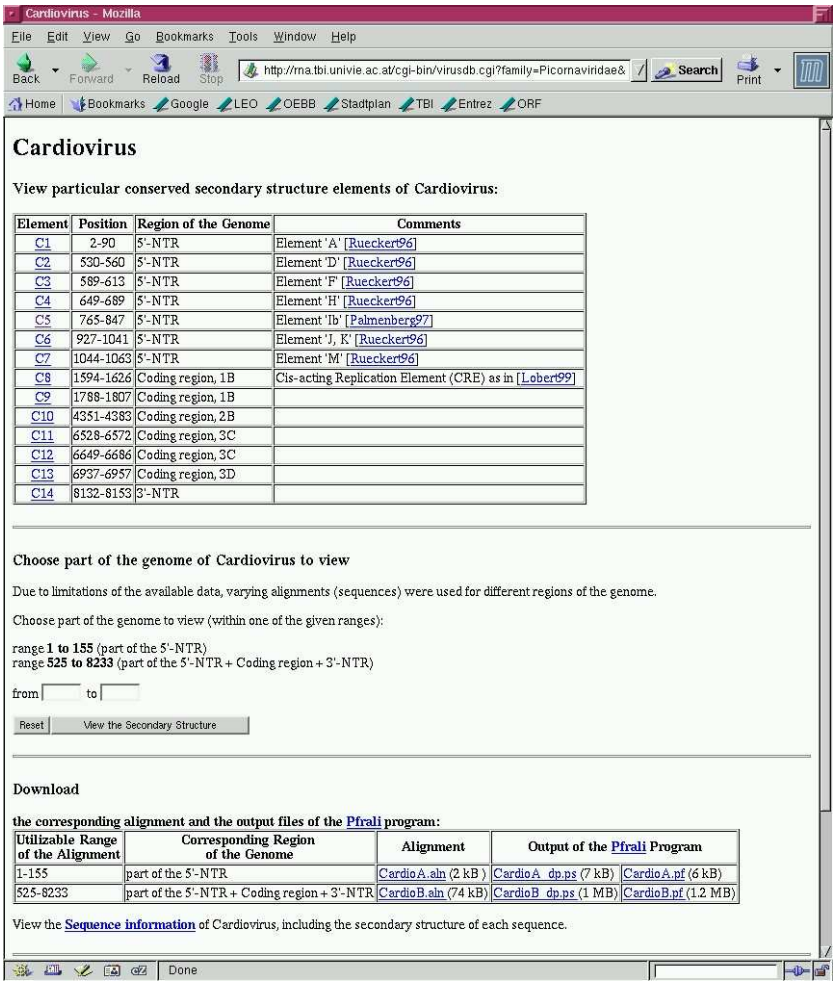


Figure 34: Example of a genus page

For each genus of the RNA viruses a *genus* page gives a list of the conserved

structural elements of that genus and provides the alignment of the complete genomes, as well as the output of the `alidot` program. This page gives access to the *result* page, either through the links of the listed structure elements, or by the completion of a form. Through the form a part of the genome can be specified, whose secondary structure is displayed on the result page. A snapshot of the genus page is shown in Figure 34.

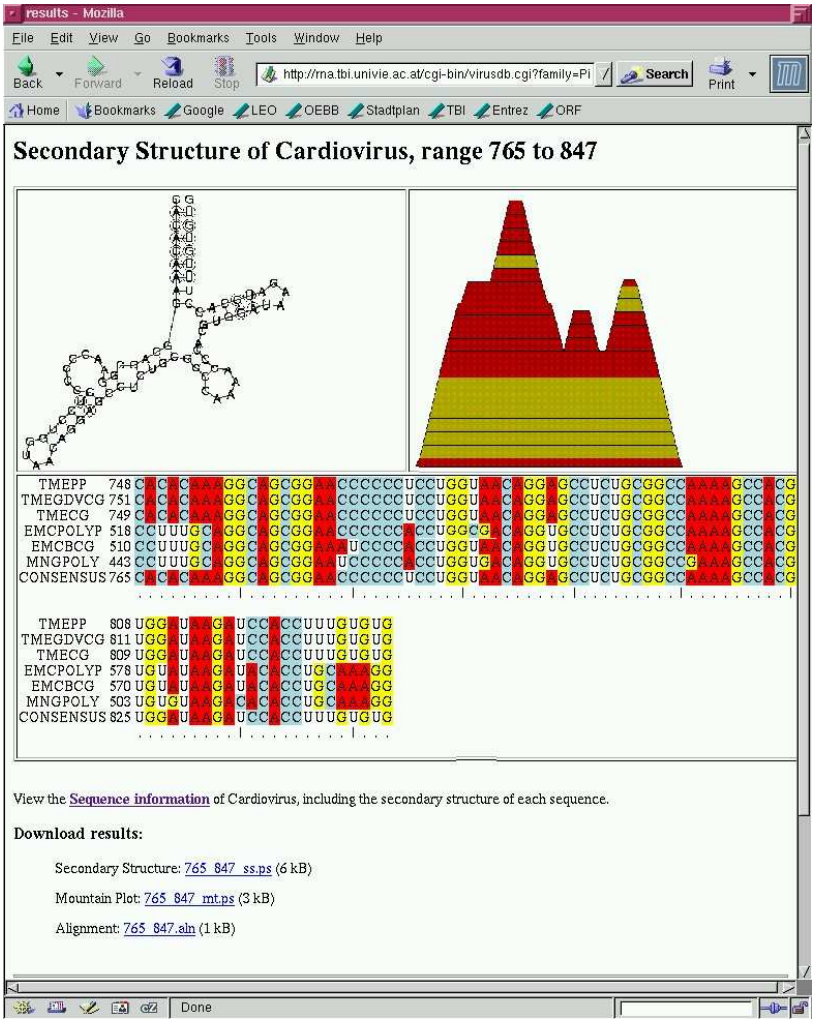


Figure 35: Example of a result page

The resultpage (Fig 35) contains the conventional secondary structure drawing and the mountain plot of the defined part of the viral RNA, which can be downloaded in `PostScript` format. In addition the corresponding alignment is shown and can be obtained in `Clustal` format.

The linked *sequence information* page enumerates the sequences used for the prediction, which are linked to the relevant NCBI/Entrez database records. The thermodynamic structure predictions for each sequence can be downloaded.

The 'Vienna Atlas of Viral RNA Structures' is based on `CGI` (Common Gateway Interface), which defines a way for a web server to interact with external content-generating programs. These are often referred to as `CGI` programs or `CGI` scripts. It is the most common way to put dynamic content on a web site. We implemented a `CGI` script that uses a code library, which is based on the `Perl` programs of the `Vienna RNA Package` [68] for displaying conserved structural elements. This allows the user to specify the part of the sequence to view, and the corresponding results are displayed in two different representations.

4.3 Results and Discussion

The putative conserved structural features that have been identified for each genus by the algorithm `alidot` [70, 72] are summarized in Figure 36. The conserved structure elements are listed in detail in A.2. Furthermore the complete data, consisting of the multiple alignments for each genus (generated by `ClustalW` [161]), the thermodynamic structure predictions for each sequence (generated by `RNAfold` [71]), the output of the `alidot` program, and the secondary structure elements listed in Fig. 36, are accessible through our web interface at `http://rna.tbi.univie.ac.at/virus/`. The largest pieces of conserved structure are located in the 5'NTR. In addition, there is a substantial number of possibly conserved RNA structures within the ORF. Table 4 gives the number of available sequences, the length of their alignment, their average pairwise sequence identity, the location of the coding region in the alignment, and the mean pairwise sequence identity in the 5′-NTR. A list of all sequences including their GeneBank accession numbers is given in A.1.

Table 4: Complete genomic RNA of picornaviruses

Genus	N	ℓ_A	η	Coding region	$\eta(5')$
Aphthovirus	9	8231	0.791	1088-8124	0.543*
Cardiovirus	6	8233	0.665	1088-8103	0.614
Enterovirus	29	7664	0.651	777-7548	0.765
Hepatovirus	10	7526	0.911	759-7449	0.901
Parechovirus	3	7391	0.774	716-7283	0.828
Rhinovirus	7	7296	0.687	652-7138	0.771
Teschovirus†	25	7135	0.893	433-7063	0.945

We list the number N of available sequences, the length ℓ_A of their alignment, their average pairwise sequence identity η, the location of the coding region in the alignment, and the mean pairwise sequence identity in the 5′-NTR. Only one complete sequence is known for both erboviruses and kobuviruses, which is not sufficient for the analysis presented here.

* From 7 of the 9 sequences since the 5'terminus is incomplete in 2 cases of the GeneBank entries.

† Teschovirus sequences do not include the S-fragment.

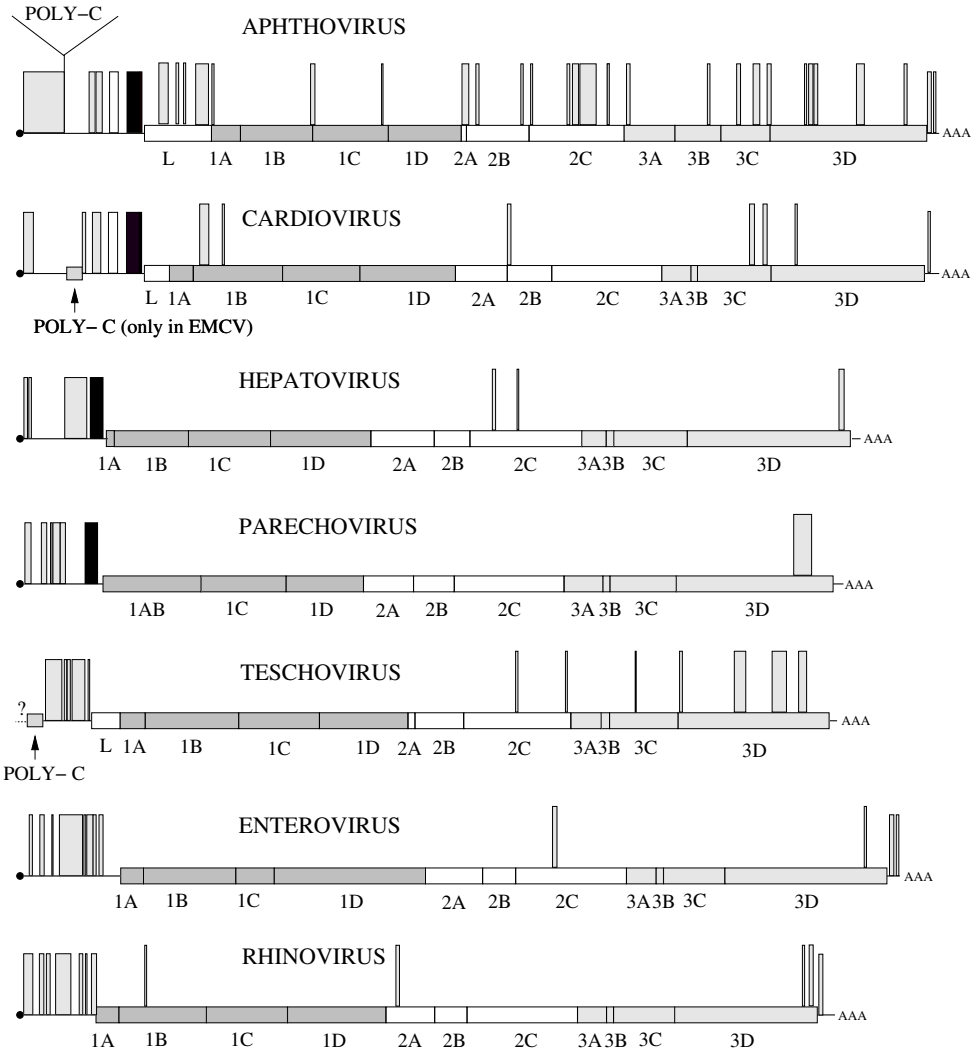


Figure 36: Overview of Picornaviruses genomes. Putative conserved RNA elements are indicated above the diagrams of the reading frames.
The black boxes indicate the J,K element, the white box is the Ib element; for details see text.

4.3.1 5'-Non-translated region

The most prominent feature in the 5'-NTR is the IRES, which consists of highly conserved RNA secondary structure domains. Figure 37 summarizes the results from our analysis which includes the genera *Teschovirus* and *Parechovirus* for the first time.

Overall, we find that the IRES structure is less conserved even within the genera than expected. Fig. 37 indicates in color those features that are conserved within a group at the level of individual base pairs. Non-shaded parts of the structure show additional structure elements of the reference sequence listed in A.1, which are not conserved within the sequences of the corresponding genus, which were predicted by folding the reference sequence using the conserved base pairs as constraints.

We recover the close structural similarities of rhinoviruses and enteroviruses. AIV (genus *Kobuvirus*) shares the cloverleaf structure (IV) with rhinoviruses and enteroviruses. A comparison of the genera *Aphthovirus*, *Cardiovirus*, *Parechovirus* and *Hepatovirus* sets *Hepatovirus* apart from the other three genera, see also the details of the J and K elements in Fig. 38. The one available complete erbovirus genome also shares the elements J and K with aphthoviruses and cardioviruses, see also [175]. Teschoviruses form a distinct fourth group of IRES structures that appears only vaguely related to the other groups.

Cardioviruses and Aphthoviruses. The secondary structure of the 5'NTR of cardioviruses is discussed elsewhere [129, 31, 92, 124]. Our results are very similar to the work of Palmenberg et al. [124] on EMCV. The main difference is that elements Ia and Ic, which flank element Ib, are not present in TMEV and therefore not conserved in the genus *Cardiovirus*. In addition, the H element is longer in our data. In comparison to the earlier studies both our results and the structures in Palmenberg et al.[124] have shorter conserved stem-loop regions.

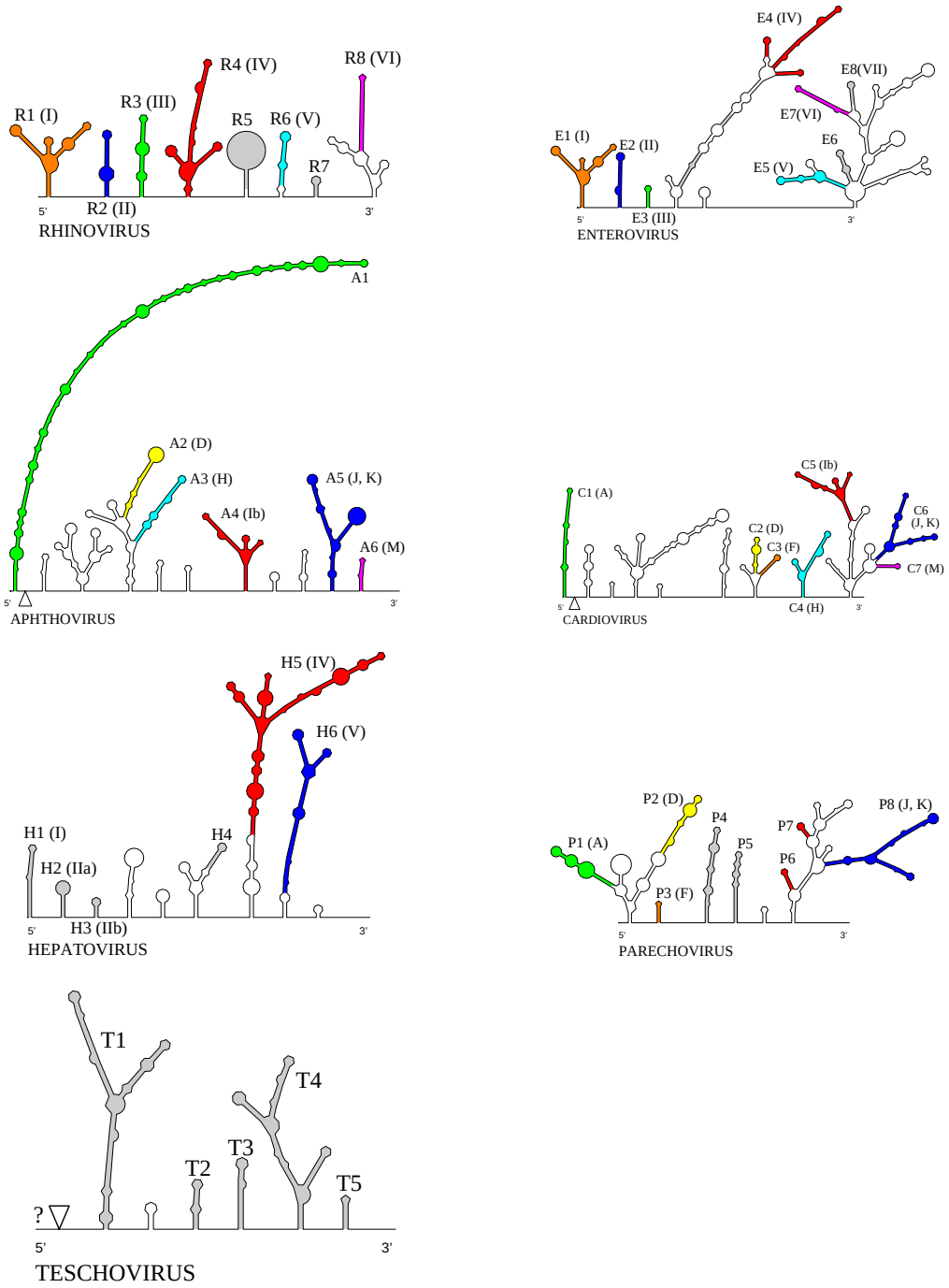


Figure 37: Schematic illustration of the 5'-NTR's of Picornaviridae. The minimum energy structure of one sequence of the respective virus genus is represented. Colored backgrounds mark regions which are conserved within all investigated sequences of that genus. The labels in brackets correspond to the notation of the 'classic' model of the IRES [144]. The sequence positions of the structure elements in a reference structure are listed in A.1. △ denotes a poly-**C** region, the ? indicates the missing data for the 5'end of the teschovirus sequences.

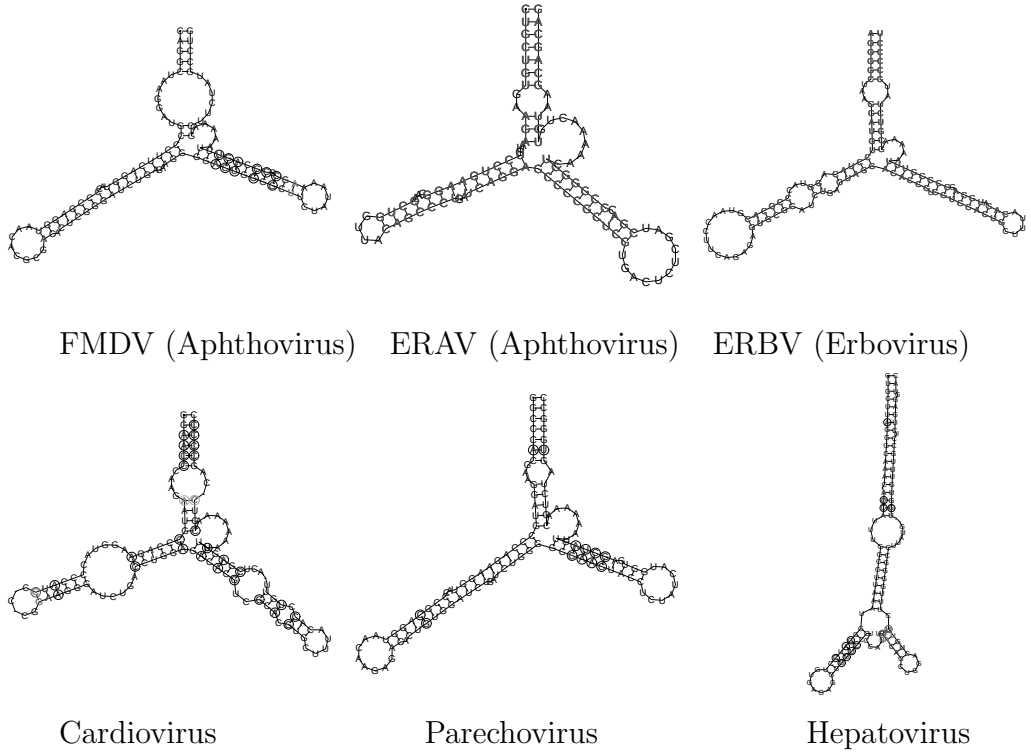


Figure 38: Consensus structures of the J and K elements. There appears to be some variation within aphthoviruses: The sizes of the loops varies between different species, FMDV and ERAV, in this example. Hepatovirus shows an analogous structure.

The 5'NTR of FMDV is discussed elsewhere [24, 129, 92]. There is only a single sequence for ERV-1 which has only marginal sequence similarity with FMDV and hence was considered separately. Our data for FMDV are similar to the earlier studies. However, we find that the conserved parts of the stem-loop regions are significantly shorter than the ones reported in Pilipenko et al. [129].

A comparison of cardiovirus and aphthovirus structures shows the following main differences: (1) The stem-loop structure A1 at the 5'end is much longer in FMDV compared to cardiovirus. (2) The D element in FMDV is enlarged at the expense of F.

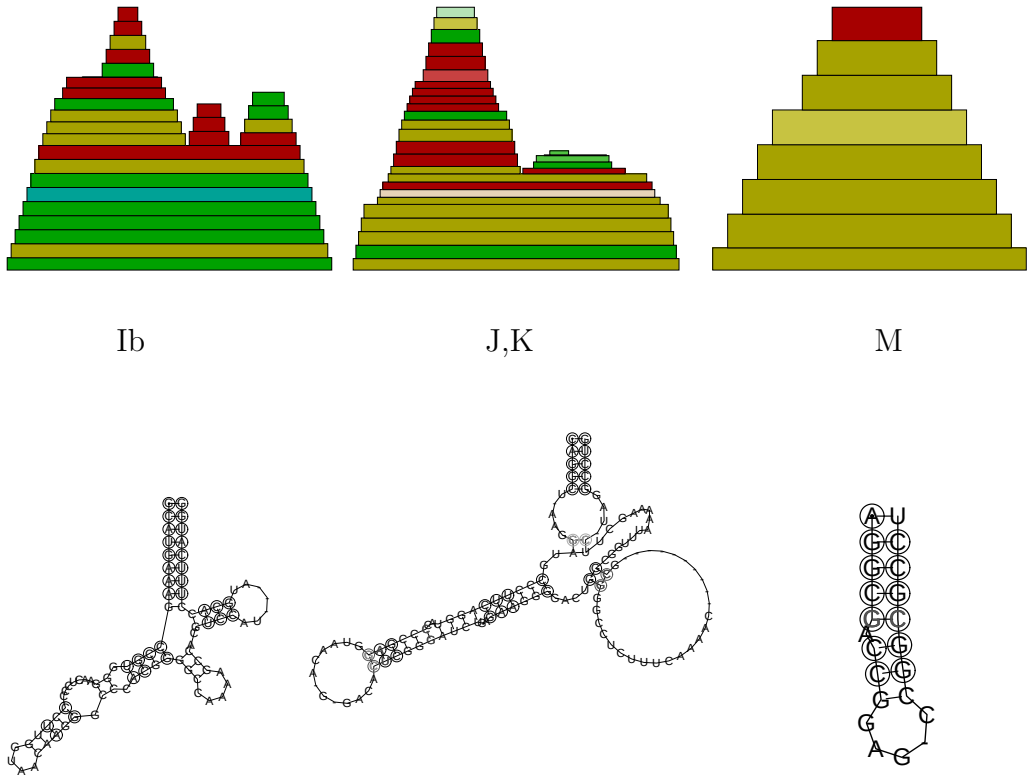


Figure 39: Common features in the IRES of Aphthovirus and Cardiovirus. Alignment manually improved.

