universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Identification and Classification of Pseudoknots and their Impact on RNA 3D Structure Prediction“

verfasst von / submitted by

Irene Katharina Beckmann, BSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2018 / Vienna 2018

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 066 834
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Molekulare Biologie
Betreut von / Supervisor:	Univ.-Prof. Dipl.-Phys. Dr. Ivo Hofacker

ACKNOWLEDGMENTS

First of all, I would like to thank my supervisor Prof. Dr. Dipl.-Phys. Ivo Hofacker for the great opportunity to enable this master thesis at the "Theoretical Biochemistry Group" at the University of Vienna.

Special thanks go to Bernhard Thiel, MSc for the advice, encouragement and the shared instructive computational input.

I would also like to thank my colleague Roman Ochsenreiter as well as the whole TBI-team for the helpful input.

Finally, I would express my very profound gratitude to my family for providing continuous support and who were a source of fabulous motivation.

CONTENTS

1 INTRODUCTION 1
- 1.1 Ribonucleic acid (RNA) 1
  - 1.1.1 RNA Structure - from primary to tertiary structures 3
- 1.2 Elucidation of RNA Structures 13
  - 1.2.1 Nuclear magnetic resonance (NMR) 13
  - 1.2.2 Biomolecular 3D X-ray crystallography 14
  - 1.2.3 Electron microscopy (EM) 14
- 1.3 RNA structure prediction 16
  - 1.3.1 Secondary structure prediction 16
  - 1.3.2 Tertiary structure prediction 17
  - 1.3.3 Prediction of pseudoknots 20

2 ISOLATION AND CLASSIFICATION OF RNA PSEUDOKNOTS - MOTIVATION AND METHOD 25

3 WORKING TOOLS 27
- 3.1 Non-redundant, representative 3D structure dataset 27
- 3.2 RNA structure prediction 32
  - 3.2.1 MC-Annotate 32
  - 3.2.2 DSSR 33
  - 3.2.3 3D structure prediction in `ERNWIN` 34

4 RESULTS AND DISCUSSION 41
- 4.1 Pseudoknots in the non-redundant, representative 3D structure dataset 41
- 4.2 Pseudoknots and the angle distribution between stems 48
- 4.3 Comparison of the dotbracket and the Ernwin identification of pseudoknots 54

5 SUMMARY AND CONCLUSION 57

BIBLIOGRAPHY 59

LIST OF FIGURES

Figure 1 Composition of the ribosome 2
Figure 2 Chemical structure elements of DNA and RNA 5
Figure 3 2D model of the tRNA 5
Figure 4 RNA secondary structures 01 6
Figure 5 RNA secondary structures 02 6
Figure 6 Examples for pseudoknots in nature 8
Figure 7 3D model of an RNA double helix 9
Figure 8 Basics of an H-type pseudoknot 10
Figure 9 From the diagram to the shadow representation of a Pseudoknot 10
Figure 10 Shadow representation of RNA 1-structures 11
Figure 11 Nested shadow representation 12
Figure 12 Basic pseudoknot representation 22
Figure 13 Representation elements in `ERNWIN` 36
Figure 14 `ERNWIN` and multiloops 38
Figure 15 `ERNWIN` and pseudoknots 39
Figure 16 Pseudoknots in the non-redundant 3D structure dataset 41
Figure 17 Number of pseudoknots versus the length of the structure 43
Figure 18 Shadow of a pseudoknot labeled as "other" 43
Figure 19 Measurement of the angles between all pairs of stem stems of all detected pseudoknots and multiloops 49
Figure 20 Measurement of the angles between all pairs of stems of all detected H-types and kissing hairpins 49
Figure 21 Measurement of the angles between all pairs of stem stems of all detected pseudoknots and multiloops 50
Figure 22 H-type: angle distributions 51
Figure 23 Kissing hairpin: angle distributions 52
Figure 24 Kissing hairpin: deflection of the structure 53
Figure 25 Consistency Ernwin | Dotbracket structure 54
Figure 26 Unconsistent `ERNWIN` counts 56

LIST OF TABLES

Table 1 RNA 1-structure classification 12

Table 2 NR-list: selection 29
Table 3 RNA: chain length 30
Table 4 `ERNWIN`: connection types 39
Table 5 Classic RNA structures and pseudoknots 45
Table 6 Stem size and their effect on pseudoknots 46

ACRONYMS

A Adenin
BGSU Bowling Green State University
C carbon
C Cytosin
cryo-EM Cryo-electron microscopy
DNA deoxy-ribonucleic acid
G Guanin
HDV hepatitis delta virus
HTM homogeneous transformation matrices
IFEs Integrated Functional Elements
INF interaction network fidelity
MCQ mean of circular quantities
MEA maximum expected accuracy
MFE minimum free energy
MFE minimum free energy
miRNA micro-RNA
mRNA messenger RNA
multiplets higher-order co-planar base associations
NMR nuclear magnetic resonance
NOEs Nuclear Overhauser Effects
NP non-polynomial time
NR non-redundant
nts nucleotides
O oxygen
P phospate
PDB Protein Data Base
RF radiofrequency

RMSD root-mean-square deviation

RNA ribonucleic acid

rRNA ribosomal RNA

siRNA small interfering RNA

snRNA small nuclear RNA

T Thymin

tRNA transfer RNA

U Uracil

1 INTRODUCTION

1.1 RIBONUCLEIC ACID (RNA)

For quite a long time ribonucleic acid (RNA) [8, 11, 38] seemed to be used for information transport only. This picture slowly but continuously changed. Nowadays it is known that RNA plays a role in many biological processes from catalysis to cellular regulation pathways. Over the years, a wide range of different RNA functions were discovered. The prefix in the shortcut marks the different RNA types. The best known ones are the ribosomal RNA (rRNA), the transfer RNA (tRNA) and the messenger RNA (mRNA), which are all related to the translation machinery.
First, I will describe the mRNA, which is translated by the ribosome. For every gene or group of genes (Prokaryotes) an mRNA is synthesized. Therefore mRNAs constitute very heterogeneous group of molecules. Especially in Eukaryotes mRNA can have structural features which can influence the translation efficiency as well as the life span of the molecule.
tRNA (figure 3) transports a single activated, bonded amino acid to the ribosome, the protein synthetic machinery of a cell. The recognition site of tRNA is a sequence of three basepairs called anticodon. The anticodon recognizes the complementary sequence (the codon) on the mRNA. The other end of the tRNA is covalent bonded to the corresponding amino acid.
Besides tRNA and rRNA (which will be discussed below), there exist many types of non-coding RNA [8, 11, 38], such as small nuclear RNA (snRNA). These are involved in the splicing process of the RNA-Exons. Another type of RNA that is involved in regulation, is the post-transcriptional micro-RNA (miRNA), which comprises around 20 nucleotides (nts), binds on the complementary mRNA, and blocks the translation. The small interfering RNA (siRNA), a class of small RNA molecules, can bind to mRNA and introduce degradation. But RNA is also part of several other processes in a cell. For example, RNA with an appropriately folded shape can serve as an enzyme (ribozymes).
Much of knowledge on RNA 3D is arrived from ribosomal RNAs [8, 11, 38] since we have several high quality structures of the ribosome. Therefore, some statistics regarding the structural composition will be presented in section 3.1 and 4.1
rRNA is one of the participants in the protein synthesis. It forms parts of the structure of a ribosome (description below) and can be distinguished by the sedimentation coefficient. The composition of the ribosme is illustrated in figure 1.
The ribosome [11, 38] is a catalytic complex made from more than 50

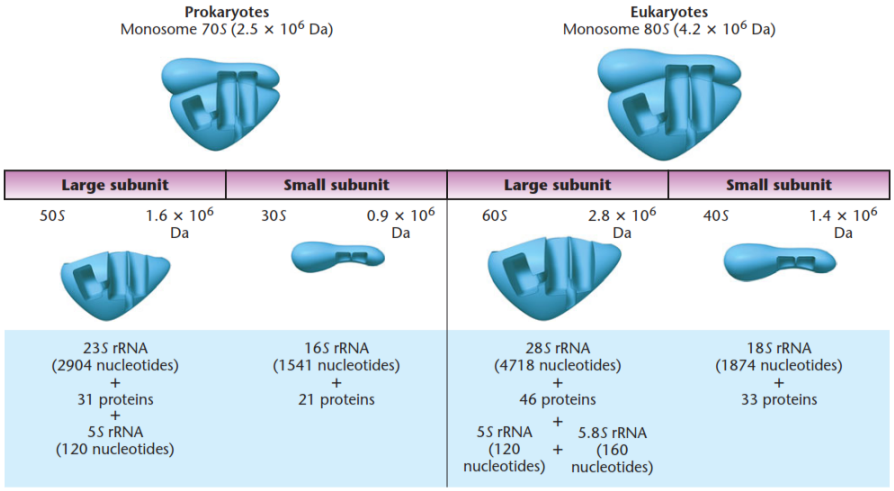


Figure 1: On the left side, the components of a prokaryotic ribosome are illustrated and on the right side, the components of a eukaryotic ribosome. A comparison of both ribosomes shows that the number of RNA chains as well as the sedimentation coefficient (S) differ between both ribosomes. Modified from [38]

different proteins and several rRNA molecules. The proteins are generally located on the surface of the ribosome, filling in the gaps of the folded RNA and stabilizing it while the conformation changes during the protein synthesis. The core of the prokaryotic and eukaryotic ribosomes have nearly the same structure, but they differ in number and in size of their rRNA and protein components (shown in figure 1. Despite the differences both have a large and one small subunit, which fit together and form the complete ribosome with a mass of several million daltons. In both prokaryotes and eukaryotes, the subunits have a similar function. The small subunit acts as a framework on which the tRNA can accurately match to the mRNA codons. The large subunit catalyses the formation of the peptide bonds. It is known that the RNA components of the ribosome perform all-important catalytic functions regarding the translation of mRNA.
The prokaryotic ribosome contains a 5S rRNA (120 nts) and a 23S rRNA (2,904 nts) at the large 50 S subunit and a 16S rRNA at the small 30S subunit. The eukaryotic large 60S subunit of the ribosome is built with a 28S rRNA (4,718 nts), a 5S rRNA (120 nts) and a 5.8S rRNA (160 nts). The small subunit contains an 18S rRNA constructed with 1874 nts.

One hypothesis is that an RNA world existed on Earth before modern cells arose.[11] This hypothesis states that RNA stored genetic information and catalysed the chemical reaction in primitive cells. The reorganization and transition out of the RNA world was not completed, as RNA is still an important part of many pathways in a cell.

1.1.1 *RNA Structure - from primary to tertiary structures*

The primary structure of RNA is largely similar to the deoxy-ribonucleic acid (DNA). [11, 38] Both molecules are assembled from four nucleotide building blocks, linked together in chains. Instead of the sugar deoxyribose, RNA nucleotides are built with ribose and the base Thymin (T) is replaced by Uracil (U). During RNA base pairing and the transcription process U is complementary to Adenin (A). The other nucleotides are Guanin (G) and Cytosin (C), summarised in figure 2. An important difference between DNA and RNA is that usually the RNA molecule is a single stranded polymer which can fold itself back to complementary bases, forming double-stranded regions by Waston Crick basepairing (AU, GC) and build complex, compact structures. Merely some viruses that store their genetic material in double stranded RNA helices are an exception.
It is well known that the function of proteins depends on well defined structure that provides surfaces with unique contours and chemical properties for catalyses. Similarly, many RNAs relay on their structure to perform their function. [11]

Therefore it is consistent that RNA molecules like for example tRNA which have to fit exactly, display a nearly identical structure, the cloverleaf model of tRNA (figure 3), and similar lengths of the analogous stems and loops. Only small parts of the primary structures are conserved like the 3'end (pCpCpA-3') where the amino acid is covalently bound to the terminal adenosine or the 5'-pG end.[38]

Even if structures are strongly conserved RNA sequences can vary widely.
Nevertheless 16S rRNA and 18S rRNA sequences are used to build phylogenetic trees. Both have ubiquitous changes in sequence during the course of evolution. They have a large number of conserved segments that can be used for comparative and metagenome analysis. As a rule 16S rRNA/18SrRNA from different organism of the same species, differ by less than 3% while rRNA sequences of the same genus differ less than 5%.[18] Notwithstanding the identification of the sequence variants of an RNA 3D motif [12] (see section 1.3.2) is of interest as it can have an impact on RNA structure prediction.

A secondary structure [12] is defined by the forming of helices. These helices are sequence parts that interacts with each other via basepairing. A helix contains a set of base-pairs {AU, UA, GC, CG, GU, UG}. Here the subset {AU, UA, GC, CG} contains the classical Watson and Crick basepairs. In addition to this variety, in typical RNA structures are also non Watson Crick basepairs. The subset {GU, UG} describes wobble basepairs. The so called G-U wobble base pair are found in RNA even if it is no classical Watson Crick base-pair. The G-U wobble base pairs conformation is very common, see section 1.3.2. Both subsets are classified as canonical basepairs in RNA. The so called Watson Crick helices in RNA are generally short and often interrupted

by single stranded regions.
More formly a secondary structure is described by the amount of basepairs within the structure. There are several condition regarding the formation of a pseudoknot free secondary structure (see section 1.3.3 for conditions including pseudoknots). If there are two basepairs i-j and k-l the nucleotide numbers (direction 5′ to 3′) has to be

$$i < j < k < l$$

or

$$i < k < l < j$$

. For the prediction of secondary structure, one addi-tional condition is the minimum loop size. This minimum loop size are often between three or four nucleotides and is defined by e.g.

$$|j - i| >= 3.$$

A look through of a huge pool of RNA secondary structures have revealed recurring building blocks. Figure 4 shows examples of short structural elements that can be found several times as parts of larger structures. A best-known motif is the loop [12]. In the secondary structure, loops can be classified into the hairpin loop, the internal loop and the multihelix loop. The different types of loops differ in size and shape, where the junctions are the most diverse ones and are attached to three or more helixes. A hairpin loop is attached to one helix and has a range in size from 3 to more than 20 nucleotides. The size of an internal loop varies within a range of 5 to 30 nucleotides and is connected to with two helices. By both loops the flanking basepairs were including in the loop size.

In structures there can occur also non canonical non Watson Crick basepairs. This phenomena can be observed in secondary structures as “loops” or as “linker strands”. In the tertiary structure, these motifs ensure stability, distinct shape and dynamic characteristics. [12] The sequence influences the RNA molecule structure by the variety of possible interactions between two nucleotides and by the RNA backbone.

The fact that one class of RNA molecules commonly displays similar secondary structures also affects the tertiary structure. Like proteins, large RNA is organized based on structural domains and subdomains. [11, 12] These may be linked to each other covalently (by double-stranded-helices or single-stranded linker’s) or non-covalently. They can join distant parts of the same RNA molecule or bring two separate ones together. Some basic examples of RNA tertiary interactions are shown in figure 5. Collectively these 3D RNA motifs [12] play an architectural and an organizational role as well as a role in providing functional binding sites for other molecules.

1.1.1.1 *Pseudoknots*

Pseudoknots [12] in RNA have been found in many organisms especially in functionally important RNA structures. An RNA pseudo-

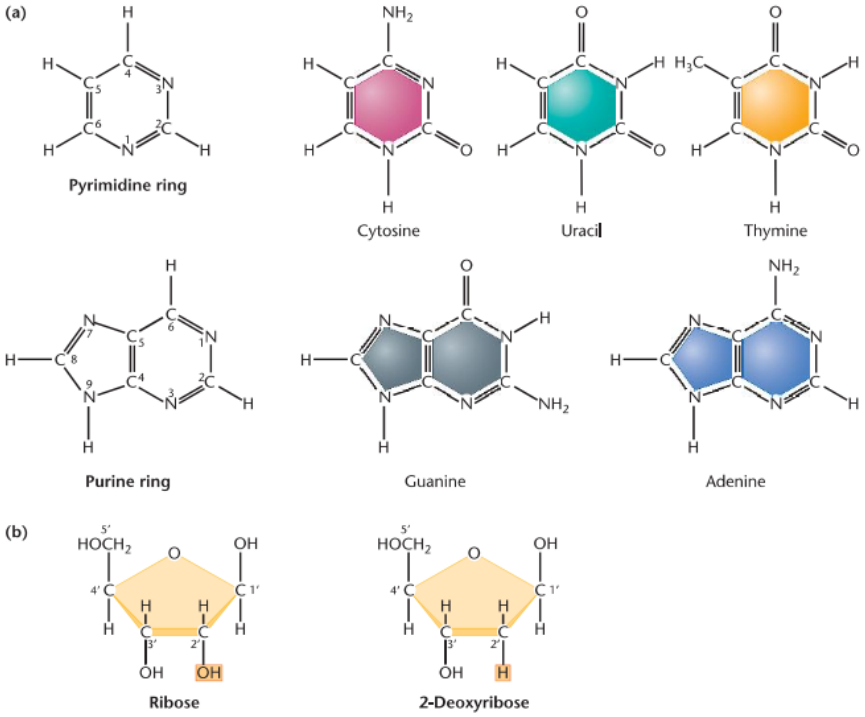


Figure 2: Chemical structures of the nucleotide bases from DNA and RNA. (a) Both, DNA and RNA, take use of the purines Adenine and Guanin as well as the pyrimidine Cytosin. In contrast to Thymin in the DNA, RNA use as second pyrimidine building block Uracil. (b) Complementary to this also the chemical structure of the pentose sugar differ between DNA (desoxyribose) and RNA (ribose). Modified from [38]

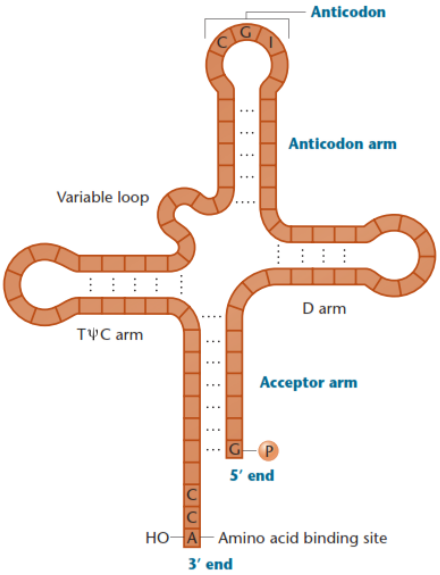


Figure 3: Secondary structure model of the tRNA that forms a cloverleaf. Modified from [38]

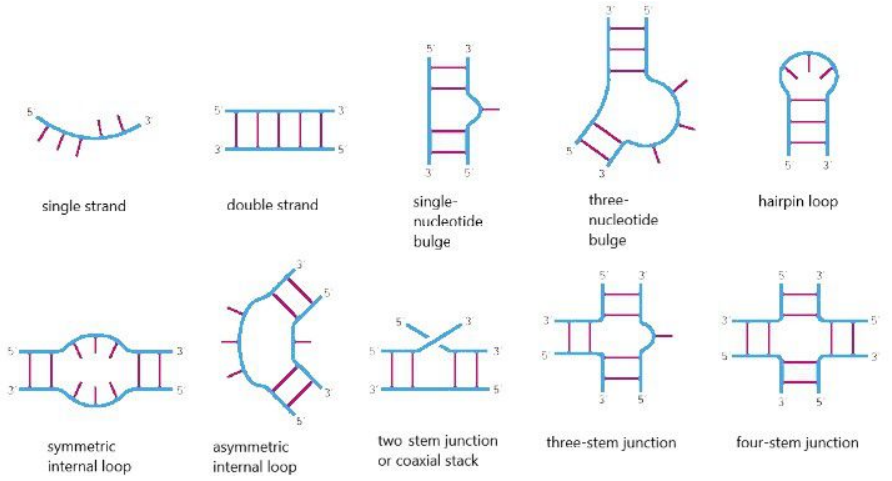


Figure 4: Conventional complementary base-pairing interaction of RNA secondary structure. Modified from [11]

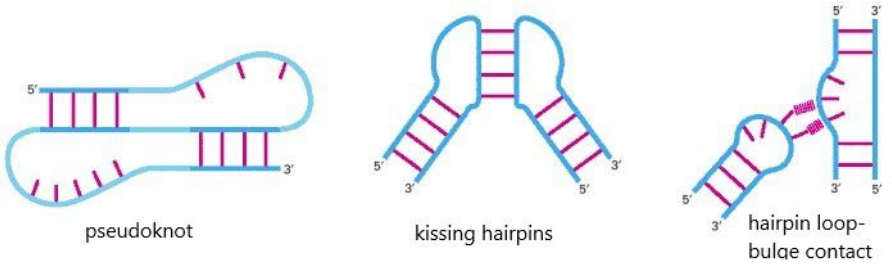


Figure 5: Commonly used examples for more complex interaction. Modified from [11]

