universität wien

DISSERTATION | DOCTORAL THESIS

Titel | Title

Prediction of RNA interaction in 3D

verfasst von | submitted by

Irene Katharina Beckmann-Schneider BSc MSc

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

Doctor of Philosophy (PhD)

Wien | Vienna, 2024

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 794 685 490
Dissertationsgebiet lt. Studienblatt | Field of study as it appears on the student record sheet:	Molekulare Biologie
Betreut von | Supervisor:	Univ.-Prof. Dipl.-Phys. Dr. Ivo Hofacker

Acknowledgements

I would like to express my gratitude to my supervisor, Univ.-Prof. Dipl.-Phys. Dr. Ivo Hofacker, for his support, guidance and mentorship throughout my Ph.D. studies. I appreciate his invaluable constructive feedback that helped shape this research and my academic journey.

I would also like to express my thanks to the members of my thesis committee, namely Univ.-Prof. Dr. Bojan Zagrovic, BA, ao. Univ.-Prof. Dr. Franz-Michael Jantsch, and ao. Univ.-Prof. Mag. Dr. Andrea Leodolter-Barta, as well as the Vienna BioCenter PhD Programme, for their support.

Furthermore, I would like to express my gratitude to the FWF for their financial support, which was essential in enabling me to access the necessary resources for the completion of this thesis.

I'd like to thank my colleagues at the Institute of Theoretical Chemistry, IDA, and my friends for all their encouragement along the way. Their input has made this research experience both worthwhile and enjoyable.

I would also like to extend special thanks to my family for their unwavering love, encouragement, and understanding during the ups and downs of this academic pursuit. Their support has been my pillar of strength and motivation.

Last but not least, I would also like to thank my coffee machine — you are definitely a Monday machine and you always provide me with new warnings and challenges. Which, looking back, reminds me a lot of the writing of this thesis. However, you always provided me with delicious, hot and especially strong coffee, which would have been indispensable during long night shifts.

I would like to thank you all for being part of this PhD journey with me and for your support in reaching this important milestone in my academic career.

Funding

This work was part of the Austrian FWF project ‘Prediction of RNA-RNA interactions´ I2874. This work was furthermore supported by the FWF projects I-4520 ‘Deciphering Complex RNA structure by probing and predictions´ and F-80 ‘RNAdeco´ and generously supported by the Austrian DK program ‘DK RNA Biology´ W-1207 and Vienna BioCenter PhD Programme – Doctoral School at the University of Vienna and the Medical University of Vienna.

Abstract

In the field of ribonucleic acid (RNA) biology, the accurate prediction of RNA structure and its intra- and intermolecular interactions play a pivotal role in understanding the complex network of cellular processes and advancing RNA-based applications in biotechnology and medicine. In many cases, knowledge of the secondary structure is sufficient for identifying interactions and understanding the function of an RNA. Although it is possible to embed secondary structures in three-dimensional space, 2D tools are insufficient to account for the impact of steric hindrance at the tertiary structural level. In particular, the inclusion of conformational essential features, such as pseudoknots or kissing hairpin interactions, can result in predicted structures that are sterically infeasible in 3D. Furthermore, they frequently over-predict interactions as they do not consider the viability of a folding pathway.

The identification of pseudoknots in experimentally determined RNA 3D structures, their classification and analysis were essential tasks in this thesis to learn more about pseudoknots and their specific properties. While 3D modelling of RNA remains a computational challenge, the developed approach to study the folding pathway by combining coarse-grained representations and modelling 3D conformational changes as a series of small steps is an efficient and effective method. It allows to identify whether a pseudoknot or kissing-hairpin interaction proposed in 2D is indeed sterically feasible and kinetically reachable. The 3D approaches developed in this thesis provide a comprehensive insight into the structural features and folding pathways of the different pseudoknot classes, emphasising the importance of 3D structural modelling and the information it provides in addition to 2D RNA predictions. The findings obtained encourage the incorporation of three-dimensional features into 2D tools, thereby markedly enhancing the reliability of predictions.

Zusammenfassung

In der RNA-Biologie spielt die genaue Vorhersage der RNA-Struktur und ihrer intra- und intermolekularen Wechselwirkungen eine zentrale Rolle für das Verständnis des komplexen Netzwerks zellulärer Prozesse und für die Entwicklung RNA-basierter Anwendungen in Biotechnologie und Medizin. In vielen Fällen reicht die Kenntnis der Sekundärstruktur aus, um Wechselwirkungen zu identifizieren und die Funktion einer RNA zu verstehen. Obwohl es möglich ist, Sekundärstrukturen in den dreidimensionalen Raum einzubetten, reichen 2D-Tools nicht aus, um die Effekte sterischer Barrieren auf der Ebene der Tertiärstruktur zu berücksichtigen. Insbesondere die Einbeziehung wichtiger Konformationsmerkmale wie Pseudoknoten oder Kissing-Hairpin-Interaktionen kann zu vorhergesagten Strukturen führen, die sterisch in 3D nicht realisierbar sind. Darüber hinaus werden Interaktionen häufig überbewertet, insbesondere in Bezug auf die Länge einer Interaktion, da die Machbarkeit eines Faltungsweges nicht berücksichtigt wird.

Die Identifizierung von Pseudoknoten in experimentell bestimmten RNA 3D-Strukturen, ihre Klassifizierung und Analyse waren wesentliche Aufgaben in dieser Arbeit, um mehr über Pseudoknoten und ihre spezifischen Eigenschaften zu erfahren. Während die 3D-Modellierung von RNA nach wie vor eine rechenintensive Herausforderung darstellt, ist der entwickelte Ansatz zur Untersuchung des Faltungspfades durch die Kombination von grobkörnigen Darstellungen und der Modellierung von 3D-Konformationsänderungen als Abfolge kleiner Schritte eine effiziente und effektive Methode. Sie ermöglicht es zu erkennen, ob eine in 2D vorgeschlagene Pseudoknoten- oder Kissing-Hairpin-Interaktion tatsächlich sterisch durchführbar und kinetisch erreichbar ist. Die in dieser Arbeit entwickelten 3D-Ansätze geben einen umfassenden Einblick in die strukturellen Eigenschaften und Faltungswege der verschiedenen Pseudoknotenklassen und unterstreichen die Bedeutung der 3D-Strukturmodellierung und der Informationen, die sie zusätzlich zu 2D-Vorhersagen liefert. Die gewonnenen Erkenntnisse ermutigen dazu, dreidimensionale Merkmale in 2D-Tools zu integrieren und dadurch die Zuverlässigkeit der Vorhersagen deutlich zu verbessern.