The stem-loop structure H is very similar in aphthovirus and cardiovirus, the loop sequence UCUUU is strongly conserved in both genera. The stem contains many compensatory mutations that `Clustal W` failed to correctly align in this region. Manual improvement of the alignment shows that the Ib-elements as well as the M-elements of both genera can be superimposed and hence are structurally almost identical, Fig. 39 (left and right). In contrast, only the J-stem of the J,K feature is structurally (almost) identical in the two genera, Fig. 39 (middle), despite the fact that the topology of the J,K elements is conserved, Fig. 38.

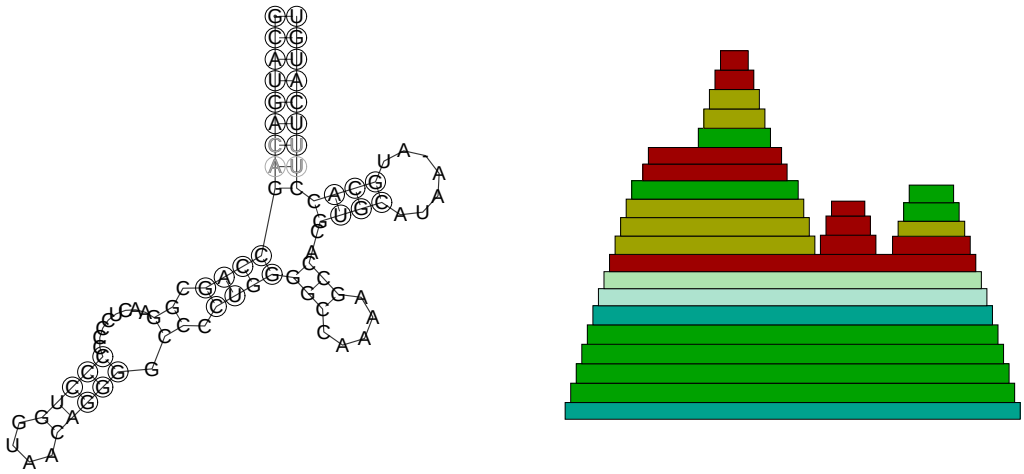


Figure 40: Common structure of element 'Ib' of Aphthovirus, Cardiovirus and Parechovirus. Alignment manually improved.

Parechoviruses. Until recently parechoviruses Echovirus 22 and Echovirus 23 were classified as members of the genus Enterovirus. The secondary structure of their 5'NTR is described in [121, 44]. As our analysis is based on only 3 sequences we might still overestimate the conserved parts, in particular in regions P1, P4, and P6 where the sequence is highly conserved. Ghazi [44] finds the same structure for Parechovirus and Cardiovirus. Our results agree in part with this analysis. In particular: Our element P1 corresponds to A in [44], but is shorter. P2 corresponds to D, P3 corresponds to F. Our element P4 is located in region of Ghazi's H. The sequence is rather conserved in this region and both Ghazi's H and our P4 have comparable thermodynamic stability, with a pairing probability of approximately $p = 1/3$ for each of the two alternatives. Both variants have little similarity with the H-element in cardiovirus and aphthovirus. The element P5 has been reported before. P8 contains the J and K elements, where J is identical to Ghazi's structures, our prediction for the K-element shows minor differences with previous studies. Our elements P6 and P7 are part of Ghazi's cloverleaf like motif I. The cloverleaf structure is thermodynamically feasible ($p \approx 1/4$), but appears to have significant structural variability in this genus so that it is not detected

as a conserved feature. However, the alignment of the corresponding region of parechovirus with cardio- and aphthovirus shows the structural conformity, see Fig. 40. The analysis in [121], which used only two sequences and Zuker's `mfold` program, agrees in part with our consensus structure.

Hepatoviruses. The secondary structure of hepatovirus RNA is considered by Brown et al. [10]. The sequences in our data set have about 90% pairwise identity, hence we have only a small number of compensatory mutations to verify structural features predicted based on the thermodynamic rules.

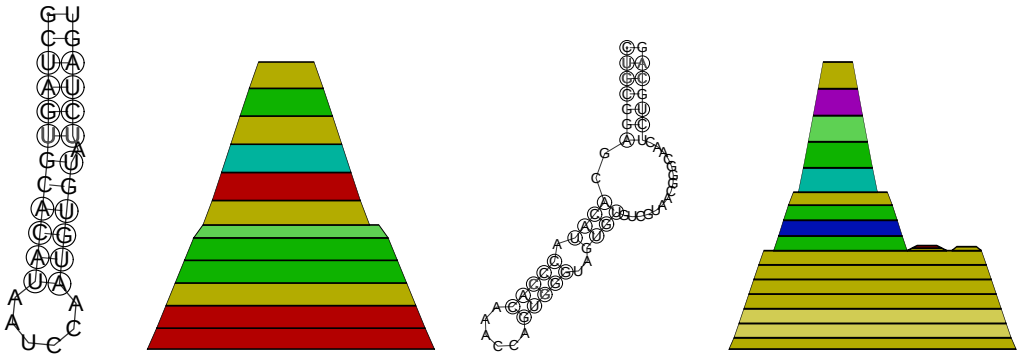
Elements I and II are identical to Brown's structure; the sequences are completely conserved in this regions, i.e., there are no co-variations to verify the thermodynamic prediction in the region. Domain III is not present in our data. We find high structural flexibility here. It is noted in the work of Brown et al. [10] already that "the structure of domain III was poorly defined." The only possibly significant structure in this region is a stem-loop with a completely conserved sequence around position 300, which does not appear in Brown's prediction.

Stem IV is significantly shortened in our analysis and the multiloop of the cloverleaf structure is slightly different. The deletion studies reported by Brown et al. [10] indicate that that domain IV is critically involved in formation of an HAV IRES element. Compared to the earlier study we find a larger element V at the expense of part of IV.

The conserved secondary structure elements of hepatoviruses cannot be compared directly to those of aphtho- and cardioviruses. But there is a structural analogy of the cloverleaf structure (Ib in cardiovirus/aphthovirus and IV in hepatoviruses) and the branching stem-loop (J,K in cardioviruses/aphthoviruses and V in hepatoviruses).

Enteroviruses and Rhinoviruses. The secondary structure of the IRES of enteroviruses and rhinoviruses has been the subject of a larger number of studies, see e.g. [130, 144, 180]. We recover elements I through VII in en-

teroviruses and I-VI in rhinoviruses (example given in Fig 41), some of them with a slightly shorter stem. In addition we found the stem-loop structures R5 and R7, and E6, respectively, see Fig. 37. Elements R5 and R7 can be detected unambiguously only in HRV-A.

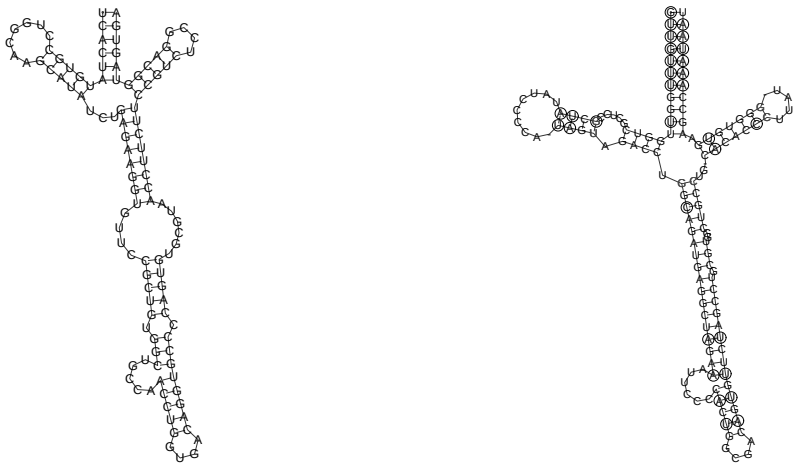


Rhinovirus: element V *Enterovirus*: element V

Figure 41: Common conserved element between enteroviruses and rhinoviruses. Element V in poliovirus is known to bind eukaryotic initiation factor 4B [9]

An attempt to extract the common structures of enteroviruses and rhinoviruses for a common multiple alignment yielded only a fraction of the structures found in each genus separately. In part this is due to small differences in the structural elements and in part the lack of structures can be attributed to the poor quality of the alignment.

Kobuviruses. There are three complete sequences available of kobuviruses. The quality of the alignment of the one bovine kobuvirus sequence with the two aichivirus sequences is insufficient for an analysis. The two aichivirus sequences show a sequence similarity of 99%, therefore their predicted common structure is not corroborated by compensatory mutations. The only feature that can be matched with a conserved structural element of other genera is the cloverleaf-like structure at position 474-582 of the alignment, which resembles element R4 of rhinoviruses, respectively element E4 of enteroviruses (see Fig. 42).



Kobuvirus *Rhinovirus*

Figure 42: Secondary structure of kobuviruses at pos. 472-582 and element R4 (pos. 275-408) of rhinoviruses

Teschoviruses. The sequences of the teschoviruses 5'NTR are not known completely. The presence of an oligo-C stretch was demonstrated for F65 [29]. The nucleotide sequence of the 5'NTR up to this C tract could be determined successfully only in 3 of the 25 assayed strains (Talfan, Bozen, and Vir-1626/89) [179]. Hence we report no structure before the oligo-C-region.

The secondary structure of the 5'NTR of teschoviruses has not been studied previously. Only element T4 shows some similarities with element V in hepatoviruses. The other conserved structures do not have obvious similarities with conserved elements in other picornavirus genera. The conserved elements T1-T5 are shown in Fig. 43.

Other Picornaviridae. According to [175] the IRES structure of ERAV (genus *Aphthovirus*) and ERBV (genus *Erbovirus*) is similar to that of FMDV (genus *Aphthovirus*) and cardioviruses. The similarity to FMDV is insufficient for a good alignment of ERAV with the FMDV sequences. The computed minimum energy structure shows an identifiable J,K-element in both species, comparable to the J,K-element of type II IRES (Fig. 38).

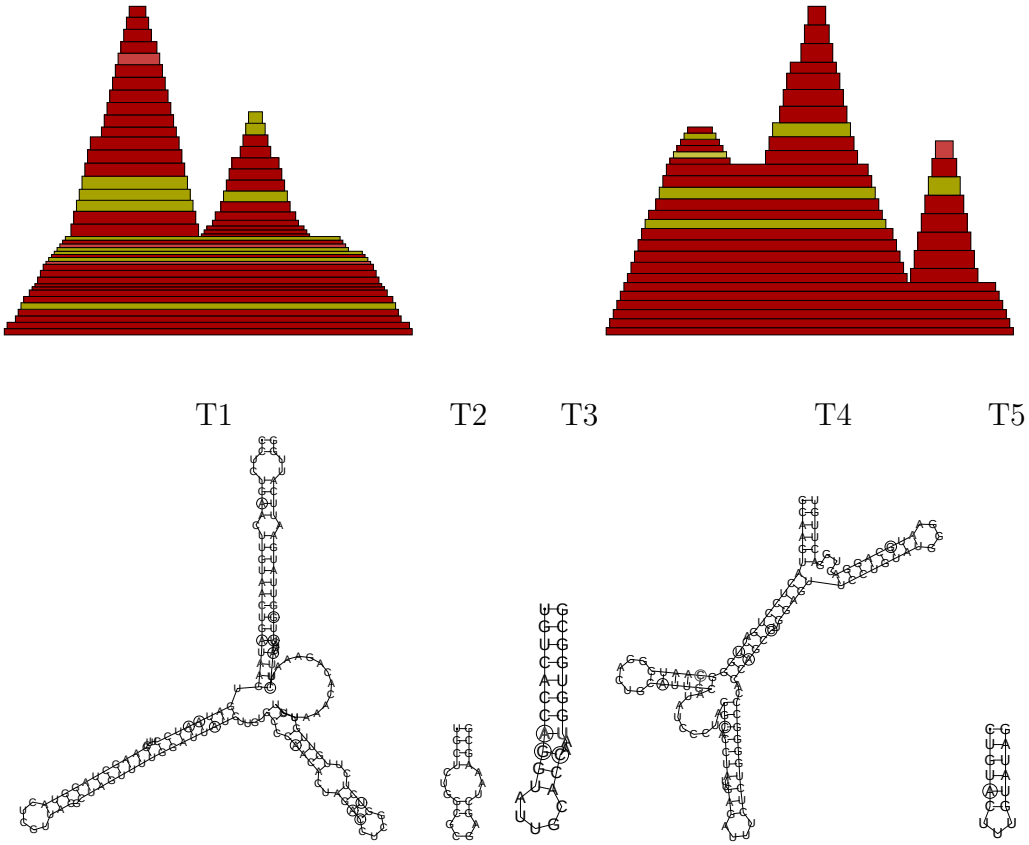


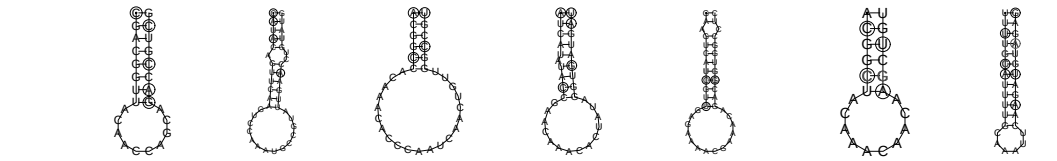
Figure 43: Conserved Structures in the 5'NTR of teschoviruses. Mountain plots are given only for T1 and T4.

4.3.2 Coding Region

Cis-acting Replication Element (CRE). A cis-acting replication element (CRE) within the coding region of several picornaviruses has been described in a number of different picornaviruses. The function of the CRE probably involves the initiation of the synthesis of the negative-sense strand template RNA during virus replication [47].

The CRE has been identified in HRV14 in region 1B of the genome [107], in Cardiovirus in region 1B [95] and in Poliovirus in region 2C [47]. Since the time we published our results [172], our prediction of the CRE in HRV-A was confirmed by experimental studies [43].

Although located within a protein-coding segment of the genomes , the CRE function is independent of its translation. Thus, this segment of the viral RNA has dual functions, both encoding the protein and participating directly in the replication of the viral genome. The existence of the computer-predicted structure was confirmed by mutational analysis [107, 95]. Furthermore, the activity of the CRE is not position dependent [47].



Aphthovirus Enterovirus Cardiovirus HRV-A HRV-B Teschovirus Hepatovirus

```
Aphto      ~~~~CGAC-GGUU------ACA-CCAAGCA------GACCGUCG~~~~~
Entero     CAUACAGU-UCAAG--------UCCAAAU-GCCGUAUUGAACCUGUAUG
Cardio     ~~~~~ACG-GCCA---CAAACACCCAAUCAACUGU-UGGCCGU~~~~~~
HRV-A      ~~~AUCAUAUACCGAACAAACA---------CUAUAGGUGAUGAU~~~~
HRV-B      GAAGUCAU-CGUUGAGAAAACG---AAACA------GACGGUGGCCUC~
Tescho     ~~~~~~AC-GGCU--ACAAACA-----ACA------AGCUGU~~~~~~~
Hepato     UUUUGCAU-UUUG---CAAA--------------UUCAAGAUGUAGAG~
           ~~~(((((-((((.....................)))))))))~~~~
           1.......10........20........30........40.........
```

Figure 44: Known and putative CREs in picornaviridae: secondary structures (top) and sequences (below). Nucleotides in the loops are highlighted to emphasize the AC-rich composition.

Table 5: Position of (putative) CRE elements.

Genus	Acc.No	Gene	Position	Remark
Aphthovirus	AJ007347	2C	4834-4859	putative
Enterovirus	V01150	2C	4456-4494	as in [47]
Cardiovirus	M81861	1B	1308-1340	as in [95]
Hepatovirus	K02990	2C	4187-4245	possible
Teschovirus	AF231769	2C	4228-4249	putative
Rhinovirus-A	M16248	2A	3325-3357	as in [43]
Rhinovirus-B	K02121	1B	1727-1764	as in [107]
Parechovirus		?		

In cardiovirus we recover the CRE in the 1B region which encodes the capsid protein VP2. For EMCV (excluding *Mengo-Virus*) and *Theilovirus* our structure agrees with the one reported in [95]. In [124] a different structure is given for *Mengo-Virus*. We find that the *Mengo-Virus* CRE-structures agree with the consensus of the other species. In enteroviruses we recover the CRE in 2C as described in [47], in HRV-B the element is found in 1B as described in [107]. Our prediction of the CRE in the 2A region of HRV-A has been confirmed later, see [43].

We find putative CRE elements in the 2C region of aphthoviruses and teschoviruses. There are three conserved elements in the coding region of hepatoviruses. The most likely candidate for a CRE is located in region 2C. For parechoviruses we were not able to identify a putative CRE. The locations of the (putative) CREs are summarized in table 5.

The loop of the CRE is relatively large in all genera and contains predominantly A and C. We note that an alignment of region 2C of all aphthovirus and teschovirus sequences shows that the CRE element is conserved between the two genera.

Other Conserved Elements. There appears to be no structural feature in the coding region that is shared among all picorna genera besides the CRE. On the other hand we find a number of structures that are conserved within a genus, see Fig. 36. There are 5 such structures in cardioviruses, a single feature in the 3D region of parechoviruses, 6 in teschoviruses, 3 in hepatoviruses, 2 in enteroviruses, 25 in aphthoviruses, and 1 in rhinoviruses. A complete list is provided in electronic form, see "supplemental material". We do not have an explanation for the large number of conserved elements in genus aphthovirus in comparison with all other picornaviridae.

The two species HRV-A and HRV-B in the genus *Rhinovirus* show significant differences at the level of their secondary structures. In HRV-A we find 4 conserved elements inside the coding region, only one of which is also conserved between HRV-A + HRV-B. These differences are emphasized by the fact that the CRE is located in different parts of the genome in these two species, see Tab. 5.

4.3.3 3' Non-Translated Region

Recent deletion studies [26, 32] show that the 3'NTR, which ends in a poly-A region of variables length in all genera, is important for RNA synthesis.

Cardiovirus. Duque and Palmenberg [32] report three conserved stem-loop motifs (I, II, III) in the 3'NTR. Deletion studies by these authors indicated that deletion of III is lethal, deletion of II resulted in marginal RNA synthesis activity but failure of transfection with genomic RNA, while stem I was found to be dispensable for viral growth. Surprisingly, we find stem I with large thermodynamic stability and a significant number of compensatory mutations, while regions II and III do not form a conserved structure in our data set. The structures reported by Duque and Palmenberg [32] are consistent with all sequences of our data-set but are not thermodynamically favorably in any one of them, Fig. 45. Neither II nor III can be recovered by

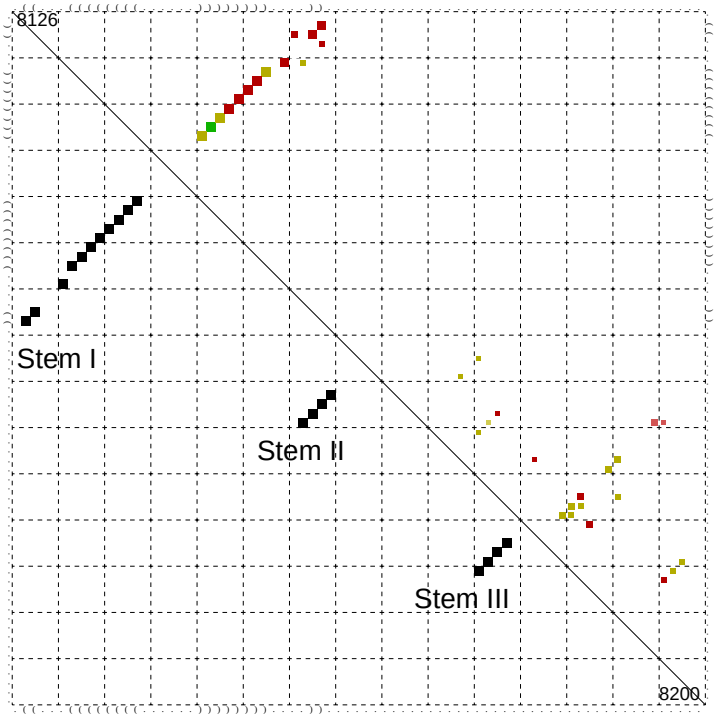


Figure 45: Dot plot of the 3'NTR of cardioviruses. Upper triangle: Output of `alidot` showing only stem I. Its prediction is well supported by a number of consistent mutations and the absence of inconsistent mutations. The size of the squares is proportional to $\log p$ here. The signals in the area of stems II and III have probabilities of only a few percent and hence are not significant. The lower triangles shows the structures reported in [32] for comparison.

considering the species Theilovirus and EMCV separately.

The secondary structures reported in [26] are completely different from both our results and the structures reported in [32]. Cui and Porter [26] suggest that a U-rich stretch, essentially region III of [32], interacts with the poly-A tail.

Enterovirus. There is ample literature on the structure of the 3'NTR of enterovirus, e.g. [131, 140, 180, 111, 167, 109]. None of the reported structures is conserved within the entire genus. Following the previous studies we have

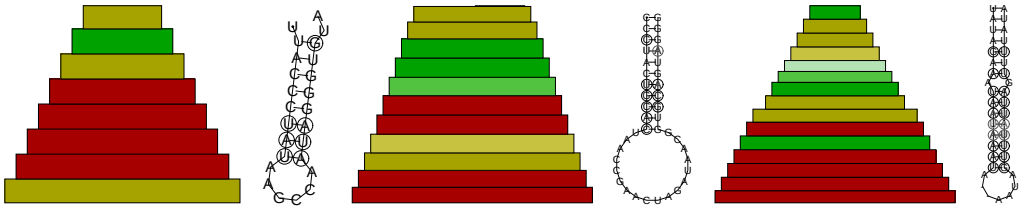


Figure 46: The most prominent features in the 3'NTR. Left: Cardiovirus region I; Middle: Region II of Enterovirus cluster 3; Right: hairpin structure of rhinovirus.

therefore split the 29 available genomes into three clusters because there are not enough sequences available for the fourth cluster described in [180], see Fig. 47.

Cluster 1 consists of *Poliovirus* and *Human Enterovirus C*, cluster 2 contains *Human Enterovirus A* [180, 85]. 3'NTR sequences are highly conserved within each of these two clusters. While we find the published structures in our data, their equilibrium probabilities are small and they appear as one of a number of thermodynamically feasible alternatives.

Cluster 3 contains *Human Enterovirus B*. We find domains I and II as reported in [131, 180]. It is interesting to note that the structure of domain II is supported by a substantial number of compensatory mutations.

Other Genera. The 3'NTR of aphthoviruses apparently has not been considered before. We find two hairpin structures, one of which has an almost conserved GAAA sequence motif in the loop. In one of the nine sequences we find GCAAA instead.