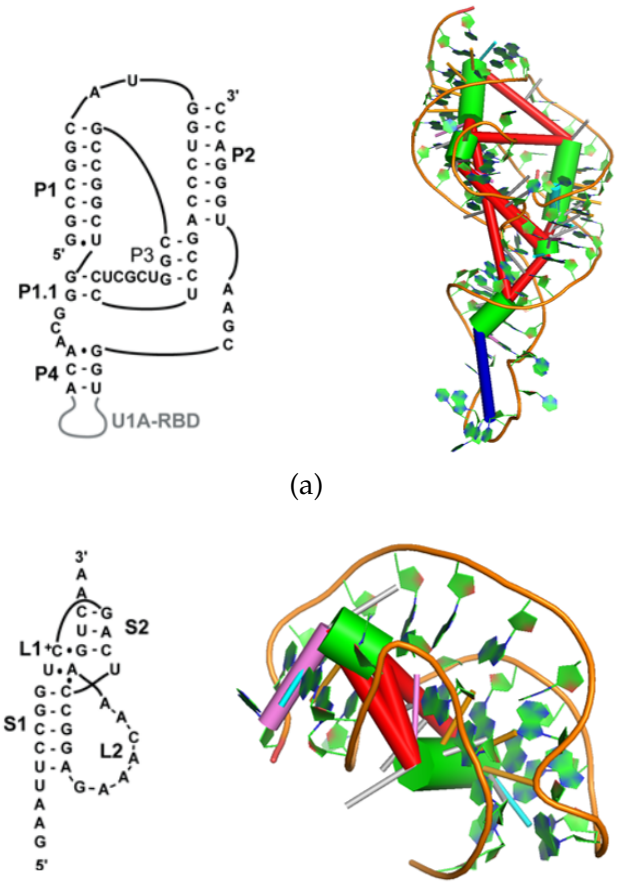
knot can result in many different and complex structures, but the basic characteristic structure follows the definition:

"A pseudoknot is an RNA structure characterized by WC base pairing between nucleotides in a loop with complementary residues outside the loop." (in collection from Leontis and Westof [45], chapter "Methods for Predicting RNA Secondary Structure" from Aigner et al. page 26 [1])

For this special conformation, most pseudoknots need a high ionic strength like divalent cations e.g Mg^{2+} for stabilization. The fundamental reason for such a stabilizing interaction is the negatively charged phosphate backbone of the stems counteracted by the loop formation. Thus pseudoknots show a dependency of ionic strength [1]. In general, Pseudoknots as a structural complex helps to form stable and very complex RNA structures and are common in many functional domains and catalytic cores, as in self-splicing introns. One example of a self-splicing intron was found in the hepatitis delta virus (HDV) [12], a satellite virus of the hepatitis B virus. The circular HDV genome is replicated by the host RNA polymerase II in a rolling cycle fashion. This generates a long RNA transcript that then has to be cleaved into individual copies of the genome. Further steps, like producing single-genome-length HDV RNA, is given by HDV ribozymes, which are self-cleaving. Therefore, the ribozyme folds a

double-pseudoknot and self-cleaves. Both pseudoknots stack on top of each other and form a nested pseudoknot architecture, visualised in figure 6 (a).
The appearance of catalytic active pseudoknots also affect the human telomerase [12]. The telomeres describe the protective ends of chromosomes. Each cycle of the DNA synthesis leads to a degradation of the telomeres. The telomerase takes care of the maintenance process. The human telomerase consists of a reverse transcriptase, some proteins and an RNA strain with a highly conserved pseudoknot at the 5′ end. This pseudoknot, more precisely an H-type pseudoknot, is essential for the functionality.
Pseudoknots are not only required for catalytic cores, they are also involved in inducing frameshifting [12]. In brief to encourage ribosomes to slide into an alternative reading frame in the +1 or the -1 direction. Many viruses and retroviruses use this programmed ribosomal frameshift functionality for replication or proliferation. Some of these structures are already detected by the nuclear magnetic resonance (NMR) technique (section 1.2.1) like the mouse mammary tumor virus or by the X-ray crystallography (section 1.2.2) like the pea enation mosaic virus, visualised in figure 6 (b).

The prediction of pseudoknots is plagued with barriers, see section 1.3.3. For a more accurate prediction of an RNA pseudoknot knowledge regarding the detailed structural composition is essential. Details concerning loop length, stem length and also distances between certain atoms are important.
One of the well analysed pseudoknots is the H-type pseudoknot [1, 13, 67]. In natural RNA structures the H-type pseudoknot is the most frequently occurring pseudoknot. It is build (building process reviewed in [1]) out of two helical segments S1 and S2 and three loop regions named L1, L2 and L3. Additionally there is the particular complement of the S1 and the S2 region, S1′ and S2′. The order of the elements of the sequence is S1, L1 ,S2, L2, S1′, L3, S2′ and the crossing order is S1<S2<S1′<S2′. It is also allowed that the L2 loop region is missing like in figure 8. Thus, but even if L2 consists of only one or two nucleotides, the two helix segments coaxially stack and form a similar structure like an uninterrupted A-form helix.
In the detailed structure (reviewed in [1]) - L1 crosses the major groove of the double helix and L3 the minor groove. The minor groove of S1 enables contacts to the L3 in form of triple pairs or hydrogen bonds. The 3D model of a helix structure in figure 7 shall give a picture about the distances of the single phosphates and a better understanding about the proportions within a helix. The distance between the phosphates, e.g. P0 and P0′, is about 1.7 nm. For a better idea, this distance can be bypassed by a loop consisting of three nucleotides. Crossing the major groove the minimal distance between a phosphate on one strand, e.g. P0′, to the phosphate on the opposite strand, e.g. P-7, is about 1 nm. Bypassing the minor groove the minimal distance between P0′ to P2 or P3 is about 1.1 nm. Back to the H-type pseudoknot, these listed distances fit very well to a stem region



(a)

(b)

Figure 6: The figure presents examples for pseudoknots in nature.
(a) They are common in catalytic cores as in self splicing introns. One example is given by the HDV ribozyme.
(b) Pseudoknots are also involved in inducing frame-shifting like in Pea enation mosaic virus RNA1
(Figures left side: modified from [68]
Figures right side: Structure coordinates obatined from PDB ID 1SJ3 [34] and 1KPZ [56], structural representation: `ERNWIN` [35, 70] mentioned in section 3.2.3 and `Pymol` [65]

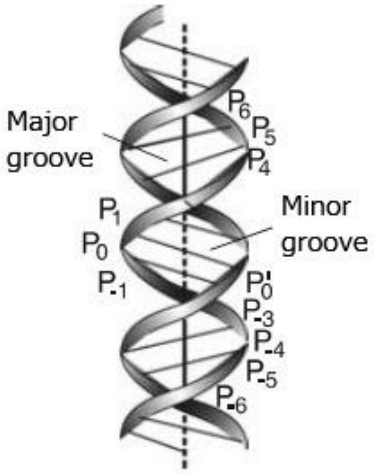


Figure 7: 3D model of an RNA double helix, P marks the phosphates Starting with Po the indices in the 3′ direction are positive and in the 5′ direction are negative. The opposite strand is marked with ‘. Modified from [1]

of 3-7 base pairs bridged by loops of a minimum of 2 nucleotides - a small pseudoknot with coaxial helix stacking. In the secondary structure diagram, a pseudoknot leads to crossing arcs, figure 8.

As mentioned, there exist different folding types of pseudoknots. For a more accurate prediction the different folding types can be classified.

We will describe the different types and genera based on the dot-bracket notation, where dots stands for unpaired bases and nested pairs are represented by matching pairs of left and right parentheses. [69] Note that for writing pseudoknots in a dot-bracket string additional types of brackets are needed, e.g.: “[,], {, }, <, >”. [12] A simple H- type will be marked with the bracket combination ([)] and a kissing hairpin with ([)(]), table 1. The more complex a pseudoknot is, the more different types of brackets are needed and the more nested is the abstract shape.
An additional way is the diagram representation [62] of a molecule, where the sequence is represented with a linear backbone and each basepair with semi circles, mentioned in figure 11. For the isolation and further on the classification of the pseudoknot only the crossing arcs and additionally only one base pair each are needed. In other words, all non-crossing arcs and dots can be removed. The extanted structures represent irreducible components. In the next step all so called multiple stacked brackets can be summarized into one single arc. The received representation is also called the “shadow” of the RNA structure, illustrated in figure 11. Note that, for the shadow that now every information about the structure regarding the stack length and non-crossing components is lost.
Now the structure can be assigned into different genera [62].

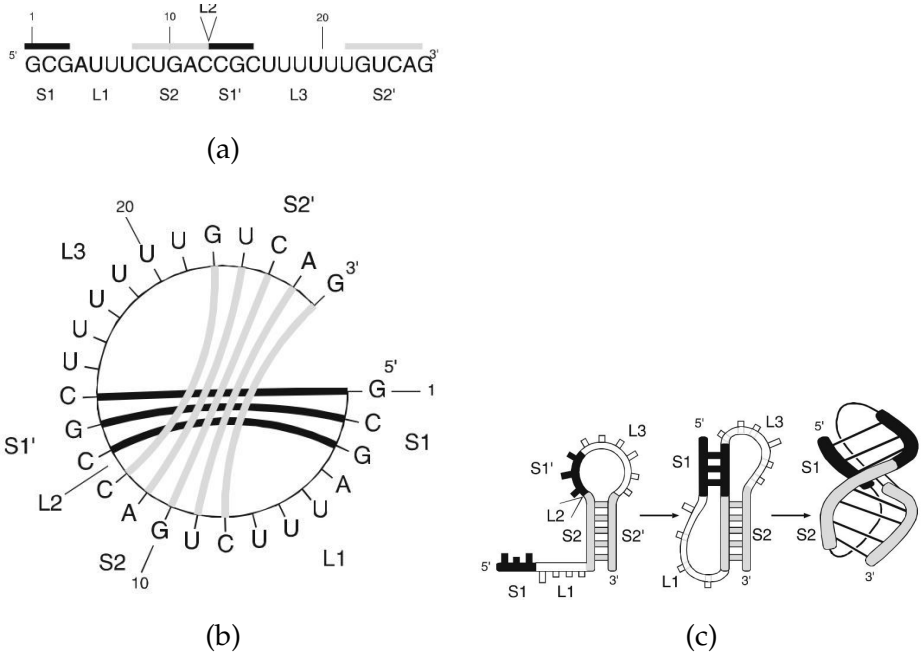


Figure 8: Basics of an H-type Pseudoknot:
(a) Sequence: the black (S1 and S1′) and the grey (S2 and S2′) boxes are complementary regions
(b) Circular graph: Circle with all used nucleotides starting with the 5′ at the left. Connecting the basepairs of S1, S1′ and S2, S2′ leads to crossing lines
(c) Formation of a pseudoknot: In the last step one of the helices rotate by 180° which results to coaxial stacking of the the helices; Modified figure from [1]

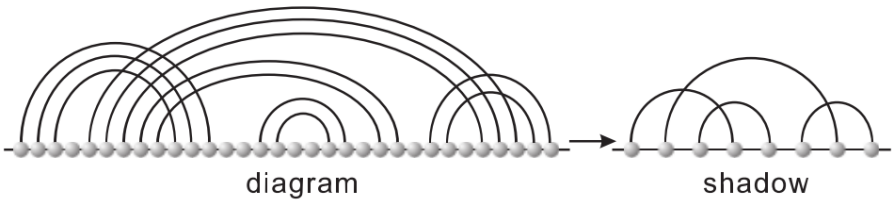


Figure 9: The Figure illustrates the difference between the diagram and the shadow representation applied on a knotted structure.
Figure modified from [62]

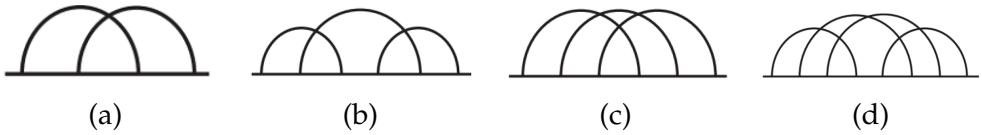


Figure 10: Shadow representation of RNA 1-structures
Classification: (a) H, (b) K, (c) L, (d) M [62]

Reidys et al. [62] used the concept of the shadow to introduce a convenient topological classification of pseudoknots. The terms genus and "γ-structure" are important for this concept.

"Topological genus means the number of handles attached to a sphere and measures topological complexity of its associated surface." (Huang/Nebel/Reidys, 2013, page 216 [32]).

The equation 1 published from Reidys et al. [62] describes where the number of genera comes from. The Genus of a set of shadows is defined by the sum of genera of its irreducible components [62].
The complexity of a structure is described by γ-structure and is determined by the maximal genus of each component within the shadows. Described in detail

"[...] S is a γ-structure if $g(\mathfrak{S}(S')) \leqslant \gamma$ holds for all irreducible components of the shadow $\mathfrak{S}(S)$." (Reidys et al., 2011, page 1079 [62]).

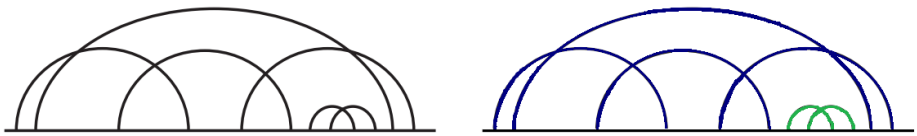
Therefore it is possible that a 1-structure contains multiple, irreducible genus 1 components.
In more detail, a 0-structure represents a simple secondary structure without a crossing arc and represent genus 0. Only if the remaining structure can be assigned into one of the resulting classes listed in table 1 and figure 10 the structure has genus 1. Additionally every 1-structure build from genus one structures without crossing arcs between elements is called a 1-structure. Which means that a 1-structure shadow can be composed of several irreducible components, which have the topological genus 2, figure 11. We look for exactly these cases like in figure 11, when we classify „genus 2 pseudoknots“. There the pseudoknot can be reduced into two components with genus 1 structures (see genus 1 shadows in figure 10). Using equation 1 the structure is defined as an 1- structure. As the pseudoknots are nested, each is defined as a genus 2 pseudoknot.

$$g(S) = g(\mathfrak{S}(S)) = \sum_{\mathfrak{S}' \in \mathbf{I}_{\mathfrak{S}_S}} g(\mathfrak{S}') \tag{1}$$

$g(S)$ ...Genus of the shadow
$\mathfrak{S}(S)$ ...Set of shadows
$\mathbf{I}$ ...Irreducible components

Abstract shape	Classification	Colloquial name	Shadow representation
([)]	H	Simple H-type	Figure 10 (a)
([)([])	K	Kissing hairpin	Figure 10 (b)
([{)]}	L		Figure 10 (c)
([{)([]})	M		Figure 10 (d)

Table 1: A Structure can only be a 1-structure if it can be assigned into one of these four groups. Every genus one structure generate one of these shadows;
Modified table from [13]



(a) (b)

Figure 11: (a) Shadow representation of an RNA 1-structures having topological genus 2. (b) The separated nested shadows represents the following pseudoknot classes: blue ... M, green ... H; H is integrated in the shadow of M. Therefore, by definition both are no genus 1 anymore. Both shadows are in complete form and do not cross each other. In other words, it is possible to separate them from each other without removing any arc.
Modified figure from [62]

1.2 ELUCIDATION OF RNA STRUCTURES

Biomolecular three-dimensional structures are models derived from techniques such as NMR, circular dichroism, X-ray scattering, electron microscopy and atomic force microscopy etc. Every structural experiment measures certain observables and comes up with one or more model that best fits to the measured data. File collections, like the Protein Data Base (PDB), publish structures as accurately as possible, therefore ensemble averaging as well as time averaging due to the chosen method plays a significant role.

1.2.1 *Nuclear magnetic resonance (NMR)*

For NMR [11, 40, 55] a small volume of concentrated purified molecule solution is placed in a strong magnetic field. Certain atomic nuclei, especially hydrogen nuclei, have a magnetic moment or spin which aligns along the strong magnetic field and can change to a misaligned state. This excited state is reached by radiofrequency (RF) pulses of electromagnetic radiation, a chemical shift. The spectrum displays the return to the aligned state – the emitted RF radiation. The chemical shift is related to the field strength at the nucleus site; additionally if one nucleus is excited it also influences the local environment. For structure determination, the basic NMR observables are Nuclear Overhauser Effects (NOEs) and J-coupling constants. When a proton is close in space to another one or to a nuclei with spin > 0 their magnetic dipoles interacts. The NOEs results from dipole–dipole cross-relaxation between nuclei, thus the measured NOE intensity depends on the interatomic distance and tumbling. J coupling is the coupling of the nuclear spins via three bonds.
The information about bond distance, angle and the connectivity is often supplemented with the orientation of internuclear vectors with respect to a reference frame. The detection of the distance between the interacting pair of hydrogen atoms combined with the information about the approximated distance between the parts of a molecule and the sequence information allows to compute a three dimensional structure of a molecule. Beside hydrogen for organic structures, also the isotope ^{13}C can be used for detection. This is because all organic material contains a specific percentage (1.1%) of ^{13}C isotope.
One advantage of NMR is that a molecule can be studied in solution, which allows studying a structure in the physiological environment. It also emphasises the issue of conformational averaging, as the molecules are susceptible to fluctuations and conformational changes. (reviewed in [11, 40, 55])

1.2.2 *Biomolecular 3D X-ray crystallography*

X-ray crystallography [11, 40, 45] has become the main routine method used for discovering three-dimensional structures at atomic resolution and delivered a vast quantity of structural information regarding serval biomolecules and cellular processes in the last years.
The key element of X-ray crystallography is the crystal. It contains the unique, smallest portion of a crystal structure with the biological unit or assembly, which is the supposed form of the molecule and the unit of interest. The resolution of the whole method completely depends on the growth of the crystal and how well the molecules are ordered and aligned in crystal.[15] The wavelengths of X-rays are typically around 0.1 nm, which corresponds to the diameter of a hydrogen bond. Most of the X-rays pass through a sample and only a small fraction is scattered by the atoms in the suitable sample crystal with a large amount of a pure molecule.
The X-ray diffraction pattern contains information about the location of the atoms based on the position and the intensity of each reflection. In general, the information concerning the wave functions includes an amplitude and a phase [7]. Both parameters are used for calculating the density map. The aim is to analyses the diffraction pattern, produce a three dimensional electron density map and to interpret this map. One disadvantage of the X-ray crystallography is that the phase information is lost by measuring the reflection intensities [7]. This influences the accuracy of the density map. Techniques like Cryo-electron microscopy (cryo-EM) (section refch:em) include both parameters, although it is plagued by noise.
The accuracy of the final atomic model depends on the resolution of the initial crystallographic data. Any excessive disorder in the crystal will result in a loss of resolved electron density. This results in a conformational averaging on phase determination in crystallography. (reviewed in [11, 40, 45])

1.2.3 *Electron microscopy (EM)*

Cryo-EM [15] is used as a single-particle cryo-EM analysis technique and can achieve resolutions from 2 to 100 A. This method allows receiving images of unprecedented quality. With the detectors from the EM two-dimensional pictures are recorded. In the single particle cryo-EM the electron dose is set very low to receive structural information at the subnanometer-resolution level. One effect of this low-dose imaging is that the pictures have a very poor signal-to-nose ratio.[15] To achieve a better signal-to-noise ratio class averaging was established. There the raw images of similar viewing angels are clustered. With this method, it is possible to reach a higher signal-to-noise ratio per each class. [40] All these projections of the structure along different viewing directions result in a 3D picture via the central selection theorem.

However, it is challenging to achieve the correct orientation of all images [40]. Due to the compositional or conformational heterogeneity, most of the sample complexes contain more than one unique 3D structure. The classification and averaging problem is not to be negligible. [7, 11, 40] Nevertheless, one advantage of the cryo-EM (reviewed in [15]) is that it does not depend on the possibility to produce a crystal and always yields with some information. One additional advantage is that because the electron micrographs are real space images, which contain amplitude and phase information, there is no phase problem, although the amplitudes are less accurate than the ones of the X-ray diffractions.

1.3 RNA STRUCTURE PREDICTION

1.3.1 *Secondary structure prediction*

The prediction of RNA secondary structure plays an essential role in research. The structure of an RNA is important for the function of the molecule. Evolutionarily conserved structures let assume how important the particular function is. [69] The RNA secondary structure consists of base stacks, built by classical Watson and Crick basepairs (AU, GC), also called canonical base pairs, or wobble base pairs (GU) and loops.