Contents

List of Tables

List of Figures

I. Introduction

Ribonucleic acids (RNAs) are best known for their role as the blueprint for protein synthesis within the ribosome. However, the majority of RNA molecules present in cells do not code for proteins. A large percentage of transcripts in organisms are non-coding RNA (ncRNA). While ncRNA do not code for proteins, they perform many functions as regulatory elements [51, 167, 195]. They play an essential role in processes such as the regulation of transcription and translation [12, 71, 194] in all kingdoms of life, particularly including bacteria [205] and extending to viruses [171, 192]. These functions of RNAs depend crucially on their structure as well as their interaction with other RNAs. The accurate prediction of RNA structure is therefore essential for our understanding of cellular biology, as well as for RNA-based applications in biotechnology and medicine. For example, the regulatory potential of RNA has led to increasing interest in RNA therapeutics, as exemplified by the novel messenger RNA (mRNA)-based vaccines used to combat the Covid-19 pandemic.

In many cases, knowledge of the secondary (2D) structure is sufficient to understand the function of a ncRNA, as they are largely evolutionarily conserved. Consequently, a number of successful computational tools for predicting 2D structures have been developed, such as those available in the `ViennaRNA Package` [107]. These tools are computationally efficient and reasonably accurate, but they typically ignore pseudoknots (pks) because they break the dynamic programming paradigm of these algorithms. Nevertheless, the 3D structure of an RNA structure largely determines the function of the molecule, and while any valid secondary structure can be embedded in a 3D tertiary structure, the addition of pseudoknots can result in structures that are sterically infeasible or potential folding pathways that are inaccessible due to steric or kinetic hindrance. The same applies to the prediction of RNA-RNA interactions (RRIs), especially when interacting structures such as kissing hairpins are considered.

Pseudoknots account for about 1.4% of RNA base pairs [117], are more frequent in regions critical for functionality [174], and play an important role in facilitating the folding process into three-dimensional shapes. There are several approaches to integrating at least some

tertiary structural elements, such as pseudoknot or a limited group of pseudoknot types, into 2D structure prediction tools. One of the main problems with pseudoknot integration is the lack of data; the lack of experimentally validated structures (ideally from different RNA types and in 3D) and the consequent lack of steric and kinetic parameters make it difficult to establish a well-founded set of rules for pseudoknot prediction. 3D RNA structure prediction tools are an excellent way to represent and explore structural data in a well-defined model. Furthermore, 3D approaches allow a better picture of the RNA structure, including its steric, kinetic and interaction properties.

The aim of this thesis is to present developed 3D-based computational approaches that allow to fill this data gap in order to provide a better insight and thus in the future a well-founded basis for the prediction of RNA structures including different types of pseudoknots. Therefore, three approaches are pursued:
1) The publication in section II.1 presents the development of a tool to identify, classify and analyse pks in experimentally determined 3D structures.
2) For an RNA-RNA interaction (or similarly a pk) to be formed, the resulting structure must be sterically feasible in 3D. Additionally, a kinetically feasible path must exist from the initial contact to the final interaction. 2D tools that ignore steric or kinetic obstacles will often over-predict interactions. To determine whether an RNA interaction is sterically feasible and reasonably achievable, the pipeline (publication in section II.2) developed for this purpose simulates a series of 3D structures along a given 2D folding path.
3) The novel combination of low-cost 2D probing data with computational 3D approaches, which provides detailed insight into structure assembly including pk formation and allows verification of whether the predicted pk in a 2D model can be embedded in 3D, is presented in the publication in section section II.3.

The tools and pipelines developed and published as part of this thesis consider different types of pseudoknots and kissing hairpins RNA-RNA interactions (RRIs) and provide valuable information on steric and kinetic constraints that can be used as parameters in 2D tools to make more accurate predictions, in addition to the plausibility of folding paths and interactions. In addition, the approaches presented allow different hypotheses to be conveniently tested and refined in a computational setup prior to costly laboratory experiments. Overall, the implementation of these approaches and the following results support the research concept that representative data collected from experimental and computational 3D models allows steric, kinetic and thermodynamic data to form a more accurate framework for 2D structure prediction.

The following chapters of the introduction provide a summary of relevant related topics, starting with the definition of an RNA pseudoknot in section I.1, including an overview of experimental extraction of pseudoknot parameters (section I.1.1) and common 2D tools with their rule sets for pseudoknot prediction (section I.1.2). Section I.2 briefly introduces the topic of RRI prediction, including an overview of representative tools, with a particular focus on kissing hairpins. Section I.3 gives an overview of common 3D structure prediction tools, including a section on deep learning approaches (section I.3.2). How predicted 3D structures are refined, analysed and evaluated is described in the section I.3.3. The final introductory section I.4 discusses the availability of RNA 3D data with and without pseudoknots and their databases. This introduction is followed by the results and published papers as part of this thesis (chapter II) and a summary & outlook (chapter III).

1. RNA Pseudoknots

RNA pseudoknots are classified as three-dimensional structures consisting of nucleotides base pairing in an unpaired loop region with a complementary sequence outside the loop, e.g. see figure I.1.

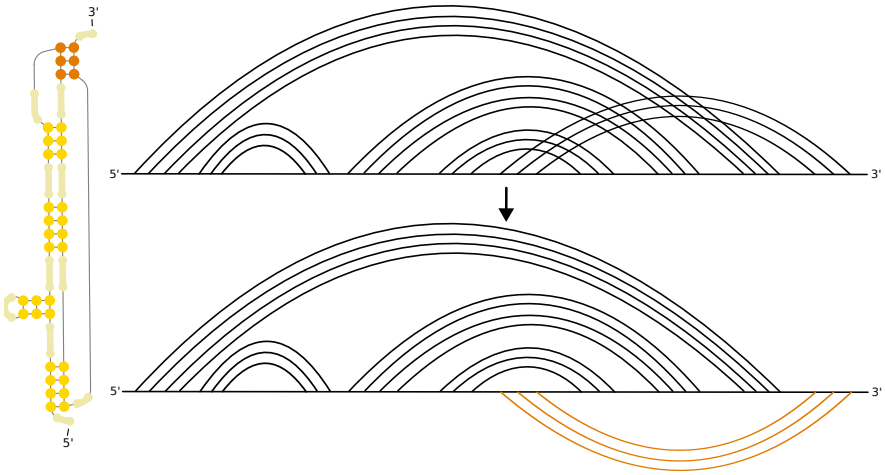


Figure I.1.
Secondary structure representation of the structure on the left, including an H-type pseudoknot (visualised with `pseudoviewer` [82]). On the top right, the linear backbone represents the nucleotide sequence and the semicircles represent the base pairs. The challenge is to reduce the minimum number of base pairs to obtain a structure with non-crossing arcs, the rest representing the pseudoknot base pairs, here the arcs folded downwards in orange (bottom right).

A pseudoknot thus consists of at least 2 helices [139, 149] (see figure I.2), which are assumed to be able to stack coaxially and thus form a quasi-continuous helix, which I can support with current research results, see publication in section II.1. While Pleij [139] speaks of 14 different types of pseudoknot (e.g. referring to the loop type in which the pseudoknot is located), Westhof and Jaeger [209] further differentiated the H-type pseudoknot into 3 types based on which of the three possible loops (L1, L2, L3) is not formed, see figure I.2.