A hairpin motif, which was already reported in [131] is detected unambiguously in rhinovirus, Fig. 45. The structure is conserved between HRV-A and HRV-B.

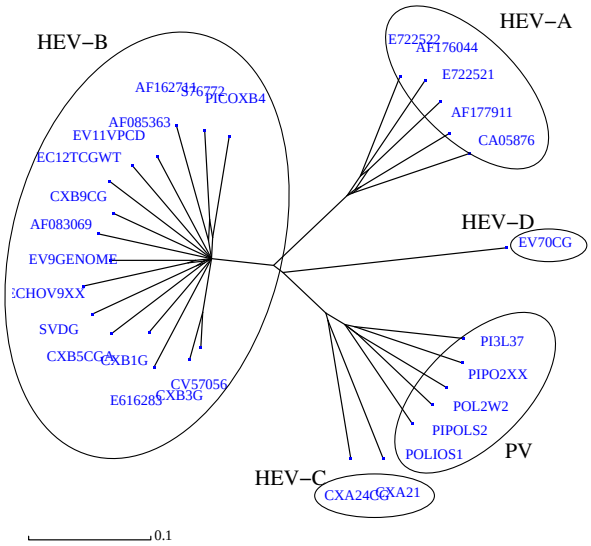


Figure 47: Phylogenetic relationship between enterovirus strains based on the alignment of the complete genome. Cluster 1: PV and HEV-C, cluster 2: HEV-A, cluster 3: HEV-B

In parechoviruses, hepatoviruses, and teschoviruses there are no significant conserved secondary structures in the 3'NTR. In particular, we could not confirm the published minimum energy structures for individual teschovirus [29, 179] and hepatovirus [140] sequences as conserved features.

5 Predicting RNA Bi-secondary Structures: A critical evaluation of the MWM approach

The prediction of RNA structures including pseudoknots is still an open problem. Since the thermodynamic structure prediction based on the standard energy model is NP-complete [100, 3], available algorithms based on dynamic programming are very restricted in sequence length and/or allowed complexity of pseudoknots. When a large number of homologous RNA sequences is available, comparative sequence analysis methods are successful in predicting the consensus structures. It would be desirable to be able to predict the consensus structure including pseudoknots based on a smaller set of sequences. Tabaska *et al.* [156] described a method based on the MWM algorithm using a rather simple scoring scheme (refer to section 3.4.2). The following chapter presents the program `hxmatch` which uses an improved scoring procedure, thus reducing the number of sequences required for a secondary structure prediction including pseudoknots.

5.1 Method

The general organization of `hxmatch` is similar to the method of Tabaska *et al.* [156], in that a scoring matrix of possible base pairs is generated, which serves as input for the MWM algorithm. Our scoring procedure differs from those used by Tabaska, in that the stacking energies of the base pairs are taken into account and the covariation score described in Hofacker *et al.* [69] is used, instead of the MI score (refer to section 3.1). In contrast to Tabaska's method, the combination of the thermodynamic score and the covariation score is implemented, as is the post-processing of the base pairs selected by the MWM algorithm.

`Hxmatch` starts with an alignment and generates a scoring matrix, where a weight is assigned to each possible base pair. This yields a weighted graph

where the nucleotides form the vertex set and the edge set contains all base pairs with positive weight. Then the maximum weighted matching algorithm finds the matching on that graph which has maximum weight. The base pairs contained in the matching include isolated base pairs and do not necessarily form a bi-secondary structure. Therefore the outcome of the MWM algorithm needs some post-processing. During post-processing several edges are deleted from the original input graph. Therefore a matching on this altered graph is determined. These two steps (MWM and post-processing) are iterated until the outcome is constant. The scoring of the base pairs is the crucial point of the method, our improved scoring procedure is described in the following section.

5.1.1 Base Pair Scoring

Starting from a RNA sequence alignment $\mathbb{A}$ of N sequences a scoring matrix Π is generated from the combination of the thermodynamic score, derived from the stacking energies of helices, and the covariation score, which is based on the number of mutations for a given alignment position.

Thermodynamic score For each sequence α all base pairs ij contained in the set of allowed base pairs $\mathcal{B} = \{GC, CG, AU, UA, GU, UG\}$ which are part of a possible helix with minimum length 3 are tabulated. The energy of each helix is calculated from the experimentally determined energy parameters for base pair stacking [104]. This requires $\mathcal{O}(Mn^2)$ time and $\mathcal{O}(n^2)$ memory, for M sequences with length n. The weight H_{ij}^{α} of a base pair in sequence α is the energy of the longest helix the base pair is part of, multiplied by (-1) to obtain positive weights. Then the gaps in the alignment are inserted into the corresponding scoring matrix H_{ij}^{α} of each sequence. The entry in the combined scoring matrix $H_{ij}^{\mathbb{A}}$ of the alignment is then

$$H_{ij}^{\mathbb{A}} = \frac{1}{N} \sum_{\alpha \in \mathbb{A}} H_{ij}^{\alpha} \tag{16}$$

Covariation score The mutual information score (Section 3.1) makes no use of RNA base-pairing rules. This allows the identification of non-canonical base pairs and tertiary interactions. On the other hand it does not account for consistent, non-compensatory mutations, i.e. if we have, for example, only GC and GU pairs at positions i and j then $M_{ij} = 0$, so this case is treated just like a pair of conserved positions. Therefore we use a covariance-like measure, which distinguishes between conserved pairs, pairs with consistent mutations and pairs with compensatory mutations [69].

It is convenient to use the notation:

$$d_{i,j}^{\alpha,\beta} = 2 - \delta(a_i^\alpha, a_i^\beta) - \delta(a_j^\alpha, a_j^\beta) \tag{17}$$

where $\delta(a', a'') = 1$ if $a' = a''$ and 0 otherwise. Thus $d_{ij}^{\alpha\beta} = 0$ if the sequences α and β coincide in both aligned positions i and j, $d_{ij}^{\alpha\beta} = 1$ if they differ in one position, and $d_{ij}^{\alpha\beta} = 2$ differ in both positions. In other words, $d_{ij}^{\alpha,\beta}$ is the Hamming distance of the restriction of the sequences α and β to the two aligned positions i and j.

We use the measure for covariation described in [69]:

$$C_{ij} = \sum_{\mathsf{XY},\mathsf{X'Y'}} f_{ij}(\mathsf{XY}) \mathbf{D}_{\mathsf{XY},\mathsf{X'Y'}} f_{ij}(\mathsf{X'Y'}) \tag{18}$$

where the 16×16 matrix $\mathbf{D}$ has entries $\mathbf{D}_{\mathsf{XY},\mathsf{X'Y'}} = d_H(\mathsf{XY}, \mathsf{X'Y'})$ if both $\mathsf{XY} \in \mathcal{B}$ and $\mathsf{X'Y'} \in \mathcal{B}$ and $\mathbf{D}_{\mathsf{XY},\mathsf{X'Y'}} = 0$ otherwise. $f_{ij}(\mathsf{XY})$ denotes the frequency of base pair XY at position ij of the alignment.

The percentage of inconsistent sequences for positions i and j, i.e. sequences that cannot form a base pair between positions i and j, is denoted by q_{ij}. They are taken into account by forming the combined score

$$B_{ij} = C_{ij} - \phi_1 q_{ij} \tag{19}$$

Including the helix score, we get

$$\pi_{ij} = H_{ij}^{\mathbb{A}} + \phi_2 B_{ij} \tag{20}$$

where ϕ_1 and ϕ_2 are scaling factors, their default values are given in Section 5.2. Note that ϕ_2 has the dimension of an energy.

The combined scoring matrix π is scanned for helices with minimum length 3, only base pairs ij with $\pi_{ij} > 0$ are considered. The weight of a helix Ψ is defined as the sum of the weights of the base pairs forming that helix:

$$\omega_\Psi = \sum_{ij \in \Psi} \pi_{ij} \tag{21}$$

The weight of each helix is then assigned to the corresponding base pairs:

$$\Pi_{ij} = \omega_\Psi \text{ for all } ij \in \Psi \text{ with } \pi_{ij} > 0 \tag{22}$$

Finally all base pairs with a score smaller then a threshold Π^* get zero weight:

$$\Pi_{ij} = 0 \text{ if } \Pi_{ij} < \Pi^* \tag{23}$$

We have therefore 3 parameters, the scaling factors ϕ_1 and ϕ_2, and the threshold Π^*. The results for different parameter sets are discussed in Section 5.2.

5.1.2 Maximum Weighted Matching

The input graph $\Gamma^{(0)}$ for the maximum weighted matching algorithm consists of the vertex set $V = \{1, \ldots n\}$, where n is the length of the alignment, and the edge set formed by all base pairs with score $\Pi_{ij} > 0$. We use the algorithm for maximum weighted matching of H. Gabow's Ph.D. thesis [41] implemented by Edward Rothberg [142]. A detailed description of the algorithm is given in Section 2.5.

5.1.3 Post-processing

The maximum weighted matching obtained for the input graph $\Gamma^{(0)}$ is not necessarily a bi-secondary structure, furthermore isolated base pairs are contained in the matching. Therefore the outcome of the MWM algorithm needs some post-processing.

All isolated base pairs and helices with length 2 are deleted from outcome, and the remaining helices are extended further, if the corresponding base pairs are contained in the graph $\Gamma^{(0)}$, compare Fig. 48.

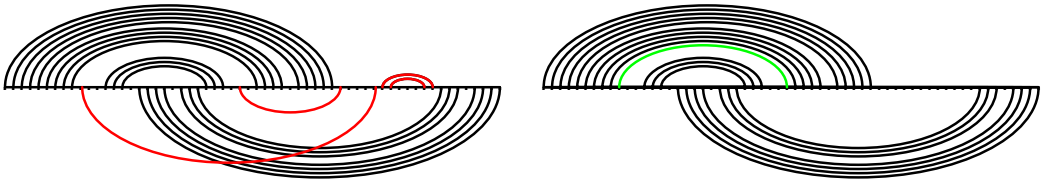


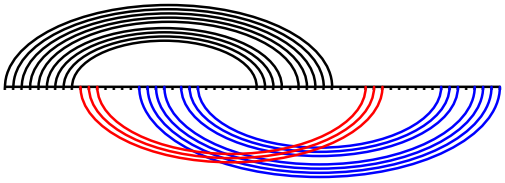
Figure 48: Post-processing: Isolated base pairs and helices with length 2 (indicated in red) are contained in the matching *(l.h.s.)*, they are deleted and the remaining helices are extended further, if possible. In our example one base pair is added to the outcome, drawn in green *(r.h.s.)*.

It would be desirable to extract that bi-secondary structure from the matching, which has maximized weight. This could be done in the following way: First we compute the inconsistency graph $I(M)$ (refer to section 2.3.3) of the weighted matching M. If $I(M)$ is bipartite there is nothing to do, since M is already a bi-secondary structure (theorem 2 in [62]). Otherwise, we consider the connected components of $I(M)$ separately. For each component C we compute its chromatic number $\chi(C)$ and then all color partitions with $\chi(C)$ colors. We choose the color partition that maximizes total weight of two color classes. The vertices in these two color classes form by construction a maximal weight subset $B(C)$ of the base pairs in the component C of $I(M)$ that can be drawn in the plane. Finally, we form the union of the base pairs $B(C)$ over all components C of $I(M)$ for the maximal planar sub-matching

of M, this gives the bi-secondary structure with maximum weight.

Unfortunately, computing $\chi(C)$ is NP-complete [74] and furthermore, we might have to optimize over exponentially many color partitions of C.

Therefore we use the following greedy procedure to derive a bi-secondary structure from the matching. The helices are ordered by descending weight, once again the weight of a helix is the sum of the weights of the base pairs forming the helix. The helix with the highest weight belongs to Ω_U, the subset of helices which are drawn in the upper half plane of the linked diagram representation (compare Section 2.3.6). Then we go through the sorted list of helices and assign all helices conflicting with a higher ranked helix (temporarily) to Ω_L. Subsequently the helices contained in Ω_L are scanned and all helices conflicting with a higher ranked helix of Ω_L are deleted from the graph. Figure 49 shows an example of the classification of the helices.



Helix	Weight	assigned to
black1	100	Ω_U
blue1	90	Ω_L
blue2	70	Ω_L
red	65	deleted
black2	45	Ω_U

Figure 49: Classification of helices: Since the red helix is inconsistent with the higher ranked helices *black1* $\in \Omega_U$ and *blue1* $\in \Omega_L$, it is deleted to obtain a bi-secondary structure.

5.2 Results and Discussion

`Hxmatch` was tested on three different types of RNA known to contain pseudoknots: Signal Recognition Particle RNA (SRP RNA), Ribonuclease P RNA (RNaseP RNA), and tmRNA. SRP RNA (Fig. 50) has a long, double helical structure with one pseudoknot structure close to the 5' end [91], which can be viewed as 'kissing hairpins'. The overall structure of RNaseP RNA (Fig. 54) is more globular, with rather short double helical domains connected by single strand stretches, and it contains two long-range pseudoknots [59]. The structure of tmRNA (Fig. 56) contains four H-type pseudoknots and is roughly globular [185].

Preliminary tests showed that the quality of the predictions, in terms of a maximum number of correct predicted helices and a minimum number of incorrect predicted helices, improves to optimum levels with the relative weight of inconsistent sequences ϕ_1 set to 0.8. This corresponds, for example, to the (negative) score of 10 % incompatible sequences balanced out by the score of consistent mutations in 9 % of the sequences. The quality of the results improves to optimal level as the scaling factor of sequence covariation is such, that the ratio of the covariation score to the thermodynamic score is approximately 3:1 ($\phi_2 = 6000$), and changes little at higher levels. All predictions in the following sections were generated with the default values of $\phi_1 = 0.8$ and $\phi_2 = 6000$. Different values for the threshold Π^* were investigated, predictions were generated with $\Pi^* = 0$, $\Pi^* = 2\pi^*$, and $\Pi^* = 3\pi^*$, where the constant $\pi^* = 2500$ is in the order of magnitude of the mean weight $\bar{\Pi}_{ij}$ of all base pairs with positive weight of the examples investigated, or about 10 % of the maximum weight, respectively.

5.2.1 Signal Recognition Particle RNA

The Signal Recognition Particle (SRP) is a phylogenetically conserved ribonucleoprotein required for the translocation of nascent secretory and mem-

brane proteins across biological membranes (for review, see e.g. [84]). SRP was first identified in mammalian cells in the early 1980s [165, 166]. Mammalian SRP is composed of six polypeptides and an RNA molecule of about 300 nucleotides (termed SRP RNA or 7SL RNA). The prokaryotic homologue comprises one polyprotein and the SRP RNA (also termed 4.5S RNA).

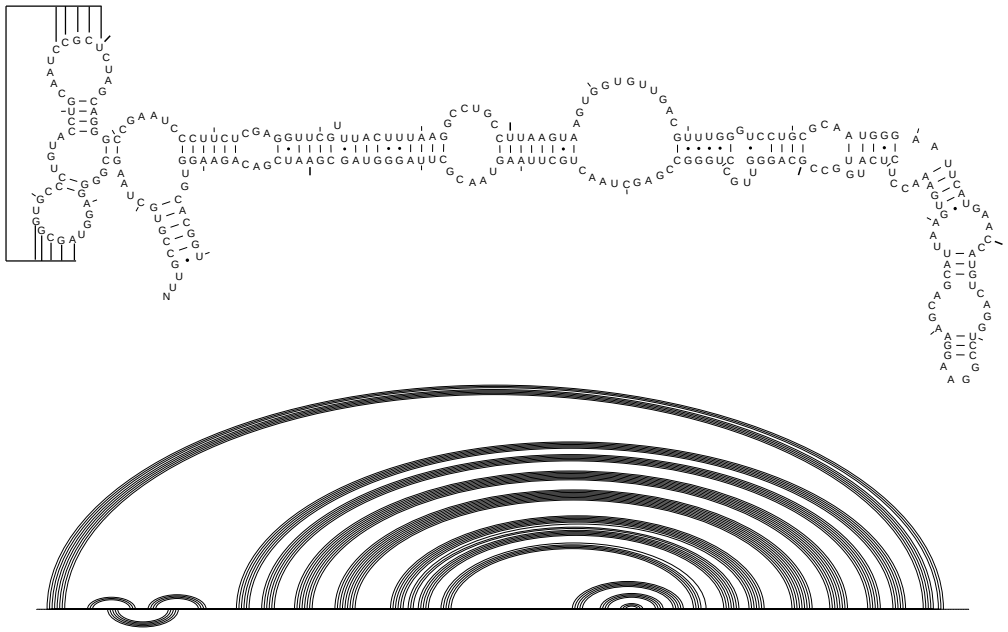


Figure 50: SRP-RNA secondary structure of *Bacillus subtilis* derived by comparative sequence analysis [91]

The reference structures for the bacterial and archaeal SRP RNAs shown in Figure 50 and Figure 51 were obtained by comparative sequence analysis [91]. The structures are based on an alignment of 39 sequences, closest relatives 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.

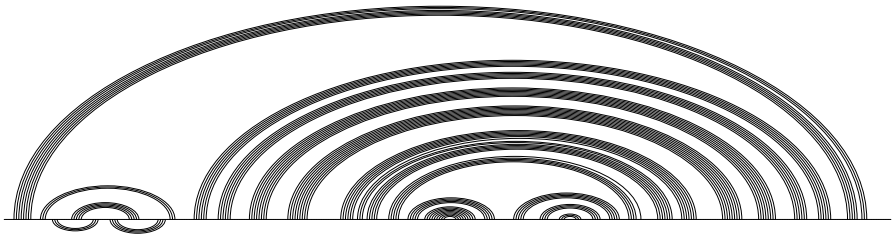


Figure 51: SRP-RNA secondary structure of *Halobacterium halobium* derived by comparative sequence analysis [91]

The consensus structure for a set of aligned sequences was computed using the program `hxmatch` as described in detail in section 5.1. Our dataset comprises 13 archaeal and 15 bacterial SRP RNA sequences, and `hxmatch` was tested on different subsets of the alignment taken from the Signal Recognition Particle Database [49]. For each set of sequences the number of sequences and the mean pairwise sequence identity of their alignment is listed in Table 6. Dataset $\mathcal{S}1$ comprises both archaeal and bacterial SRP RNA sequences, while $\mathcal{S}2$ contains only bacterial SRP RNA sequences, and $\mathcal{S}3$ and $\mathcal{S}4$ include mere archaeal SRP RNA sequences. The sequences used for the prediction of the consensus structure are listed in A.3. The structure predictions of the alignments containing bacterial SRP RNA sequences ($\mathcal{S}1$ and $\mathcal{S}2$) are compared to the structure of *Escherichia coli (E.coli)*, while for the structure predictions of archaeal sequences (derived from the alignments $\mathcal{S}3$ and $\mathcal{S}4$) the structure of *Halobacterium halobium (Hal.hal.)* is used for reference.

Figure 52 and Figure 53 show the results of the different test settings for SRP RNA. The predicted structures are drawn in the linked diagram representation and show the predicted structure together with the reference structure. Base pairs contained in both structures are colored black, base pairs present only in the reference structure are colored green, and base pairs which are contained only in the prediction of `hxmatch`, but not in the reference structure, are colored red.

Table 6: Prediction for each test case of SRP RNA

SRP RNA									
aln	N	μ	Π^*	c	m	f	c/r	ch/rh	fh
$\mathcal{S}1$	28	0.55	0	71	12	2	85.5	13/16	0
			$2\pi^*$	68	15	2	81.9	12/16	0
			$3\pi^*$	61	22	2	73.5	11/16	0
$\mathcal{S}2$	15	0.81	0	61	22	22	73.5	12/16	5
			$2\pi^*$	59	24	8	71.1	12/16	2
			$3\pi^*$	56	27	5	67.5	9/16	1
$\mathcal{S}3$	13	0.53	0	86	11	20	88.7	17/18	0
			$2\pi^*$	83	14	20	85.6	16/18	0
			$3\pi^*$	83	14	20	85.6	16/18	0
$\mathcal{S}4$	6	0.51	0	85	12	18	87.6	17/18	0
			$2\pi^*$	82	15	18	84.5	16/18	0
			$3\pi^*$	75	22	17	77.3	12/18	0

For each alignment the prediction of the consensus structure was generated with different values of the threshold Π^*. $\pi^* = 2500$ is a constant which is in the order of magnitude of the mean weight $\bar{\Pi}_{ij}$ of all base pairs with positive weight. N is the number of sequences contained in the corresponding alignment, and μ is the average pairwise sequence identity. $\mathcal{S}1$ and $\mathcal{S}2$ are compared to the structure of *Bac.sub.* SRP RNA, $\mathcal{S}3$ and $\mathcal{S}4$ are compared to the structure of *Hal.hal.* SRP RNA. The table gives the number of correctly predicted base pairs c, the number of missing base pairs m, the number of false predicted base pairs f, the percentage of correctly predicted base pairs c/r, the number of correct helices identified in relation to the number of helices of the reference structure ch/rh, and the number of false predicted helices fh.

In examining the predicted and the reference structures, most prediction errors correspond to the assignment of the end of helices. This problem can, in part, be attributed to the comparison of the consensus structure of a set of sequences with the reference structure of a single sequence. The algorithm `hxmatch` allows for a certain number of inconsistent sequences for a given base pair, since on the one hand possible alignment errors have to be taken into account, and on the other hand real variation in the structures of individual sequences may exist. If, for instance, the reference structure contains shorter helices then most of the other sequences of the alignment, the predicted helices are somewhat longer compared to the reference. Therefore the results given in Table 6 compare the prediction with the reference structure not only at the level of base pairs, but the number of correctly and false predicted helices is given as well.

In Figure 52, *l.h.s.*, we compare the `hxmatch` consensus structure of dataset $\mathcal{S}_1$, which contains 28 (archaeal and bacterial) SRP RNA sequences, with the phylogenetically derived structure of *E.coli*. Dependent on the threshold Π^*, 11 to 13 helices out of 16 are correctly predicted, and there are no false positives. While the prediction of dataset $\mathcal{S}_2$ (Figure 52, *r.h.s.*) contains an incorrect helix even with the high threshold $\Pi^* = 3\pi^*$. This is due to the relatively high average pairwise sequence identity, $\mu = 0.81$, of the alignment, so there are less mutations contradicting or confirming base pairs.

For the homologous datasets $\mathcal{S}_3$ and $\mathcal{S}_4$, which contain only archaeal sequences, but have a relatively low average pairwise sequence identity of about 50 %, the `hxmatch` consensus structure (Fig. 53) gives nearly all helices correct (about 85 % base pairs correctly predicted), just one or two helices are missed. Only for dataset $\mathcal{S}_4$, which comprises 6 sequences, using the high threshold $\Pi^* = 3\pi^*$ results in missing 6 helices out of 18, the percentage of correct predicted base pairs in this case is 77.3 %. The predictions of the different test cases for these two datasets do not contain false positive helices.