The first ingredient for prediction is based on the energy function. The energy function is derived from the loop decomposition. [31, 69]. The decomposition of the secondary structure results into loops, which are aligned into different degrees regarding the number of base pair that delimit a loop. The degree 1 is assigned to hairpin loops, degree 2 is assigned to internal loops/interior loop and multiloops, e.g. four-stem junction/four way junction, are assigned to a degree >2, figure 4. Additionally the number of unpaired nucleotides within the loops characterizes the loop itself. The total free energy is approximated as the sum of all loop energy that occur in the structure. It is mentionable that the loop type, size and the sequence affect the loop energy. Loops as well as unpaired regions have a destabilizing effect and on the other hand stacked pairs have a stabilizing effect.
Most commonly used computational tools, like `mfold` [74] or `the Vienna RNA Package` [26, 30], can predict the minimum free energy (MFE) [13, 26, 30, 49] secondary structure of a single sequence using the dynamic programming algorithm developed by Zuker and Stiegler [75]. Generally, the standard RNA energy parameters of the Turner group [52] are mostly used by default for the calculations. The parameters are known within certain error limits and the default settings are determined at the smallest error, which is near a temperature of 37°C. [13] The provided MFE structure is also called the optimal structure, the structure with the highest probability and with the lowest free energy.
It has to be pointed out that even if the condition and the model parameters fit together and the thermodynamic model is perfect, the MFE structure will only present a fraction of all possible structures in the Boltzmann ensemble. [26, 30, 49] Additionally a minimal error regarding the thermodynamic parameter set can cause an incorrect predicted structure. Another restriction is that the folding of the correct tertiary structure is not limited to the most favourable secondary structure. As long as the free energy decreases, it is feasible to start from a suboptimal one. The effect of site-specific binding of small molecules, ions, like Mg^{2+} or macromolecules, such as proteins or other RNAs, can affect the folding process. The influence of metal ions stabilize intermediate structures and can lead to long-range interaction. [9] The neglection of these effects can be the reason for a

limited accuracy. [1, 49]
Some of the drawbacks of MFE prediction can be alleviated by calculating the equilibrium base pairing probability via John McCaskills' partition function algorithm [53] . The pairing probabilities can be visualized using a dot plot [13, 31]. A squared grid $(n x n)$ represents the contact matrix of the structure. Each base pair (i, j) is represented by a dot or a box with an area proportional to its probability in row i and column j. In general, dot plots are well suited to represent structures and thermodynamic ensembles on the equilibrium between all possible structures. In general the pairing probabilities allows to calculate the centroid structure and the maximum expected accuracy (MEA). Both are alternatives to the MFE. Besides the MFE secondary structure and only if a single optimal structure is requested, the centroid structure [49] is an additional available choice. It represents the structure with minimal total base-pair distance to all structures in the thermodynamic ensemble. A prediction can be indicated as reliable if the MFE secondary and the centroid structure show a high similarity to each other.
One measurement of the prediction reliability is the ensemble diversity [13, 26], which is the average base pair distance between all structures in the thermodynamic Boltzmann ensemble. It is possible to calculate the ensemble diversity directly from the base pair probabilities.
An alternative type of an optimal structure is the MEA structure [1, 49] extracted from the partition function. The weighted sum of single-stranded nucleotides and base pairs is maximized and for the balance a weighted factor is introduced.
Especially for longer RNA sequences chemical probing experiments, like SHAPE, can increase the accuracy of the predicted RNA structure and can support the structural hypothesis. [1, 49]

1.3.2 *Tertiary structure prediction*

The tertiary RNA structure offers a more detailed picture about molecular details and at the end complements the deficiency of the secondary RNA structure. But the prediction of a tertiary RNA structure is still a challenging task. One way to extend the secondary structure prediction is to use a so called 2.5 D prediction. A 2.5D prediction is an extension of a 2D prediction but no fully-fledged 3D prediction. There exist several ways to extend the secondary structure prediction.
One example is to include non-canonical interactions [54, 69]. Two nucleotides can interact at three different edges: Watson-Crick, Hoogsteen and sugar edge. Additionally they can interact in cis or in trans. Combining all these possibilities together leads to 12 different pairing families with at least two hydrogen Bonds, including the traditional Watson and Crick base pair and 11 non-canonical interactions. On the other hand in RNA there is the possibility to interact with only one hydrogen, forming base triples or multiplets or even G-quadruplex.

These non-canonical interactions are mainly found in helix-helix interactions and are specific for 3D RNA structure elements.

Another approach is to use 3D structure motifs [12, 69]. The unit of ordered non-canonical base pairs, which are regularly found in tertiary structures, is called a motif. A motif is characterized by a well defined base paring and/or geometry. As the geometrical similarity of a motive is related to similar functions, this knowledge helps to support structure analyses by identifying structural conserved motifs. This approach is limited to already detected motifs in known tertiary structures.

Modelling an RNA molecule is associated with several degrees of freedom, therefore and to facilitate the performance, coarse-graining is used by modelling an RNA molecule. For describing nucleotides several dihedrals are used. Thus, the sampling of the structural space is difficult. Coarse-graining helps to deal with this issue.
In some models, the secondary structure can be used as a base for a coarse-grained tertiary structure prediction (comparison of the programs reviewed in [69]). Programs like `Vfold3D` [73], `RNAComposer` [59], `RAGTOP` [36] or `ERNWIN` [35, 70] (section 3.2.3) use this approach. To give a short overview, both `Vfold3D` and `RNAComposer` select the best-matching template for every helix and loop region, based on sequence similarity, given from a library of fragments. `Vfold3D` uses the AMBER all-atom force filed, to relax the full atom model, `RNAComposer` uses the `CHARMM` force field and `CYANE` for the refinement.
`Vfold3D` [69] works more precise the more motifs are found in the structure or the more subsegments are homolog to the target RNA. In addition, both prediction methods offer the possibility to create a random suboptimal structure, with the restriction that there is no provided sampling method or mention energy function for an exploration of the conformational space.
In Contrast `RAGTOP` and `ERNWIN` use a given secondary structure and make use of the level of loops and helices to explore the conformational space. `ERNWIN` is explained in section 3.2.3 in detail. Both methods, `ERNWIN` and `RAGTOP` differ regarding the sampling and the energy evaluation. `ERNWIN` requests the conformation information of local loops and helices from a fragment library and `RAGTOP` works with the angles form continuous space. `RAGTOP` shows that a more aggressive coarse graining can lead to a degraded prediction accuracy. [69]
Compared to all these approaches, `MC-Sym` ([58] and reviewed in [43, 69]) uses a different approach, by sampling adjacent nucleotide cyclic motif fragments from a library of 3D fragments. Within a given time, the algorithm offers as many structures as possible, considering the correct base paring and base stacking given by the input.
Beside the information about the secondary structure `NAST` [33] also can make use of user-supplied tertiary contacts within the structure, which considerably increase the prediction accuracy. This option allows `NAST` to model larger RNA molecules than other computational models. In `NAST` each nucleotide is represented by a single point,

stacked together by the backbone but without any information about the orientation of the residue. Using the provided information, parameters like the energy potentials on the length, angles and dihedrals are computed.

In all the above-mentioned models, the secondary structure is the core for the prediction of the tertiary structure [69]. There are also approaches that can use the secondary constraints only to amplify their predictions [44].
A model that does not necessarily require the secondary structure is `iFoldRNA` [39, 66]. Thus, only the prediction of small RNA structures or fragments are possible. The extended version of `iFoldRNA` integrates secondary and tertiary contacts based on experimental SHAPE chemistry. Thereby `iFoldRNA` was able to improve the prediction. `iFoldRNA` uses three points per nucleotide – the phosphate, the sugar and the base, where the direction of the hydrogen bonding is defined by the relative position between the sugar and the base. Local terms, like the bond length, angles or base pairing and stacking, are included into the energy function. Beside these, bias regarding the loop entropy shall be compensated by an additional energy term. [43, 44, 69] In contrast to `iFoldRNA`, `SimRNA` [10, 51] works with a given secondary structure. Additionally, three points represent each base and for the energy model, non-canonical base pairs are completely integrated. Both, `SimRNA` and `iFoldRNA` make use of discrete, grid-based statistical potentials. `SimRNA` utilizes Monte Carlo simulations and `iFoldRNA` discrete molecular dynamics.
A common used model is `FARNA` [2] which uses a Monte Carlo search for the assembling of the fragments. `FARNA` works with a library of fragments with three nucleotides each and uses a low energy function. Including all-atom Rosetta energy function `FARNA` evolved to `FARFAR`, as reviewed in [43, 57, 69]. `FARFAR`s energy function takes base-paring potentials, planarity of base pairs and steric clashes into account. Also the radius of gyration is included. Favourable energy contributions, if base pairs occur more often in already solved structures, are also incorporated. The concept is that, out of the randomly selected three nucleotide window, one suitable fragment of the library pool is randomly selected. Replacements of the selected fragments leads to an energy change. If the energy of the newly formed structure is lower than the old energy the structure will be accepted.
The mentioned classification of programs is based on the time point when the secondary structure is taken into account and in which range.[69] Another possibility to sort the methods is according to their coarse-graining degree [17, 57] or by considering the chosen potential energy model. Based on this `NAST` and `Vfold` are classified as backbone models. `NAST` is an one-bead model and `Vfold` a two-bead model because each nucleotide is represented by two points (C_4 sugar atom, phospate (P)). `FARNA` belongs to the methods with focus on the nucleobases. The third section are backbone-nucleobase hybrid models.

A possibility to give a short overview of the performance about the prediction models is RNA-Puzzles, reviewed in [54, 69]. It is a community-wide series of CASP-like blind experiments. Therefore sequences of, unpublished, determined, especially crystal structures are provided to computational groups. Within a period of time it is possible to calculate various informations like the prediction of the tertiary structure and the chemical probing. Three offered categories can be attended: first of all, webservers without additional human assistance; second, the opposite, the prediction from experts, without knowing the (pre-) experimental data; and third, the prediction from experts with knowledge about the (post-) experimental date. For the scoring root-mean-square deviation (RMSD), interaction network fidelity (INF) and from the second round on, the mean of circular quantities (MCQ) were measured and compared. RMSD is used for a comparison of the atom coordinates of predicted structure and the native one – an all-atom similarity. Because the base paring nature isn't taken into account with this measurement, INF enables to calculate different interaction types, like stacking or (non) canonical hydrogen-bonding. An additional measurement for the angular coordinates is summarized in the MCQ. RNA puzzles shall facilitate the development of RNA prediction by identifying potential, as well as the limitations of existing methods and drive the development of automated predictions.

Results show that there are still some limitations regarding RNA 3D structure prediction and the accuracy varies with the structure composition. For instance, the influences of the binding of ligands or ions in RNA structures or the conformational changes of the RNA which are essential for the functionality. Besides these limitations the inclusion of non-canonical base pairs is one of the main possibilities to provide an enlarged set of information to enhance the RNA 3D structure prediction.
Two additional facts are essential for the accuracy of RNA folding [44, 54]. On the one hand the information about RNA loop-loop interaction and on the other hand the correct chirality is needed for optimization. If an RNA loop-loop interaction appears within a single RNA strand and contains Watson-Crick or GU-Wobble pairs this is defined as pseudoknot, see section 1.1.1.1.

1.3.3 *Prediction of pseudoknots*

As mentioned, pseudoknots (see section 1.1.1.1) occur in many functionally important RNA structures. The prediction of structures with pseudoknots is, in both 2D and 3D, plagued with barriers. Many tools are not capable of predicting pseudoknots. It is still a nearly computational unsolved problem. The main problem is the limited information about pseudoknot formation. Studying free energy landscapes, folding pathways, dynamics, rearrangements and the final structure including pseudoknots is a central point to facilitate the

prediction of RNA structures. Via coarse-grained based simulations it can be shown that a sequence dependency is often evident by topological similar folding landscapes. Furthermore, the stability of the secondary structure is decisive for the folding of pseudoknots. In several experiments it has been observed that the folding of a pseudoknot is a multiple pathway mechanism especially at low forces or a hierarchical or cooperative mechanism. [9]. Nevertheless there are too few experimental data for the huge number of possible pseudoknots to precise thermodynamic parameters as well as more comprehensive datasets about topologies and sizes. Current models for example do not consider the interdependence of loop length and the stem length in pseudoknots.

Most of the prediction algorithms use simple models for the composition of the energy parameters of a pseudoknot. They are often similar to the composition of the energy parameters of a multiloop.
I will now describe a more complex composition of the energy parameters of a pseudoknot (discussed by [1, 28, 67]) to illustrate common problems with the pseudoknot prediction. The common assumption regarding the energy parameters of a pseudoknot is that the total free energy of a pseudoknot is the sum over all free energies of stems, coaxial stacking (stabilizing negative values), the loop lengths and the sequences and the tertiary interactions within a pseudoknot as well as the assembling that describe the entropy changes within a pseudoknot; equation (2) [1]. Both the values for the stems and for the coaxial stacking are derived from obtained experimental data. Here the most common method to calculate $\Delta G^0_{\text{stems}}$ is the nearest neighbour model. This model is based on the empirical thermodynamic parameters for base stacks.
For the other parameters, like $\Delta S^0_{\text{loops}}$, there are established computational, physical models - equation (3) [67]. $\Delta G^0_{\text{assemble}}$ describes the entropy change of the two subunits, containing the two stems with the attached loops, into the pseudoknots. For the remaining parameters, like $\Delta G^0_{\text{loop sequence}}$ or $\Delta G^0_{\text{tertiary interactions}}$, which describe the possible interactions between loops and stems, there exist neither precise experimental nor computed models.

$$\begin{aligned}\Delta G^0_{\text{pseudoknot}} = &\sum \Delta G^0_{\text{stems}} + \Delta G^0_{\text{coaxial stacking}} - T \sum \Delta S^0_{\text{loops}} \\ &+ \Delta G_{\text{loop sequence}} + \Delta G^0_{\text{tertiary interactions}} \\ &+ \Delta G^0_{\text{assemble}}\end{aligned} \tag{2}$$

T ... temperature

$$\Delta S^0_{\text{loops}} = -k_B \ln \Omega_{coil}/\Omega \tag{3}$$

Ω ... number of the 3D conformation of the loops

Ω_{coil} ... number of the corresponding coil conformations

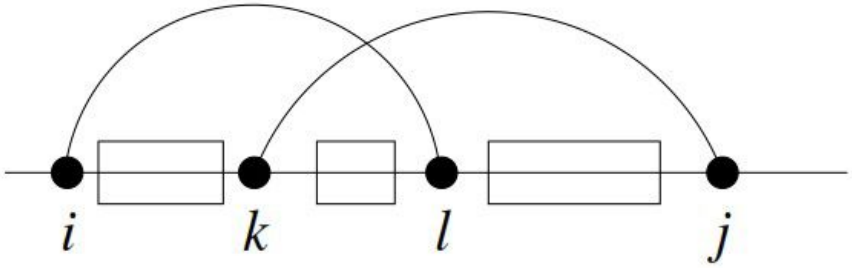


Figure 12: The evaluation of the contribution of pseudoknots required additional computing cost. There for the pseudoknot on [i,j] it is needed to iterate over all combination of the positions k < l. [31]

One approach (discussed by Gultyaev et al.) is to estimate the free energy parameters for H-type pseudoknots by the general theory of polymer loop thermodynamics [28] using available experimental data complemented with phylogenetically proven datasets regarding pseudoknots. The nearest-neighbour model for calculating the stacking energy remains the same. The limitation is the estimation of the loop energies.

An additional, above mentioned problem is the computational complexity. The prediction of pseudoknots is computationally expensive and the prediction of an RNA structure with unrestricted pseudoknots by using a loop based energy model is a non-polynomial time (NP) complete problem [13, 31], as illustrated by figure 12. Therefore various heuristic approaches were developed. Most of the results are not consistent with each other. This occurs due to the different classes of pseudoknots and the used heuristics. New experimental data (reviewed in [1]) show that summing up the calculated nearest-neighbor interactions is inadequate to obtain the free energy for a specific pseudoknots. Despide this, the gotten parameters for the conformational entropy of loops tend to be a good approximation of the entropies.
In addition, the lack of determinant regarding tertiary interaction can be compensated by using as many Watson and Crick base pairs in cis conformation as possible, even if the isolated structure indicates that with this proceeding favourable tertiary structures are replaced.
In conclusion the more specialized and focused on a number of motives or sequences the program is, the less suitable it is for the all-encompassing approach.

There are some programs specialized to predict RNA structures including pseudoknots (as reviewed in [1, 13]). They work fine as they produce a similar structural output compared to the true original structure, but between each other, they produce quite different results. Many of the programs are specialized on a specific class of pseudoknots like `PKNOTSRG` [1, 13] (computing effort: N = length of sequence: $O(N^4)$) which can fold simple H-type pseudoknots. Another program specialised in a pseudoknot class, kissing hairpin structures, is `pKISS`

[13] ($O(N^4)$ or $O(N^5)$), which is an extension of PKNOTSRG. PROBKNOT works with a heuristic regarding the base pair probabilities of unknotted structures ($O(N^3)$) [13].
Due to the fact, that only the most likely base pairs are added to the structure and are allowed to form crossing interaction this program works fast.
A decisive disadvantage is that there is no assertion regarding the energetic optimality. The more precise the program and the higher the possibility to find an optimal, pseudoknotted structure the more the computational effort increases and finally results in an restriction to short sequences like with PKNOTS [1] with a computational effort of $O(N^6)$.
Tools like ILM [1] ($O(N^3)$ or $O(N^4)$) work with no restriction on the type of pseudoknot. The parameters for the thermodynamic approach of the helix stacking are covered by the the Vienna RNA Package [26, 30]. An expansion of ILM is HotKnots [1] which considers alternative secondary structures.

2

ISOLATION AND CLASSIFICATION OF RNA PSEUDOKNOTS - MOTIVATION AND METHOD

The present work deals with the identification of pseudoknots and their further analysis. The aim is to develop a tool that identifies pseudoknots in the 3D structure of noncoding RNA and assign them to different classes regarding their topological genus. Another central theme is the analyses of the frequency and the description of the structural features of pseudoknots in general as well as in certain classes of pseudoknots.

To give an overview about the background of this issue chapter 1 deals with the general structure of RNA and further on with the prediction of RNA structures.
Every knowledge regarding the folding process, the interactions and the energy parameters relative to the RNA structure helps to make any kind of prediction more accurate and more reliable. Thereby not only the overall structure plays an essential role but also information about fragments like for example the use of 3D structure motifs, which help to facilitate a more precise prediction. As mentioned, essential parts that often occur in RNA structures are the pseudoknots, see section 1.1.1.1.
We currently know that there exist several types of pseudoknots in nature, which could be classified into classes and genera. Predicting a pseudoknot, described in section 1.3.3, is still a nearly unsolved computational issue and depends on the lack of information as well as the computational complexity.
Nevertheless, several tools are capable to reproduce the secondary structure including pseudoknots. 3D prediction programs like `ERNWIN` (see section 3.2.3) use the coarse grain model together with a fragment library for the prediction of the tertiary structure of an RNA.

The aim of this work was to find a way to identify the boundaries of a pseudoknot in a given knotted dotbracket notation. The next step was to reduce further structures within a pseudoknot to enable the classification combined with `ERNWIN`. Therefore, the main idea was to use the already known dotbracket strings of special classes of pseudoknots as basis for further computational steps to classify pseudoknots regarding their topological genus.
Another essential point was the analyses of the frequency of pseudoknots within a structures as well as the frequency of certain pseudoknot classes. Furthermore data regarding the composition of pseudoknots were detected and analysed. In addition, the objective was to find a way to show that it definitely makes sense to handle pseudoknots differently compared with normal multiloops and one step further, that it makes sense to assign them into different classes.

To give a more detailed overview about the pipeline of the developed tool the further steps will be described.
The computational pipeline of the developed tool starts with the usage of all PDB files selected by the non-redundant, representative RNA dataset and further constraints (see section 3.1). The next step was to commit them to `MC-Annotate`, tool described in section 3.2.1 (respectively `DSSR`, section 3.2.2) to receive a dotbracket structure based on the already detected tertiary coordinates, including pseudoknots.
Besides the 3D prediction of RNA, `ERNWIN` offers a huge amount of extended tools. Some of them were used in the next computational steps.
From an applied secondary structure, first of all the developed tool skips all occurring dots. On the one hand, this is a safety check that only folded structures were selected from the dataset, on the other hand the interest lies on the brackets only.
The second step includes the removal of all so called multiple stacked brackets. For the isolation and further on the identification of a pseudoknot only one basepair each is needed.
Third and essential step is the isolation of the full pseudoknotted structure from the beginning to the end. Therefore, it is important to consider the correct order of bracket types. A traditional basepair is marked with '(,)', for a pseudoknot additional ones are needed: '[,], {, }, <, >'.
The last step is the removal of all insertions within the pseudoknots.
The developed tool is able to identify pseudoknots of genus 1. For all not assignable ones a revision of the tool was conducted to detect genus 2 pseudoknotted components within irreducible components in the knotted structures.

Besides the identification, the classification and the frequency of pseudoknots, one aim of this work was to analyse the composition of pseudoknots.
Moreover, this thesis also deals with the difference between a classical multiloop and a pseudoknot, which is used to be modelled as a kind of a multiloop in `ERNWIN`. Therefore, a detailed look at the angle distribution between the stems of the respective structure was done and the results were analysed. For this step, the collection of data concerning the structural `ERNWIN` elements also extended `ERNWIN` tools were embedded. The received data enables further data analysis and differentiation between multiloops and pseudoknots. The combination of the data regarding the distribution of the angles within a pseudoknots and the information about classification allows further analyses of the composition of certain pseudoknot classes.
Finally the method for pseudoknot prediction in `ERNWIN` is confronted with the results of the dotbracked based prediction of pseudoknots.

3

WORKING TOOLS

3.1 NON-REDUNDANT, REPRESENTATIVE 3D STRUCTURE DATASET

The number of isolated RNA structures (for methods see section 1.2) listed in the PDB database are huge. However, a detailed look shows that many of the structures are copies of each other or that they are rather similar to each other. Therefore, the information content is less than expected. For that reason, we why we work with a nonredundant list of representative RNA structures.

The non-redundant (NR) 3D structure dataset for RNA is a continually updated, widely automated, list of NR 3D RNA structures provided from the Bowling Green State University (BGSU), Ohio [6, 45].
The source of the introduced dataset is the PDB/NDB database, containing every published protein, DNA or RNA 3D structure. As a result, the dataset of the PDB contains more than one file representing a given macromolecule.
The aim of the NR list is to increase the knowledge extraction, data mining and benchmarking. Therefore, it is essential to implement methods to identify the best-resolved and representative model of each structure.