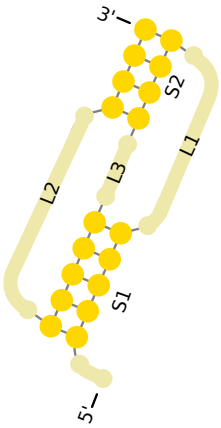


Figure I.2.
Basic pseudoknots, known as a classic (genus 1) H-type pseudoknot, consisting of two helices/stems (S) and two to three loop segments (L).

This is later contrasted by Reidys et al. [149] concept of classification. It is based on a reduced form of dotbracket notation in which each helix (regardless of length) is represented by a single open and closed bracket. Using this approach, he isolates 4 different pseudoknot classes (H/H-type, K/Kissing Hairpin, L, M) that can predominate in an RNA structure, see figure I.3. There are also different 'genera' of each class, which describe the degree of knotting with other pseudoknots. Thus, a single H-type pseudoknot would be in genus 1, and two H-type pseudoknots that are knotted together would each be in genus 2. This classification concept allows a targeted search within existing structures and facilitates the analysis of class-specific parameters that can subsequently be used in prediction models, as shown in the publication in section II.1.

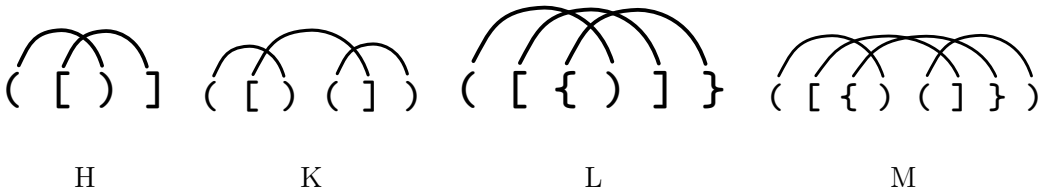


Figure I.3.
Reidys et al. [149] classified 4 main classes of simple, non-nested (genus 1) pseudoknots, derived from their irrudicible components (arcs). All non-crossing elements are decomposed and the crossing elements are merged. This results in the four main pseudoknots: H-type (H), kissing hairpin (K), L-type and M-type.

To calculate the free energy of a secondary structure for structure prediction, it is sufficient to add up the individual components (loops, stems, dangling segments) in a very simplified way. In the case of a pseudoknot these components interact with each other and thus

influence each other, so that a "simple addition" is no longer possible. It is therefore important to obtain parameters such as general structural, thermodynamic, steric or kinetic properties in order to correctly predict pseudoknots.

1.1. Formation and Thermodynamic Characteristics

The difficulty in finding an unified definition is also reflected in efforts to isolate thermodynamic parameters specific to pseudoknots. The most common approach to obtain thermodynamic values for pseudoknots is to study the unfolding pathway of RNA by melting the helices under different conditions (mutations in the sequence, different Mg^{2+} concentrations, etc.). The expected values corresponding to secondary structures based on the nearest neighbour model are determined. The difference reflects non-nearest neighbour and tertiary interactions, which include pseudoknots.

As shown in the case of $tRNA^{fMet}$[1] [48, 60], see figure I.4. In the unfolding experiment (see table I.1 for the respective thermodynomic energies), it was demonstrated that the tertiary structures unfold first at low temperatures, while the secondary structure elements unfold only at higher temperatures. This would have the advantage that one could easily distinguish between tertiary and secondary structure with targeted experiments, since the tertiary interaction must be unfolded before the secondary structure can unfold. In this experiment it has also been shown that the higher the Mg^{2+} concentration, the higher the free energy to form the tertiary structure compared to the secondary structure (see table I.1).

[1] $tRNA^{fMet}$ interacts with the AUG start codon to provide the first amino acid in the form of N-formylmethionine during the initiation of prokaryotic protein synthesis [110]

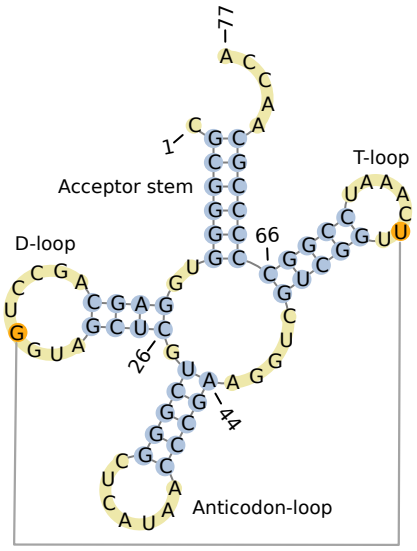


Figure I.4.
$tRNA^{fMet}$ secondary structure [48, 23] showing the typical cloverleaf structure with labelled helices and stems. The tertiary interaction between the D-loop and the T-loop is shown in orange. Visualised using `pseudoviewer`

Unfolding step	ΔG (25°C) kcal/mol	Structure element
1	3 mM Mg^{2+}: 11 no Mg^{2+}: 3.4	Tertiary + D-loop
2	5.8	T-loop
3	7.6	Anticodon-loop
4	10.4	Acceptor stem

Table I.1.
Free energy changes required for the unfolding steps of $tRNA^{fMet}$ shown in figure I.4. It unfolds in 4 steps where Mg^{2+} has a stabilising effect on the tertiary structure. The ΔG values represent the Gibbs free energy change associated with the unfolding steps of RNA tertiary structures. Table modified from Draper [60].

In fact, the tertiary structure can be Mg^{2+} sensitive and result in a stabilising effect during folding, but to a varying degree and is not necessarily required for the formation of tertiary structures [144].

This is demonstrated by unfolding experiments on the *Escherichia coli* alpha operon mRNA fragment (length: 112 nucleotides (nt)) and its equilibrium between the pseudoknot state and the hairpin structure [73, 60]. A high Mg^{2+} concentration or a low pH shifts the equilibrium of its functional structural states, which in turn affects translation. Thus, only the conformation formed at high Mg^{2+} recognises the translational repressor, ribosomal protein *S4*. More importantly, the difference in free energy between these two states is relatively small, making it hard to distinguish clearly between secondary and tertiary structural elements. In fact, secondary structure and tertiary interactions can also be more tightly linked or interdependent, suggesting that certain secondary structures can only form a stable structure with the help of tertiary interactions. An equilibrium between the pseudoknot state and the hairpin structure, and the energy minima close to each other, has been consistently demonstrated in melting experiments and further leads to the conclusion that it is essential to consider the sup-optimal as well as the transitional

structures [143, 213, 14]. As a side effect, potential intermediate structures can also be used for phylogenetic research.