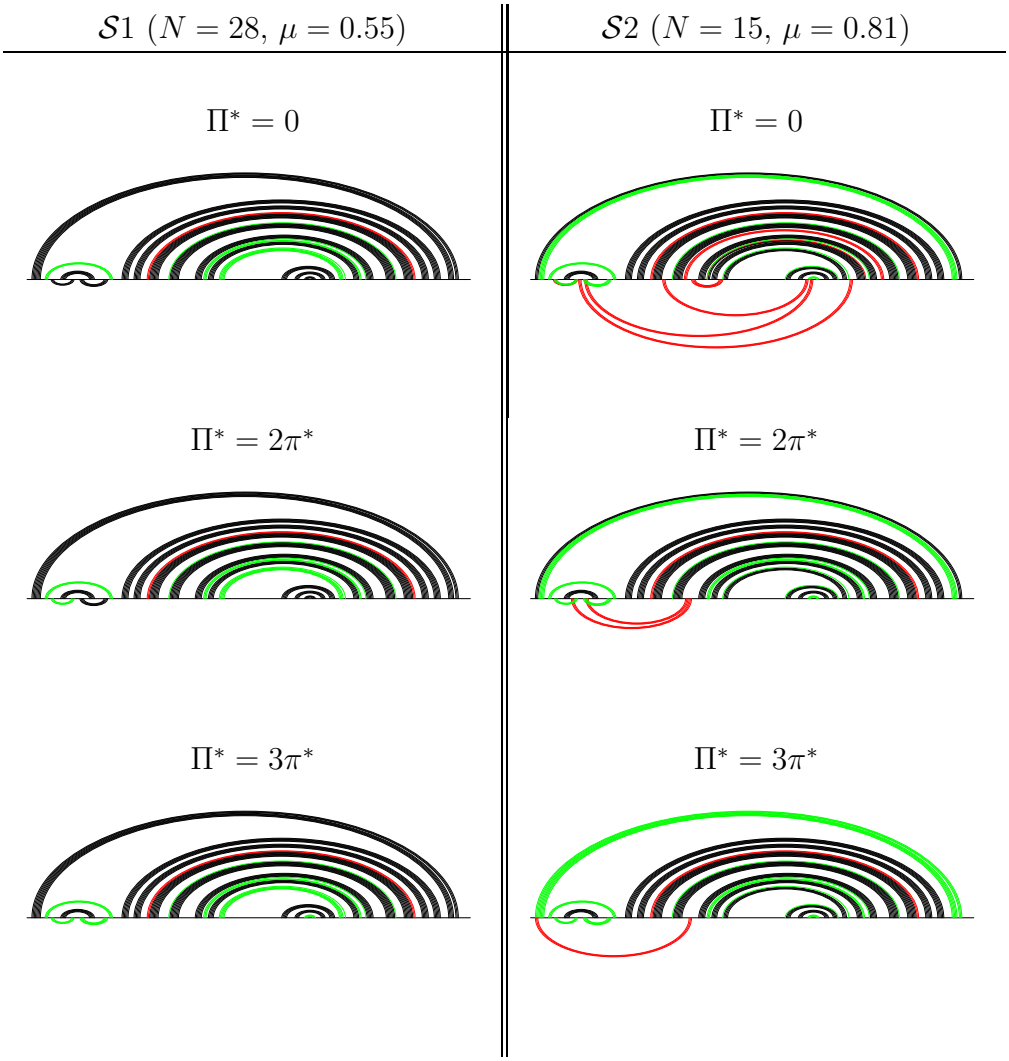


Figure 52: Prediction of the consensus structure for $\mathcal{S}1$ (l.h.s.) and $\mathcal{S}2$ (r.h.s.) generated with different values of the threshold Π^*. The predicted structures are compared to the phylogenetically derived structure of *Bac.sub. SRP RNA* [91]. We use this example to explain the representation of the results: Base pairs predicted by `hxmatch` which are part of the reference structure are colored black, the predicted base pairs not contained in the reference structure are colored red, and base pairs of the reference structure not contained in the prediction are colored green.

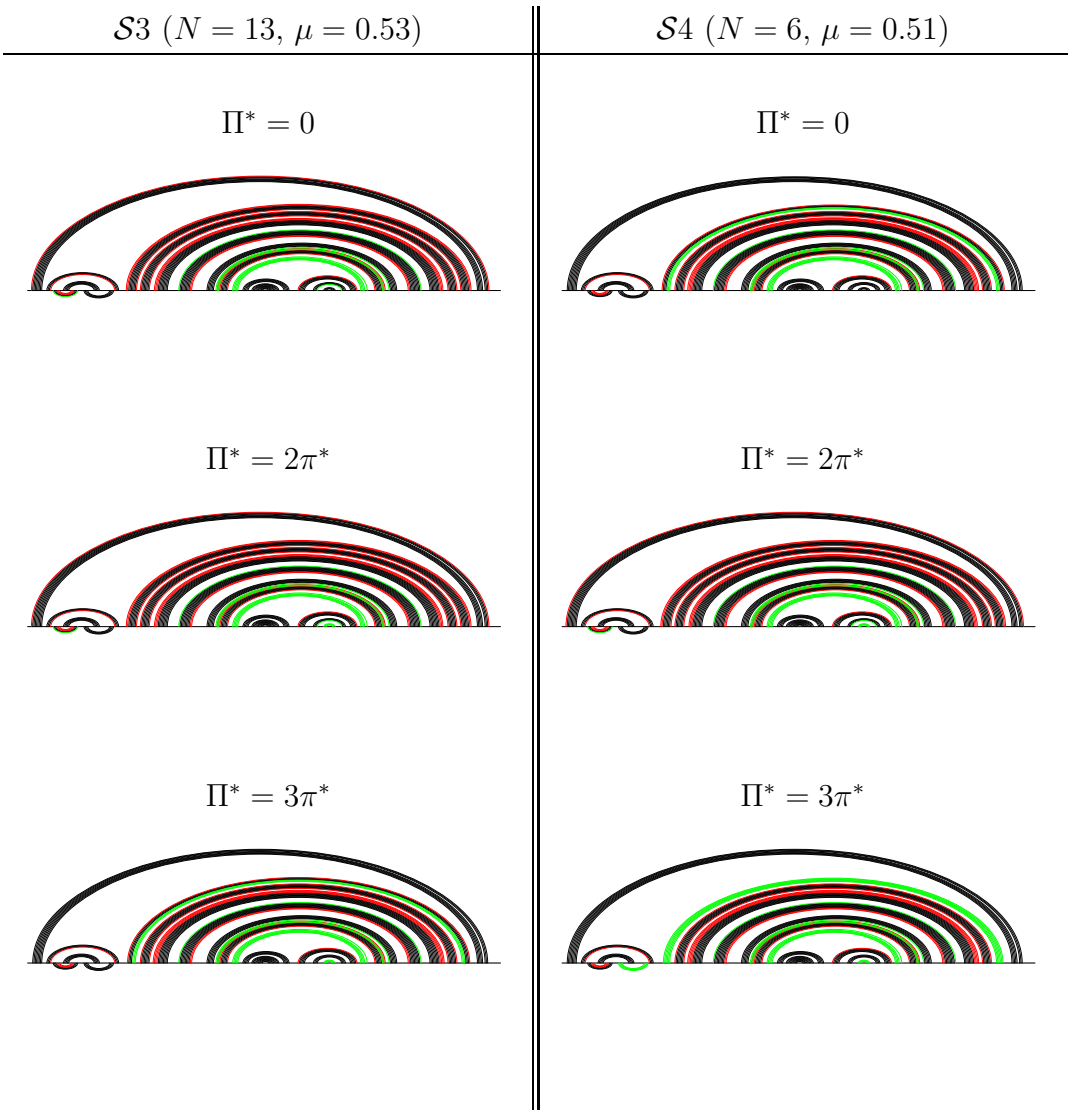


Figure 53: Prediction of the consensus structure for $\mathcal{S}3$ (l.h.s.) and $\mathcal{S}4$ (r.h.s.) generated with different values of the threshold Π^* (for notation refer to Figure 52). The predicted structures are compared to the phylogenetically derived structure of *Hal.hal. SRP RNA* [91].

5.2.2 Ribonuclease P RNA

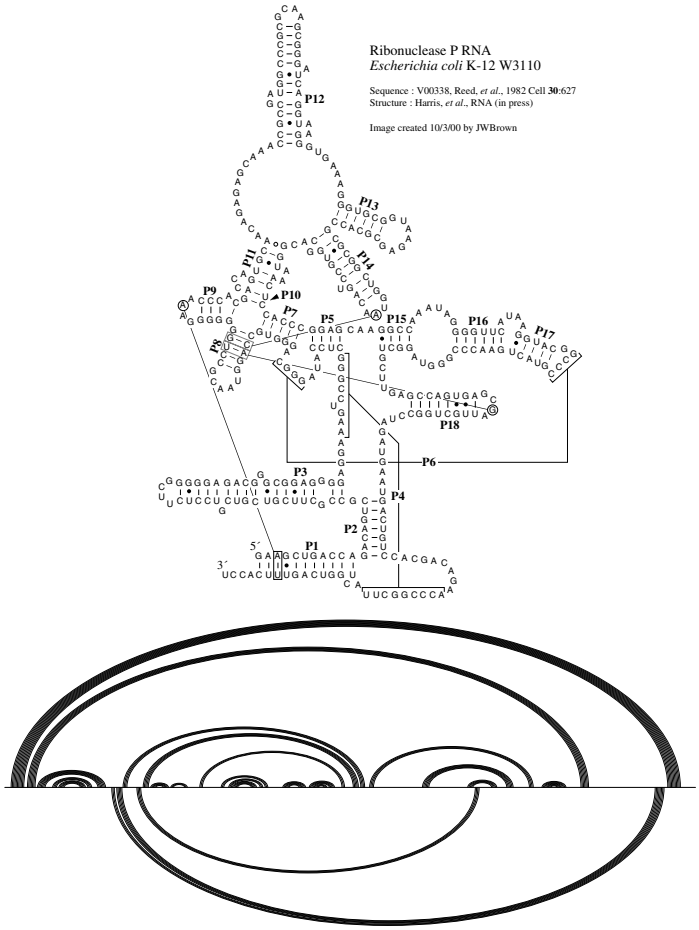


Figure 54: RNaseP RNA structure of *E. coli* derived by comparative sequence analysis [13, 59]

Ribonuclease P (RNase P) is a ribonucleincomplex that catalyzes the removal of leader sequences from precursor tRNA (for review, see e.g. [40]). 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 [51].

The reference structures of *E.coli*, shown in Figure 54, and *Agrobacterium tumefaciens* 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 [13, 59].

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 the each structure type [56]. Dataset $\mathcal{R}1$ comprises sequences of both structure types, while $\mathcal{R}2$ contains only sequences of type A. Due to the structural variation, the structure of one single sequence is only an estimate of the consensus structure. Therefore the designation of correctly predicted base pairs/helices by comparing the `hxmatch` prediction to one reference structure gives only a lower limit of the actual situation.

Table 7: Prediction for each test case of RNaseP RNA

RNaseP RNA									
aln	N	μ	Π^*	c	m	f	c/r	ch/rh	fh
$\mathcal{R}1$	20	0.63	0	80	34	15	70.2	17/23	2
			$2\pi^*$	77	37	8	67.5	15/23	0
			$3\pi^*$	77	37	8	67.5	15/23	0
$\mathcal{R}2$	9	0.61	0	74	36	20	67.3	15/20	2
			$2\pi^*$	74	36	17	67.3	15/20	1
			$3\pi^*$	71	39	11	64.5	14/20	0

For each alignment the prediction of the consensus structure was generated with different values of the threshold Π^* (for notation refer to Table 6). The predicted structure of $\mathcal{R}1$ is compared to the phylogenetically derived structure of *E.coli* RNase P RNA [13, 59], the reference of dataset $\mathcal{R}2$ is *A.tumefaciens* RNase P RNA [12, 59].

The consensus structures predicted by `hxmatch` are based on two different subsets of the alignment taken from the RNase P Database [11] containing 20 and 9 bacterial RNase P sequences, respectively. The sequences contained in the two subsets $\mathcal{R}1$ and $\mathcal{R}2$ are listed in A.3, the average pairwise identity of their alignment is given in Table 7.

The results are summarized in Table 7 and Figure 55. The predictions with lower threshold contain incorrectly predicted helices, whereas in the case of $\Pi^* = 3\pi^*$ no false positives appear, and the two long range pseudoknots are still identified. For dataset $\mathcal{R}1$, which contains 20 sequences, 65 % of the base pairs of the reference structure are identified, the prediction using 9 sequences ($\mathcal{R}2$) assigns 65 % base pairs correctly.

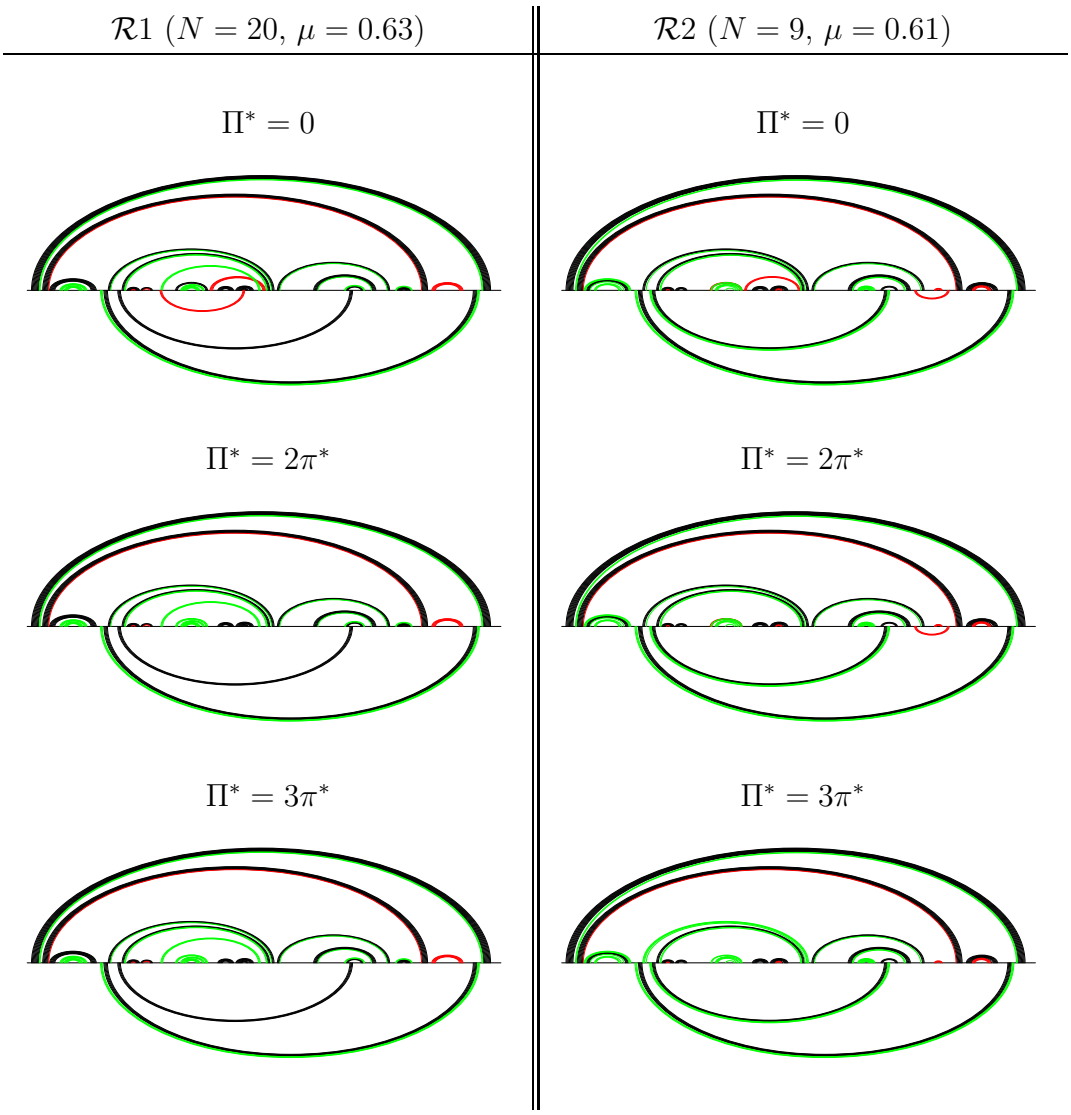


Figure 55: Prediction of the consensus structure for $\mathcal{R}1$ (l.h.s.) and $\mathcal{R}2$ (r.h.s.) generated with different values of the threshold Π^* (for notation refer to Figure 52). The predicted structure of $\mathcal{R}1$ is compared to the phylogenetically derived structure of *E.coli* RNase P RNA [13, 59], the reference of dataset $\mathcal{R}2$ is *A.tumefaciens* RNase P RNA [12, 59].

5.2.3 Transfer-messenger RNA

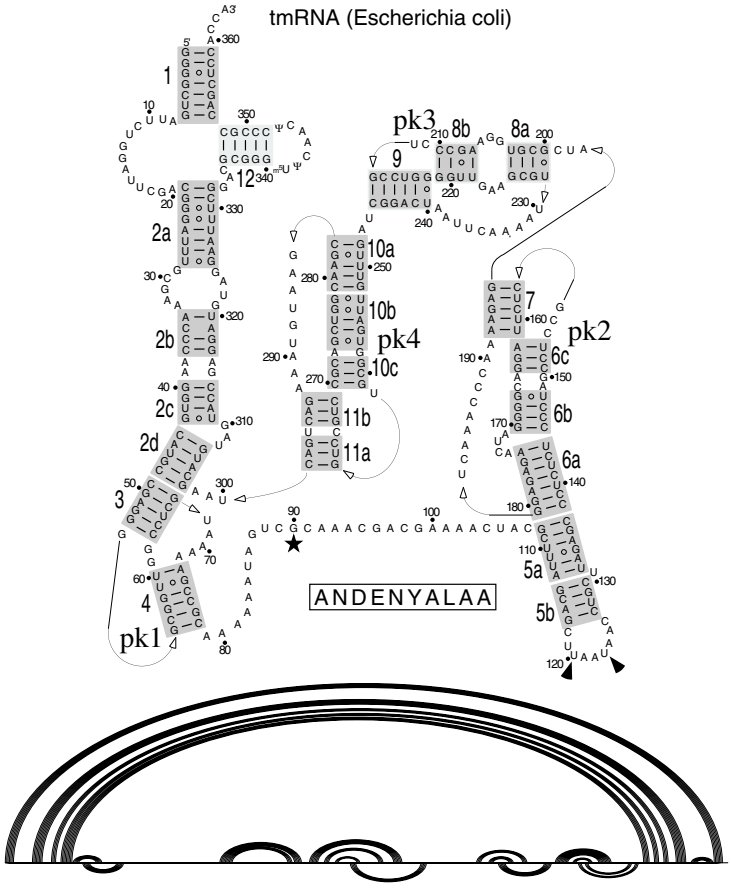


Figure 56: tmRNA structure of *E. coli* derived by comparative sequence analysis [185]

Transfer-messenger RNA (tmRNA), also known as 10SRNA or SsRNA, is a cytoplasmic RNA found in bacteria with properties both of tRNA and mRNA combined in one molecule, reviewed in [113, 174, 171]. No homologous RNA could be identified in archaea and eukaryota. tmRNA is involved in a translational mechanism, termed *trans*-translation, where 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, which signals the proteolytic destruction of the defective protein.

The reference structure shown in Figure 56 was obtained by comparative sequence analysis [185] in an analogous manner as the SRP RNA structure (section 5.2.1), and is based on 50 tmRNA sequences.

The consensus structures predicted by `hxmatch` are based on two different subsets of the alignment taken from the tmRNA Database [87], containing 22 and 8 bacterial tmRNA sequences, respectively. The sequences used are listed in A.3, and Table 8 gives the mean pairwise sequence identity of the alignments and the number of correctly predicted base pairs and helices.

Table 8: Prediction for each test case of tmRNA

tmRNA									
aln	N	μ	Π^*	c	m	f	c/r	ch/rh	fh
$\mathcal{T}1$	22	0.66	0	84	22	17	79.2	18/21	2
			$2\pi^*$	75	31	11	70.8	16/21	1
			$3\pi^*$	68	38	7	64.2	14/21	0
$\mathcal{T}2$	8	0.60	0	80	26	19	75.5	16/21	3
			$2\pi^*$	79	27	11	74.5	16/21	1
			$3\pi^*$	65	41	3	61.3	12/21	0

For each alignment the prediction of the consensus structure was generated with different values of the threshold Π^* (for notation refer to Table 6). The predicted structures are compared to the phylogenetically derived structure of *E.coli* tmRNA[185].

Figure 57 shows the comparison of the `hxmatch` results with the reference structure for the two subsets of the alignment with different settings of the threshold Π^*. While the prediction with zero threshold identifies all four pseudoknots correctly, there are also incorrect helices present. With threshold $\Pi^* = 3\pi^*$ no incorrect helices are contained in the `hxmatch` prediction, and 14 (dataset $\mathcal{T}1$) and 12 helices (dataset $\mathcal{T}2$) out of 21 are identified, including three of the four pseudoknots. This corresponds to about 63 % correctly assigned base pairs.

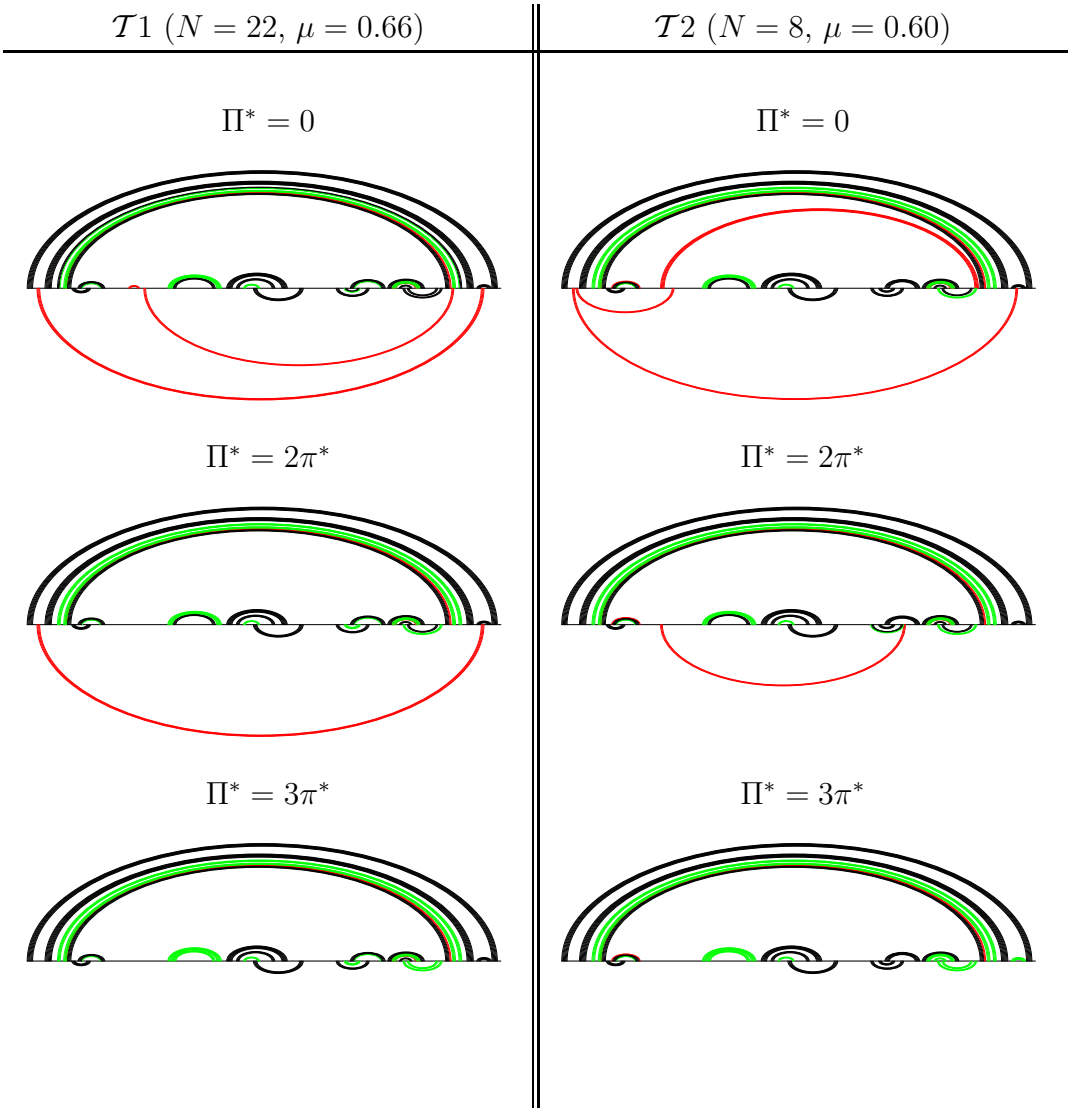


Figure 57: Prediction of the consensus structure for $\mathcal{T}1$ (l.h.s.) and $\mathcal{T}2$ (r.h.s.) generated with different values of the threshold Π^* (for notation refer to Figure 52). The predicted structures are compared to the phylogenetically derived structure of *E. coli* tmRNA[185].