The majority of the PDB entries are X-ray structures in different resolution thresholds followed by some NMR (described in section 1.2.1) and cryo-EM structures (described in section 1.2.3). The NR-list comes in different resolution thresholds. The resolution thresholds goes from 1.5 A to 4.0 A with a threshold step size of 0.5 A. An additional resolution pool is mentioned with a threshold of 20 A which is a nominal value to include all X-ray (described in section 1.2.2) and cryo-EM structures [45].
The NR 3D RNA dataset contains about 1,479 RNA structures with a resolution threshold up to 4.0 A selected from 7,953 different PDB data files (2017-12-15).

The selection process to get the best-resolved and most representative model of each structure is described in detail in the review of Leontis & Westof, 2012 [45].

The steps to receive a non redundant list are sequence comparison, structural superposition and geometric analysis and the relationship between the chains.

The first step is the sequence comparison by determining the length of each chain and the number of identical bases in an alignment.
The former versions [45] focus on the longest RNA segment in each file. Even if different chains within a file stand for different RNA structures (e.g. 16S rRNA and tRNA), only the longest chain in each file is selected. If two chains have the same length, the first one is chosen. There is only one exception. If these RNA chains in a file are from different species both are selected.
Next, the selected chains are compared with each other. For comparison a Needleman-Wunsch global alignment algorithm with a certain gap penalty is used.
Chains up to a length of 18 nucleotides are only compared with chains of the same length. Chains => 19 nucleotides are compared with chains up to twice their length.
Every chain longer than 80 nucleotides is assigned to be redundant if more than 95% of the nucleotides are identical. For chains with a length of <= 80 nucleotides up to four base differences are allowed. For chains with a length of <19 nucleotides an exact match is necessary to label a sequence as redundant.

A sequence can build different structural shapes even if the sequence shows similarity. Thus, the second step deals with the structural superposition and geometric analysis. Here all chain pairs that are labelled as redundant in the first step are checked for structural similarity.
The geometric discrepancy is defined by geometric measurement that incorporates the general shape of a candidate motif and the orientation of its bases. When the calculated average geometric discrepancy exceeds a certain threshold the structure is considered to be non-redundant.

The third step deals with the similarity between multiple pairs of chains. One concrete example is that we have three chains (A, B, C) and two pairs of these chains (AB, BC) are labelled as redundant. The third pair (AC) is non-redundant. As the other pairs are redundant, also the relationship between (AC) is labelled as redundant. In an updated versions (changes published on [6]) the procedure was extended.
Each file is separated into its Integrated Functional Elements (IFEs) (e.g. 16S rRNA and tRNA). These elements are grouped by equivalence classes as sequence, structure, and species. For each of these equivalence classes one non-redundant representative structure is chosen.

The database is updated weekly, including the download and the annotation of new structures, the rerun of the above mentioned procedure as well as the removal of deprecated structures. The aim of the NR 3D RNA structure dataset is to provide the best copy of each distinct homolog in the PDB/NDB.

PDB ID	Declaration
1E8S	Only P atoms, no C1
1M8W	Chain E,F: Ignored by `MC-Annotate`
1Y1Y	Only P atoms, no C1
3CIY	Stem larger than 40 nts
3J71	Obsolete PDB ID
3J73	Obsolete
3OK4	Chain P: Ignored by `MC-Annotate`
4JZU	Only P and oxygen (O) atoms, no C1
4NIA	Chain o: Sequence missing in `MC-Annotate`
4OQ9	Chain U: Ignored by `MC-Annotate`
4X4T	Different molecules at alt-loc
5ELS	Chain I: Ignored by `MC-Annotate`

Table 2: Overview over the missing NR-entries. References to chapter `MC-Annotate` 3.2.1

To receive this non-redundant dataset first all related PDB files were downloaded – in sum 1502 PDB files. These are more than the 1,479 non-redundant entries, as a PDB-bundle contains multiple files. If a structure contains more than 62 chains and/or 99999 atoms lines it is not possible to reproduce the full structure in a single classical PDB-file, which is still common for many softwares. Therefore, the structural data are split over several files. All these files and an additional mapping file are grouped together to a TAR file [61], which then contains the entire structure .
An additional file format conversion to a .cg file results in 2805 files. The .cg format is a internal `ERNWIN` file format that stores coarse grained helix coordinates. This is because many files contain multiple chains and `ERNWIN` (section 3.2.3) places each connected component of chains into a single file. From these files, all the ones that do not occur in the NR dataset were removed. This leaves 1481 remaining files. 12 files from the NR-list are missing (table 5) due to problems with the PDB entry.
From all these received structures, all the ones with less than two stems were deleted. This is on the one hand due to the fact, that to form any kind of a multiloop a certain amount of stems is needed and on the other hand the angles between the stems are needed for further calculations, section 3.2.3. The obtained structures undergo one additional restriction – a stem has to have a minimum length of two canonical base pairs. A stem with only one base pair indicates that it is more an "on-off" interaction than a stable bound between the nucleotides. There are 617 remaining representative RNA structures.

RNA-type	Compared structures [count]	NR 3D RNA database [average length in nts]	Literature value [average length in nts]
tRNA	84	76	75-90
16S rRNA	8	1,511	1,542
23S rRNA	9	2,241	2,904
18S rRNA	11	1,725	1,874
5.8S rRNA	10	154	160
28S rRNA	1	3776	4,718
5S rRNA	21	116	120

Table 3: Comparison of the chain length of tRNA and rRNAs within the NR 3D RNA database and the literature values [38]

To give a short overview over the structures listed in the NR list, all structures have at least two representatives at the PDB. About half have at least three PDB files, one quarter four, followed by a constant decrease. The structures with the most representatives have the 16S rRNA of Thermus Thermophilus with 381 representatives.

As mentioned in section 1.1 tRNA and rRNA represent a big part of non-coding RNA. Table 3 compares the average lengths of the common known RNA classes between the data from the NR 3D RNA database and the average literature values [38].

The most isolated structures were detected for the tRNA (84). The average length of the tRNA structures from the NR list are within the range given by the literature. For the 28S rRNA there are only one structure in the NR list. There are twenty-one 5S rRNA structures with an average length of 116 nts (literature value 12 nts) within the dataset. Each of the other structure classes have between 8 and 11 representatives. Although this is a small dataset for each structure class, the values are close to the literature values.

In some of the isolated files, there are more than one single structural element included. One example is the 5.8S rRNA. The 5.8S rRNA is a short RNA with around 160 nts (literature value). All of the nine listed structures in table 3 contains an additional 25S rRNA or a 28S rRNA. A manual isolation and detection of all the 5.8S rRNA strains shows an average length of 154 nts. For the comparison of the lengths (see table 3), files whose description listed several rRNA structures or only fragments of a structure were excluded. One exception was the above-mentioned 5.8S rRNA that are analysed manually.

In 2012 Leontis and Westof [45] published that for instance the NR 3D structure dataset contains representative 16S rRNA structures from Escherichia coli, T. thermophiles, D. radiodurans and the homologous 18S rRNA from yeast. In 2017, the NR 3D RNA structure list offered 16S rRNA structures from eight different species and from

the 18S rRNA eleven structures belonging to different species.
In general, the NR 3D RNA structure list includes structures from more than 87 species. For most of them, there were only 12 or fewer entries. The major part of files/structures has no registered classification of species within the PDB (299 registered structures) and one structure has the species attribution undentified (5KPY) [60]. The main part of classified and dedicated structures were isolated from Homo sapiens (36 non-redundant RNA structures), Escherichia coli (35 structures) followed by Saccharomyces cerevisiae with 26 non-redundant structures and Thermus thermophiles with 23 different selected structures.

3.2 RNA STRUCTURE PREDICTION

The aim is to present a short overview about the different prediction methods, especially ERNWIN [35, 70], the different structure elements and how to enhance the computational accuracy. ERNWIN is a program for coarse grain tertiary RNA structure prediction. It uses representative sets of RNA structures, the NR 3D structure dataset [45], as knowledgebase, see section 3.2.3.
For a prediction ERNWIN requests for the sequence and the secondary structure of this sequence.
For de-novo prediction of an unknown structure the secondary structure can be provided for example by the ViennaRNA tool RNAfold [26, 30, 48].

One of ERNWINs dependencies to generate a coarse-grain model out of a already known tertiary structure is MC-Annotate [3, 22, 24, 27]. As ERNWIN can use a PDB file as input, it requires a tool that provides the extraction of the secondary structure from the offered PDB file which only includes the coordinates of the atoms within the tertiary environment. [70] MC-Annotate supports a Linux interface as well as a MAC™ OS X 10.5 interface. There is DSSR [29, 50] as a substitute for MS Windows™ user. Both tools are not only able to handle RNA structures but also DNA.

3.2.1 *MC-Annotate*

MC-Annotate [24], can analyse geometrically and quantitatively a given tertiary structure of any RNA and search for structural motives and patterns.
As mentioned in table 5 there are some structures ignored by MC-Annotate. This is commonly due to irregular format specifications, like misidentified or incomplete nucleotides.
MC-Annotate itself makes use of the MC-Sym library. The motif search is based on a subgraph isomorphism algorithm operating with homogeneous transformation matrices (HTM). These include the relative position and the orientation of two base pairs. Base-base interactions as well as specific nucleotide interactions are identified via distinctive feature factors. The range of these factors is between 0 and 1, where a high value indicates high confidence. The basis for these computations are metrics with information regarding RMSD and torsion angles.
In more detail, MC-Annotate processes the information embedded in the tertiary structure. Therefore, a generated structural graph with information concerning coordinates, conformation, interactions and the torsion angles is used to assign symbols and numbers to each of this geometric information.
The symbols of the nucleotide conformation are based on the relative position of the atoms of the sugar ring related to the general plane and the relative position of the nitrogen base related to the sugar. Thus the symbols can be summarized by the nucleotide type, the

sugar puckering mode ($C1^0$ to $C4^0$, $O4^0$; and in each case "-endo" or "exo") and the orientation around the glycosidic bond (anti, syn). In the next step, the base-base interactions are divided into five different classes combining the pairing, the adjacency and the stacking information.
As the force between non-paired, non-adjacent and non-stacked pairs is not measurable, they are disregarded. The other combinations are adjacent-stacked, non adjacent-stacked, adjacent-paired, non adjacent-paired and last adjacent-non stacked. The base pairing is further classified by symbols corresponding to the types of H-bonds. `MC-Annotate` uses `MC-Sym` as knowledge base for the classification of the base pairing types. One additional factor that should be considered in more detailed is the base stacking. Similar to H-bonds, the base stacking facilitates folding, helps to stabilize a structure and enables a complex structure.
Following the classification of symbols in the graphical representation, each node represents a nucleotide and each edge an interaction. At the end, these set of information's, geometrically and symbolic, is the basis of the list of motifs.

3.2.2 *DSSR*

The second available tool for analysing and annotating, by accepting 3D coordinates from a PDB file, is `DSSR` [29, 50]. `DSSR` stands for "Dissecting the Spatial Structure of RNA" and is a tool of the 3DNA software system, available for Linux, Mac™ or MS Windows™ users, as well as web service combined with `Jmol/JSmol`.
The tool starts with the identification of a nucleotide via the atomic coordinates, combined with the atom names of a base ring. For the validation of a nucleotide at least three base ring atoms have to be assigned to one residue and additional a certain cut-off (by default < 0.1 A) concerning the RMSD has to be passed. The recognition of modified nucleotides is possible and is distinguished with lower case letters instead of the traditional upper case ones (A, C, G, U). For an accepted nucleotide the position and the orientation in the structure is defined. In addition, the information about geometric features regarding the base edges is stored. In further steps hydrogen bonds are detected by using a geometrical approach and base stacking within the structure is identified.
For base pairs, five parameters have to be applied: the distance between the residues, the vertical separation between the base planes, as well as the angles between the base vectors and last, parameters regarding the basepair stacking and the presence of hydrogen bonds. The values are set by default and it is possible to detect canonical and non-canonical basepairs. Besides the common base pair names like Watson and Crick, the base pairs received several describing characters like the information about the relative base orientation ("-" normal, "+" flipped) or three translation parameters and the respective three rotation parameters. The rotation parameters describe

the nonplanarity. It is mentionable that the translation parameters for a canonical base pair are close to zero. Three nomenclature approaches regarding the base pairs are available – first concerning the 28 hydrogen-bonding types (Sanger), second the 12 classes by Leontis-Westhof regarding the basic geometric classes and third a DSSR specific one including the base orientation by defining three base-centric interaction edges.
DSSR can identify higher-order co-planar base associations (multiplets), helices, stems, isolated canonical basepair and coaxial stacking by certain definitions. Here a helix has a minimum length of two basepairs and in contrast, a stem is built up with a continuous backbone and only contains canonical base pairs. The reason for the special treatment concerning isolated canonical or wobble base pairs is that they can play a crucial role in folding, structure stabilization and characterization of pseudoknots. Structural elements like loops can be assigned into three different classes: hairpin loop, internal loop and multiloop. For the representation of pseudoknots DSSR uses the dot-bracket notation, see section 1.1.1.1. The offered secondary structure is provided in three common file formats (dbn, ct, bqseq).

3.2.3 *3D structure prediction in ERNWIN*

The tertiary structure of an RNA molecule can provide essential and additional information about its function as well as for the regulation. With the RNA sequence and the 2D dot-bracket structure, which is described before (section 1.1.1.1), ERNWIN can provide a coarse-grain 3D structure using only a few degrees of freedom by using a Markov Chain Monte Carlo simulation. The predicted RNA structure is parameterized by the angles and shifts between helices and is sampled by using a simplified representation. In general, a helix is described by the starting and the end position or respectively by the direction of the helix presented as an vector. In addition an angle describes the twist of the helix and thus determines the position of the major and the minor groove.

Despite this simple representation ERNWIN achieves comparable prediction accuracy to other programs. However due to the strong coarse-graining it achieves much better sampling of the conformation space.

The primary output is a sample of many structures.
The output includes a list of all used structural motives and the number of nucleotides in the structure. Beside the 3D structure, ERNWIN offers various values of structural parameters such as the asphericity, the relative shape anisotropy and the radius of gyration. As RNA molecules tend to form compact structures, the radius of gyration is a characteristic measurement for the compactness of a structure and can be complimented with the asphericity and the anisotropy.[5] A provided sampling trajectory gives an idea what structures where sampled in what order.

ERNWIN is based on the Forgi [71] library for manipulating RNA as a graph-like structure. Forgi provides the functionality to parse various representations of secondary RNA structures.

First of all a dot-bracket notation (with or without pseudoknots) such as delivered by RNAfold, for unknown structures, or MC-Annotate/DSSR ,if a prediction of an already known model is requested, is needed . This representation is commonly used for the Webservice ERNWIN. Therefore, it classifies and stores 2D RNA structures in a specific data structure. It stores information about the connections between elements and the nucleotide composition of these elements.
ERNWIN distinguishes six different types of structural elements. First, there are the unpaired nucleotides at the 5′ or the 3′end of a molecule. The ending/starting position is indicated by the first/last stem. Both have separate sets of parameters. The second structural element is the stem. It is a region of contiguous canonical Waston-Crick base-paired nucleotides and there can be many in the structure. Third, there are the different types of loops: interior loop, hairpin- and multiloop.
The interior loop is characterized by a double-stranded unpaired region flanked by stems at either side. The hairpinloop is characterized by two regions of the same strand base paired and end in an unpaired region – the loop. The last structural element is the multiloop, which is a single-stranded unpaired region. Each of these mentioned structures are visualized as cylinders with different diameter and colour, figure 13. The length of one cylinder is the length of the particular secondary structure. It is mentionable that during the whole predicting process the secondary structure is never modified.

Sampled orientation parameters for each of these elements are used for the insertion of the element into the coarse grain structure. Additionally the length as well as the twist parameters of a stem are taken into account for the sampling. The sampling is a step wise mechanism, thereby the sampling starts with first element and places the following in relation to the one before by determining the orientation of the previously, already built structure. For each step an element of the correct type and the correct length is randomly picked from the library database. Here the sampling of the elements is independent from the sequence. The parameters are the length of the cylinder and the connection type between two cylinders (described in detail in section 3.2.3.1 and in table 4).
The sampling offers locally correct structures but does not include long-range interactions and the prevention of clashes. The energy of the resulting structure is calculated, validated and accepted or rejected.
For evaluating the energy of the structure five terms are used. Two terms should prevent the integration of physically impossible structure elements, like preventing helices that occupy the same physical space. One term is also related to the building of multiloops. Multiloops present a problem in this scheme because of the cycle dependencies within the loop. Multiloop modelling is described in section 3.2.3.1.

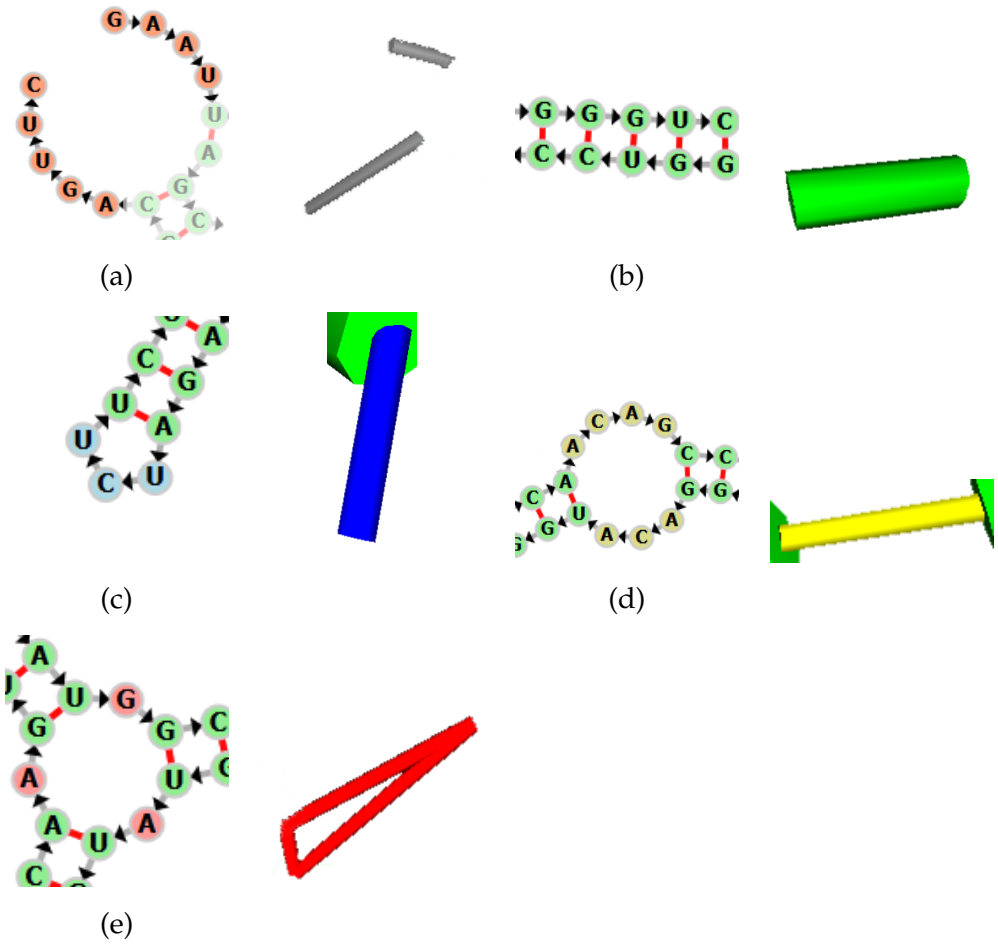


Figure 13: Elements in the secondary structure versus the representation in ERNWIN (a) 5′ and the 3′end (b) Stem (c) Hairpinloop (d) Interiorloop (e) Mulitloop
Structure elements from PDB file 1GID chain A [14], illustrations modified from ERNWIN [70]

The three remaining terms are derived from known structures, like the energy based on the radius of gyration. The second one of these three terms is related to the A-minor motif which is defined by the insertion of adenines into the minor groove of a stem. Generally the A-minor as a long range interaction often occurs in several RNA molecules. For the energy term the probabilities of A-minor interactions concerning a particular helix-loop pair are calculated and each loop is scored by the number of interactions. Third the loop-loop interaction energy is conserved with particular aspect to hairpin-hairpin interaction.
To compare the similarity between two structures as well as structures from other tools the RMSD is computed.