In summary, it can be said that it is not so straight forward to obtain parameters specific for the tertiary structure, and that in addition to the temperature, the ion concentration plays a role in the folding process and significantly influences the final structure. Furthermore, it is important to consider the folding path and its sub-optimums in order to detect functional intermediates and to prevent the formation of infeasible structures. The approach of stepwise RNA folding and identification of RNA sub-optimums has been attempted integrated into some 2D algorithms [2, 78, 14] and in the context of my work I have developed a tool at the 3D level that deals specifically with the stepwise extension of RRIs — in particular kissing hairpin structures and pseudoknot formations, see section II.2 and II.3.

An open question is the differentiation between tertiary interactions and pseudoknots. Most tertiary interactions are pseudoknots, but it is preferable to define them separately as this makes parameterisation easier. The definition of tertiary interactions in RNA is not standardised and often differs from paper to paper.

One definition is that RNA pseudoknots, with their hydrogen bonds connecting loops, are tertiary RNA structure [40]. Another is to define tertiary structure as all types of interactions that must first be unfolded before a secondary structure element can be unfolded [60], including only interactions that serve as an additional linker within the RNA structure and not uncommon hydrogen bonds, which in this definition count as non-canonical structure. This definition considers tertiary interactions such as hydrogen bonding between irregular complementary surfaces as seen in the tertiary structure of transfer RNA (tRNA), monovalent and divalent ions binding to specific sites as seen in ribosomal RNA (rRNA), and pseudoknots as factors leading to greater stability and compactness of RNA. However, this definition of tertiary interaction includes simple Watson-Crick helical base pairing pseudoknots, which do not fulfil the thermodynamic parameters of a tertiary structure.

Hence the approach of Gluick and Draper [73], who define base pairing during pseudoknot formation as part of the secondary structure, since it forms standard helices as known from simple 2D structural elements, as shown in figure I.1. This is particularly true for simple RNA pseudoknots consisting of canonical base pairs. Other important working assumptions around the definition of pseudoknots are to determine whether only Watson-Crick base pairs are considered or the entire range of possible base pairs [184], and at what

helix length it is considered to be an interaction and thus a pseudoknot (usually at least 2-3 base pairs) and at what point it is considered to be only a tertiary contact.

As more and more pseudoknots were discovered in RNA structure, the desire to integrate pseudoknots into the constantly developing structure prediction tools became increasingly important in order to make more accurate predictions. As Pleij [139] acknowledged, the presence of a complementary sequence does not necessarily imply the formation of a pseudoknot within an RNA structure. Given the constraints of stereochemistry and thermodynamics, it is not certain that such a structure can be formed. Conversely, pseudoknots have also been identified in structures that do not appear feasible according to our current steric and thermodynamic knowledge. These observations demonstrate the potential of tertiary structure elements and the current limitations of our understanding regarding their parameterisation, which in turn affect our ability to develop and refine prediction models.

1.2. Prediction Rule Set and associated 2D Tools

In order to establish reference values for pseudoknot integration in RNA 2D structure prediction, restrictions are placed on the pseudoknot properties depending on the parameter set implemented in the tool. These specify whether only certain classes of pks, such as only H-types or only K-types, can be determined, or whether they are preferred because only a parameter set tailored to these classes is available. The implemented rules can also specify the allowed lengths of stems and loops in relation to their position to the major/deep or minor/shallow groove, or the allowed pseudoknot genus. These restrictions make it possible to provide highly specialised tools which, at the same time, only provide valuable structure predictions within certain structural constraints.

In addition to the parameters specific to pseudoknot structures, the approaches presented here employ the conventional energy parameters for the remaining secondary structures, such as the Mathews-Turner parameter [166, 119, 118] for the calculation of the stems and the sequence-dependent values of Walter and Turner [199] for the coaxial stacking.

Groove Dependent Parameter Models

The following parameter sets are based on the relationship between the lengths of the pseudoknot loops and whether they have to cross the minor or the shallow groove, see Figure I.5.

The underlying concept is mainly designed for simple H-type pseudoknots (see Figure I.2), where the pseudoknot construction requires one loop to cross the deep groove and the other to cross the shallow groove, resulting in a different minimum loop length relative to the stem length (a regular RNA-A helix) and thus defining different penalties depending on loop/stem length and "groove position".

Abrahams et al. [2] made one of the first attempts to incorporate energy references for predicting H-type pseudoknots in his genetic algorithm `STAR`. The general algorithm is based on a list of stems that can be added to the structure with a certain probability. This probability is made up of the free energy of the stems and the loops formed by their assembly. For a pseudoknot, based on research by Wyatt et al. [213], if Mg^{2+} is present, a stable pseudoknot can be formed if the first loop has a minimum length of 3 nts and the second loop region has a minimum length of 4 nts. There is no maximum limit. A standard positive energy of 4.2 kcal/mol was set for each intersecting groove for all permitted loops (see table A.AA.1), and the rest were prohibited by an extremely high energy penalty. 4.2 kcal/mol is based on the assumption that the loop energy of short connecting loops is not very different from that of medium hairpin loops and on experimental pseudoknot data. However, due to the lack of data, the range of possible pseudoknots is very limited.

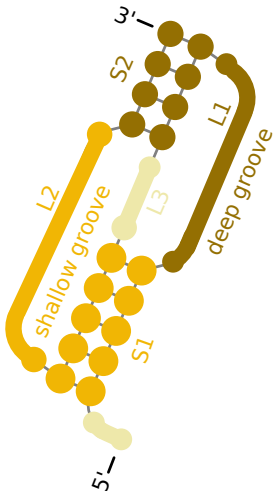


Figure I.5.
Depending on whether the pseudoknot loops cross the deep or shallow groove, different penalties are applied for the respective stem/loop combination (L1/S2 or L2/S1).

This approach was later refined and extended by van Batenburg et al. [14] and Gultyaev et al. [79], see table A.AA.2. It includes more loop/stem/groove combinations, allowing variability in pseudoknot forms, and instead of an uniform penalty it weights different combinations. For example, the loop penalty for a minimal loop size (L1 = 1 nt) and optimal stem length (S2 = 6 or 7 bps), $\Delta G_{37°C}$ is 3.5 kcal/mol. For the same penalty by crossing the shallow groove, you have to choose L2 = 1 nts and S1 = 3 bps. For all loops longer than 4 nts the value is in some cases considerably higher than the 4.2 kcal/mol estimated by Abrahams et al. [2], which is a better distinction with respect to the variable structure of an H-type pseudoknot.

The Cao-Chen parameters [32] are computational estimates based on a virtual bond model

mapped in a diamond lattice, see table A.AA.4 for the used pseudoknot loop entropy. The model also distinguishes between deep and shallow grooves, and the virtual bond model provides almost continuous values up to a loop length of ≤ 12 bps. For larger loops, due to computational cost, a simple formula with weighting factors is used. L3, which is assumed to be a short and flexible element (see figure I.2 for description), is neglected here. The virtual bond model developed by [32] is also the basis of many `Vfold` approaches, see section I.3.