5.2.4 Prediction Based on ClustalW Alignment

All examples presented so far used the manually edited alignments taken from databases. In addition we tested the algorithm on automatically generated alignments, which were calculated using `ClustalW` [161]. While the number of correctly predicted helices drops significantly, no incorrect helices are predicted with threshold $\Pi^* = 3\pi^*$ for most of the examples. Only for two datasets, $\mathcal{S}4$ and $\mathcal{R}2$, the prediction contains one false positive helix, even with the high threshold $\Pi^* = 3\pi^*$.

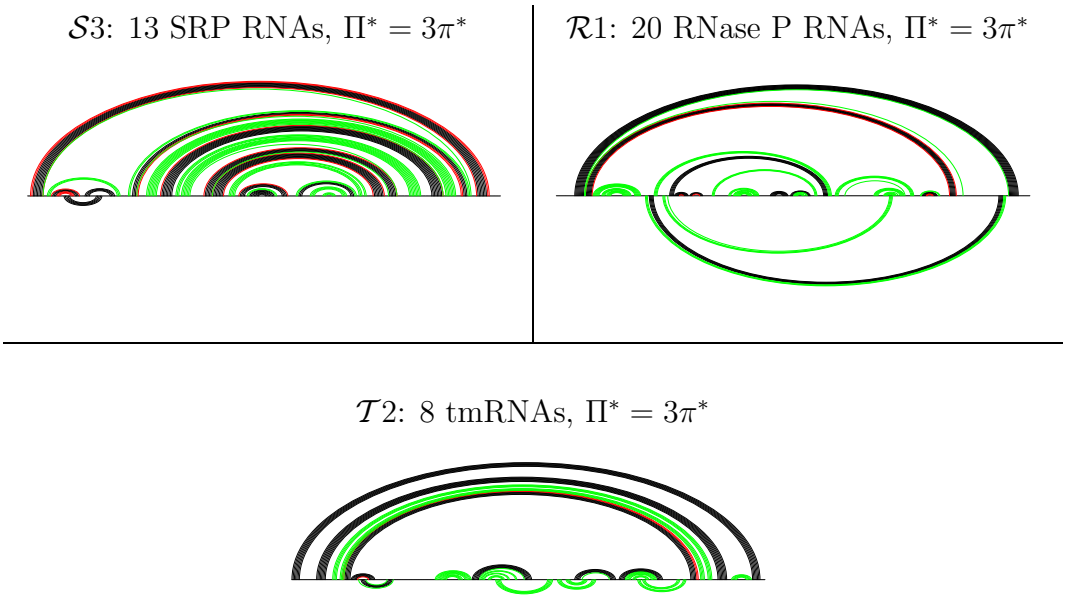


Figure 58: Prediction of the consensus structure for datasets $\mathcal{S}3$ (*top, l.h.s.*), $\mathcal{R}2$ (*top, r.h.s.*), and $\mathcal{T}2$ *(bottom)*, based on the automatically alignment generated using `ClustalW` (for notation refer to Figure 52)

Figure 58 shows the results of three examples: The prediction based on 13 SRP RNA sequences identifies 11 helices out of 18, compared to 16/18 based on the SRP-DB alignment. The `hxmatch` structure derived of the ClustalW alignment of dataset $\mathcal{R}1$, which contains 20 RNase P RNAs, finds 9 helices out of 23, while the structure derived from the RNaseP-DB alignment

assigns 15 helices correctly. The automated alignment based on dataset $\mathcal{T}2$ (8 tmRNAs) leads to the correct prediction of 8 helices out of 21, whereas 14 correct helices are found with the manually generated alignment of the tmRNA-DB.

5.2.5 The Role of the MWM Algorithm

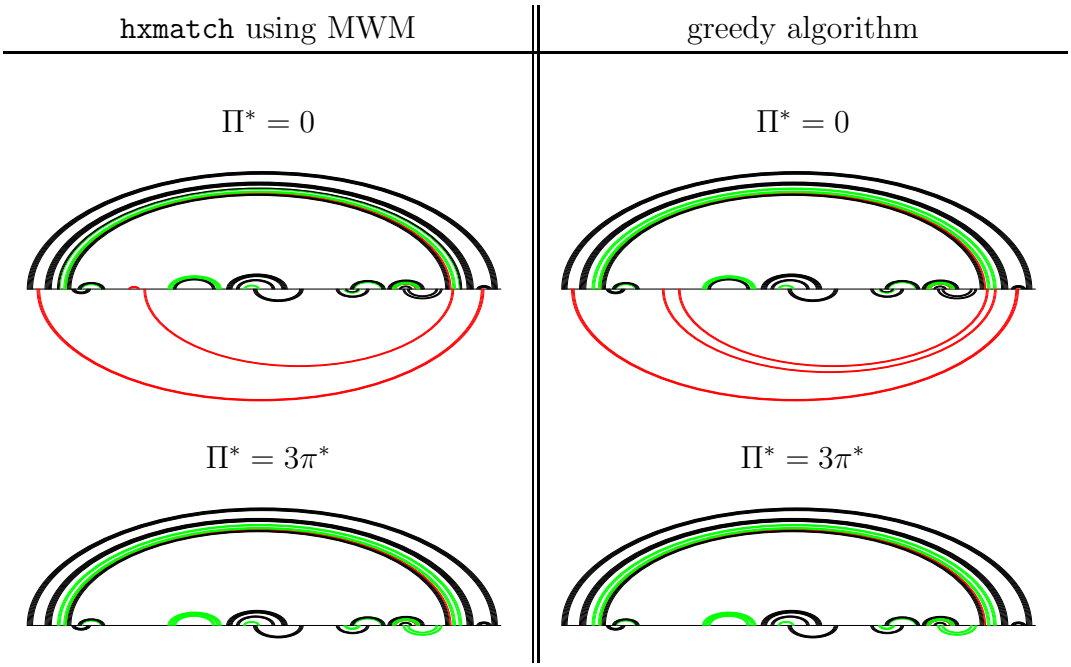


Figure 59: Prediction of the consensus structure of tmRNA (dataset $\mathcal{T}2$) using `hxmatch` including the MWM algorithm *(l.h.s)* and the greedy algorithm *(r.h.s)* with two different settings of the threshold Π^* (for notation refer to Figure 52). The consensus structures are compared to the phylogenetically derived structure of *E.coli* tmRNA.

To analyze the effect of the MWM algorithm, the core of `hxmatch` was modified. Instead of selecting the base pairs with the help of the MWM algorithm, a greedy algorithm was implemented. The scoring of base pairs is identical to the original program. Subsequently the base pairs are sorted by descending weight, and we run through the sorted list and remove all base pairs that

conflict with a higher ranked base pair with respect to the conditions for a bi-secondary structure.

In Figure 59 the structure of tmRNA (dataset $\mathcal{T}2$, 8 sequences) predicted by the original `hxmatch` program is compared to the prediction using the greedy algorithm, with two different values of the threshold Π^*. As can be seen easily, there are minor differences in the predictions using zero threshold, however, the predicted structures with threshold $\Pi^* = 3\pi^*$ are identical. This result is also found in all other examples of the preceding sections (data not shown).

Figure 60 shows the outcome of `hxmatch` using the MWM algorithm (*l.h.s*) compared to the greedy algorithm (*r.h.s*), based on the tmRNA sequence of *E.coli*. The score of a base pair, when using a single sequence, is reduced to the stacking energy of the longest helix the base pair is part of. It is clear, that both variants of the algorithm fail, because the information contained in the stacking energies is not sufficient for a prediction.

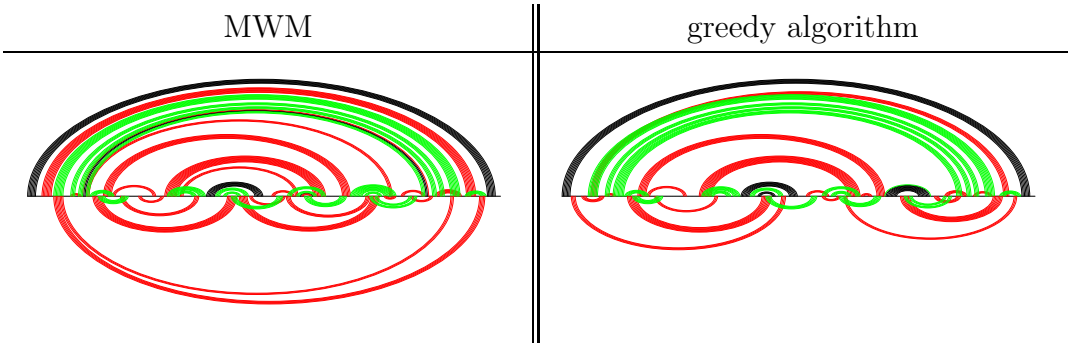


Figure 60: Structure prediction of `hxmatch` using the MWM algorithm (*l.h.s*) compared to the greedy algorithm (*r.h.s*), based on the tmRNA sequence of *E.coli*. Both results are compared to the phylogenetic generated structure of *E.coli* tmRNA (for notation refer to Figure 52).

The MWM algorithm finds the optimal solution in terms of maximizing the total score, which is 6548 (arbitrary units) using the variant based on MWM. The total score using the greedy algorithm is 6456. However, comparing the results of the two variants with the accepted structure of *E.coli* tmRNA,

shows, that the quality of the results is very similar in correctly predicting three and four helices, respectively.

5.2.6 Comparison to Other Methods

The paper of Tabaska *et al.* [156] describes an algorithm for secondary structure prediction based on an alignment. Their program `imatch` [2] has been applied to an alignment of 33 eubacterial and archaebacterial SRP RNA sequences, and almost complete agreement with the phylogenetic derived structure has been found.

To compare our results to the method of Tabaska *et al.* we calculated the consensus structure using `hlxplot` and `imatch`. The predictions are based on the same datasets as used for the `hxmatch` predictions using the alignment of the corresponding databases. The first variant of output filtration included removal of helices with length smaller than three and restriction to a bi-secondary structure.

The percentage of correctly predicted base pairs for dataset $\mathcal{S}1$, containing 28 SRP RNA sequences, obtained by `imatch` (84.5%) is comparable to the result of `hxmatch` (85.5%), but the first pseudoknot helix, which is present in the `hxmatch` result using zero threshold, is not contained in the outcome of `imatch`, see Figure 61.

Another example analyzed is based on dataset $\mathcal{T}2$, containing 8 tmRNA sequences. With the output filtration given above, 61.3% base pairs are predicted correctly, and 2 incorrect helices are predicted, versus 75.5% correct base pairs and 3 incorrect helices in the `hxmatch` outcome. A different output filtration method suggested by Tabaska *et al.* removes rematched vertices (refer to Section 2.5), which results in 48.1% correct base pairs and 2 incorrect helices predicted, and application of an offset, respectively. Ap-

[2] `ftp://cshl.org/pub/science/mzhanglab/tabaska/`

plying an offset of -80 to the initial weight of each base pair gives 19.8% base pairs correct, and 1 incorrect helix is still contained in the prediction. The corresponding results are presented in Figure 61.

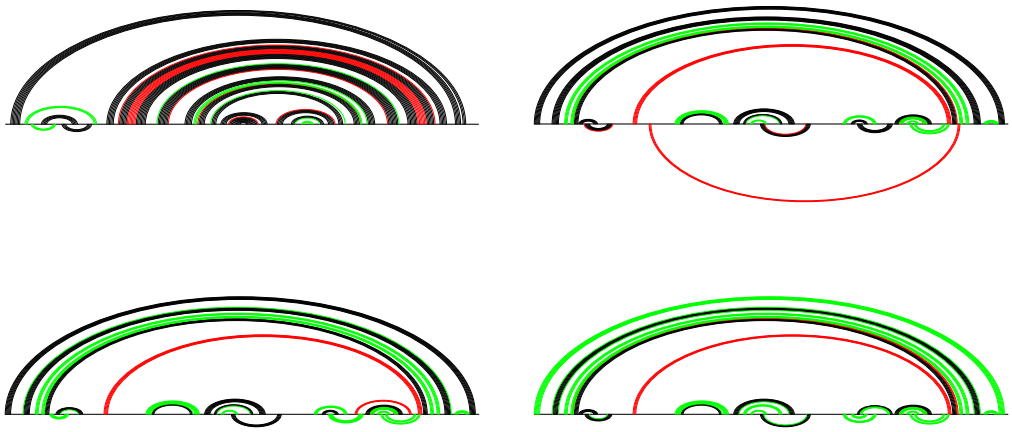


Figure 61: Prediction of the consensus structure by `imatch`: The predicted consensus structure of SRP RNA (dataset $\mathcal{S}1$) using the first variant of output filtration (*top, l.h.s.*). The predicted consensus structures of tmRNA (dataset $\mathcal{T}2$) using the first variant of output filtration (*top, l.h.s.*), additionally excluding rematched vertices (*bottom, l.h.s.*), and applying an offset of -80 (*bottom, r.h.s.*). The predicted structure of SRP RNA is compared to the phylogenetically derived structure of *Hal.hal. SRP RNA*, the predicted structures of tmRNA are compared to the phylogenetically derived structure of *E.coli*.

Another method for prediction of consensus structure including pseudoknots has been developed by Juan and Wilson [83], refer to section 3.4.2. Our dataset $\mathcal{T}2$ is identically to the tmRNA test set used in the work of Juan and Wilson. Applying their method, 12 out of 19 helices are correctly predicted for this dataset, which is the same as for `hxmatch` using threshold $\Pi^* = 3\pi^*$. However, their prediction contains a number of false positive helices, and identifies only one of the four H-pseudoknots contained in the accepted structure of tmRNA, while `hxmatch` identifies all four pseudoknots correctly, when using zero threshold, and three pseudoknots, when threshold $\Pi^* = 3\pi^*$ is used.

5.2.7 Alternative Thermodynamic Score

Another score based on thermodynamic models would be, for instance, the base pairing probabilities calculated by RNAfold, since base pairing probabilities often contain hints to the existence of pseudoknots in the form of competing stems. However, base pairs contained in pseudoknots appear, in general, only with very low probabilities, and often nucleotides involved in pseudoknot formation are part of a base pair of the predicted mfe secondary structure. For instance, the base pairing probability of the terminal base pair in the first pseudoknot of *E.coli* tmRNA (sequence positions 53-63) is 0.003, and the mfe secondary structure contains a base pair at sequence positions 36-53 whose base pairing probability is 0.95. Therefore an algorithm based on the base pairing probabilities could not identify this pseudoknot.

6 Conclusion and Outlook

Structural genomics, the systematic determination of all macro-molecular structures represented in a genome, is at present focused almost exclusively on proteins. Over the past two decades it has become clear, however, that a variety of RNA molecules have important, and sometimes essential, biological functions beyond their roles as rRNAs, tRNAs, or mRNAs. To comprehensively understand the biology of a cell, it will ultimately be necessary to know the identity of all encoded RNAs, the molecules with which they interact and the molecular structures of these complexes [30]. Viral RNA genomes, because of their small size and the strong selection that acts upon them, are an ideal proving ground for techniques that aim at identifying functional RNA *structures*.

A combination of structure prediction based on the thermodynamic rules of the "standard energy model" for nucleic acid secondary structures and the evaluation of consistent and compensatory mutations can be employed for scanning complete viral genomes for functional RNA structure motifs. This work gives a detailed, comprehensive survey of such structural features for those seven (out of nine) genera of the family *Picornaviridae* for which sufficient sequence information is currently available: *Aphthovirus*, *Cardiovirus*, *Enterovirus*, *Hepatovirus*, *Parechovirus*, *Rhinovirus*, and *Teschovirus*.

The 5'-region 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. 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'NTR.

On the other hand, there is a large number of secondary structure elements that have not been described before, most importantly within the coding region. Most notably, we have been able to identify likely or at least possible candidates for the CRE region in *Aphthovirus*, *Hepatovirus*, *Rhinovirus-A* and *Teschovirus*, apart from recovering the known locations of the CRE in *Enterovirus*, *Cardiovirus*, and *Rhinovirus-B*. Only for *Parechovirus* we did not find a significant signal.

The approach used here goes beyond search software such as `RNAMOT` [42] in that it does not require any *a priori* knowledge of the functional structure motifs and it goes beyond searches for regions that are thermodynamically especially stable or well-defined [80] in that it returns a specific prediction for a structure if and only if there is sufficient evidence for structural conservation. The results collected here (see A.2 and 4.2.2) could be used to refine descriptors e.g. for the CRE that can then be used for structure-specific scans in other RNAs.

While there exist fully developed algorithms for the prediction of consensus secondary structures when pseudoknots are forbidden, the prediction of secondary structures including pseudoknots still relies on comparative sequence analysis which requires a large set of related sequences. In the second part of this thesis we described a computational method which combines covariational and thermodynamical information to predict a consensus bi-secondary structure from a smaller set of homologous sequences.

The program `hxmatch`, developed during this thesis, is based on the MWM algorithm, like the method of Tabaska *et al.* [156], but uses an improved scoring function and an elaborated postprocessing. `Hxmatch` relies on the contribution from each base pair being independent, since the MWM algorithm does not consider any dependencies between edges. Therefore the thermodynamic information is rudimentary, since only the stabilizing effect of stacking pairs can be included, but the destabilizing effects of interior loops, hairpin loops and multiloops are neglected completely. So the infor-

mation obtained from consistent and compensatory mutations, reflected by the covariation score, is of utmost importance. The thermodynamic score as implemented in `hxmatch` brings a minor contribution to the total score.

`Hxmatch` was tested on three different types of RNA known to contain pseudoknots: Signal Recognition Particle RNA (SRP RNA), Ribonuclease P RNA (RNaseP RNA), and tmRNA. Given an alignment of homologous sequences, whose mean pairwise sequence identity is smaller than 65%, the consensus structure of these RNA sequences is predicted. With a sufficiently large threshold applied to the scores of the individual base pairs, no false positive helices are predicted, and 60-85% of the base pairs of the phylogenetically derived reference structure are identified, even from datasets containing only six sequences.

At least with our scoring procedure, the usage of the MWM algorithm does not improve the quality of the results compared to selecting the base pairs by a greedy algorithm.

A conceivable improvement of the thermodynamic score could include (small) bulges and interior loops in the calculation of helix energies, as well as tetraloop bonuses. Furthermore a penalty for long range interactions, approximating the entropic cost could enhance the quality of the thermodynamic score.

A Appendix

A.1 List of Picornavirus Sequences

The schematic drawings in Fig. 37 are obtained from a typical strain with the strictly conserved structural features indicated by shadings. Here we give the reference sequences, alternative nomenclature where available, and the exact sequence positions of the outermost base pair of each of the indicated elements.

Genus *Aphthovirus*: FMDV, strain C3Arg85 (Acc.No. AJ007347) [144]:
A1 2-367, A2 (D) 587-640, A3 (H) 648-703, A4 (Ib) 769-846, A5 (J,K) 924-1033, A6 (N) 1037-1058.
Genus *Cardiovirus*: TMEV, strain DA (Acc.No. M20301) [31, 144, 124]:
C1 (A) 1-86, C2 (D) 524-554, C3 (F) 580-602, C4 (H) 610-680, C5 (Ib) 749-831, C6 (J,K) 909-1020, C7 (M) 1023-1042.
Genus *Parechovirus*: HPeV-1, strain Harris (Acc.No. S45208 L00675) [44]:
P1 (A) 14-67, P2 (D) 157-205, P3 (F) 239-253, P4 261-325, P5 327-373, P6 416-431, P7 452-464, P8 (J,K) 550-661. P6 and P7 are part of Ib.
Genus *Teschovirus*: PTV-11, strain Dresden (Acc.No. AF296096), no S-fragment:
T1 19-166, T2 187-208, T3 212-242, T4 257-372, T5 401-415.
Genus *Hepatovirus*: HAV, strain MBB (Acc.No. M20273) [10]:
H1 (I) 5-37, H2 (II) 49-72, H3 (IV) 349-545, H4 (V) 577-688.
Genus *Enterovirus*: Coxackievirus B, strain 1 Japan (Acc.No. M16560), [94, 144, 180]:
E1 (I) 2-86, E2 (II) 127-165, E3 (III) 200-215, E4 (IV) 240-443, E5 (V) 477-534, E6 535-559, E7 (VI) 583-622, E8 (VII) 625-641.
Genus *Rhinovirus*: strain HRV89 (Acc.No. M16248, A10937), [94, 144, 180]:
R1 (I) 3-85, R2 (II) 128-166, R3 (III) 183-229, R4 (IV) 272-405, R5 422 462, R6 (V) 479-511, R7 536-548, R8 (VI) 582-624.