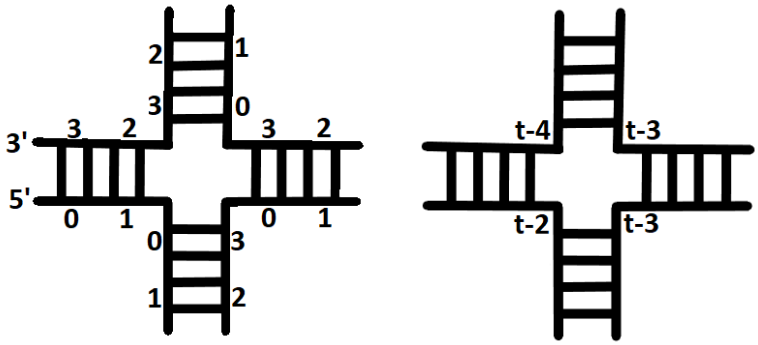
Beside the ERNWIN Webserver an upgraded version of ERNWIN and Forgi via Github and a detailed instruction of the system is available via direct download [70]. The download version of both ERNWIN and Forgi are equipped with several additional features. One feature

is to show a graph like representation of a structure, like in figure 26. Another one is the possibility to use different file formats within the system. Beside the internal ERNWIN file format "cg" that stores coarse-grained helix coordinates, ERNWIN can also work with PDB files and other additional formats. One example are BPSEQ formatted files that use one line for each nucleotide in the sequence. Each line is separated in three columns providing the index of the nucleotide, its identity and its pairing partner. The common known "fasta" file format containing a structure ID, the sequence and the dotbracked notation can be used as well.
Besides the kind of the input and output files, the download version offers many additional features and calculated values and data, like information about the energy status or the number of iterations. In addition, there are a number of possibilities to get more information about the structure e.g. finding the partner of a base pair, extracting a pair table, finding the longest stem or getting the sequence of an element and its neighbours. For the download version the computed structure can be visualized via Pymol [65].

Note that the structures used for the benchmarks were collected and filtered from the non-redundant 3D structure (section 3.1). For the parametrization of the individual elements (figure 13), the dataset Version 3.0 with the date 2017-12-15 (1,466 entries filtered from 7,953 PDB files with a resolution threshold up to 4.0 A), is used. The parameters are collected in a “stats”-list. The ERNWIN package provides different stats-lists and in addition, there is the possibility to use more than one list in one calculation, where the files will be used in the order specified. It is also advantageous that the download version of ERNWIN can work with a list of fallback stats with additional structural parameters. This can be useful if insufficient stats are found in the normal stats file for a coarse-grained element. The fallback stats lists are based on the Rosetta RNA3D structure prediction.

3.2.3.1 *ERNWIN and the prediction of pseudoknots*

While the published version of ERNWIN [35] deals only with pseudoknot free structures, a crude method for handling pseudoknots was added since. This method treats pseudoknots like multiloops.
Therefore, the strategy how a structure is built, especially a multiloop, is needed.
In section 3.2.3 the sampling process, which works like a skeleton graph, was introduced. For building a multiloop the multiloop segment with the highest number of nucleotides has to be broken down and will be recomposed again in the last step within the reconstruction of the multiloop. Within the energy function there is a included term regarding the junction closure detection. As after each building step the energy of the new structure is calculated, for impossible "last" multiloop segments a large energy penalty is imposed. If a segment is inconvenient or one step further if there is no segment left that com-



(a) (b)

Figure 14: ERNWIN and the classification of connection type due to the stem sides of multiloops:
(a) Stem sides (b) Conntection types

plies with the parameters, the previous structure will be stepwise removed and rebuild.

ERNWIN can handle pseudoknots - more or less – it can identify the middle segment of a pseudoknot due to its connection type (described in table 4) which will be described in detail below and for the prediction it uses the same parameters like for a multiloop. The only difference is, that it can identify a pseudoknot.

In the following, we will number the four sides of a helix 0,1,2,3. Taking a look through figure 14 and table 4 the unpaired segements of a multiloop can be labeled. Each number is related to one side of a stem starting from the 5 prime to the 3 prime end and the connection type is defined by the sides which are connected. Where a classical multiloop only needs connection types 2 to 4, a pseudoknot needs connection type 5 instead of connection type 3. Connection type 5 marks a middle segment of a pseudoknot. A simple interior loop is labelled with 1. As an additional sampling parameter, for the structure composition the randomly selected stem from the library has to belong to this kind of connection type.
Additionally ERNWIN can count all segments of a loop containing a segment of type 5 forwards a pseudoknot, not only middle pseudoknot segment. This is important for huge pseudoknotted structures or even simple ones like a kissing hairpin which contain two middle segments of a pseudoknot, see figure 15. However this is not used for sampling.

With this rather simple method ERNWIN can detect pseudoknots in 215 structures in the dataset mentionrd above in section 3.1. Most of these structures, 148, contain one single pseudoknot. 19 structures contain two pseudoknots and 17 three. The number of pseudoknots per structure go up to 10, but in most structures only one pseudoknot was detected. Note that ERNWIN does not differentiate between any class or genus of a pseudoknot.

stem from - to	connection type	declaration
1 - 0	t-2	first segment of normal multiloops and most pseudoknots
3 - 0	t-3	middle segment of multiloop
3 - 2	t-4	last segment of normal multiloops and most pseudoknots
1 - 2	t-5	middle segment of pseudoknots

Table 4: ERNWIN: Connection types for multiloops and pseudoknots.
For examples see:
figure 14 for a multiloop
figure 15 for pseudoknots

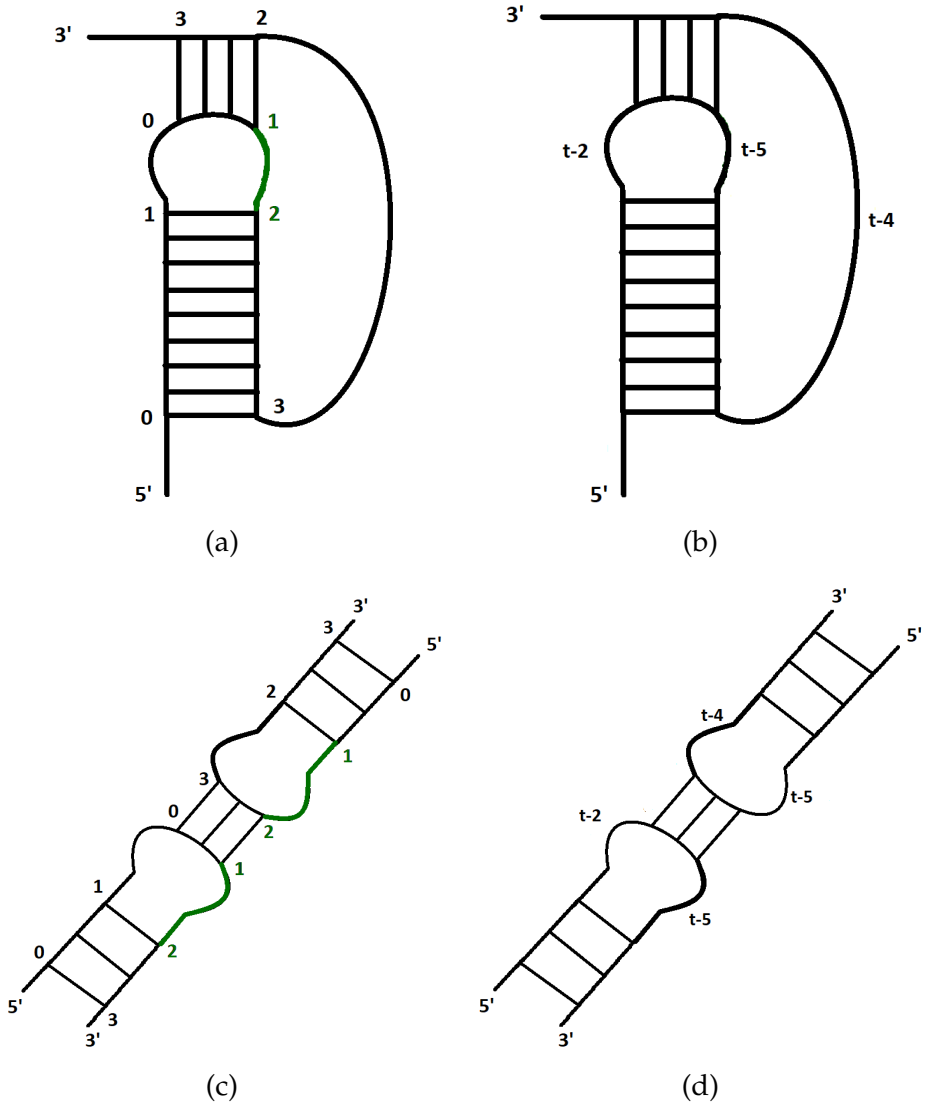


Figure 15: ERNWIN and Pseudoknots: Identifying of pseudoknots via the connection type defined by the sides of a stem which are connected.
Green marks the middle segment of a pseudoknot
H-type pseudoknots: (a) Stem sides (b) Conntection types
Kissing hairpins: (a) Stem sides (b) Conntection types

4

RESULTS AND DISCUSSION

In this section three kinds of results will be discussed. The first part presents an overview about pseudoknots in general - which classes and genera are detected and in which quantity. Are they found in classical RNA structures and are they specific for one of them? How does the minimum stem length influence the occurrence of pseudoknots? In the second part a detailed analysis regarding the angle distribution between the stems of the different classes and genera of pseudoknots is presented. The third part deals with the already implemented detection of pseudoknots in `ERNWIN` and their accuracy compared with the detection via the dotbracket structure. The used dataset is described in section 3.1. For the detection and measurement of the analysis `ERNWIN` with the subtool `MC-Annotate` was used.

4.1 PSEUDOKNOTS IN THE NON-REDUNDANT, REPRESENTATIVE 3D STRUCTURE DATASET

From 617 selected representative structures, 200 contain one or more pseudoknots. Most of these structures (129) contain only one pseudoknot. As a few structures contained many pseudoknots, in total there were 187 pseudoknots of genus 1 (for a detailed description see table 1 in section 1.1.1.1) detected. There the number of the classes are more or less width scattered. In 107 structures there were a H-type found, in 59 a kissing hairpin and the other ones are distributed between L (3) and other pseudoknotted structures (72), see figure 16.
29 structures form two pseudoknots. Within these structures there are mainly found a combinations of H-types and other knotted structures. For genus 1 kissing hairpins the maximum number is only one per structure. The maximimum number of genus 1 H-type pseudoknots within a structure is three (PBD ID: 5jte.AY, descripiton: E-site

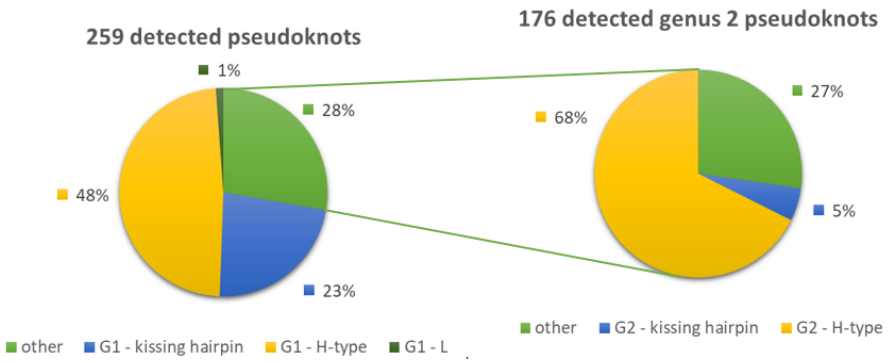


Figure 16: Pseudoknots in the non-redundant 3D structure dataset

tRNA Valine [4]) but in the main subset there is a mix of H-types, kissing hairpins and other pseudoknots. All the other remaining 72 knots are inspected for genus 2 pseudoknotted components within of irreducible components in the knotted structurests (description in section 1.1.1.1, example figure 11.
We will call an H-type pseudoknot that is nested in another pseudoknot such that a genus 2 pseudoknot is formed a "genus 2 H-type pseudoknot". The same we will also do for all the other shadows defined in table 1. Out of this dataset 119 genus 2 H-types and 9 genus 2 kissing hairpins were detected, mentioned in figure 16. The count of genus 2 H-types per structure goes from none up to four. Genus 2 kissing hairpins occur on their own and "other" only twice per structure.

A detailed look within the "other" structures shows, that they can be assigned into 21 different shadows. Most of them (14) occur only once, but there are also some recurring ones.
One knot with the shadow ([])(([])({})])(])({}) has to be highlighted as it was detected nine times in different structures. The shadow of the pseudoknot is illustrated in figure 18. Additionally, each of the nine structures that contains this special pseudoknot is a knotted one, with at least three up to six pseudoknots per structure. All structures are a kind of rRNA - 5x 23S rRNA, 1x 26S rRNA, 1x 4.5S rRNA an 23SrRNA complex, 1x 21S rRNA, and 1 x 28S rRNA.
Another mentionable pseudoknot is, ([])((([])({})])(])({})). It occurs five times in different species and always in 5.8S rRNA complexes. In four cases it is a 5.8S rRNA + 25S rRNA complex and in one case it is a 5.8S rRNA + 28S rRNA complex. All five rRNA complexes are part of the large eukaryotic ribosme subunit.

The comparison of the structure length with the number of pseudoknots shows that the longer the structure is the more knots are within the structure, as illustrated in figure 17. Within the dataset, structures without a pseudoknot have an average length of 69 [nts]. The length of these structures varies from 13 [nts] to 500 [nts]. Only four structures show a length of more than 500 [nts]. The length of structures that contain one pseudoknot have range from 17 [nts] to 1972 [nts] with an average of 145 [nts]. Furthermore, structures with two pseudoknots have a length of 40 to 2022 nucleotides. There is a huge difference of structure length between structures with three pseudoknots (average length 1505 [nts]) and those with four (average length 3359 [nts]) or more pseudoknots. The structure length remains nearly the same for structures with five, six and seven pseudoknots.

From the 617 selected representative structures, two cg files gave an error while analysing the structural components, PDB ID 4K4S [25] and 2RFK [46]. This error occurs due to an error in `ERNWIN` that is related to the multiple chain characteristic of the cg files.
Additional four structures contain an additional pseudoknots that are erroneously detected as an "other" pseudoknoted structure although they have the shadow of an H-type or a kissing hairpin. By a manual analysis of three of the structures, PDB ID 2NZ4 [16], 2Z75 [37] and

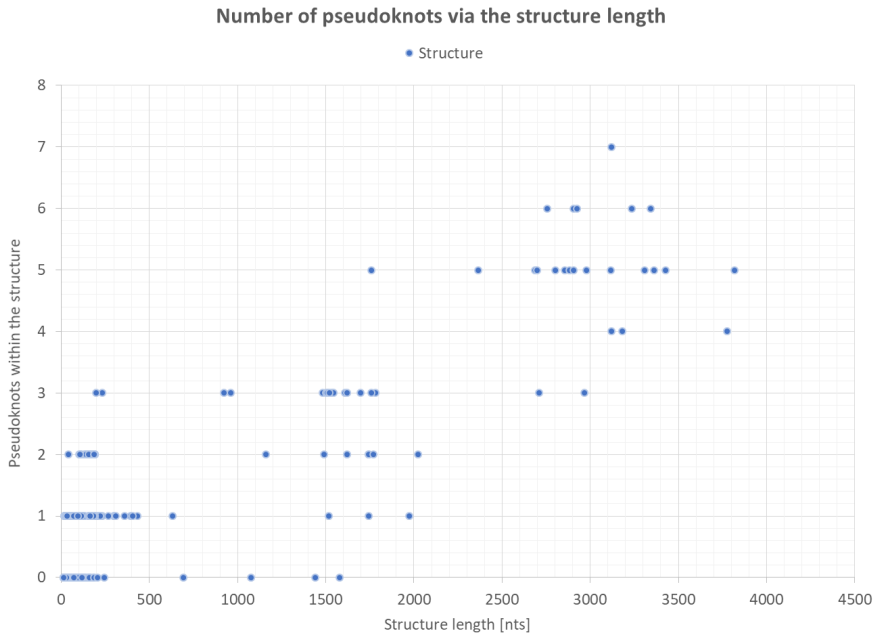


Figure 17: The plot compares the number of pseudoknots versus the length of the corresponding structure. Each dot represents one structure.

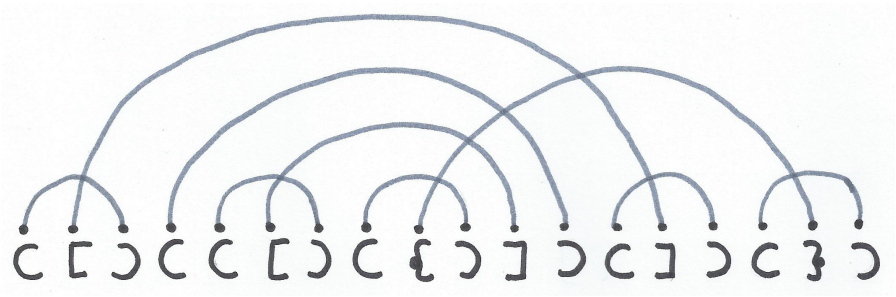


Figure 18: Shadow of a pseudoknot labeled as "other". This special pseudoknot occurs in nine different structures.

5lzd [21], contain an additional genus 1 H-type, one structure (3j7q [72]) an additional genus 2 kissing hairpin and the structure PDB ID 5t5h [47] an additional genus 2 H-type. This can be based on a wrong condensed structure, which is caused by the multiple chain characteristics. With the new unbundled PDB format mmCIF and the possibility to use it with `DSSSR` and respectively `ERNWIN` this error will probably be rectified. From two structures (5MPS [20], 5LJ3 [23]) the isolated condensed structure is also classified as an other but on first sight they look like a classical H-type. The classification is correct as the H-Type is partially closed, then interrupted by some stems, interior loops and also knotted with another H-type, until the first H-type is completely closed. This remaining shadow looks like an H-Type but is neither a genus 1 nor a genus 2 type. The "multiple chain"-handicap only affects the deletion of one remaining nested pair.

As described in section 1.1 ribosomal RNA and tRNA play an essential role within the RNA world. One additional important group

are riboswitches. I will now describe our results with respect to these three classes. Table 5 gives an overview about the distribution of pseudoknots and its classes through these well known classes of RNA. From 200 detected structures with pseudoknots, 12 are tRNA and therefrom the majority (11) contain a genus 1 H-type pseudoknot.

In a large part of the 16s rRNA as well as in the 23s RNA there are also pseudoknots found. It is interesting that all of the seven 16S rRNA structures contain three genus 2 H-types per structure. It is mentionable that the 23S rRNA seems to be a multiple knotted structure.
In all eleven offered 18S rRNA at least one pseudoknot is found.
Also in many riboswitches pseudoknots were found. Compared to the other listed structures riboswitches manly contain genus 1 pseudoknots. As a result, riboswitches are less knotted than other structures. This is also based on the fact, that they have a short structure length.

There were three separated 5.8S rRNA chains within the dataset. Each of them build a kissing hairpin. The remaining of 5.8S rRNA chains are combined with 25S rRNA or a 28S rRNA. Within these complexes, there occur many different pseudoknots. This can be one result of the interaction between these two rRNA chains. It is apparent that in these complexes not all pseudoknots contain only long-range interactions. Based on this the occurrence of pseudoknots with a default helix length within a pseudoknot were analysed.

Within our dataset a stem has to have a minimum length of at least two canonical base pairs. In general, the longer the interaction side is the more stable is this interaction within the stem. In order to obtain an overall picture about the stability of pseudoknots also pseudoknots with a minimum stem length of three and four basepairs were evaluated.
It has to be mentioned that all unpaired basepaires were removed first, and afterwards this restriction becomes effective. More specific, within a stem there can appear unpaired nucleotides, in form of bulges (e.g. figure 4: “single-nucleotide bulge”, “three nucleotide bulge”) or interior loops with only one unpaired base pair. Such a bulge destabilizes a helix. However, if the sum of nucleotide basepairs of the total helix is high enough, it can still be a very stable helix. If the restriction of a minimum stem length of three basepares is applied to a dot-bracket structure including unpaired nucleotides, e.g. “((.((.....)).))”, some structural elements will be excluded even if they are able to form a stable stem. Excluding the unpaired nucleotides before the minimum stem length come into effect, prevents this possibility, e.g. “((.((.....)).))” results in “(((())))”.
Table 6 gives an overview how the number of structural elements within a structure, especially concerning pseudoknots, changes with such a simple restriction. The more basepairs are required the more structures have no stems and the fewer pseudoknots are detected. By doubling the number of required basepairs, the structures with pseudoknots decrease more than half.

	Representative structures	Containing pseudoknots	G1 H	G1 KH	G2 H	G2 KH	others
Total set	617	200	107	59	52	9	47
tRNA	84	12	11	0	0	0	0
16S rRNA	8	7	0	0	7	0	0
23S rRNA	9	7	6	6	7	3	5
18S rRNA	11	11	1	0	10	1	4
5.8S rRNA	3	3	0	3	0	0	0
5.8S rRNA + 25S/28S rRNA	7	7	7	6	7	1	7
28S rRNA	1	1	1	0	1	0	1
5S rRNA	21	0	0	0	0	0	0
Riboswichtes	54	38	23	14	0	0	1

Table 5: Allocation of RNA of the classic RNA structures and the distribution of pseudoknots within these classes.
Within this table, "G" stands for genus, "H" for H-type and "KH" for kissing hairpin.
The column "containing pseudoknots" represents a count of all representative structures with pseudoknots.
The following columns shows the number of how many structures contains a certain genera/class of pseudoknots.

This effect is also notable with genus 2 pseudoknots where by a minimum of four basepairs almost all pseudoknots dissolve. The result regarding the genus 2 pseudoknots also gives information about their composition. Genus 2 pseudoknots are built with shorter helices than genus 1. In general short helixes are harder to predict.
The identification of the pseudoknots includes subtools from `ERNWIN`. `ERNWIN`s' proceedings are based on stems with a minimum size of two basepairs. This specification combined with the fact that `ERNWIN` does not yet fully support multiple chains causes in the implementation that some structures are not detectable anymore. In case of a limitation of three basepairs 27 structures are affected and of four basepairs 35 structures.