Groove Independent Parameter Models

In these parameter sets, penalties are assigned to the different structural elements associated with the pseudoknot, regardless of type. The simplest concept by Gultyaev [78], only distinguishes between the different loop lengths within a pseudoknot. Based on the loop and stacking values presented in the work of Freier et al. [69], a bifurcation loop destabilisation energy of 0.02 kcal/mol per base was introduced. In his definition, pseudoknot loops have a minimum length of 2 nts and a maximum of only one 'knotting' helix is involved. This approach gives comparatively good predictions for long-range interactions such as pseudoknots in rRNA. Furthermore, the approach of slightly adjusted energy values based on loop size gives better prediction results than the constant value of 4.2 kcal/mol [80].

This is consistent with the observation that the relatively high values used for loop destabilisation energy are not ideal for predicting pseudoknots and their properties. For example, the mRNA H-type pseudoknot in the research of Theimer et al. [184] differs very little from the Turner parameter [190] and is only slightly more stable than the structure without pseudoknot. Too high penalties would significantly affect the structure prediction.

In contrast, the `pknot` approach published by Rivas and Eddy [154] introduces a much larger set of pseudoknot parameters. It distinguishes where a pseudoknot is formed, and counts any bp and nt within the pseudoknot with a penalty, see table A.AA.3. The parameters for the pseudoknots

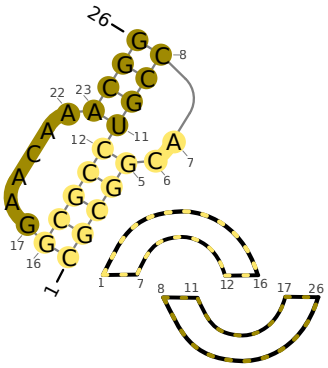


Figure I.6.
`pknot` [154] approach to the use of pseudoknots in implementation. The algorithm splits a pseudoknot into gap matrices, as shown here for a simple H-type pseudoknot represented by two gap matrices.

are based on non-nested thermodynamic parameters [69, 166, 199], which have been adapted in the course of tests on several small RNA structures (with and without pseudoknots). To compute a pseudoknot, the algorithm splits it into so-called 'gap matrices', see figure I.6, which in turn allows a certain variability of pseudoknot predictions, including differentiation into nested pseudoknots, as well as the introduction of regular structural elements within a pseudoknot.

While this dynamic programming method has $O(n^6)$[2] in time and $O(n^4)$ in memory, the pragmatic, updated approach, `pknotsRG` [146, 147] has $O(n^4)$ in time and $O(n^2)$ in memory. However, this improvement also results in a restriction of the allowed pseudoknot conformations. For example, a simple pseudoknot as in Figure I.2 may contain another pseudoknot in L1, L2, L3, but restrictions apply such as; (a) both sides of a pseudoknot stem must have the same length, so e.g. interior loops are not allowed. (b) L3 must be as short as possible, so that no further bp are possible for S1 or S2; if overlap is possible, then (c) S1 and S2 are splited at an arbitrary point. The more restrictive rule set allows a more efficient calculation, but in particular excludes e.g. kissing hairpins or pseudoknots, where a large interaction between L1 and L2 takes place (if necessary, these restrictions can be turned off, increasing the runtime back to $O(n^6)$). Another innovation are the slightly adapted free energy values for the pseudoknots, see table A.AA.3.

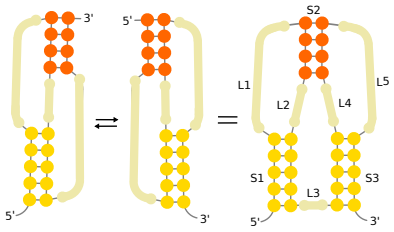
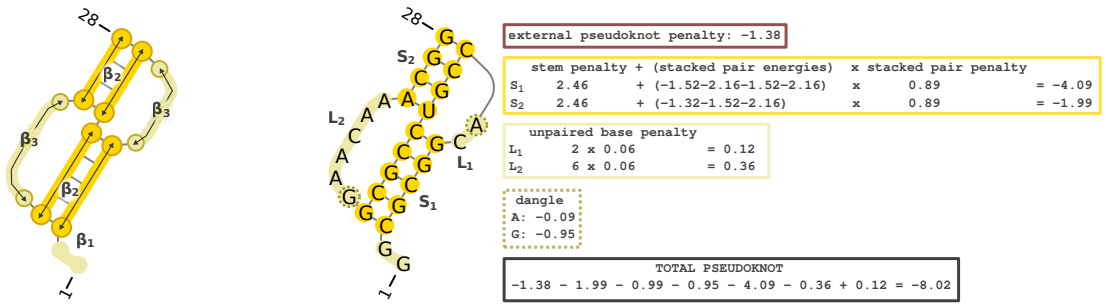


Figure I.7.
`pkiss` [185] provides a strategy for incorporating kissing hairpins into RNA prediction tools: if two H-type pseudoknots overlap perfectly, they are merged into a kissing hairpin.

An extension of the `pknotsRG` approach is provided by `pkiss` [185], which also allows kissing hairpin pseudoknots with $O(n^4)$ in time and $O(n^2)$ in memory. Simplified, if two locally found pseudoknots overlap perfectly, they are superimposed to form a simple kissing hairpin, see figure I.7.

The Dirks-Pierce model [57, 58], implemented in `NUPACK`, is an extension of previous models. The energy model for the pseudoknot is similar to that of Rivas and Eddy [154] in `pknot`, with a penalty for the formation of the pseudoknot, plus a penalty for all bps and all unpaired bases (loops) of the pseudoknot, see figure I.8. The dynamic programming approach requires only $O(n^5)$ in time and $O(n^4)$ in memory.

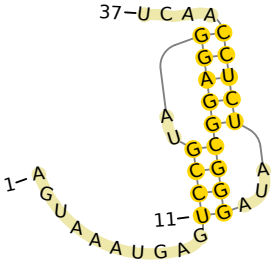
[2] n = sequence length



(a) (b)

Figure I.8.
Simple external pseudoknot with the energy expression based on the Dirks-Pierce model. The model assigns a corresponding penalty to different pseudoknot elements ($\beta_1 = pk - start$, $\beta_2 = pk - stem$, $\beta_3 = pk - loop$) without distinguishing between the deep and shallow groove. See the table A.AA.5 for more details on the penalties. (b) Using this model, the free energy of the pseudoknot was calculated to be 8.02 kcal/mol.