Table 9: List of aphthoviruses sequences

Aphthovirus				
ID	Acc.No.	Length	Virus	Strain
Species: *Foot-and mouth disease virus* (FMDV)				
AF189157	AF189157	6996	FMDV-O	Geshure (G), Israel
FAN133359	AJ133359	8115	FMDV-C_1	Santa Pau/Spain/70 (rp146)
AF154271	AF154271	7739	FMDV-0	Tau-Yuan TW97
FMV7347	AJ007347	8161	FMDV-C_3	Argentina/85
PIFMDV2	X00871	7804	FMDV-O_1	Kaufbeuren/FRG/66
PIFMDV1	X00429	7107	FMDV-A_{10}	Argentina/61 (A61)
FMDVALF	X74812	7820	FMDV-A_{22}	Azerbaijan/USSR/65
APHA12CDR	M10975	7712	FMDV-A_{12}	119/Kent/UK/32
FDI251473	AJ251473	7774	FMDV-SAT2	Kenya/3/57
APHA10SA	M14409	368	FMDV-A_{10}	Argentina/61
APHO1BFSA	M14408	373	FMDV-O_1	BFS 1860/UK/67
APHSFRAG	L11360	368	FMDV-A_{12}	119/Kent/UK/32
FAMD18531	Y18531	365	FMDV-O	BAK/90
FMDVASF	X74811	380	FMDV-A_{22}	Azerbaijan/USSR/65
FMDVSF5	X83209	370	FMDV-Asia1	India/63/72
Species: *Equine rhinitis A virus* (ERAV)				
ERPROTYP1	X96870	7734	ERAV	PERV

Table 10: List of cardioviruses sequences

Cardiovirus				
ID	Acc.No.	Length	Virus	Strain
Species: *Encephalomyocarditis virus* (EMCV)				
MNGPOLY	L22089	7761	EMCV	Mengo
EMCPOLYP	M81861	7835	EMCV	Ruckert
EMCBCG	M22457	7825	EMCV	B
EMC5ES	K01410	149	EMCV	'Russian'
Species: *Theilovirus*				
TMEPP	M16020	8098	TMEV	BeAn 8386
TMEGDVCG	M20562	8105	TMEV	GDVII
TMECG	M20301	8093	TMEV	DA
TMENCRE	M80887	1100	TMEV	Vl

Table 11: List of hepatoviruses sequences

Hepatovirus				
ID	Acc. No.	Length	Virus	Strain
Species: *Hepatitis A virus* (HAV)				
HAVRNAGBM	X75214	7421	HAV	GBM
AB020569	AB020569	7477	HAV	FH3
AB020567	AB020567	7477	HAV	FH1
AB020565	AB020565	7477	HAV	AH2
AB020564	AB020564	7477	HAV	AH1
SHVAGM27	D00924	7400	SHAV	AGM-27
HAVCOMPL	X83302	7421	HAV	FG
HPACG	M20273	7474	HAV	MBB
HPAACG	K02990	7478	HAV	LA
HPA18F	M59808	7423	HAV	HM-175

Table 12: List of parechoviruses sequences

Parechovirus				
ID	Acc.No.	Length	Virus	Strain
Human parechovirus (HPeV)				
NC-001897	AJ005695	7348	HPeV-2	Williamson
AF055846	AF055846	7352	HPeV-2	86-6760
S45208	L02971	7339	HPeV-1	Harris

Table 13: List of rhinoviruses sequences

Rhinovirus				
ID	Acc.No.	Length	Virus	Strain
Species: *Human rhinovirus A* (HRV-A)				
HRV85		7140	HRV-85	50-525-CV54
HRV89	M16248	7152	HRV-89	41467-Gallo
HRV9		7128	HRV-9	211-CV13
HRVACG	D00239	7133	HRV-1B	B632
HRVPP	L24917	7124	HRV-16	11757
PIHRV2G	X02316	7102	HRV-2	HGP
Species: *Human rhinovirus B (HRV-B)*				
HRV14	K02121	7212	HRV-14	1059 (South Carolina/59)

Table 14: List of erboviruses sequence

Erbovirus				
ID	Acc.No.	Length	Virus	Strain
ERPROTYP2	X96871	8828	ERBV	P1436/71

Table 15: List of kobuviruses sequences

Kobuvirus				
ID	Acc.No.	Length	Virus	Strain
Kobuvirus				
AB010145	AB010145	8251	Aichi virus	A846/88
AB040749	AB040749	8280	Aichi virus	
NC_004421	NC_004421	8374	Bovine kobuvirus	

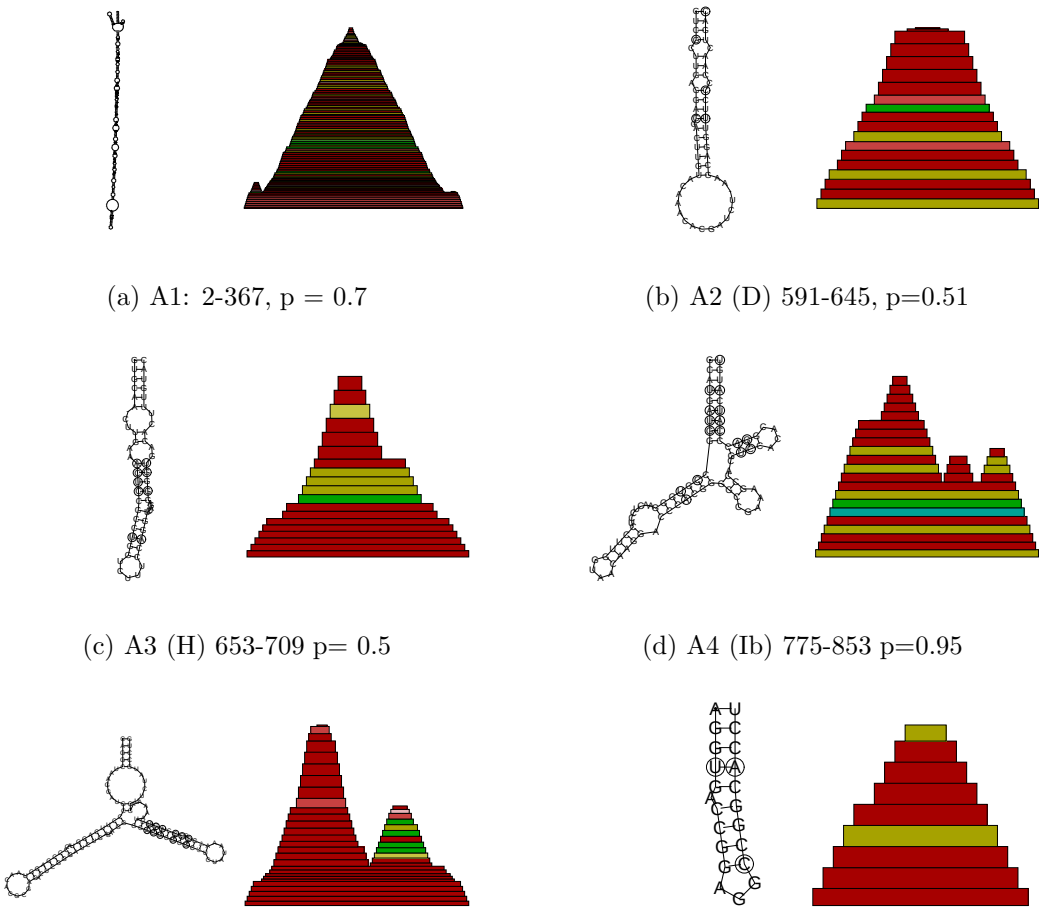
Table 16: List of teschoviruses sequences

Teschovirus				
ID	Acc.No.	Length	Virus	Strain
Species: *Porcine teschovirus* (PTV)				
PEN011380	AJ011380	7117	PTV-1	F65
AF231769	AF231769	7108	PTV-1	Talfan
AF231768	AF231768	7013	PTV-1	Teschen-Konratice
AF296104	AF296104	7112	PTV-1	Vir 1627/89
AF296100	AF296100	7008	PTV-1	DS562/91
AF296102	AF296102	7009	PTV-1	Vir 2236/99
AF296087	AF296087	7017	PTV-2	T80
AF296107	AF296107	7017	PTV-2	Vir 6711-12/83
AF296108	AF296108	7017	PTV-2	Vir 6793/83
AF296109	AF296109	7019	PTV-2	Vir 480/87
AF296088	AF296088	7012	PTV-3	O2b
AF296089	AF296089	7014	PTV-4	PS36
AF296111	AF296111	7015	PTV-4	Vir 918-19/85
AF296112	AF296112	7015	PTV-4	Vir 3764/86
AF296113	AF296113	7015	PTV-4	Vir 2500/99
AF296090	AF296090	7008	PTV-5	F26
AF296091	AF296091	7018	PTV-6	PS37
AF296115	AF296115	7017	PTV-6	Vir 3634/85
AF296117	AF296117	7018	PTV-6	21-SZ
AF296092	AF296092	7014	PTV-7	F43
AF296093	AF296093	7017	PTV-8	UKG/173/74
AF296118	AF296118	7020	PTV-8	25-T-VII
AF296094	AF296094	7006	PTV-9	Vir-2899/84
AF296119	AF296119	7009	PTV-10	Vir 461/88
AF296096	AF296096	7111	PTV-11	Dresden

Table 17: List of enteroviruses sequences

Enterovirus				
ID	Acc.No.	Length	Virus	Strain
Species: *Poliovirus* (PV)				
POLIOS1	V01150	7441	PV-1	Sabin
PIPOLS2	X00595	7439	PV-2	Sabin
POL2W2	D00625	7434	PV-2	W-2
PI3L37	K01392	7431	PV-3	Leon
PIPO3XX	X04468	7435	PV3	23127
Species: *Human enterovirus A* (HEV-A)				
CAU05876	U05876	7413	CV-A16	G-10
AF177911	AF177911	7410	CV-A16	Tainan/5079/98
AF176044	AF176044	7433	EV-71	1245a/98/tw
E722521	U22521	7408	EV-71	BrCr
E722522	U22522	7411	EV-71	MS/7423/87
Species: *Human enterovirus B* (HEV-B)				
CXB9CG	D00627	7452	CV-A9	Griggs
CXB1G	M16560	7389	CV-B1	Japan
AF085363	AF085363	7411	CV-B2	Ohio-1/Ohio/US/47
CV57056	U57056	7400	CV-B3	Woodruff
CXB3G	M16572	7396	CV-B3	Nancy/Connecticut/US/49
PICOXB4	X05690	7395	CV-B4	JVB/New York/US/51
S76772	S76772	7397	CV-B4	E2
CXB5CGA	X67706	7402	CV-B5	1954/UK/85
AF083069	AF083069	7433	E-5	Noyce/Maine/54
E616283	U16283	7417	E-6	Charles
ECHOV9XX	X92886	7451	E-9	Barty
EV9GENOME	X84981	7420	E-9	Hill
EV11VPCD	X80059	7438	E-11	Gregory
EC12TCGWT	X79047	7501	E-12	Travis
AF162711	AF162711	7440	E-30	Bastianni
SVDG	D00435	7401	SVDV	H/3'76
Species: *Human enterovirus C* (HEV-C)				
CXA21	D00538	7401	CV-A21	Coe
CXA24CG	D90457	7461	CV-A24	EH 24/70
Species: *Human enterovirus D* (HEV-D)				
EV70CG	D00820	7390	EV-70	J670/71

A.2 Conserved Structure Elements in Picornaviruses



(a) A1: 2-367, p = 0.7

(b) A2 (D) 591-645, p=0.51

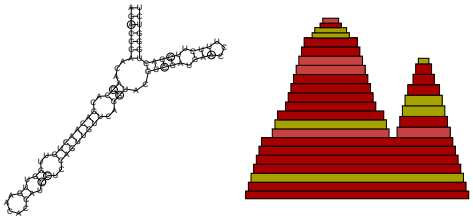
(c) A3 (H) 653-709 p= 0.5

(d) A4 (Ib) 775-853 p=0.95

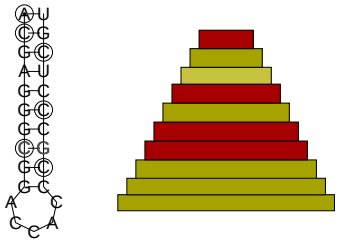
(e) A5 (J, K) 932-1041 p=0.45

(f) A6 (M) 1045-1066 p=0.90

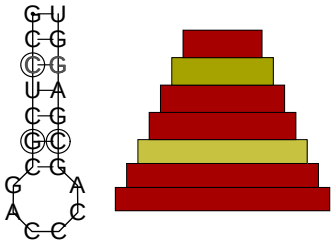
Figure 62: Conserved secondary structure elements in the 5'-noncoding region of aphthoviruses



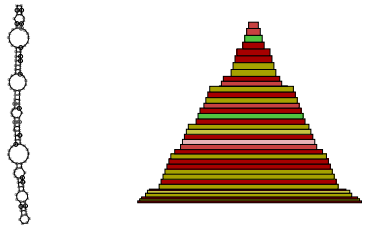
(a) A7: 1218-1302, p = 0.6/0.6/0.8



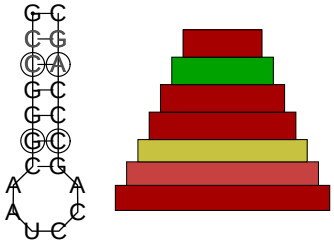
(b) A8: 1372-1396, p = 0.97



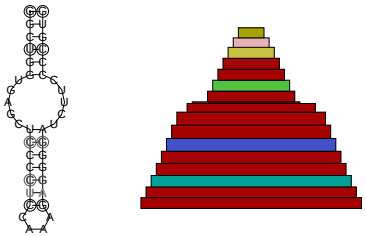
(c) A9: 1440-1459, p = 0.9



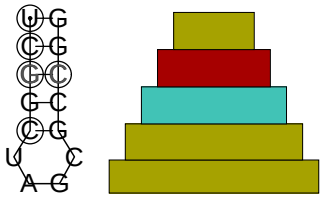
(d) A10: 1549-1664, p = 0.65



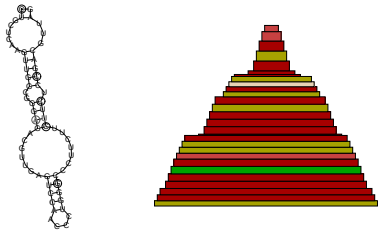
(e) A11: 1695-1714, p = 0.9



(f) A12: 2582-2621, p = 0.8

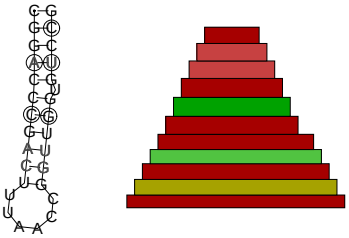


(g) A13: 3220-3233, p = 0.7

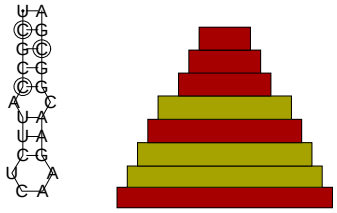


(h) A14: 3943-4004, p = 0.5

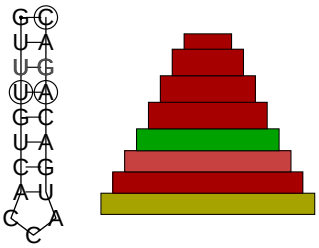
Figure 63: Conserved secondary structure elements in the coding region of aphthoviruses



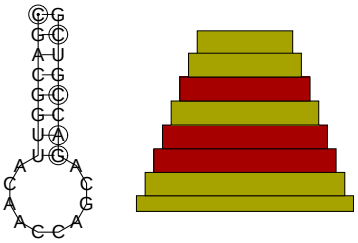
(a) A15: 4068-4096, p = 0.75



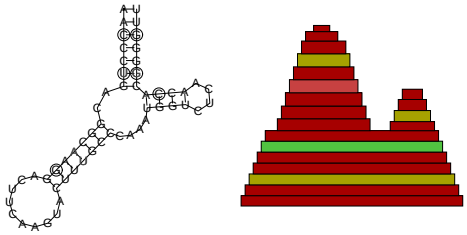
(b) A16: 4472-4493, p = 0.9



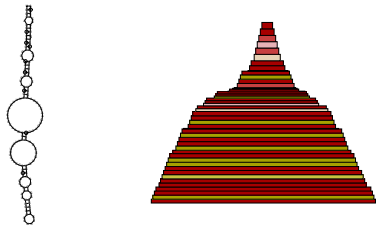
(c) A17: 4560-4578, p = 0.7



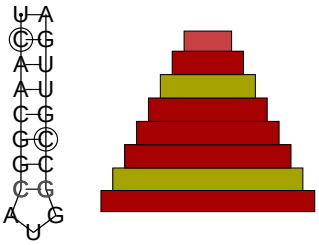
(d) A18: 4887-4912, p = 0.83



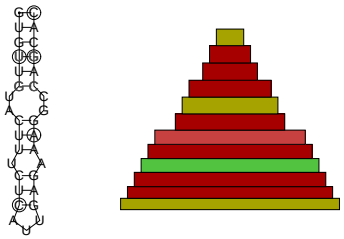
(e) A19: 4936-4991, p = 0.7



(f) A20: 5003-5149, p = 0.43

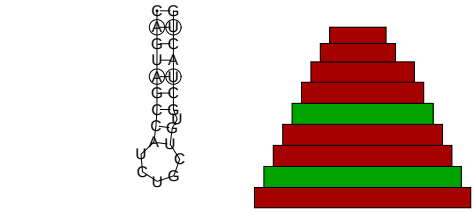


(g) A21: 5249-5267, p = 0.93

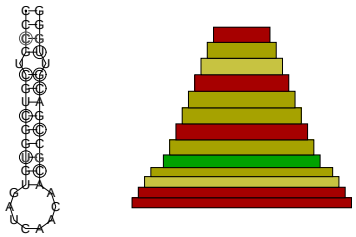


(h) A22: 5422-5454, p = 0.55

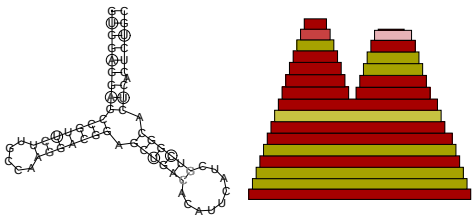
Figure 64: Conserved secondary structure elements in the coding region of aphthoviruses



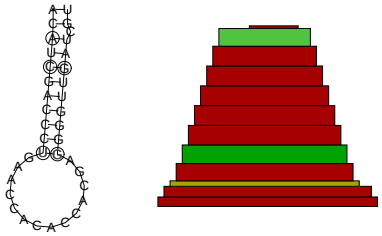
(a) A23: 6149-6172, p = 0.63



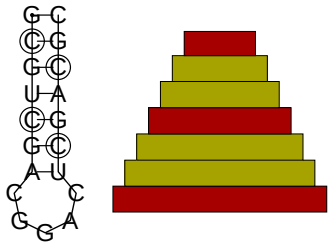
(b) A24: 6412-6447, p = 0.85



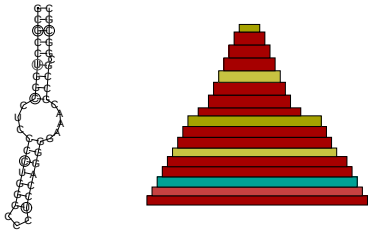
(c) A25: 6560-6620, p = 0.83



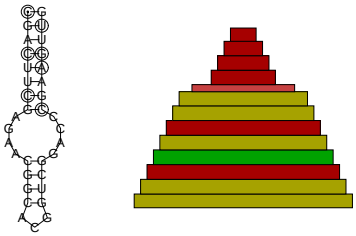
(d) A26: 6686-6722, p = 0.9



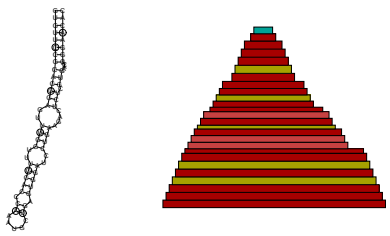
(e) A27: 7022-7041, p = 0.83



(f) A28: 7059-7098, p = 0.7/0.97

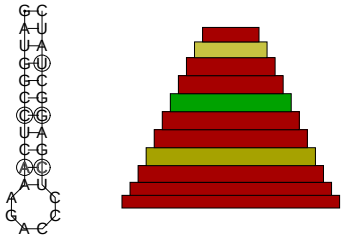


(g) A29: 7107-7141, p = 0.75



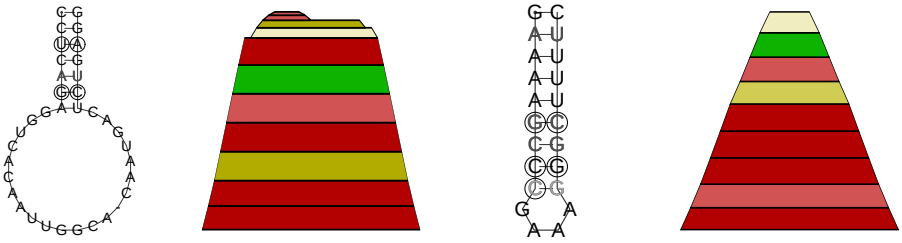
(h) A30: 7489-7560, p = 0.8

Figure 65: Conserved secondary structure elements in the coding region of aphthoviruses



(a) A31: 7917-7944, p = 0.95

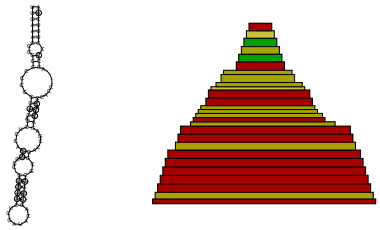
Figure 66: Conserved secondary structure elements in the coding region of aphthoviruses



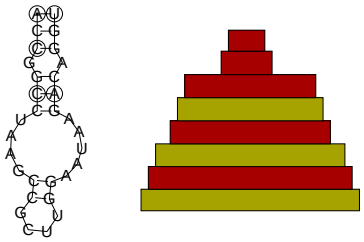
(a) A32: 8127-8162 , p = 0.9

(b) A33: 8127-8162, p = 0.9

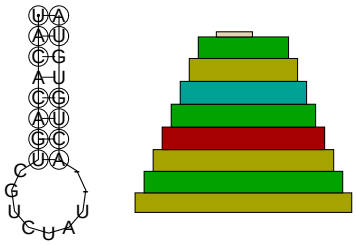
Figure 67: Conserved secondary structure elements in the 3'NTR of aphthoviruses



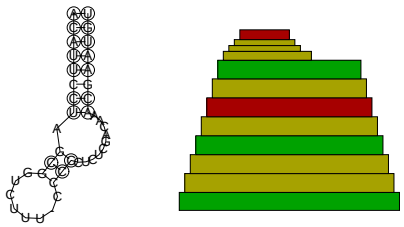
(a) C1 (A) 2-90 p = 0.8



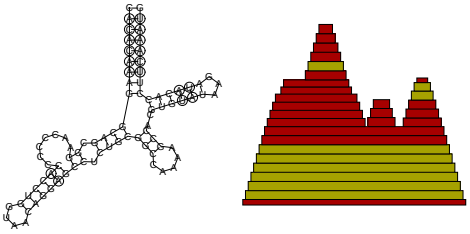
(b) C2 (D) 530-560 , p=0.45



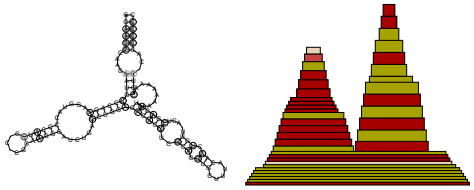
(c) C3 (F) 589-613 p=0.9



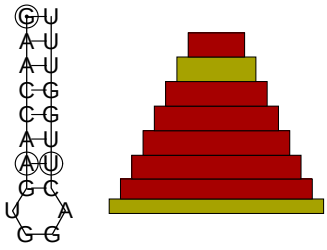
(d) C4 (H) 621-696 p=0.9



(e) C5 (Ib) 765-847 p=0.9

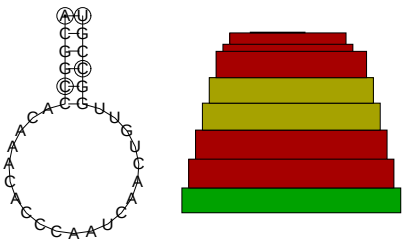


(f) C6 (J, K) 927-1041 p=0.2, p=0.6, p=0.9

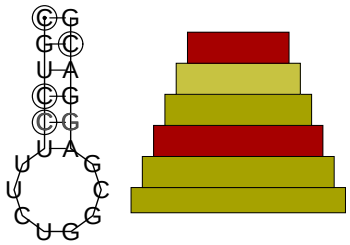


(g) C7 (M) 1044-1063 p= 0.6

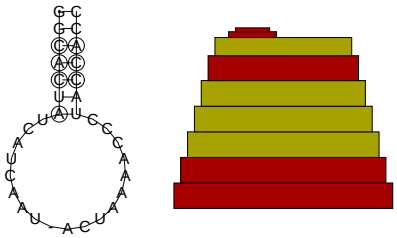
Figure 68: Conserved secondary structure elements in the 5'-ntr of cardioviruses