To conclude the basic facts, the longer a structure the more pseudoknots are within a structure. The results demonstrate that pseudoknots are not only a structural element for particular RNA structures but pseudoknots occur in many well known rRNA structures as well as tRNA.

As mentioned in section 1.3.2 there are prediction programs that start from a secondary structure. `ERNWIN` also starts from a given sequence and a given secondary structure. Based on the secondary structure specific requirements are to be applied for dealing with pseudoknots. One fundamental requirement is to differentiate between an internal

	2 basepairs per structure	3 basepairs per structure	4 basepairs per structure
Unfolded structures	0	5	11
Structures without pseudoknot	417	451	485
Structures with pseudoknots	200	134	86
GENUS 1		[count of pseudoknot]	
H-type	125	76	56
Kissing hairpin	59	22	27
L	3	1	1
GENUS2		[count of pseudoknots]	
H-type	119	74	4
Kissing hairpin	9	13	0
OTHERS		[count of pseudoknots]	
	48	25	8

Table 6: The basis dataset (mentioned in section 3.1) has the restriction of a minimum stem length of two basepairs. Therefore there are no unfolded structures admitted. The increase of the basepairs leads to the occurrence of unfolded structures within this basis dataset. Simultaneously the number of structures with pseudoknots and the associated total number of pseudoknots within a class decreases.

interaction and a long-range interaction that leads to a pseudoknot. Thus one of the main points is to differentiate between an internal interaction and a pseudoknot with high confidence. A single basepair even if it is a Watson Crick basepair should be counted as an internal interaction. To form a helix at least two basepairs are needed. In contrast to a single basepair interaction, this interaction is more stable. However, in total this interaction is too unstable and only affects the structure to a limited extend. The longer the helix is the higher is the level of confidence that a structure contains a stabile pseudoknot. In addition, the analyses of the dataset with a minimum helix length of three or four basepairs shows that the longer the helix is the less complex are the pseudoknots. It can be seen that there are less structures with pseudoknots. Furthermore, helices within a pseudoknot with a helix length less than three/four are labelled as internal interaction. As a result pseudoknots initially labelled as "other" are now classified as one of the simple classes, like H-types and kissing hairpins.

Based on the frequency of different classes of pseudoknots we can conclude that the most frequently detected class of pseudoknots is the H-type pseudoknot (genus 1 and genus 2). Although a higher genus indicates for a more complex structural surface, both genera of an H-type pseudoknot have the same shadow (see figure 10 (a)). This leads us to another question: Do genus 1 H-type pseudoknots behave similar compared with genus 2 H-type pseudoknots? For that reason,

it makes sense to look more closely at this class (see section 4.2). Another important class are the kissing hairpins. Kissing hairpins occur especially in riboswitches as well as in rRNA. To established a well-founded library for the sampling there has to be enough data. The more differentiations within a dataset exist the merrier the higher is the risk to get non-representative libraries. In contrast, it is not reasonable to combine data concerning one structural component in one library if they behave different. The kissing hairpin is the second most common shadow within the pseudoknots. Therefore, we had a detailed look within the composition of a kissing hairpin and compared it with the composition of an H-type (see section 4.2).

4.2 PSEUDOKNOTS AND THE ANGLE DISTRIBUTION BETWEEN STEMS

To treat a pseudoknot different from a normal other multiloop a dataset that differentiates between these two groups is needed. One possibility is to evaluate the angles between stem pairs within a structure. These angles provide further information about the composition of a structure. Furthermore, the analyses of the angles in certain pseudoknot classes facilitate the differentiation of these classes.

For the sampling each loop is defined by six parameters (see section 3.2.3). Three parameters describe the distance to the next sampling element, one parameter describes the twist and the last two parameters describe the direction of an element. `ERNWIN` takes the original helix and replaces it with an ideal helix. The direction is measured based on the ideal helix. In `ERNWIN` a helix is illustrated as a stem.
A stem is described by the length and the twist. `ERNWIN` samples fragments with these six parameters for each single stranded region in multiloops and pseudoknots (section 3.2.3). The angles between stems are used as a proxy for the mentioned six describing parameters. It has to be mentioned that the angles are measured from the outermost to the innermost basepair.

In general it is plausible to put angles from different classes together in one fragment pool if the angle distribution is similar. In contrast, if they differ, it makes sense to separate them and treat them differently. This thesis uses histograms for the illustration of the angle distribution between stem pairs. Each angle is measured within a range of 0° and 180° and plotted in a histogram. The histogram is separated into 60 parts were the angles are grouped and counted. The red line represents the kernel density estimate. It is important to mention that angles around 0° and 180° are associated with coaxial stacking.

One important aspect regarding the coarse-grained prediction are the angles between the stems. The histogram 19 (a) shows the distribution of the angles between all stem pairs of detected pseudoknots. In sum, there were 2,088 angles within all pseudoknots measured. It is visible that the angle distribution is not distributed evenly between the whole fields of 0 to 180 angular degrees. The majority of angles are acute or obtuse ones. There are less right or straight angles found within the pseudoknot structure. The results show that on the one hand the stems within pseudoknots tend to coaxial stack with each other but on the other hand this is not a complete coaxial stack.

Complementary to the angle distribution of pseudoknots, the angles between the stems were also measured within a multiloop, figure 19 (b). In sum this set of angles includes 1702 datapoints. Due to the composition of a multiloop the arrangement of stems is different from pseudoknots. It has to be pointed out that both do not show complete coaxial stacking within their structural composition. The analysis shows that the distribution of angles differ between multiloops

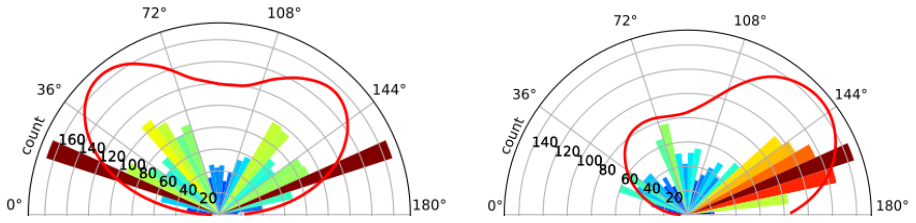


Figure 19: (a) The plot shows the distribution of the measured angles between all pairs of stems of all detected pseudoknots. In comparison to this plot (b) illustrates the angle distribution measured between the stems in a multiloop.

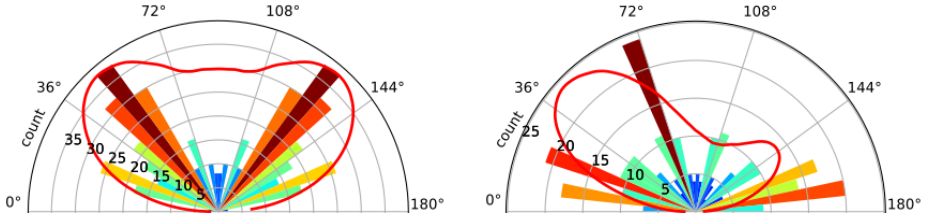


Figure 20: Plot (a) shows the distribution of the measured angles between all pairs of stems of all detected genus 1 H-types.
Plot (b) shows the distribution of the measured angles between all pairs of stems of all detected genus 1 kissing hairpin without the angle ϵ mentioned in 24 (a).

and pseudoknots. This result indicates that a different treatment of multiloops and pseudoknots is obvious.

In addition it is important to answer the question: "Does the two mean classes of pseudoknots differ in their angle distribution". To answer this question all angles between stem pairs within the class H-type and kissing hairpin were calculated and analysed. There were to less L and M pseudoknots detected for further analyses of these structures.
For the analyses only genus 1 structures were taken into account. Figure 20 shows the distribution of both classes in a histogram. While the histogram illustrating the distribution within the H-types shows a symmetry axis at 90° there is no symmetry within the angle distribution of the kissing hairpins. The distribution of angles within the kissing hairpins shows a conspicuous accumulation of angles at about 70°. Nevertheless, the kissing hairpins show more angles around 0° and 180°, which presume coaxial stacking whereas the H-type angles are either around 40° or 140°.
From these distributions, we can conclude that the angle distribution of an H-type and a kissing hairpin behave different. For that reason, it make sense to differentiate between not only pseudoknots and multiloops but also between different classes of pseudoknots.

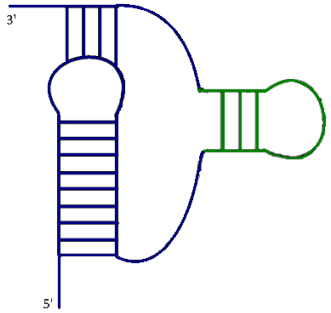


Figure 21: The Figure illustrates an example of an inserted element within a pseudoknot. Blue marks the pseudoknotted structure and green an inserted interior loop.

The more data are available that describe the composition of a pseudoknot the more precise the following prediction. Furthermore, a more detailed description of the composition leads to a higher accuracy of a prediction. To enable a detailed picture inside the structure of a pseudoknot and to demonstrate that the composition concerning the angle distribution differs, two classes (H-type and kissing hairpin) were further analysed.

The information which angles are analysed are found in figure 22 (a) and figure 23 (a). It has to be pointed out that within an H-type the sum of the angles α and β has to be 180°. Within a kissing hairpin the angles α and β have to sum up to 180° as well as the angles γ and δ.

Regarding the sampling process, it makes sense to take a detailed look at the angle distribution between stem pairs without an inserted element. Figure 21 illustrates an example of a pseudoknot with an inserted element (here an interior loop). An inserted element influences the composition of the stems within the pseudoknot and induce a shift within the angles. For this reason, only H-type pseudoknots and kissing hairpins without any inserted element were further analysed.

The distribution of the angles illustrated in the histogram 20 (a) are reflected in the angle distribution of α and β mentioned in figure 22. The sum of both angles have to be 180°.

The disadvantage of a very detailed sampling library is that the more separation within a dataset (e.g. pseudoknot -> H-Type, kissing hairpin) the less the remaining available data are, to build a representative sampling library. To extend the sampling library, also genus 2 H-types were further analysed regarding their angle distribution.
The outcome is that the angles of genus 2 H-types are distributed within the same range as genus 1 H-types. One exception is that the majority of angles indicates that there is rather coaxial stacking in genus 2 than in genus 1.
Either in the angle distribution of genus 1 H-type nor in the angle distribution of genus 2 H-type perfect coaxial stacking are found. This can be due to a kink in the helix. Such a kink can pervades through the whole helix structure. An additional reason can be that two short

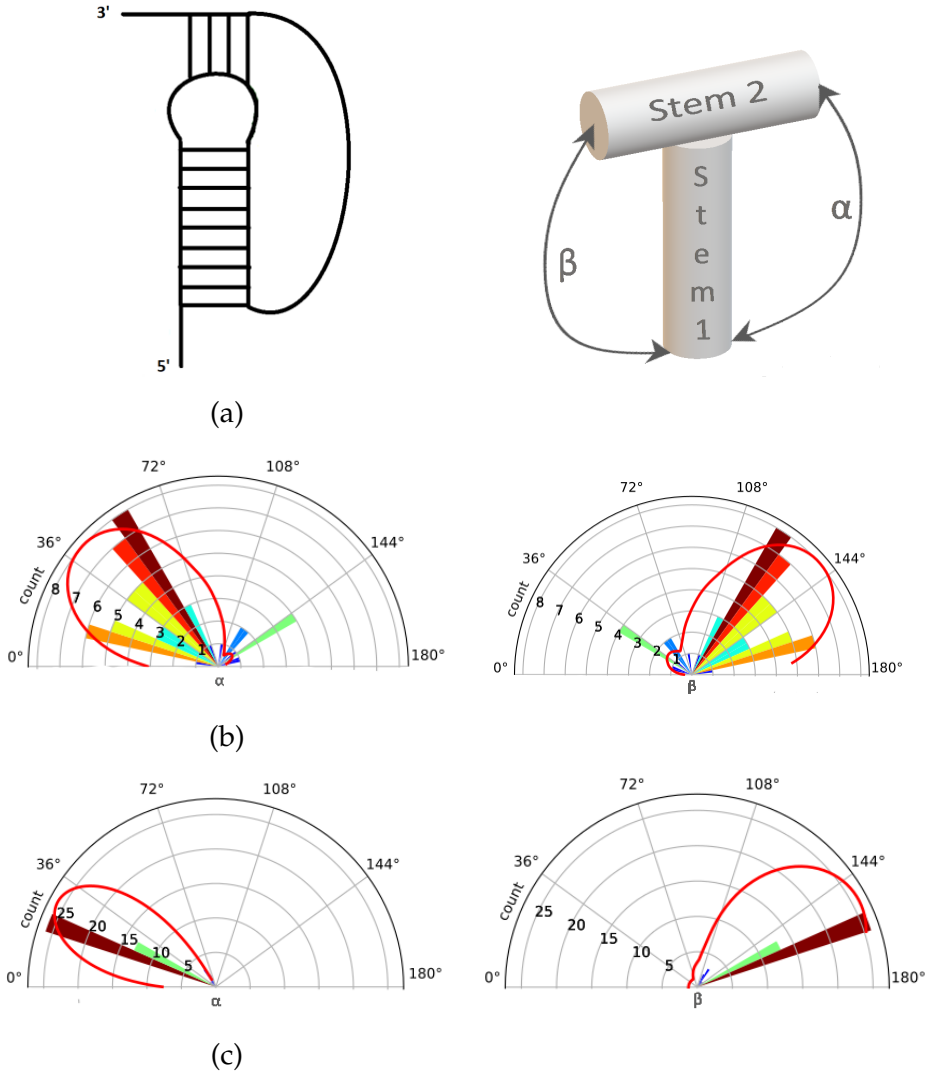


Figure 22: (a) Description which angles between the stems have been measured in a classical H-type pseudoknot and (b) the distribution of the angles (0° to 180°) between its two stems evaluated for genus 1 and (c) by genus 2

helixes that coaxial stack, are always more flexible than long helices. This is another result that can be influenced by the helix length within a pseudoknot and is important to take into account by building a sampling library.

In contrast to the H-types there is only an insufficient number of detected genus 2 kissing hairpins to enable a representative overview about the angle distribution within this structure. Thus, only genus 1 kissing hairpins and their angle distribution within stem pairs were analysed. As by the H-types only kissing hairpins without any structural elements were taken into account for the analyses. Figure 23 (a) shows where the angles were measured. In brief, the angles are measured as a kind of circle around the structure, where the angles α and β has to sum up to 180° as well as the angles γ and δ. All angles of a kissing hairpin sum up to 360°. As well as in all other histograms an angle of 0° or 180° describes coaxial stacking between two stems.

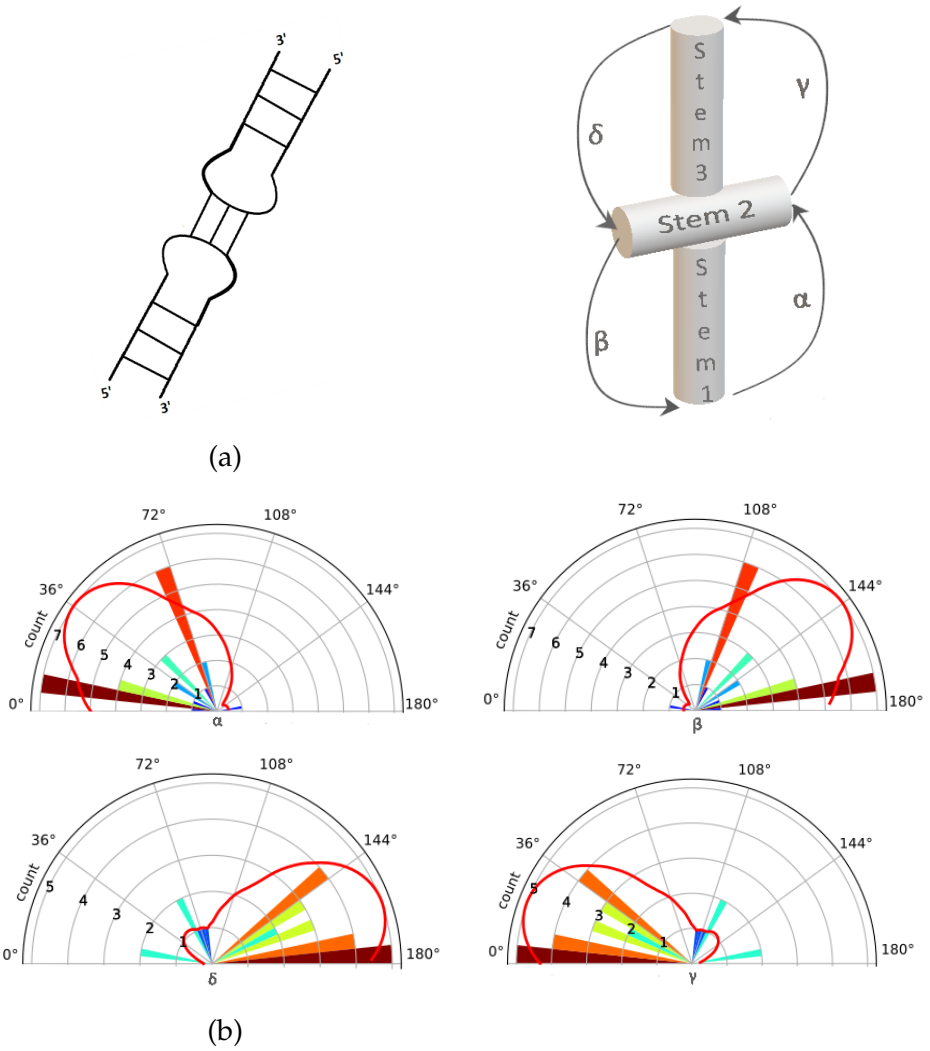
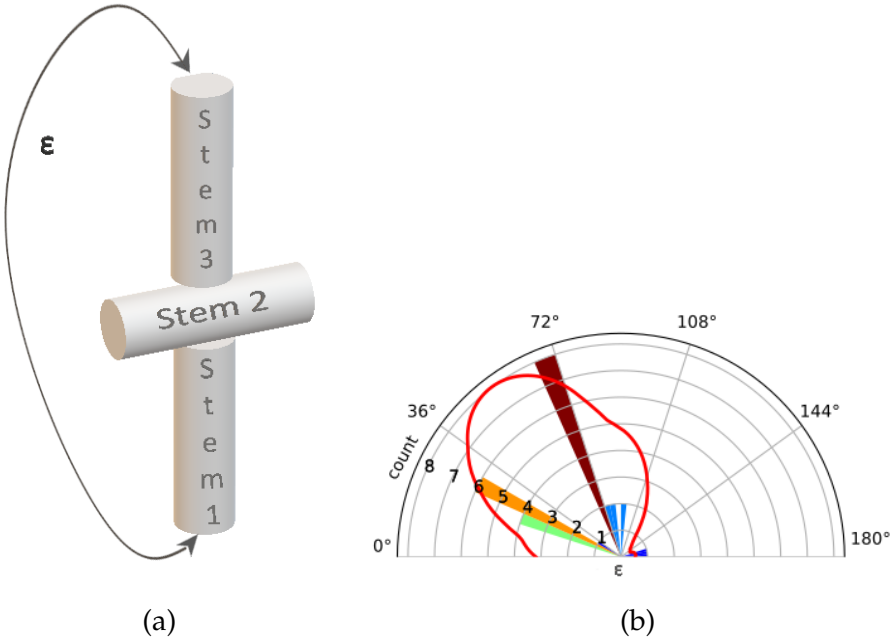


Figure 23: (a) Description of the angles between the stems measured in a traditional genus 1 kissing hairpin and (b) the distribution of the angles (0° to 180°) between its three stems. The angles are measured as a kind of circle once around the structure

Figure 23 (b) shows the distribution of the angles between the certain stems.
To sum up there are more coaxial stacking between stem two and stem three (described via the angles γ and δ) as between stem one and stem two (described via the angles α and γ). This fact is shown by the accumulation at around 70° in the α histogram and the accumulation at around 110° in the β histogram. A look at the structure of a kissing hairpin shows that it is apparent that not all three helices can perfectly coaxial stack with each other. An additional kink within a helix can also affect the coaxial stacking.

The analyses of a very special angle type allows an extended insight into the tertiary structure of a kissing hairpin. In this case ϵ stands for the deflection between stem 3 and 1, figure 24. There are mainly two possibilities, the main one is a deflection around 70 °, and the second, more acute one is around 30° .



(a) (b)

Figure 24: A traditional genus 1 kissing hairpin and the deflection of the structure measured between (a) stem 1 and 3 (b) from 0° to 180°

This angles can exemplify the comportment of a kissing hairpin in the three dimensional space.

Considering the different angle distribution between the two classes of pseudoknots, it can be concluded that it makes sense to treat them separately. As a result, the sample library should not only differentiate between H-types and kissing hairpins. There should be also a differentiation between the angles within a certain class of pseudoknots.