The Dirks-Pierce model allows more flexibility in how a predictable pseudoknot can be constructed, it also distinguishes between external pseudoknots, which roughly corresponds to a genus 1 pseudoknot according to Reidys et al. [149], pseudoknots within a multiloop, and pseudoknots nested with another, such as a genus 2 pseudoknot. However, unlike the Cao-Chen model, which focuses on H-type pseudoknots, it does not include the deep/shallow groove ratio — stem and loop length.



	original	trained
Dirks-Pierce	-5.60	-6.70
Cao-Chen	-11.66	-6.66

Figure I.9.
Depending on the Dirks-Pierce, Cao-Chen or Andronescu trained model, the free energy (kcal/mol) for this simple H-type pseudoknot can vary considerably, see the comparison table below (parameters from Andronescu et al. [7]).

The heuristic approach `FlexStem` by Chen et al. [42] uses the pseudoknot energy model from Dirks and Pierce [57], but also allows overlapping pseudoknots and some of the penalties used are partly different, see table A.AA.5.

Andronescu et al. [7] trained, fitted and tested the accuracy of the parameters of RNA secondary structures with and without pseudoknots for the Dirks-Pierce and Cao-Chen models, see table A.AA.5. A comparison of how different the free energies can be for the

same structure depending on the parameter set is shown in figure I.9

Further Parameter Set Applications

The presented parameter sets for the energy calculation of pseudoknots are used not only in the original approaches, but also in others e.g. the energy parameters of Dirks and Pierce [57] are also used in the `Hfold` approach [85]. `Hfold` is a dynamic programming algorithm that predicts the minimum free energy (MFE) structure based on the hierachical folding hypothesis — first the folding of the structure free of pseudoknots and in a second step the folding of the pseudoknots.

An approach that uses a combination of several approaches is `DotKnot` [172, 173]. The `DotKnot` structure prediction is based on the base pair probability dotplot calculated by `RNAfold` [107], from which the resulting stems are extracted and secondary structure elements are sorted. Based on this secondary structure element dictonary, potential pseudoknots are recursively constructed. The energy parameters of Cao and Chen [32] and Cao and Chen [33] are used for H-type pseudoknots. For the prediction of kissing hairpins the `pkiss` representation is applied, see figure I.7, and for their energy calculation the stacking energy of the 3 stems involved, an initiation penalty of 9 kcal/mol and on all loop lengths except L3 a penalty of 0.5 kcal/mol is applied.

In addition to thermodynamic models, which provide the mathematically optimal MFE structure, several comparative, statistical and heuristic approaches have been developed and tested. Comparative approaches predict a structure based on common base pairings among all sequences in a given data set. One of the requirements for a valid prediction is a sufficiently large dataset of related RNA. Statistical sampling has the advantage that clusters of structures and their respective centroids can be identified based on the set of collected structures and the probability of the resulting base pairs. This provides an overview of sub-optimal structures, but the centroid structure does not necessarily have to be the MFE structure. Heuristic methods such as `STAR`, the Iterated Loop Matching approach (`ILM`) [157, 156], `CyloFold` [24], `HotKnots` [151, 7, 86] or `Flexstem` [42] try to overcome the high computational and time complexity and the structural limitations due to the complex underlying energy models in dynamic programming approaches. However, they cannot guarantee to find the MFE. In `ILM`, stems within a sequence are predicted in a pseudoknot free manner and assigned a score. The scores for each stem are composed of the respective standard pseudoknot free thermodynamic parameters including stacking energies from the `ViennaRNA package` [107]. The final structure

including pseudoknots is assembled by repeated iteration to find additional helices based on the remaining free nucleotides. `Cylofold` also follows the principle of maximising the number of matching helices and consists of an estimated free energy of the assembled stems using the `ViennaRNA package` [107]. It is worth noting that although it does not use pseudoknot specific parameters, it at least tries to check the 3D steric feasibility in the course of building the overall 2D structure. To do this it uses a representative cylinder for each helix and any unpaired nucleotides represent spacers between the cylinders. If a cylinder collapses or the distance is too large, the cylinder is discarded. As described for the other heuristic approaches, `HotKnots` also predicts secondary structures by adding substructures. These substructures can consist of stacked pairs, bulge loops of size 1 nt and inner loops of size 2 nts on each side, and are sorted by their free energy value. The algorithm uses the Turner parameters [166, 119] plus penalties for loops and bulges to prevent substructures from becoming too large, and various pseudoknot-based parameters available for selection: Dirks and Pierce [57] , Cao and Chen [32] and Andronescu et al. [6]. Thresholds for the maximum free energy per substructure, as well as the maximum number of substructures, determine the selection for the initial list available for structure building. This is done in a tree-like fashion, under the condition that a substructure used must not overlap with any bp of a substructure already used for this final structure.

In general, dynamic programming algorithms such as `pknots`, `pknotsRG` or `NUPAK` are computationally expensive and are therefore usually limited in some way: A sequence length restriction, a restriction on the types of pseudoknots, such as only H-types or nested H-types, and in addition the prediction is usually only provided on the simple MFE and only a few low energy suboptimal structures, although we already know that these suboptimal intermediate structures can often lead to an even better structure. Within these approaches, there are different strategies to work around these limitations, such as models like Dirks-Pierce's, which have the advantage of allowing more variability in the shape of the pseudoknot, but neglect the specific parameters related to individual types, such as the H-type pseudoknot, of which we know that a distinction should be made between the loop that crosses the deep or shallow groove, as well as the ratio of loop to stem length. In particular, this ratio helps to avoid predicting inveasible structures, such as pseudoknots with loops that are too short or unrealistically long, and the same applies to the interacting stems. Groove-dependent models, such as the Cao-Chen model, are more concerned with this, but tend to focus only on simple H-type pseudoknots and are not compatible with dynamic programming. The fact that the energy models are limited in several ways, and that the allowed pseudoknot conformations are very tightly

constrained, significantly reduces the usability of the available approaches and the validity of the predictions.

2. RNA-RNA Interactions & Kissing Hairpins

In addition to pseudoknots in the context of RNA interactions, kissing hairpins in RNA-RNA interactions also play an essential role in the functionality of a cell, in particular as regulators of the transcription and translation process, as well as RNA processing. Identifying these and the framework conditions will provide a valuable understanding of this function.

There are several biophysical, biochemical and cellular experimental methods that can be used to study RNA interactions [74, 218]. Low-throughput methods such as electrophoretic mobility shift assay (EMSA), surface plasmon resonance (SPR), Forster resonance energy transfer (FRET), co-sedimentation assays or cellular methods such as RNA hybrid systems are particularly suitable for testing interaction partners previously determined by computational methods at the molecular level or for testing functionality with mutation experiments. One of the first targeted high-throughput methods was crosslinking, ligation, and sequencing of hybrid (CLASH), which has the disadvantage of limited sensitivity and protein-mediated interactions. Like RNA interactome analysis followed by deep sequencing (RIA-seq) and RNA antisense purification (RAP), it works by cross-linking, the difference being that these two methods use single-stranded deoxyribonucleic acid (DNA) or RNA molecules to bind to the target RNA via complementary base pairing. These approaches have the advantage of finding target sites across the entire length of RNA segments, but are limited to their target RNA. This is where transcriptome-wide approaches such as psoralen analysis of RNA interactions and structures (PARIS), sequencing of psoralen crosslinked, ligated, and selected hybrids (SPLASH) and ligation of interacting RNA followed by high-throughput sequencing (LIGR-seq) come in. All three differ in their methodology, their degree of restriction in terms of crosslinking or ligation efficiency. The fact that RRI is a dynamic process and not always specific also influences the results and their interpretation.