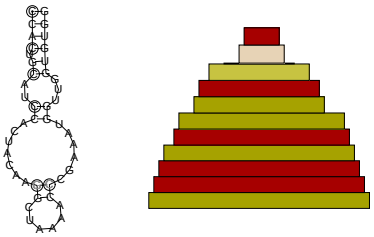
(a) C8: 1594-1596 p=0.65 , CRE



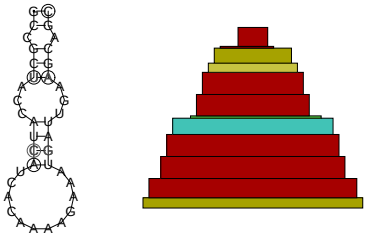
(b) C9: 1788-1807 p=0.5



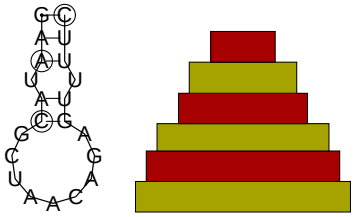
(c) C10: 4351-4383 p=0.65



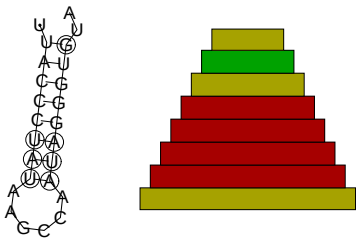
(d) C11: 6528-6572 p=0.43



(e) C12: 6649-6686 p=0.4

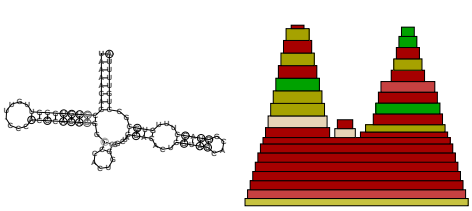


(f) C13: 6937-6957 p=0.45

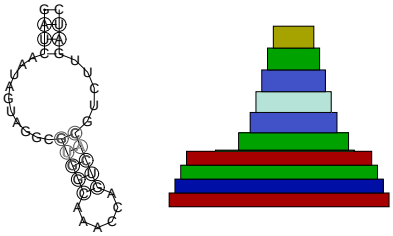


(g) C14: 8132-8153 p=0.9

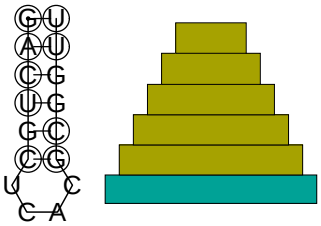
Figure 69: Conserved secondary structure elements in the coding region and 3'-utr (fig. 69(g)) of cardioviruses



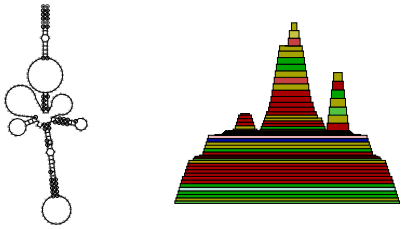
(a) E1 (I) 3-90, p = 0.6/0.9



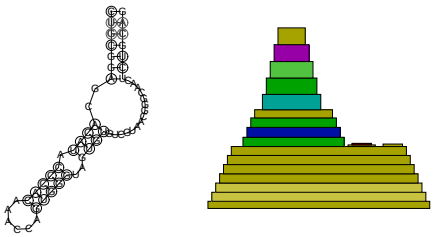
(b) E2 (II) 140-178, p = 0.5/0.7



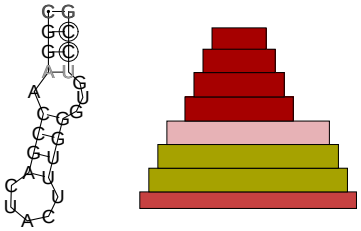
(c) E3 (III) 213-228, p = 0.45



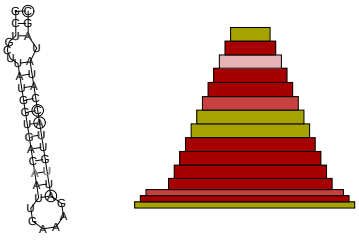
(d) E4 (IV) 253-461, p = 0.4/0.4/0.7/0.85



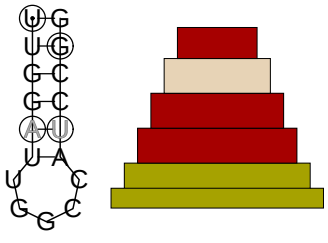
(e) E5 (V) 495-552, p = 0.3/0.7



(f) E6 553-577, p = 0.45

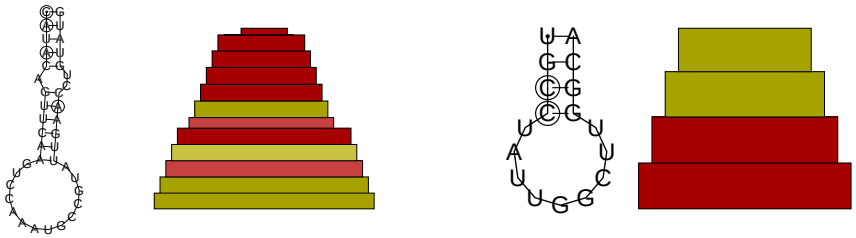


(g) E7 (VI) 601-640, p = 0.9



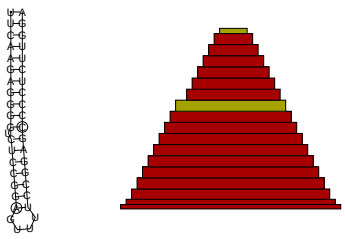
(h) E8 (VII) 644-660, p = 0.2/0.3

Figure 70: Conserved secondary structure elements in the 5'NTR of enteroviruses

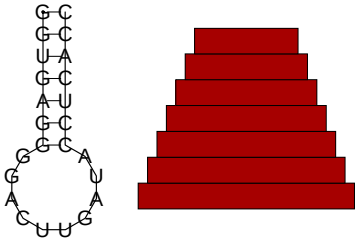


(a) E9: 4617-4618, p = 0.9, CRE (b) E10: 7417-7430, p = 0.45

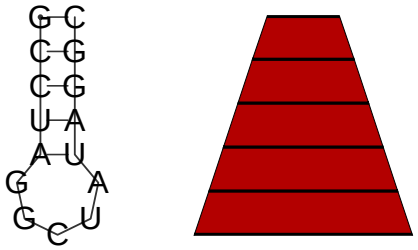
Figure 71: Conserved secondary structure elements in the coding region of enteroviruses



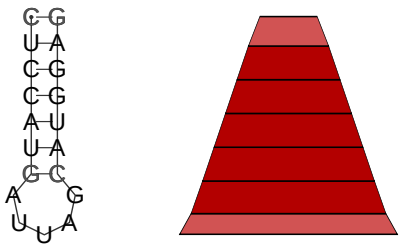
(a) H1 (I) 1-41 p=0.6



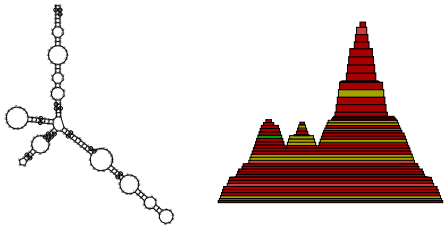
(b) H2 (IIa) 49-72 p=0.8



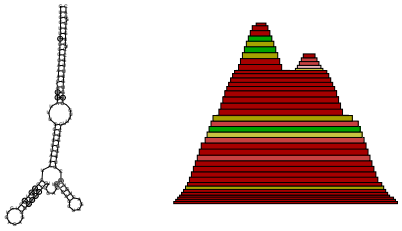
(c) H3 (IIb) p=0.8



(d) H4 p=0.6

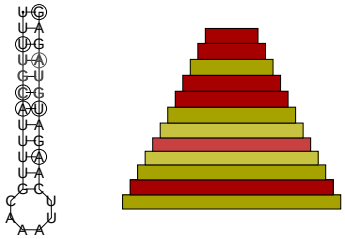


(e) H5 (IV) 372-570 p=0.35, p=0.9

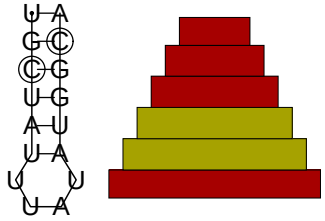


(f) H6 (V) 602-716 p= 0.3, p= 0.6, p=0.9

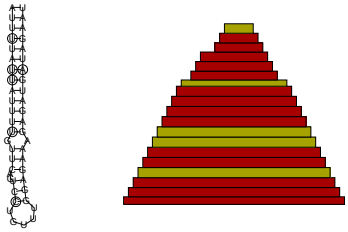
Figure 72: Conserved secondary structure elements in the 5'-noncoding region of hepatoviruses



(a) H7: 4216-4245, p = 0.75, possible CRE

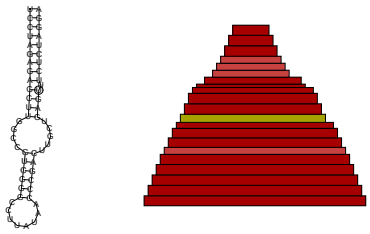


(b) H8: 4438-4453, p = 0.6

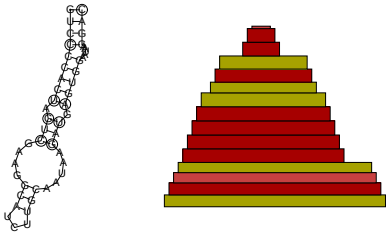


(c) H9: 7332-7378, p = 0.8

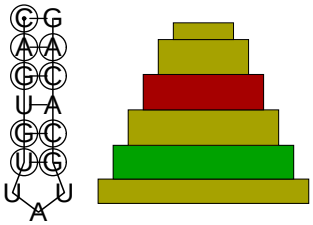
Figure 73: Conserved secondary structure elements in the coding region of hepatoviruses



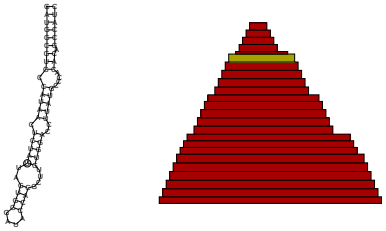
(a) P1 (A) 14-69, p = 0.9



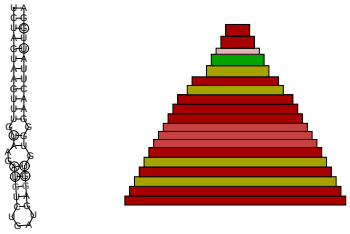
(b) P2 (D) 162-210, p = 0.4



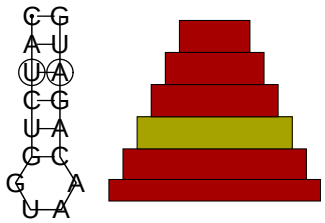
(c) P3 (F) 244-258, p = 0.6



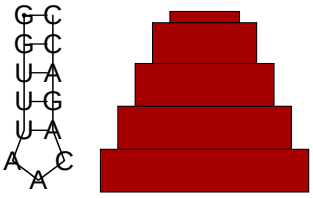
(d) P4 266-330, p = 0.37



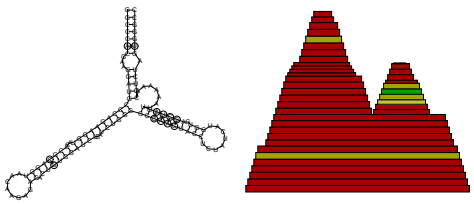
(e) P5 332-378, p = 0.5



(f) P6 (part of Ib) 421-436, p = 0.9

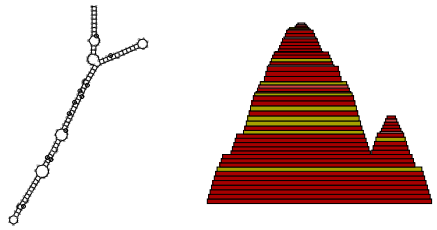


(g) P7 (part of Ib) 457-469, p = 0.4



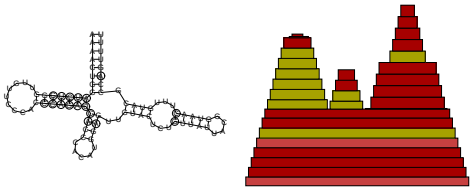
(h) P8 (J, K) 555-667, p = 0.9

Figure 74: Conserved secondary structure elements in the 5'-noncoding region of parechoviruses

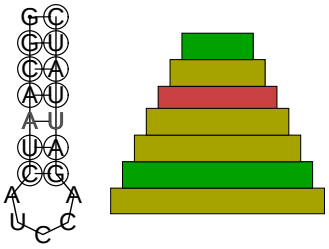


(a) P9: 6925-7086, p = 0.9/0.6

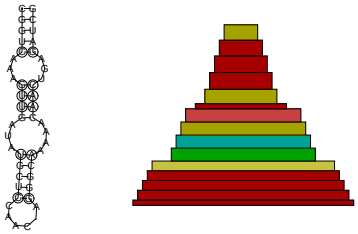
Figure 75: Conserved secondary structure elements in the coding region of parechoviruses



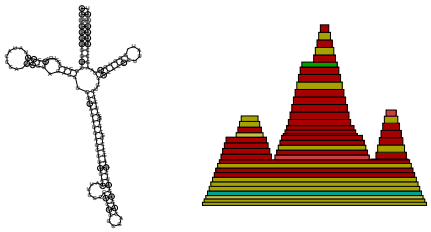
(a) R1 (I) 3-85, p = 0.9



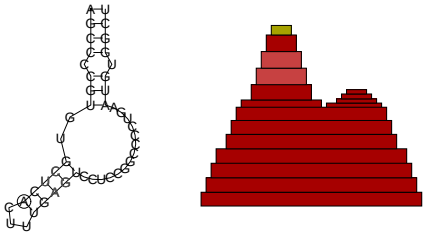
(b) R2 (II) 142-160, p = 0.9



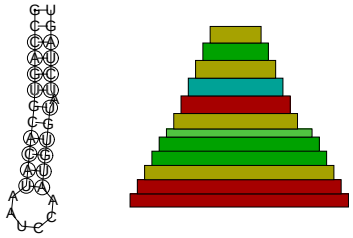
(c) R3 (III) 185-231, p = 0.5/0.8



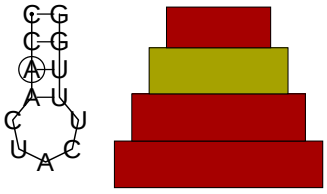
(d) R4 (IV) 275-408, p = 0.6/0.8



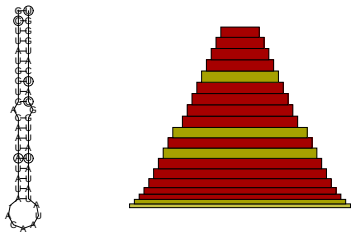
(e) R5: 424-468, p = 0.35



(f) R6 (V): 484-514, p = 0.8

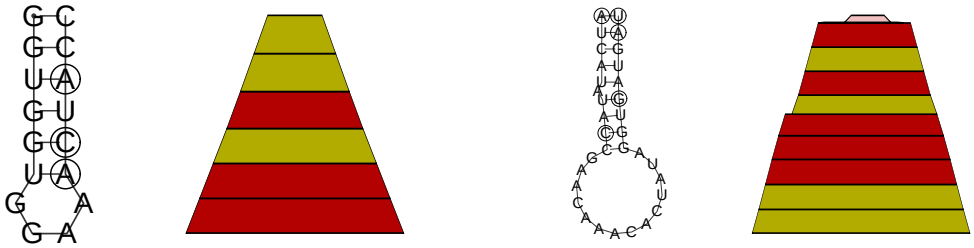


(g) R7: 540-552, p = 0.6



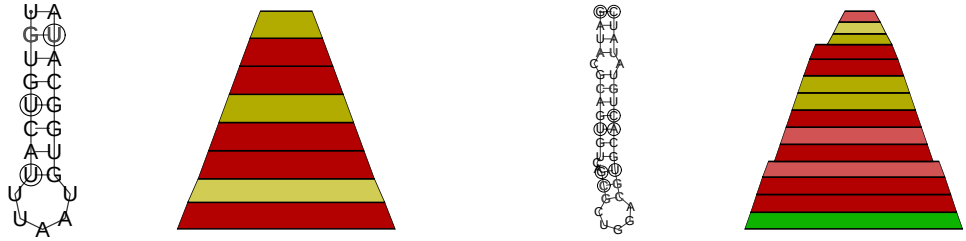
(h) R8: (VI) 593-639, p = 0.9

Figure 76: Conserved secondary structure elements in the 5'NTR of HRVA



(a) R9: 1073-1088

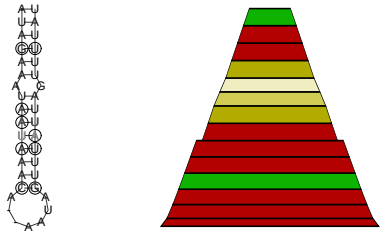
(b) R10: 3345-3377



(c) R11: 7003-7023

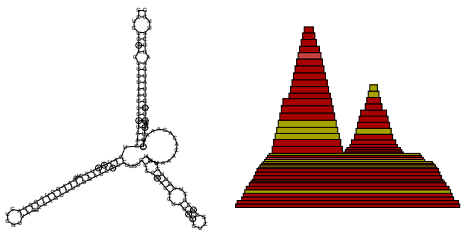
(d) R12: 7058-7094

Figure 77: Conserved secondary structure elements in the coding region of HRVA

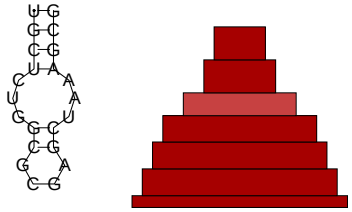


(a) R13: 7151-7187

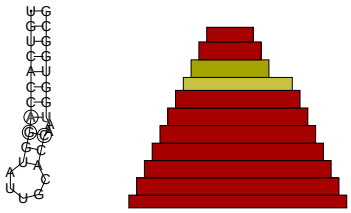
Figure 78: Conserved secondary structure elements in the 3'NTR of HRVA



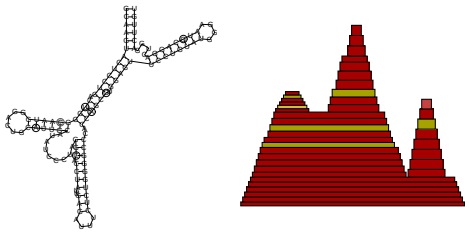
(a) T1 19-166, p = 0.4/0.8/0.7; only 4 seq



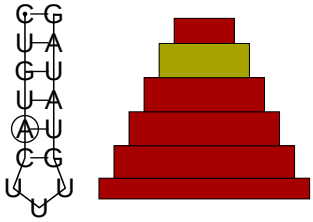
(b) T2 188-209, p = 0.5



(c) T3 214-242, p = 0.5

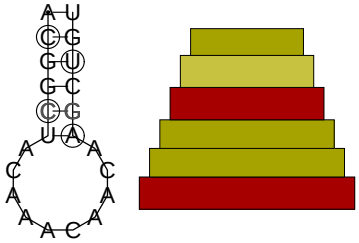


(d) T4 258-373, p = 0.4/0.5/0.3/0.7/0.8

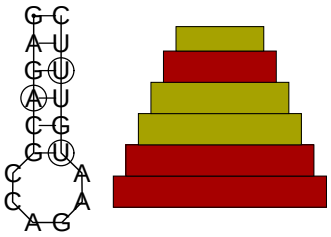


(e) T5 402-416, p = 0.3

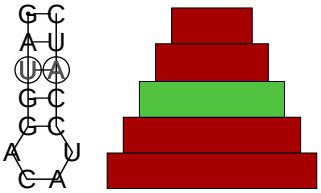
Figure 79: Conserved secondary structure elements in the 5'-ntr of teschoviruses



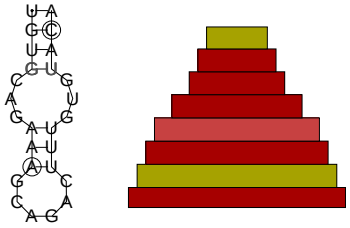
(a) T6: 4245-4266, p = 0.8, putative CRE



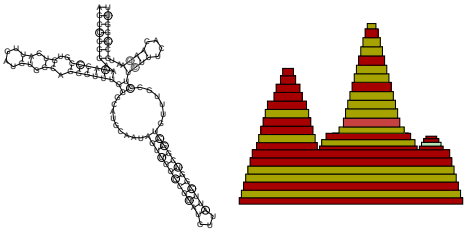
(b) T7: 4693-4710, p = 0.4



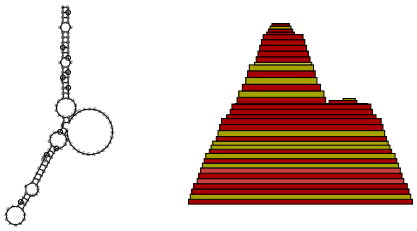
(c) T8: 5320-5329, p = 0.8



(d) T9: 5719-5744, p = 0.6

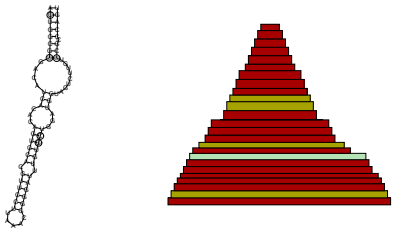


(e) T10: 6211-6314, p = 0.7



(f) T11: 6551-6680, p = 0.7

Figure 80: Conserved secondary structure elements in the coding region of teschoviruses



(a) T12: 6789-6861, p = 0.8

Figure 81: Conserved secondary structure elements in the coding region of teschoviruses

A.3 Sequences Used for Hxmatch Predictions

Bi-secondary structure predictions were calculated using different sets of RNA sequences. Here we list the full species names of the RNA sequences contained in each dataset.