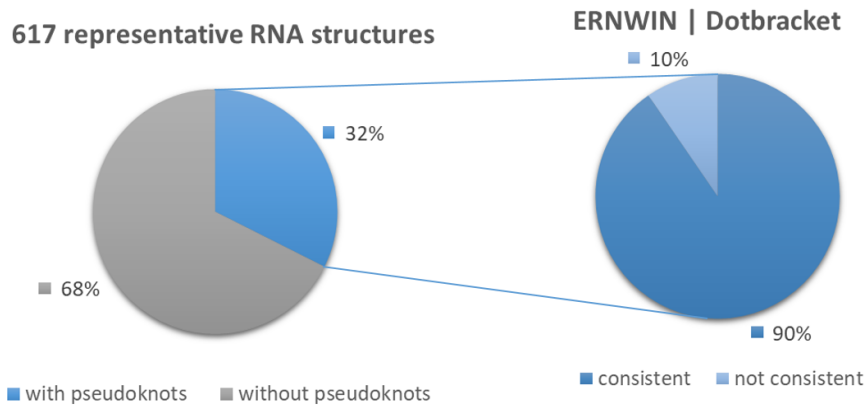


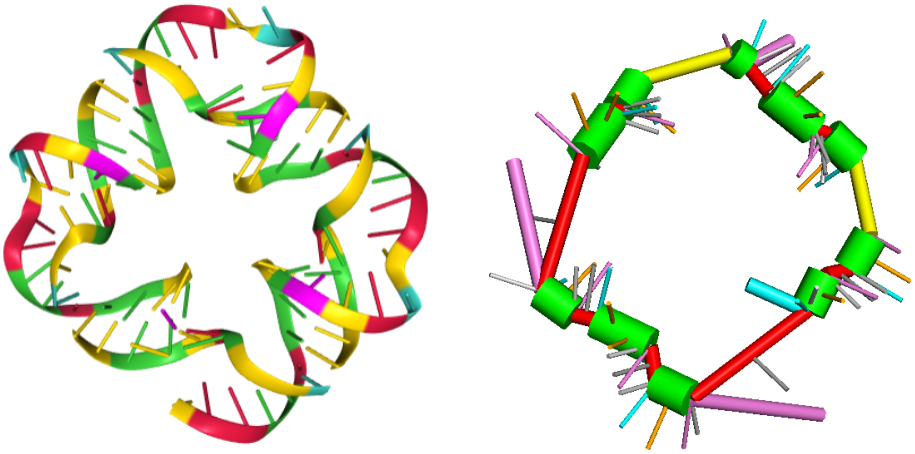
Figure 25: Consistent of the pseudoknot detection between Ernwin and the dotbracket structure

4.3 COMPARISON OF THE DOTBRACKET AND THE ERNWIN IDENTIFICATION OF PSEUDOKNOTS

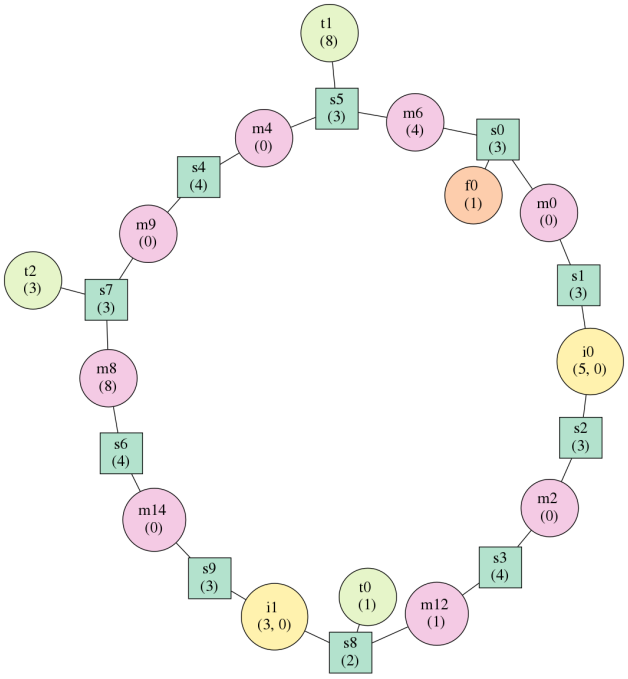
From 617 filtered structures, one third contains pseudoknots. If we compare all predictions of both methods - the prediction via the dotbracket structure as well as ERNWIN's heuristic, a slight difference between both is noticeable. Figure 25 compares the sum of pseudoknots per structure between these two methods. In only 10% of the cases there is an inequality. It is impressive, as the calculated sum also includes all genus 2 pseudoknots. It is mentionable that in most cases ERNWIN counts one to three pseudoknots more than in the dotbracket structure. Most of these structures contain several types of pseudoknots including very knotted "other" structures. There is the increased risk that ERNWIN disassembles one "other" pseudoknot into several ones. From the remaining ones, where ERNWIN predicts more than existing ones, one structure contains one of the rare L-types and the others only one genus 1 type each. This can be based on several or huge insertions within the pseudoknots which impede ERNWIN's prediction. There is only a single case were the pseudoknot prediction is far away from the dotbracket prediction. ERNWIN predicts seven pseudoknots although there is actually only one knotted "other" pseudoknot, illustrated in figure 26. In sum there are just three structure predictions where ERNWIN detects less pseudoknots than requested. All of them contain a mix of genus 1 H-types, genus 2 H-types and/or "other" ones. To summarize, most of the inaccurate counts occur within the range of genus 2 predictions.

Like many other published 3D prediction tools also the published version ERNWIN [35] only works with pseudoknot free structures. Since then a crude method were implemented to handle pseudoknots (described in section 3.2.3.1). Although the number of predicted pseudoknots per structure in ERNWIN does not differ much from the identification via the dotbracket structure, the implemented identification system has to be enhanced. For a prediction ERNWIN uses a given secondary structure. This structure should be used to get the position

of a pseudoknot instead of an imprecise identification via the implemented connection type. The usage of the secondary structure allows a precise detection of the whole pseudoknotted structure as well as the class of a pseudoknot and information about its composition. This information allows a more accurate prediction.
Additionally, combined with a sample library that includes the angle information of e.g. H-types and kissing a certain class of pseudoknots can be predict more accurate.



(a) (b)



(c)

(d) ......((((((.....(((&[[[[]]]]]]&[[[[[[[[........&))).....]]]&((((........(((&]]]]]))...&]]]].[[...[[[&))))]]]]].
(e) fsssssiiiiiisssssssssssssssssssssttttttttsssmmmmsssssssmmmmmmmmssssssssssttttsssssmssiiissssssssssssst
(f) 000011100000222333322211144445551111111100066665556666888888887774444777222333328811199966669998800

Figure 26: The following illustration shows one of ERNWIN's unconsistent counts - PDB structure 3P59 [19]. 3P59 is a crystal structure of an RNA nanosquare.
(a) The illustration from NGL Viewer [63, 64] exemplify the structure composition, where each residue is separated by a different color.
(b) shows the common ERNWIN structure prediction illustration (section 3.2.3) , whereas (c) is the graph like representation (section 3.2.3) that gave a better overview about the connections between the stems in knotted structures. In the graph like representation the fiveprime is defined with "fo" and the threeprime with "to". Stems have the signature "s", the interior loop has "i", the multiloop has "m" and a hairpin is marked with "h" each followed by a consecutive number starting with "0". The numbers indicate the nucleotides that are present in each element.
(d) Dotbracket structure
(e) and (f) coarse-grained-elements (described in section 3.2.3) and their corresponding numbers

5

SUMMARY AND CONCLUSION

Predicting a pseudoknot is still a nearly unsolved issue due to the lack of information as well as the computationally complexity. There exist several types of pseudoknots in nature, which could be classified into classes and genera. Besides the identification and the classification, differences of the angle distribution between the stems of different classes of pseudoknots give a detailed look concerning the composition of pseudoknots. The objective was the identification and the classification of pseudoknots combined with `ERNWIN`. Therefore the dotbracket structure translated from the tertiary structure supplied from a non-redundant list of PDB files was analysed.

The pseudoknots were separated regarding their boundaries according to its dotbracket structure; all additional internal structures within a pseudoknot were excluded to receive only the cleaned pseudoknot structure and assigning it to the different classes of pseudoknots.

The obtained dataset gives an overview about the frequency of pseudoknots within already isolated RNA structures. In general, pseudoknots are present in about one third of the isolated structures even in well-known, regulatory components in the cell. Within the world of pseudoknots, H-types, independent of the genera, are the most represented class of pseudoknots, followed by the genera of kissing hairpins and other types of pseudoknots. The shadow L, mentioned by Reidys [62], were only detected on rare occasions and shadow M in no case.

Surprisingly, there is a high number of highly complex pseudoknots. Some of them seem to be specific for some rRNA types.
Further analyses regarding the helix lengths within pseudoknots show that many of them vanish the longer the required minimum helix length. A short helix indicates for an internal interaction and a longer helix for a long-range interaction.
One of the restrictions of our tested dataset was a minimum stem length of two basepairs. Regarding the composition of pseudoknots previous test cases show that with a restriction of only one basepair per stem RNA structures contain much more stems in general. Such a mild restriction is not expedient, as supporting the stability of the whole structure is an adverse condition for only one basepair. Hence a restriction to a minimum of two basepairs per stem is reasonable and state of the art for related predictions, e.g. [41]. Each intensification of the restriction reduces the number of stems per structure and subsequently the number of detected pseudoknots. Nevertheless, fact that many pseudoknots contain long stems illustrates that pseudoknots can form strong bonds and interact a lot, if necessary. Additionally, a

sampling library that facilitates the prediction has to be well-founded. Therefore, at this point a certain threshold regarding the helix length is essential. This threshold ensures that only pseudoknots that are based on long-range interaction are included into this library and are permissible for the prediction.

Another essential point was to find a way to show, that it makes sense to handle pseudoknots different than normal multiloops, and to differentiate between each individual class of pseudoknots. The analysis of the angles of a pseudoknot and a comparison with the angles within a multiloop shows a difference regarding the distribution. The assumption that the structural elements within a pseudoknot have a defined configuration and differ between the variable types of pseudoknots is also corroborated by Laing and Schlick [42] relating to four-way junctions or Lescoute and Weshouf [41] concerning three-way junctions. It has to be pointed out that different classes of pseudoknots (H-type and kissing hairpins) offer a different set of angles to each other. In conclusion, this shows that it make sense to treat H-type different than kissing hairpins.
The analysis shows that for building a sample library it is possible to assign genus 2 H-type angles to the genus 1 H-types as they are both within a similar range within the distribution. Such combination of datasets enlarges the library and helps to build a well-founded library.

Considering that there are differences between the angle distributions of H-types pseudoknots and kissing hairpins structures, it can be concluded that a more detailed and class separated dataset of angles can support the RNA 3D coarse-grained structure prediction. Furthermore, the detailed demonstration of the angles measured between the single elements let emphasis that a dataset separated by the angle can facilitate an accurate prediction. This encourages the assumption that it make sense to handle pseudoknots differently. And if you take one step further, to deal differently with each class of pseudoknots.

Finally, it has to be mentioned that although the RNA 3D prediction tool `ERNWIN` is not capable to separate genus 1 from genus 2 pseudoknots the count of the number of pseudoknots inside of a structure is very accurate. Nevertheless the dotbracked based identification of pseudoknots is more accurate and offers additional and essential information for the prediction The identification of pseudoknots based on the dotbracket structure combined with the future possibility to assign this knowledge to the `ERNWIN`'s representative elements will result in a better overall picture of predicted structure.

BIBLIOGRAPHY

[1] Kornelia Aigner, Fabian Dreßen, and Gerhard Steger. "Methods for Predicting RNA Secondary Structure." In: *RNA 3D Structure Analysis and Prediction*. Springer-Verlag Berlin Heidelberg, 2012. Chap. Methods for Predicting RNA Secondary Structure, pp. 19–41. ISBN: 978-3-642-25739-1. DOI: 10.1007/978-3-642-25740-7_3. URL: http://link.springer.com/10.1007/978-3-642-25740-7{_}3.

[2] Tanvir Alam, Mahmut Uludag, Magbubah Essack, Adil Salhi, Haitham Ashoor, John B. Hanks, Craig Kapfer, Katsuhiko Mineta, Takashi Gojobori, and Vladimir B. Bajic. "FARNA: Knowledgebase of inferred functions of non-coding RNA transcripts." In: *Nucleic Acids Research* 45.5 (2017), pp. 2838–2848. ISSN: 13624962. DOI: 10.1093/nar/gkw973.

[3] G. Alterovitz, R.M. Benson, and M. Ramoni. *Automation in Proteomics and Genomics: An Engineering Case-Based Approach*. John Wiley & Sons, 2009, p. 340. ISBN: 9780470741177. URL: https://books.google.at/books?id=0EYHLzTsEtwC.

[4] Stefan Arenz, Lars V. Bock, Michael Graf, C. Axel Innis, Roland Beckmann, Helmut Grubmüller, Andrea C. Vaiana, and Daniel N. Wilson. "A combined cryo-EM and molecular dynamics approach reveals the mechanism of ErmBL-mediated translation arrest." In: *Nature Communications* 7 (2016), p. 12026. ISSN: 2041-1723. DOI: 10.1038/ncomms12026. URL: http://www.nature.com/doifinder/10.1038/ncomms12026.

[5] Handan Arkin and Wolfhard Janke. "Gyration tensor based analysis of the shapes of polymer chains in an attractive spherical cage." In: *Journal of Chemical Physics* 138.5 (2013). ISSN: 00219606. DOI: 10.1063/1.4788616. arXiv: NIHMS150003.

[6] BGSU RNA group. *All Representative Set Releases*. URL: http://rna.bgsu.edu/rna3dhub/nrlist (visited on 12/15/2017).

[7] Xiao-chen Bai, Greg McMullan, and Sjors H.W Scheres. "How cryo-EM is revolutionizing structural biology." In: *Trends in Biochemical Sciences* 40.1 (2015), pp. 49–57. ISSN: 09680004. DOI: 10.1016/j.tibs.2014.10.005. URL: http://dx.doi.org/10.1016/j.tibs.2014.10.005http://linkinghub.elsevier.com/retrieve/pii/S096800041400187X.

[8] Jeremy M. Berg, John L. Tymoczko, and Lubert Stryer. *Stryer Biochemie*. 7. Edition. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013. ISBN: 978-3-8274-2988-9. DOI: 10.1007/978-3-8274-2989-6. arXiv: arXiv:1011.1669v3. URL: http://link.springer.com/10.1007/978-3-8274-2989-6.

[9] Yunqiang Bian, Jian Zhang, Jun Wang, Jihua Wang, and Wei Wang. "Free energy landscape and multiple folding pathways of an H-Type RNA pseudoknot." In: *PLoS ONE* 10.6 (2015), pp. 1–16. ISSN: 19326203. DOI: 10.1371/journal.pone.0129089.

[10] Michal J. Boniecki, Grzegorz Lach, Wayne K. Dawson, Konrad Tomala, Pawel Lukasz, Tomasz Soltysinski, Kristian M. Rother, and Janusz M. Bujnicki. "SimRNA: A coarse-grained method for RNA folding simulations and 3D structure prediction." In: *Nucleic Acids Research* 44.7 (2015). ISSN: 13624962. DOI: 10.1093/nar/gkv1479.

[11] Tim Hunt Bruce Alberts, John Wilson. *Molecular biology of the cell*. Ed. by Marjorie Anderson and Sherry Granum. 5th ed. New York: Graland Science, 2008, p. 1725. ISBN: 9780815341055.

[12] Song Cao and Shi-Jie Chen. "Statistical Mechanical Modeling of RNA Folding: From Free Energy Landscape to Tertiary Structural Prediction." In: *RNA Structure and Folding : Biophysical Techniques and Prediction Methods*. Ed. by De Gruyter. Berlin, 2012. Chap. 15, pp. 185–212. ISBN: 9783110284591. DOI: 10.1007/978-3-642-25740-7_10. URL: http://search.ebscohost.com/login.aspx?direct=true{\&}db=e000xat{\&}AN=661695{\&}site=ehost-livehttp://link.springer.com/10.1007/978-3-642-25740-7{_}10.

[13] Song Cao and Shi-Jie Chen. "Statistical Mechanical Modeling of RNA Folding: From Free Energy Landscape to Tertiary Structural Prediction." In: *RNA Structure and Folding : Biophysical Techniques and Prediction Methods*. De Gruyter, 2012. Chap. 14, pp. 185–212. ISBN: 9783110284591. DOI: 10.1007/978-3-642-25740-7_10. URL: http://search.ebscohost.com/login.aspx?direct=true{\&}db=e000xat{\&}AN=661695{\&}site=ehost-livehttp://link.springer.com/10.1007/978-3-642-25740-7{_}10.

[14] Jamie H. Cate, Anne R. Gooding, Elaine Podell, Kaihong Zhou, Barbara L. Golden, Craig E. Kundrot, Thomas R. Cech, and Jennifer A. Doudna. *Crystal structure of a group I ribozyme domain: Principles of RNA packing*. 1996. DOI: 10.1126/science.273.5282.1678. URL: http://www.rcsb.org/structure/1GID (visited on 03/29/2018).

[15] Yifan Cheng. "Single-particle Cryo-EM at crystallographic resolution." In: *Cell* 161.3 (2015), pp. 450–457. ISSN: 10974172. DOI: 10.1016/j.cell.2015.03.049. arXiv: 15334406. URL: http://dx.doi.org/10.1016/j.cell.2015.03.049.

[16] Jesse C. Cochrane, Sarah V. Lipchock, and Scott A. Strobel. "Structural Investigation of the GlmS Ribozyme Bound to Its Catalytic Cofactor." In: *Chemistry & Biology* 14.1 (2007), pp. 97–105. ISSN: 10745521. DOI: 10.1016/j.chembiol.2006.12.005. URL: http://linkinghub.elsevier.com/retrieve/pii/S1074552106004601.

[17] Wayne K. Dawson, Maciej Maciejczyk, Elzbieta J. Jankowska, and Janusz M. Bujnicki. "Coarse-grained modeling of RNA 3D structure." In: *Methods* 103 (2016), pp. 138–156. ISSN: 10462023. DOI: 10.1016/j.ymeth.2016.04.026. URL: http://linkinghub.elsevier.com/retrieve/pii/S1046202316301050.

[18] Petra Dersch, Bernhard Eikmanns, Marcella Eikmanns, Reinhard Fischer, and Dieter Jahn. *Mirkobiologie*. 2008. URL: https://www.thieme.de/de/biologie/systematik-phylogenie-13900.htm (visited on 06/01/2017).

[19] S. M. Dibrov, J. McLean, J. Parsons, and T. Hermann. "Self-assembling RNA square." In: *Proceedings of the National Academy of Sciences* 108.16 (2011), pp. 6405–6408. ISSN: 0027-8424. DOI: 10.1073/pnas.1017999108. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1017999108.

[20] Sebastian M. Fica, Chris Oubridge, Wojciech P. Galej, Max E. Wilkinson, Xiao-Chen Bai, Andrew J. Newman, and Kiyoshi Nagai. "Structure of a spliceosome remodelled for exon ligation." In: *Nature* 542.7641 (2017), pp. 377–380. ISSN: 0028-0836. DOI: 10.1038/nature21078. URL: http://www.nature.com/articles/nature21078.

[21] Niels Fischer et al. "The pathway to GTPase activation of elongation factor SelB on the ribosome." In: *Nature* 540.7631 (2016), pp. 80–85. ISSN: 0028-0836. DOI: 10.1038/nature20560. URL: http://www.nature.com/articles/nature20560.

[22] Ph.D. François Major. *MC-Tools*. URL: http://major.iric.ca/MajorLabEn/Home.html (visited on 03/18/2018).

[23] Wojciech P. Galej, Max E. Wilkinson, Sebastian M. Fica, Chris Oubridge, Andrew J. Newman, and Kiyoshi Nagai. "Cryo-EM structure of the spliceosome immediately after branching." In: *Nature* 537.7619 (2016), pp. 197–201. ISSN: 0028-0836. DOI: 10.1038/nature19316. URL: http://www.nature.com/articles/nature19316.

[24] Patrick Gendron, Sébastien Lemieux, and François Major. "Quantitative analysis of nucleic acid three-dimensional structures." In: *Journal of Molecular Biology* 308.5 (2001), pp. 919–936. ISSN: 00222836. DOI: 10.1006/jmbi.2001.4626. URL: http://linkinghub.elsevier.com/retrieve/pii/S0022283601946261.

[25] Peng Gong, Matthew G. Kortus, Jay C. Nix, Ralph E. Davis, and Olve B. Peersen. "Structures of Coxsackievirus, Rhinovirus, and Poliovirus Polymerase Elongation Complexes Solved by Engineering RNA Mediated Crystal Contacts." In: *PLoS ONE* 8.5 (2013). Ed. by Luis Menéndez-Arias, e60272. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0060272. URL: http://dx.plos.org/10.1371/journal.pone.0060272.

[26] Andreas R. Gruber, Ronny Lorenz, Stephan H. Bernhart, Richard Neuböck, and Ivo L. Hofacker. "The Vienna RNA websuite." In: *Nucleic acids research* 36.Web Server issue (2008), pp. 70–74. ISSN: 13624962. DOI: 10.1093/nar/gkn188.

[27] J. Gu and P.E. Bourne. *Structural Bioinformatics*. Methods of biochemical analysis. Wiley, 2009. ISBN: 9780470181058. URL: https://books.google.at/books?id=4H_ai7ivRIcC.