The prediction of RRI can be simplified as an extension of regular RNA structure prediction as simple added base pairs. In a more complex representation, the entire structure of RRI is composed of both intramolecular and intermolecular base pairs. Predicting the accessible target site where RRI may occur is also a challenging task. These challenges are often divided into separate tasks and addressed using different tools.

Tools such as `TargetRNA` [188] or `RNApredator` [64] address the issue of finding an accessible target site. The basic idea here is often the biological scenario of finding possible interaction sites for small RNA (sRNA) on the mRNA or microRNA target prediction. For example,

as with TargetRNA, the modified alignment algorithm RNAplex [107, 177, 25] is used and the accessibility profile is precalculated using RNAplfold [22, 20][3]. The search for complementary sequences is even faster if the binding motif is known, which is often the case for sRNA, as these often have multiple targets all served by the same motif.

In general, RRI tools can be categorised as follows [100, 11, 191, 66, 55]:

(I) Alignment like methods are covered by tools such as RIsearch [208] and Guugle [72]. Here only the intermolecular base pairs play a role in the calculation and the final result.

(II) MFE based methods, which in turn can be divided into four categories:

  (i) Approaches that predict the intermolecular structure without considering the intramolecular structure with the intention of achieving prediction speed, such as RNAduplex [107], TargetRNA, DuplexFold [152], RNAhybrid [148].

  (ii) Concatenation-based approaches, such as Pairfold [8], RNAcofold [21], bifold [152]. They consider both intra- and intermolecular base pairs in their prediction and output. They consist of a classical RNA secondary structure prediction of the two input sequences and a prediction of the interaction base pairs.

  (iii) Accessibility based approaches such as Interacting RNAs (IntaRNA) [30, 212, 114] and RNAup [125]. The pairing probabilities for all bases are used, so the stability of the interaction is calculated from the stability of the stacking base pair and the target accessibility probability.

(III) Comparative (homology) approaches, such as RNAaliduplex [107], TargetRNA2 [95] or PETcofold [165], aim to improve predictions by incorporating evolutionarily conserved information. However, even when conserved information is included, prediction results are often variable because the interaction sites are not always directly conserved. The CopraRNA software uses conserved interactions between two RNAs from different species and combines this with the accessibility-based approach of IntaRNA to improve prediction accuracy [131]. The accuracy of comparative methods depends on the type of RNA between which the interaction occurs and how conserved it is [100, 153, 106].

[3] An extension of the RNAplfold algorithm by Waldl et al. [197] with the addition of stochastic sampling of local structures, including the calculation of maximum expected accuracy (MEA) structures and the use of mutual information measures to quantify the sensitivity of individual sequence positions, provides a more robust and efficient framework for local RNA folding predictions.

(IV) Probabilistic approaches, including the tool RNA-RNA interACTion prediction (RactIP) [92]. Using base pairing probabilities and hybridisation probabilities, scores are calculated for intra- and intermolecular base pairs. For the former, the tool CONditional TRAining for RNA Secondary Structure Prediction (CONTRAFOLD) [59], which can be considered an early machine learning tool for secondary structure prediction, is used. For the latter, a modified code of the thermodynamic model RNAduplex is used. Using a threshold cut technique, it predicts the MEA structure using integer programming.

(V) Machine learning approaches such as the miRNA target filtering system MiRTif [220].

Tools can also be divided according to which RRI they specialise in, for example tools such as PLEXY [93] and RNAsnoop [178], are specialised in interactions with small nucleolar RNA (snoRNAs), while others, such as TargetRNA, focus on interactions with bacterial sRNA, as reviewed in Backofen and Hess [11] and Pain et al. [131].

As with RNA structure prediction, which excludes certain structural elements such as pseudoknots, the same is true for RRI. Simple models such as RNAcofold can efficiently predict RRI using standard energy parameters from Mathews et al. [118] and Turner and Mathews [190], but are limited to nested interactions and pseudoknots, as kissing hairpin structures are excluded due to their computational complexity. This also applies to the RNAplex and RNAup tools. IntaRNA allows interactions between single hairpin loops, but not more knotted structures such as kissing hairpin pseudoknots. A limitation of some interaction prediction tools, such as RNAup and intaRNA, is that they only consider a short local interval within the sequence. Nevertheless, they do allow for an extended set of non-nested interactions in their predictions, but often only consider continuous interaction sites with bulges and gaps of a few base pairs at most [66, 100]. Longer interaction sites or more complex interaction structures may be incorrectly predicted, see publication in section II.2.

2.1. Kissing Hairpin Interactions

Kissing-hairpin interactions belong to the basic types of RNA tertiary structural elements, see section I.1, and are a highly functional, widely used structural motif due to their dynamics [29].

An important aspect of kissing hairpin pseudoknots is that the formation of their tertiary structure can influence the (re)formation of the rest of the structure and may even be necessary for local or global structural changes to occur [102, 5, 27, 89], see also

publication in section II.2 and II.3. The kissing hairpin can also represent both an 'end state' and an initial structure leading to a longer kissing hairpin formation up to a full duplex (see publication in section II.2 and [97, 98, 10, 126]). However, in contrast to these kissing hairpin complexes, there are those that are formed without further structural rearrangement at the 2D level [96, 39]. The length of a functional kissing hairpin therefore varies greatly from system to system and can almost exclusively be related to the loop structure itself [96, 145, 39]. In particular, kissing hairpin interactions with short loop segments (∼1 nt) tend to form coaxial stacking interactions between the stems, [65] and publication in section II.1. The observed stacking patterns suggest that these structures can be classified as a classical RNA A-helix form, but studies have shown that they can differ from this classical form [39, 115].

In addition to loop complementarity and conserved sequence motifs that facilitate the formation of the initial interaction, the loop-closing pairs [62], the stem sequence [63, 68, 29] or the type of base pairing [27, 37] also play a role for a stable interaction. However, 3D structural factors such as the orientation of the loops or the internal loop structure [76, 130] are crucial for the stability of a kissing hairpin interaction. Depending on their properties, kissing hairpin complexes also differ in their sterically and kinetically feasible 3D structure and consequently in their feasible 2D structure, see publication in section II.2.