SRP RNA:
$\mathcal{S}1 = \{\mathcal{S}2, \mathcal{S}3\}$
$\mathcal{S}2 = \{$*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*$\}$
$\mathcal{S}3 = \{\mathcal{S}4,$ *Methanococcus voltae, Pyrococcus abyssi, Pyrococcus horikoshii, Pyrodictium occultum, Sulfolobus solfataricus, Sulfolobus solfataricus, Thermococcus celer*$\}$
$\mathcal{S}4 = \{$*Aeropyrum pernix, Archaeoglobus fulgidus, Methanosarcina acetivorans, Methanothermus fervidus, Methanococcus jannaschii, Methanobacterium thermoautotrophicum*$\}$

RNase P RNA:
$\mathcal{R}1 = \{\mathcal{R}2,$ *Acidithiobacillus ferrooxidans, Bacteroides thetaiotaomicron, Clostridium acetobutylicum, Clostridium difficile, Carboxydothermus hydrogenoformans, Campylobacter jejuni, Chromatium vinosum, Escherichia coli, Mycobacterium avium, Prochlorococcus marinus, Vibrio cholerae*$\}$
$\mathcal{R}2 = \{$*Alcaligenes eutrophus, Agrobacterium tumefaciens, Caulobacter crescentus, Corynebacterium diphtheriae, Desulfovibrio desulfuricans, Erwinia agglomerulans, Mycobacterium bovis, Mycobacterium leprae*$\}$

tmRNA:
$\mathcal{T}1 = \{\mathcal{T}2$, *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*$\}$
$\mathcal{T}2 = \{$*Actinobacillus actinomycetemcomitans, Aeromonas salmonicida, Escherichia coli, Haemophilus influenzae, Marinobacter hydrocarbonoclasticus, Pseudomonas aeruginosa, Pseudoalteromonas haloplanktis, Vibrio cholerae*$\}$

A.4 Manual Pages

Hxmatch

NAME

`hxmatch`

SYNOPSIS

`hxmatch [-T` *thresh*`]`

DESCRIPTION

`Hxmatch` reads an alignment file from stdin and calculates the consensus bi-secondary structure. The multiple sequence alignment must be contained in a single file in `ClustalW` format. In particular, the first line of the alignment file must begin with the word CLUSTAL. It writes the consensus structure in bracket notation to a file termed 'name_db.hx'. For each base pair $i - j$ contained in the outcome of the MWM algorithm there is a line of the form

$$i\ j\ wt\ xx\ col\ rm\ comp\ g - g\ mut$$

written to stdout, the abbreviations are as follows:

wt	...	weight of base pair $i - j$
xx	...	0, if the base pair has been removed during postprocessing
		2, if the base pair is the beginning of a helix
		1, if the base pair is in the middle of a helix
		3, if the base pair is the end of a helix
col	...	0, if the base pair belongs to the part of the structure drawn in the upper half plane (secondary structure)
		1, if the base pair belongs to the part of the structure drawn in the lower half plane (pseudoknots)

rm	...	the number of rematching events for base pair $i-j$ during the run of the MWM algorithm
comp	...	the number of sequences compatible with base pair $i-j$
$g-g$	...	the number of sequences containing gaps at positions $i-j$
mut	...	the number of different pairing combinations at position $i-j$ of the alignment

OPTIONS

`-T` *thresh*

Calculate the consensus structure applying threshold $thresh \times \pi^*$, where $thresh \in N$, π^* is in order of magnitude of the average weight of all base pairs, default value for *thresh* is zero.

VERSION

Version 1.0 of `hxmatch`

REFERENCES

The MWM algorithm, which is part of `hxmatch`, has been implemented by Edward Rothberg 7/85 [142], based on H. Gabow's Ph.D. thesis, Stanford Univ. 1973 [41].

AUTHOR

Christina Witwer

Iscolor

NAME

`iscolor`

SYNOPSIS

`iscolor FN`

where `FN` is the path and name of the input file.

DESCRIPTION

`iscolor` finds the IS-coloring on a graph Γ given in `GML` format, and constructs the gel $\mathsf{Gel}(\Gamma)$. The gel and its constituting nets and stems are written in `GML` format.

OPTIONS

`-gel`

The output contains only the gel.

VERSION

Version 1.0 of `iscolor`

REFERENCES

The reading and writing of the files is based on the scanner and parser for the GML file format by M. Himsolt [66, 67] available at: `http://infosun.fmi.uni-passau.de/Graphlet/GML/`.

AUTHOR

Christina Witwer

Splot

NAME
`splot`

SYNOPSIS
`splot FN`
where `FN` is the path and name of the input file.

DESCRIPTION
`Splot` reads a secondary structure graph in `GML` format, and produces a `PostScript` file displaying the schematic drawing of the secondary structure.

OPTIONS
none

VERSION
Version 1.0 of `splot`

REFERENCES
The reading of the input file is based on the scanner and parser for the GML file format by M. Himsolt [66, 67] available at: `http://infosun.fmi.uni-passau.de/Graphlet/GML/`.

AUTHOR
Christina Witwer

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] R. Archarya, E. Fry, and D. Stuart. The tree dimensional structure of foot-and-mouth disease virus at 2.9 Å resolution. *Nature*, 337:709–716, 1989.

[6] M.S. Bazaraa and J.J. Jarvis. *Linear Programming and Network Flows*. John Wiley and Sons, New York, 1977.

[7] C Berge. Two theorems in graph theory. In *Proceedings of the National Academy of Sciences*, volume 43, pages 842–844, USA, 1957.

[8] 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.

[9] A. M. Bormann, M.M. Michel, and K.M. Kean. Detailed analysis of the requirement of heatitis A virus internal ribosome entry segment for the eukaryotic initation factor complex eIF4F. *J. Virol.*, 75(17):7864–7871, 2001.

[10] E. A. Brown, S. P. Day, R.W. Jansen, and S. M. Lemon. The 5' nontranslated region of hepatitis A virus RNA: Secondary structure and elements required for translation in vitro. *J. Virol.*, 65:5828–5838, 1991.

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

[12] J.W. Brown, E.S. Haas, B.D. James, D.A. Hunt DA, J.S. Liu, and N.R. Pace. Phylogenetic analysis and evolution of RNase P RNA in proteobacteria. *J. Bacteriol.*, 173(12):3855–3863, 1991.

[13] 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.

[14] R.E. Bruccoleri and G. Heinrich. An improved algorithm for nucleic acid secondary structure display. *Comput. Appl. Biosci.*, 4:167–173, 1988.

[15] 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.

[16] 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.

[17] 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.

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

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

[20] 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.

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

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

[23] V. Chvátal. *Linear programming*. W.H. Freeman and Company, New York, 1983.

[24] B.E. Clarke, A.L. Brown, K.M. Currey, S.E. Newton, D.J. Rowlands, and A. R. Carroll. Potential secondary and tertiary structure in the genomic RNA of foot and mouth disease virus. *Nucleic Acids Res.*, 15:7067–7079, 1987.

[25] R. B. Couch. Rhinoviruses. In N.R. Fields, D.M. Knipe, and P.M. Howley, editors, *Virology*, volume 1, pages 713–734. Lippincott-Raven Publishers, Philadelphia, New York, third edition, 1996.

[26] T. Cui and A.G. Porter. Localization of the binding site for encephalomyocarditis virus RNA polymerase in the 3'-noncoding region of the viral RNA. *Nucleic Acids Res.*, 23:377–382, 1995.

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

[28] Peter De Rijk and Rupert De Wachter. Rnaviz, a programm for the visualization of RNA secondary structure. *Nucleic Acids Res.*, 25:4679–4684, 1997.

[29] M. Doherty, D. Todd, N. McFerran, and E. M. Hoey. Sequence analysis of a porcine enterovirus serotype 1 isolate: relationships with other picornaviruses. *J. Gen. Virol.*, 80:1929–1941, 1999.

[30] Jennifer A. Doudna. Structural genomics of RNA. *Nature Struct. Biol.*, 7:954–956, 2000.

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

[32] Hernando Duque and Ann C. Palmenberg. Phenotypic characterization of three phylogenetically conserved stem-loop motifs in the mengovirus 3' untranslated region. *J. Virol*, 75:3111–3120, 2001.

[33] J. Edmonds. Maximum matching and a polyhedron with(0,1) vertices. *Journal of Research of the National Bureau of Standards*, 69B:125–130, 1965.

[34] J. Edmonds. Paths, trees and flowers. *Canadian Journal of Mathematics*, 17:449–467, 1965.

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

[36] Ramon M. Felciano, Richard O. Chen, and Russ B. Altmann. RNA secondary strucutre as a reusable interface to biological information resources. *Gene*, 190:GC59–GC70, 1997.

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

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

[39] 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.

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

[41] H.N. Gabow. *Implementation of algorithms for maximum matching on nonbipartite graphs*. PhD thesis, Stanford University, 1973.

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

[43] K. Gerber, E. Wimmer, and A.V. Paul. Biochemical and genetic studies of the initiation of human rhinovirus 2 RNA replication: identification of a cis-replication element in the coding seuence of 2Apro. *J. Virol*, 75(22):10979–10990, 2001.

[44] F. Ghazi, P. J. Hughes, T. Hyypiä, and G. Stanway. Molecular analysis of human parechovirus type 2 (formerly echovirus 23). *J. Gen. Virol.*, 79:2642–2650, 1998.

[45] 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.

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

[47] I. Goodfellow, Y. Chaudhry, A. Richardson, J. Meredith, J. W. Almond, W. Barclay, and D. J. Evans. Identification of a cis-acting repli-

cation element within the poliovirus coding region. *J. Virol.*, 74:4590–4600, 2000.

[48] 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.

[49] 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.

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

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

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

[57] Y. Hahn. http://hsc.virginia.edu/med-ed/micro/images/vir6/fig1.gif, 2002.

[58] K. Han, D. Kim, and H.J. Kim. A vector-based method for drawing RNA secondary structure. *Bioinformatics*, 15:286–297, 1999.

[59] 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.

[60] 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.

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

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

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

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

[65] M. J. Hewlett, J. K. Rose, and D. Baltimore. 5' terminal structure of poliovirus polyribosomal RNA is pUp. *Proc. Natl. Acad. Sci. USA*, 73:327–330, 1976.

[66] M. Himsolt. GraphEd: A graphical platform for the implementation of graph algorithms. In R. Tamassia and I.G. Tollis, editors, *Graph*

Drawing, Lecture Notes in Computer Science, volume 894, pages 182–193, 1994.

[67] M. Himsolt. GML: graph modelling language. http://www.fmi.uni-passau.de/archive/archive.theory/graphlet/GML.ps.gz, 1997.

[68] 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.

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

[70] 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.

[71] 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.

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

[73] P. Hogeweg and B. Hesper. Energy directed folding of RNA sequences. *Nucl. Acids Res.*, 12(67-74), 1984.

[74] I. Holyer. The NP-completeness of edge coloring. *SIAM J. Comput.*, 10:718–720, 1981.

[75] 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.

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

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

[78] R. J. Jackson. Comparative view of initation site selection mechanism. In N. Sonenberg, J. W. B. Hershey, and M. B. Mathews, editors, *Translational Control of gene expression*, pages 127–184. Cold Spring Harbor, NY, 2000.

[79] Richard J. Jackson, Michael T. Howell, and Ann Kaminski. The novel mechanism of initiation of picornavirus RNA translation. *Trends in Biochemical Sciences*, 15:477–483, 1990.

[80] A. B. Jacobson and M. Zuker. Structural analysis by energy dot plot of large mRNA. *J. Mol. Biol.*, 233:261–269, 1993.

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

[82] 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.

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

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

[85] A.M.Q. King, F. Brown, P. Christian, T. Hovi, T. Hyypiä, N.J. Knowles, S.M. Lemon, P.D. Minor, A.C. Palmenberg, T. Skern, and G. Stanway. Picornaviridae. In M. H. V. Van Regenmortel, C.M. Fauquet, D. H. L. Bishop, C. H. Calisher, E. B. Carsten, M. K. Estes, S. M. Lemon, J. Maniloff, M. A. Mayo, D. J. McGeoch, C. R. Pringle, and R. B. Wickner, editors, *Virus Taxonomy. Seventh Report of the International Committee for the Taxonomy of Viruses*, pages 657–673. Academic Press, New-York, San Diego, 2000.

[86] Rolf Knippers. *Molekulare Genetik*. Thieme, Stuttgart, New York, 6th edition, 1995.

[87] 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.

[88] V. G. Kolupaeva, T. V. Pestova, C. U. Hellen, and I. N. Shatsky. Translation initiation factor 4g recognizes a specific structural element within the internal ribosome entry site of encephalomyocarditis virus RNA. *J. Biol. Chem.*, 273:18599–18604, 1998.

[89] 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.

[90] A. Krumbholz, M. Dauber, A. Henke, E. Birch-Hirschfeld, N. J. Knowles, A. Stelzner, and R. Zell. Sequencing of porcine enterovirus groupe ii and iii reveals unique features of both virus groups. *J. Virol.*, 76(11):5813–5821, 2002.

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

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

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

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

[95] Pierre-Emanuel Lobert, Nicolas Escriou, Jean Ruelle, and Thomas Michiels. A coding RNA sequence acts as a replication signal in cardioviruses. *Proc. Natl. Acad. Sci. USA*, 96:11560–11565, 1999.

[96] F. Löffler and P. Frosch. Berichte der Kommission zur Erforschung der Maul- und Klauenseuche bei dem Institut für Infektionskrankheiten in Berlin. *Zbl. Bakter. Abt. 1. Orig.*, 23:371–391, 1898.

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

[98] 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.

[99] M. Luo, G. Viend, and G. Kamer. The atomic structure of mengo virus at 3 Å resolution. *Science*, 235(182-191), 1987.

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

[101] 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.

[102] P. Marvil, N.J. Knowles, A.P.A. Mockett, P. Britton, T.D.K. Brown, and D. Cavanagh. Avian encephalomyelitis virus is a picornavirus and is most closeley related to hepatitis A virus. *J. Gen. Virol.*, 80:653–662, 1999.

[103] 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.

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

[105] 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.

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

[107] 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.

[108] J. L. Melnick. Enteroviruses: Polioviruses, coxsackieviruses, echoviruses, and newer enteroviruses. In N.R. Fields, D.M. Knipe, and P.M. Howley, editors, *Virology*, volume 1, pages 655–712. Lippincott-Raven Publishers, Philadelphia, New York, third edition, 1996.

[109] Janet M. Meredith, Jonathan B. Rohll, Jeffrey W. Almond, and David J. Evans. Similar interactions of the poliovirus and rhinovirus 3d polymerases with the 3' untranslated region of rhinovirus 14. *J. Virol.*, 73:9952–9958, 1999.

[110] K. Meyer, A Peterson, M. Niepmann, and E. Beck. Interaction of eukaryotic initiation factor eIF-4B with a picornavirus internal translation initiation site. *J. Virol.*, 69:2819–2824, 1995.

[111] Mohammad H. Mirmomeni, Pamela J. Hughes, and Glyn Stanway. An tertiary structure in the 3' untranslated region of enteroviruses is necessary for efficient replication. *J. Virol.*, 71:2363–2370, 1997.

[112] G. Muller, C. Gaspin, A. Etienne, and E. Westhof. Automatic display of RNA secondary structures. *Comput. Appl. Biosci.*, 9:551–561, 1993.

[113] 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.

[114] B. Niklasson, L. Kinnunen, B. Hornfeldt, J. Horling, C. Benemar, K.O. Hedlund, L. Matskova, T. Hyypiä, and G. Winberg. A new picornavirus isolated from bank voles (clethrionomys glareolus). *Virology*, 255(1):86–93, 1999.

[115] 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.

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

[117] A. Nomoto, Y.F. Lee, and E. Wimmer. The 5'-end of poliovirus mRNA is not capped with m7G(5')pppG(5 ')Np. *Proc. Natl. Acad. Sci. USA*, 73:375–380, 1976.

[118] 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.

[119] 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.

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

[121] M. S. Oberste, K. Maher, and M. A. Pallansch. Complete sequence of echovirus 23 and its relationship to echovirus 22 and other human enteroviruses. *Virus Res.*, 56:217–223, 1998.

[122] K. Ochs, R. C. Rust, and M. Niepmann. Translation initiation factor eIF4B interacts with a picornavirus internal ribosome entry site in both 48S and 80S initiation complexes indepenently of initiator AUG location. *J. Virol.*, 73:7505–7514, 1999.

[123] K. Ochs, L. Saleh, G. Bassili, V. H. Sonntag, A. Zeller, and M. Niepmann. Interaction of translation initiation factor eIF4B with the poliovirus internal ribosome entry side. *J. Virol.*, 76(5):2113–2122, 2002.

[124] Ann C. Palmenberg and J. Sgro. Topological organization of picornaviral genomes: Statistical prediction of RNA structural signals. *Seminars in Virology*, 8:231–241, 1997.

[125] C.H. Papadimitriou. *Combinatorial Optimization: Algorithms and Complexity*. Prentice-Hall Inc., New Jersey, 1982.

[126] 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.

[127] J. Perochon-Dorisse, F. Chetouani, S. Aurel, N. Iscolo, and B. Michot. RNA-d2: a computer programm for editing and display of RNA secondary structures. *Bioinformatics*, 11:101–109, 1995.

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

[129] E. V. Pilipenko, V. M. Blinov, B. K. Chernov, T. M. Dmitrieva, and V. I. Agol. Conservation of the secondary structure elements of the 5'-untranslated region of cardio- and aphthovirus RNAs. *Nucleic Acids Res.*, 17:5701–5711, 1989.

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

[131] Evgeny V. Pilipenko, S. V. Maslova, A.N. Sinyakov, and V. I. Agol. Towards identifaction of cis-acting elements involved in the replication of enterovirus RNAs - a proposal for the existence of tRNA-like terminal structures. *Nucleic Acids Res.*, 20:1739–1745, 1992.

[132] C.W. Pleij and L. Bosch. RNA pseudoknots: structure, detection, and prediction. *Methods Enzymol.*, 180:289–303, 1989.

[133] 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.

[134] www.ks.uiuc.edu/Research/vmd/images/polio.gif, 2002.

[135] www.hhmi.org/grants/lectures/images/99/polio.gif, 2002.

[136] J. R. Putnak and B. A. Philips. Picornaviral structure and assembly. *Microbiol. Rev.*, 45:287–315, 1981.

[137] 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.

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

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

[140] J. B. Rohll, D. H. Moon, D. J. Evans, and J. W. Almond. The 3' untranslated region of picornavirus RNA: Features required for efficient genome replication. *J. Virol.*, 69:7835–7844, 1995.

[141] M. G. Rossmann, E. Arnold, and J. W. Ericson. Structure of a human common cold virus and functional relationship to other picornaviruses. *Nature*, 317:145–153, 1985.

[142] E. Rothberg. Solver for the maximum weight matching problem. ftp://dimacs.rutgers.edu/pub/netflow/matching/weighted/solver-1, 1985.

[143] David J. Rowlands. Foot-and-mouth disease viruses (picornaviridae). In R.G. Webster and A. Granoff, editors, *Encyclopedia of Virology*, pages 586–575. Academic Press, 2nd edition, 1999.

[144] R. R. Rueckert. Picornaviridae: The viruses and their replication. In N.R. Fields, D.M. Knipe, and P.M. Howley, editors, *Virology*, volume 1, pages 609–654. Lippincott-Raven Publishers, Philadelphia, New York, third edition, 1996.

[145] L. Saleh, R. C. Rust, R. Füllkrug, E. Beck, G. Bassili, K. Ochs, and M. Niepmann. Functional interaction of translation initiation factor eIF4G with the foot-and-mouth-disease virus internal ribosome entry side. *J. Gen. Virol.*, 82:757–763, 2001.

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

[147] G. Schäfer. Weighted matchings in general graphs. Master's thesis, Fachbereich Informatik, Universität des Saarlandes, Saarbruecken, 2000.

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

[149] 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.

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

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

[152] 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.

[153] R. Stocsits, I.L. Hofacker, and P.F. Stadler. Conserved secondary structures in hepatitis b virus RNA. In Giegerich et al, editor, *Proceedings of the GCB'99*, Computer Science in Biology, pages 73–79, Hannover, 1999. Univ. Bielefeld.

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

[155] L. Su, L. Chen, M. Egli, J.M. Berger, and A. Rich. Minor groove RNA triplex in the crystal structure of a ribosomal frameshifting viral pseudoknot. *Nature Structural Biology*, 6(3):285–292, 1999.

[156] 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.

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

[158] 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.

[159] 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.

[160] 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.

[161] 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.

[162] C. Thurner, C. Witwer, I.L. Hofacker, and P.F. Stadler. Conserved RNA secondary structures in flaviviridae genomes. *In preparation*, 2003.

[163] 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.

[164] Lori M. Vance, Nicola Moscufo, Marie Chow, and Beverly A. Heinz. Poliovirus 2C region functions during encapsidation of viral RNA. *J. Virol.*, 71:8759–8765, 1997.

[165] P. Walter and G. Blobel. Purification of a membrane-associated protein complex required for protein translocation across the endoplasmic reticulum. *Proc. Natl. Acad. Sci. USA*, 77(12):7112–6., 1980.

[166] P. Walter and G. Blobel. Signal recognition particle contains a 7s rna essential for protein translocation across the endoplasmic reticulum. *Nature*, 299:691–698, 1982.

[167] Jinhua Wang, Judith M.J.E. Bakkers, Joep M.D. Galama, Hilbert J. Bruins Slot, Evgeny V. Pilipenko, Vadim I. Agol, and Willem J.G. Melchers. Structural requirements of the higher order RNA kissing element in the enteroviral 3'utr. *Nucl. Acids Res.*, 27:485–490, 1999.

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

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

[170] B. Weiser and H. Noller. XRNA. *ftp://fangio.ucsc.edu/pub/XRNA/*, 1995.

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

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

[173] 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.

[174] 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.

[175] G. Wutz, N. Nowotny, B. Grosse, T. Skern, and E. Küchler. Equine rhinovirus serotypes 1 and 2: relationship to each other and to aphthoviruses and cardioviruses. *J. Gen. Virol.*, 77:1719–1730, 1996.

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

[177] 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.

[178] T. Yamashita, K. Sakae, H. Tsuzuki, Y. Suzuki, N. Ishikawa, N. Takeda, T. Miyamura, and S. Yamazaki. Complete nucleotide sequence and genetic organization of aichi virus, a distinct member of the picornaviridae associated with acute gastroenteritis in humans. *J. Virol.*, 72(10):8408–8412, 1998.

[179] R. Zell, M. Dauber, A. Krumbholz, A. Henke, E. Birch-Hirschfeld, A. Stelzner, D. Prager, and R. Wurm. Porcine teschoviruses comprise at least eleven distinct serotypes: molecular and evolutionary aspects. *J. Virol.*, 75:1620–1631, 2001.

[180] R. Zell and A. Stelzner. Application of genome sequence information to the classifation of bovine enteroviruses: the importance of 5'- and 3'-nontranslated regions. *Virus Res.*, 51:213–229, 1997.

[181] Roland Zell. Vorlesung Virologie für Biochemiker: Picornaviren. http://www.personal.uni-jena.de/~i6zero/start.html, Inst. f. Virologie d. Friedrich-Schiller-Universität Jena, 2002.

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

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

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

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