[28] Alexander P Gultyaev, F. H D Van Batenburg, and C. W A Pleij. "An approximation of loop free energy values of RNA H-pseudoknots." In: *RNA* 5.5 (1999), pp. 609–617. ISSN: 13558382. DOI: 10.1017/S135583829998189X. URL: https://ww.cambridge.org/core/product/D3983E32A7A9A037BC11DC84EE25FD33http://www.ncbi.nlm.nih.gov/pubmed/10334330http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC1369787.

[29] Robert M. Hanson and Xiang Jun Lu. "DSSR-enhanced visualization of nucleic acid structures in Jmol." In: *Nucleic Acids Research* 45.W1 (2017), W528–W533. ISSN: 13624962. DOI: 10.1093/nar/gkx365.

[30] Ivo L. Hofacker. "Vienna RNA secondary structure server." In: *Nucleic Acids Research* 31.13 (2003), pp. 3429–3431. ISSN: 1362-4962. DOI: 10.1093/nar/gkg599. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkg599.

[31] Ivo L Hofacker and Peter F Stadler. "RNA Secondary Structures." In: *Bioinformatics-From Genomes to Therapies*. Weinheim, Germany: Wiley-VCH Verlag GmbH, 2007, pp. 439–489. ISBN: 9783527312788. DOI: 10.1002/9783527619368.ch14. URL: http://doi.wiley.com/10.1002/9783527619368.ch14.

[32] Fenix W.D. Huang, Markus E. Nebel, and Christian M. Reidys. "Generation of RNA pseudoknot structures with topological genus filtration." In: *Mathematical Biosciences* 245.2 (2013), pp. 216–225. ISSN: 00255564. DOI: 10.1016/j.mbs.2013.07.014. arXiv: arXiv:1304.7397v1. URL: http://dx.doi.org/10.1016/j.mbs.2013.07.014http://linkinghub.elsevier.com/retrieve/pii/S0025556413001788.

[33] M. A. Jonikas, R. J. Radmer, A. Laederach, R. Das, S. Pearlman, D. Herschlag, and R. B. Altman. "Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters." In: *Rna* 15.2 (2009), pp. 189–199. ISSN: 1355-8382. DOI: 10.1261/rna.1270809. URL: http://rnajournal.cshlp.org/cgi/doi/10.1261/rna.1270809.

[34] Ailong Ke, Kaihong Zhou, Fang Ding, Jamie H. D. Cate, and Jennifer A. Doudna. "A conformational switch controls hepatitis delta virus ribozyme catalysis." In: *Nature* 429.6988 (2004), pp. 201–205. ISSN: 0028-0836. DOI: 10.1038/nature02522. URL: http://www.nature.com/doifinder/10.1038/nature02522.

[35] Peter Kerpedjiev, Christian Höner zu Siederdissen, and Ivo L. Hofacker. "Predicting RNA 3D structure using a coarse-grain helix-centered model." In: *RNA* 21.6 (2015), pp. 1110–1121. ISSN: 1355-8382. DOI: 10.1261/rna.047522.114. URL: http://rnajournal.cshlp.org/lookup/doi/10.1261/rna.047522.114.

[36] Namhee Kim, Mai Zahran, and Tamar Schlick. "Chapter Five - Computational Prediction of Riboswitch Tertiary Structures Including Pseudoknots by RAGTOP: A Hierarchical Graph Sampling Approach." In: *Computational Methods for Understanding Riboswitches*. Ed. by Shi-Jie Chen and Donald H. Burke-Aguero. Vol. 553. Methods in Enzymology. Academic Press, 2015, pp. 115–135. DOI: https://doi.org/10.1016/bs.mie.2014.10.054. URL: http://www.sciencedirect.com/science/article/pii/S007668791400055X.

[37] Daniel J. Klein, Sara R. Wilkinson, Michael D. Been, and Adrian R. Ferré-D'Amaré. "Requirement of Helix P2.2 and Nucleotide G1 for Positioning the Cleavage Site and Cofactor of the glmS Ribozyme." In: *Journal of Molecular Biology* 373.1 (2007), pp. 178–189. ISSN: 00222836. DOI: 10.1016/j.jmb.2007.07.062. URL: http://linkinghub.elsevier.com/retrieve/pii/S0022283607010285.

[38] William S. Klug, Michael R. Cummings, Charlotte A. Spencer, and Michael A. Palladino. *Concepts of Genetics*. 10th ed. San Francisco, Calif [u.a]: Pearson Education, 2011, p. 896. ISBN: 9780321754356.

[39] Andrey Krokhotin, Kevin Houlihan, and Nikolay V. Dokholyan. "iFoldRNA v2: Folding RNA with constraints." In: *Bioinformatics* 31.17 (2015), pp. 2891–2893. ISSN: 14602059. DOI: 10.1093/bioinformatics/btv221.

[40] Daniela Kruschel and Bojan Zagrovic. "Conformational averaging in structural biology: issues, challenges and computational solutions." In: *Molecular BioSystems* 5.12 (2009), p. 1606. ISSN: 1742-206X. DOI: 10.1039/b917186j. URL: http://xlink.rsc.org/?DOI=b917186j.

[41] A. LESCOUTE. "Topology of three-way junctions in folded RNAs." In: *Rna* 12.1 (2006), pp. 83–93. ISSN: 1355-8382. DOI: 10.1261/rna.2208106. URL: http://www.rnajournal.org/cgi/doi/10.1261/rna.2208106.

[42] Christian Laing and Tamar Schlick. "Analysis of Four-Way Junctions in RNA Structures." In: *Journal of Molecular Biology* 390.3 (2009), pp. 547–559. ISSN: 00222836. DOI: 10.1016/j.jmb.2009.04.084. URL: http://linkinghub.elsevier.com/retrieve/pii/S0022283609005695.

[43] Christian Laing and Tamar Schlick. "Computational approaches to 3D modeling of RNA." In: *Journal of Physics Condensed Matter* 22.28 (2010). ISSN: 09538984. DOI: 10.1088/0953-8984/22/28/283101.

[44] Christian Laing and Tamar Schlick. "Computational Approaches to RNA Structure Prediction, Analysis and Design." In: *Current opinion in structural biology* 21.3 (2012), pp. 306–318. DOI: 10.1016/j.sbi.2011.03.015.Computational.

[45] Neocles B. Leontis and Eric Westof. *RNA 3D Structure Analysis and Prediction*. Ed. by Neocles Leontis and Eric Westhof. Vol. 27. Nucleic Acids and Molecular Biology. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. ISBN: 978-3-642-25739-1. DOI: 10.1007/978-3-642-25740-7. arXiv: 1103.3032. URL: http://link.springer.com/10.1007/978-3-642-25740-7.

[46] Bo Liang, Song Xue, Rebecca M Terns, Michael P Terns, and Hong Li. "Substrate RNA positioning in the archaeal H/ACA ribonucleoprotein complex." In: *Nature Structural & Molecular Biology* 14.12 (2007), pp. 1189–1195. ISSN: 1545-9993. DOI: 10.1038/nsmb1336. URL: http://www.nature.com/articles/nsmb1336.

[47] Zheng Liu et al. "Structure and assembly model for the Trypanosoma cruzi 60S ribosomal subunit." In: *Proceedings of the National Academy of Sciences* 113.43 (2016), pp. 12174–12179. ISSN: 0027-8424. DOI: 10.1073/pnas.1614594113. URL: http://www.pnas.org/lookup/doi/10.1073/pnas.1614594113.

[48] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. "ViennaRNA Package 2.0." In: *Algorithms for Molecular Biology* 6.1 (2011), p. 26. ISSN: 1748-7188. DOI: 10.1186/1748-7188-6-26. URL: http://almob.biomedcentral.com/articles/10.1186/1748-7188-6-26.

[49] Ronny Lorenz, Michael T. Wolfinger, Andrea Tanzer, and Ivo L. Hofacker. "Predicting RNA secondary structures from sequence and probing data." In: *Methods* 103 (2016), pp. 86–98. ISSN: 10959130. DOI: 10.1016/j.ymeth.2016.04.004. URL: http://dx.doi.org/10.1016/j.ymeth.2016.04.004.

[50] Xiang-Jun Lu, Harmen J. Bussemaker, and Wilma K. Olson. "DSSR: an integrated software tool for dissecting the spatial structure of RNA." In: *Nucleic Acids Research* 43.21 (2015), gkv716. ISSN: 0305-1048. DOI: 10.1093/nar/gkv716. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv716.

[51] Marcin Magnus, Michał J. Boniecki, Wayne Dawson, and Janusz M. Bujnicki. "SimRNAweb: a web server for RNA 3D structure modeling with optional restraints." In: *Nucleic acids research* 44.W1 (2016), W315–W319. ISSN: 13624962. DOI: 10.1093/nar/gkw279.

[52] David H. Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H. Turner. "Expanded sequence dependence of thermodynamic parameters improves prediction of {RNA} secondary structure1." In: *Journal of Molecular Biology* 288.5 (1999), pp. 911–940. ISSN: 0022-2836. DOI: 10.1006/jmbi.1999.2700.

[53] J. S. McCaskill. "The equilibrium partition function and base pair binding probabilities for RNA secondary structure." In: *Biopolymers* 29.6-7 (1990), pp. 1105–1119. ISSN: 0006-3525. DOI: 10.1002/bip.360290621. URL: http://doi.wiley.com/10.1002/bip.360290621.

[54] Zhichao Miao and Eric Westhof. "RNA Structure: Advances and Assessment of 3D Structure Prediction." In: *Annual Review of Biophysics* 46.1 (2017), pp. 483–503. ISSN: 1936-122X. DOI: 10.1146/annurev-biophys-070816-034125. URL: http://www.annualreviews.org/doi/10.1146/annurev-biophys-070816-034125.

[55] Charles E. Mortimer and Ulrich Müller. *Chemie*. Ed. by Charles E. Mortimer and Ulrich Müller. Stuttgart: Georg Thieme Verlag, 2010, p. 207. ISBN: 9783134843101. DOI: 10.1055/b-002-54081. arXiv: 9809069v1 [arXiv:gr-qc]. URL: http://www.thieme-connect.de/products/ebooks/book/10.1055/b-002-54081.

[56] Paul L Nixon, Anupama Rangan, Y-G Kim, Alexander Rich, David W Hoffman, Mirko Hennig, and David P Giedroc. "Solution structure of a luteoviral P1-P2 frameshifting mRNA pseudoknot." In: *Journal of molecular biology* 322.3 (2002), pp. 621–33. ISSN: 0022-2836. URL: http://www.ncbi.nlm.nih.gov/pubmed/12225754.

[57] G.A. Papoian. *Coarse-Grained Modeling of Biomolecules*. Series in Computational Biophysics. CRC Press, 2017. ISBN: 9781466576179. URL: https://books.google.at/books?id=wmlQDwAAQBAJ.

[58] Marc Parisien and François Major. "The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data." In: *Nature* 452.7183 (2008), pp. 51–55. ISSN: 14764687. DOI: 10.1038/nature06684.

[59] Mariusz Popenda, Marta Szachniuk, Maciej Antczak, Katarzyna J. Purzycka, Piotr Lukasiak, Natalia Bartol, Jacek Blazewicz, and Ryszard W. Adamiak. "Automated 3D structure composition for large RNAs." In: *Nucleic Acids Research* 40.14 (2012), pp. 1–12. ISSN: 03051048. DOI: 10.1093/nar/gks339.

[60] Ely B Porter, Jacob T Polaski, Makenna M Morck, and Robert T Batey. "Recurrent RNA motifs as scaffolds for genetically encodable small-molecule biosensors." In: *Nature Chemical Biology* 13.3 (2017), pp. 295–301. ISSN: 1552-4450. DOI: 10.1038/nchembio.2278. URL: http://www.nature.com/articles/nchembio.2278.

[61] Protein Data Bank Japan. *PDB bundle*. URL: https://pdbj.org/help/pdb-bundle (visited on 02/27/2018).

[62] Christian M. Reidys, Fenix W.D. Huang, Jørgen E. Andersen, Robert C. Penner, Peter F. Stadler, and Markus E. Nebel. "Topology and prediction of RNA pseudoknots." In: *Bioinformatics* 27.8 (2011), pp. 1076–1085. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr090.

[63] Alexander S. Rose and Peter W. Hildebrand. "NGL Viewer: a web application for molecular visualization." In: *Nucleic Acids Research* 43.W1 (2015), W576–W579. ISSN: 0305-1048. DOI: 10.1093/nar/gkv402. URL: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkv402.

[64] Alexander S. Rose, Anthony R. Bradley, Yana Valasatava, Jose M. Duarte, Andreas Prlić, and Peter W. Rose. "Web-based molecular graphics for large complexes." In: *Proceedings of the 21st International Conference on Web3D Technology - Web3D '16*. New York, New York, USA: ACM Press, 2016, pp. 185–186. ISBN: 9781450344289. DOI: 10.1145/2945292.2945324. URL: http://dl.acm.org/citation.cfm?doid=2945292.2945324.

[65] Schrödinger, LLC. "The PyMOL Molecular Graphics System, Version 1.8." 2015.

[66] Shantanu Sharma, Feng Ding, and Nikolay V. Dokholyan. "IFoldRNA: Three-dimensional RNA structure prediction and folding." In: *Bioinformatics* 24.17 (2008), pp. 1951–1952. ISSN: 13674803. DOI: 10.1093/bioinformatics/btn328.

[67] Cao Song and Chen Shi-Jie. "Statistical Mechanical Modeling of RNA Folding: From Free Energy Landscape to Tertiary Structural Prediction." In: *RNA 3D Structure Analysis and Prediction*. Department of Physics and Department of Biochemistry University of Messouri - Columbia MO 65211 USA, 2012. Chap. 10, pp. 185–212. ISBN: 9783110284591. DOI: 10.1007/978-3-642-25740-7_10. URL: http://link.springer.com/10.1007/978-3-642-25740-7{_}10.

[68] David W Staple and Samuel E Butcher. *Pseudoknots: RNA structures with diverse functions*. 2005. DOI: 10.1371/journal.pbio.0030213.

[69] Bernhard C Thiel, Christoph Flamm, and Ivo L Hofacker. "RNA structure prediction: from 2D to 3D." In: *Emerging Topics in Life Sciences* 1.3 (2017), pp. 275–285. ISSN: 2397-8554. DOI: 10.1042/ETLS20160027. URL: http://www.emergtoplifesci.org/lookup/doi/10.1042/ETLS20160027.

[70] Bernhard Thiel, Peter Kerpedjiev, and Ivo Hofacker. *Ernwin (GitHub)*. URL: https://github.com/pkerpedjiev/ernwin.

[71] Bernhard Thiel, Peter Kerpedjiev, and Ivo Hofacker. *Forgi (GitHub)*. URL: https://github.com/ViennaRNA/forgi.

[72] Rebecca M. Voorhees, Israel S. Fernández, Sjors H.W. Scheres, and Ramanujan S. Hegde. "Structure of the Mammalian Ribosome-Sec61 Complex to 3.4 Å Resolution." In: *Cell* 157.7 (2014), pp. 1632–1643. ISSN: 00928674. DOI: 10.1016/j.cell.2014.05.024. URL: http://linkinghub.elsevier.com/retrieve/pii/S0092867414006679.

[73] Xiaojun Xu, Peinan Zhao, and Shi Jie Chen. "Vfold: A web server for RNA structure and folding thermodynamics prediction." In: *PLoS ONE* 9.9 (2014). ISSN: 19326203. DOI: 10.1371/journal.pone.0107504.

[74] Michael Zuker. "Mfold web server for nucleic acid folding and hybridization prediction." In: *Nucleic Acids Research* 31.13 (2003), pp. 3406–3415. ISSN: 03051048. DOI: 10.1093/nar/gkg595. arXiv: NIHMS150003.

[75] Michael Zuker and Patrick Stiegler. "Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information." In: *Nucleic Acids Research* 9.1 (1981), pp. 133–148. ISSN: 03051048. DOI: 10.1093/nar/9.1.133. arXiv: arXiv:1011.1669v3.

"Ich habe mich bemüht, sämtliche Inhaber der Bildrechte ausfindig zu machen und ihre Zustimmung zur Verwendung der Bilder in dieser Arbeit eingeholt. Sollte dennoch eine Urheberrechtsverletzung bekannt werden, ersuche ich um Meldung bei mir."

"I have endeavored to locate all copyright holders and obtained their consent to use the images in this work. Should nevertheless a copyright infringement become known, I ask for message with me."

ABSTRACT

The present work deals with noncoding RNA and gives an overview about the structure, structure prediction and their limitations. The functionality of noncoding RNA has a high variety from regulatory to functional processes in a cell. Therefore, the prediction of structures in 2D and 3D plays an increasing role to provide information on the function and regulation of noncoding RNA, which is not possible having just the sequence. One tool to predict RNA 3D structures is `ERNWIN`, which uses a coarse-grained sampling method.

A special structural element within the noncoding RNA is the pseudoknot. Pseudoknots help to form stable structures or support reaction pathways. As the prediction of pseudoknot is a computational hard issue, many available tools disallow nested basepairs and handle them as normal multiloops.

The aim of this thesis was to identify pseudoknots in the 3D structure of noncoding RNA, to classify these pseudoknots according to their topological genus, to analyse the frequency of different classes of pseudoknots and to describe their structural features. Therefore, the non-redundant 3D structure dataset for RNA is used as knowledge base. This dataset provides a non-redundant list of PDB files at a selected resolution threshold, using the best structure to represent each equivalence class of structures.

A newly developed tool enables the detection of pseudoknots within a RNA structure. Further on the tool is capable to classify an isolated pseudoknot which enables further analysis of the composition of pseudoknots. Based on the results of this analyses it could be shown that definitely makes sense to handle pseudoknots differently from other multiloop structures. Results concerning structural features of pseudoknots also indicates that it makes sense to assign them into different classes.

In summary, the knowledge about the exact position of a pseudoknot within a structure can be used for the improvement of pseudoknot detections within `ERNWIN`. Furthermore, these results can support the improvement of RNA 3D modelling and the prediction of the whole RNA structures.

ZUSAMMENFASSUNG

Die hier präsentierte Arbeit befasst sich mit der Vorhersage der Struktur von noncoding RNA. Im Zuge dessen wird ein kurzer Überblick über die unterschiedlichen Arten von noncoding RNA präsentiert, bevor auf die Strukturvorhersage mit einen Vergleich der verschiedenen Ansätze und ihren Stärken und Schwächen eingegangen wird. Generell umfasst die Klasse der noncoding RNA eine sehr große Bandbreite von Funktionen, die bei der Regulation von Prozessen, aber auch als Unterstützung in funktionalen Prozessen in einer Zelle zum Tragen kommen. Auch der strukturelle Aufbau von noncoding RNA ist sehr vielfältig.

Hierbei ist entscheidend, dass bei einem wiederkehrenden strukturellen Teilelement innerhalb einer noncoding RNA auf eine ähnliche Funktion geschlossen werden kann. Daher spielt die Vorhersage der Struktur, sowohl als 2D Modell, als auch als 3D Modell, eine immer wichtigere Rolle, um Informationen über ein Biomolekül und seine Funktion und Regulierung zu erlangen. Ein 3D Model bietet einen größeren Informationsgehalt als ein 2D Model und dieses wiederum mehr Information als die reine noncoding RNA Sequenz. Für die Vorhersage von RNA Structuren, sowie deren Analyse stehen eine Reihe von Computerprogrammen zu Verfügung. Eines hiervon ist `ERNWIN`.

Für die Vorhersage muss ein Tool imstande sein mit unterschiedlichen strukturellen Elementen innerhalb einer Struktur umgehen zu können. Eines dieser Elemente ist der Pseudoknoten. Der Pseudoknoten wird aufgrund seiner Komplexität von vielen Programmen mit denselben Parametern wie ein Multiloop berechnet. Das Ziel dieser Arbeit ist es, eine Möglichkeit zu finden Pseudoknoten zu identifizieren und eingefügte Strukturen innerhalb eines Pseudoknoten zu entfernen, um eine Klassifizierung nach Genus 1 und Genus 2 zu ermöglichen.

Es wird ein repräsentatives, nicht redundantes Set von RNA Strukturen als Datenset verwendet. Dieses Datenset besteht aus einer redundanzfreie Liste von PDB Dateien, die einer spezifischen Selektion unterzogen wurden, um von jeder isolierten Struktur einen repräsentativen Datensatz anzuführen zu können.

Das hier entwickelte Computerprogramm ermöglicht die Identifizierung eines Pseudoknoten innerhalb einer Struktur. Des Weiteren kann es die identifizierten Pseudoknoten klassifizieren.
Die Identifizierung und die nachfolgende Klassifizierung ermöglichen eine weitergehende und genauere Analyse des Aufbaues von Pseudoknoten und deren Unterordnungen. Die Analyses dieses Datensets ermöglicht des weiteren die Bestätigung der Hypothese, dass Pseudoknoten sich vom elementaren Aufbau von normalen Multi-

loops unterscheiden und auch die verschieden Klassen von Pseudoknoten Unterschiede aufweisen. Die gewonnen Daten und Erkenntnissen können in weiterer Folge dazu beitragen die Vorhersage von 3D Modellen von RNA Strukturen zu verbessern.

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography *"The Elements of Typographic Style"*. https://bitbucket.org/amiede/classicthesis/