A RRI tool that allows kissing hairpin interactions in its prediction is RNA-RNA interaction problem (`RIP`) [3, 83]. In this approach the RNA interaction complex is calculated by a combination of the total free energy minimisation of the sum of the energies of the individual base pairs [129], the stacking pair energies [119] and the loop energies [119, 226], where for kissing hairpins the `pknot` approach by Rivas and Eddy [154] is used, for a more detailed description see also section I.1.2 and table A.AA.3. Although the `RIP` algorithm allows kissing hairpin interactions, it prohibits all intramolecular and all other types of intermolecular pseudoknot structures for computational reasons. An equivalent approach is also offered by `piRNA` [45], which uses the Dirks and Pierce [57] approach for kissing hairpin interactions. See also section I.1.2 and table A.AA.5. Detailed descriptions of the two algorithms can be found in the respective publications. Unfortunately, both tools are no longer available via the links provided at the time of writing.

In addition to specific entropy parameters for H-type pseudoknots, see section I.1.2, the team around Cao and Chen [35] has also evaluated entropy parameters specifically for variations of kissing hairpin interactions using the `Vfold` approach, see table A.AA.6. However, the parameters are limited to small RRI kissing hairpin complexes and are not

yet implemented in Vfold at the time of writing this dissertation.

Nevertheless, the possibility of a single kissing hairpin interaction state does not necessarily imply the feasibility of the entire pathway from the initial binding site to a potential final interaction structure. This includes considerations of steric and kinetic feasibility, as well as the ability to occur within a biologically relevant time frame.Concurrently, for instance, sRNA possess numerous binding partners, thereby regulating the entirety of pathways [132, 68]. For this to be feasible, the interactions must be stable and flexible enough to allow multi-step structural refolding to further metastable intermediates [159].

Tools such as RRIkinDP [196] in the 2D domain, or the tool RRI-3D, in the 3D domain, see publication in section II.2, are comprehensive approaches to gain a better understanding of the stability and diversity of interaction intermediates, from the starting point to the best possible complete interaction, in detail and on a larger scale.

3. 3D Structure Prediction

Compared to the availability of RNA sequence data, there is comparatively less secondary structure data and many times less RNA tertiary structure data available. However, important for understanding the function of an RNA. Tools for predicting the 3D structure have been developed with which the 3D structure can be determined without time-consuming and costly experiments. The determination of a 3D structure can be based on the sequence alone, on the sequence with a secondary structure as a constraint, or on an initial 3D structure to detect variations. In addition, some approaches allow for additional constraint information of experimental probing data e.g. SHAPE[4] [120] or microarrays for chemical mapping, or the determination of distances between specific nucleotides or atoms. The strength of the constraints based on the provided secondary structure can also be varied by some tools. The use of experimental 2D RNA structural data as constraints in structure prediction tools can lead to a significant improvement in the accuracy of secondary structure prediction tools [61] as well as to more targeted, reliable and faster 3D modelling, see publication in section II.3. A positive effect of structure prediction in 3D, besides the observation of steric or kinetic properties, is that there are fewer structural constraints than in 2D. For example, depending on the approach, possible pseudoknots and kissing hairpin complexes are more easily included in the predicted structure ensembles, regardless of their complexity.

To create a 3D structure ensemble, a generator is used to generate multiple candidate structures — usually implemented in a Monte Carlo (MC) (e.g. YUP [113, 181]) or molecular dynamicis (MD) (e.g. NAST [88]) simulation framework or a variant such as discrete molecular dynamics (DMD) (e.g. iFoldRNA [56, 168, 99]), replica exchange Monte Carlo (REMC) (e.g. SimRNA), replica exchange molecular dynamics (REMD) (e.g. HireRNA [46, 47, 135]), Markov chain Monte Carlo (MCMC) (e.g. Ernwin [94, 187]) simulation, followed by a discriminator a scoring/energy function. The energy functions are largely based on native structures, which in turn should allow the most natural prediction of 3D structures for which no native structure is yet available. Ideally, the combination of generator and discriminator ensures that a local energy minimum is found efficiently. A categorisation of 3D approaches can be based on their scoring/energy function, which include physics-based energy functions (e.g. YUP, NAST, iFoldRNA, SimRNA), knowledge-based statistical potentials (e.g. FARNA/FARFAR, MCFold/MC-Sym, Vfold, RNAComposer) and deep learning-based scoring functions (see section I.3.2). A combination as in NAST,

[4] The selective 2-hydroxyl-acylation-and-primer (SHAPE) experiment allows a high-throughput determination of RNA secondary structures. [120]

which uses RNA-specific knowledge-based potential energy, is also possible. As with 2D prediction tools, the output consists of either the lowest energy structure (e.g. YUP, MCFold/MC-Sym, Ernwin) or low energy clusters and the resulting centroid structure (e.g. NAST, iFoldRNA, SimRNA, FARNA/FARFAR, VFoldLA).

Knowledge-based statistical potential methods can also be called fully automated fragment assembly methods. In general, they predict the 3D structure based on the sequence and secondary structure, and the derived 3D fragments from previously determined structures. This approach, including the fragments used, can also be performed at the atomic level, as in FARNA/FARFAR [53, 54]. Physics-based energy functions that mimic an all-atom force field as in Amber [207, 137] or CHARMM [109, 28] also allow all-atom prediction. As this is computationally expensive, it is often limited to extremely short and simple RNA structures or fragments.

Consequently, 3D approaches can be divided into two categories depending on their resolution: all-atom and coarse-grained approaches, where pseudoatoms/beads are used to represent parts of the structure. Depending on the resolution, the elements that make up an RNA structure are simplified to some extent, such as nucleodites, which are represented by a certain number of beads, or entire structural elements, such as entire helices and loops, as is the case in Ernwin. They also use only a subset of the many physical terms involved in the formation of the structure, such as energy terms for stacking or the various possibilities for base pairing. Nevertheless, they provide detailed information about a structure, such as the position of certain helices or nuclodites. In general, the 'bigger' the coarseness, the larger the structures that can be predicted with comparatively little computational effort. Coarse-grained tools often allow the reconstruction of the resulting 3D structures into all-atom models using in-house tools, and may also include a refinement step. Alternatively, external tools can be used, see section I.3.3.

As a further classification, the available approaches can be divided into those that can only generate a 3D structure de novo based on the RNA sequence alone, such as SimRNA, and those that require a secondary structure in addition to the sequence, such as Ernwin. Tools such as MCFold/MC-Sym or VFold3D can be started from the sequence alone as they contain a 2D structure prediction on which the following 3D structure prediction is based.

3.1. The Tools

This section presents 3D structure prediction tools that have been mentioned in the context of presenting the different approaches in the 3D introduction, that have been mentioned in other chapters, or that have played an important role in the course of this thesis and its related publications.

`FARFAR/FARNA`
: The Fragment Assembly of RNA with Full-Atom Refinement (`FARFAR`) [54] structure modelling algorithm is a two-step process for assembling RNA structures. In the first step, existing RNA crystal structures are used to assemble the target RNA by matching sub-sequences. This process, known as Fragment Assembly of RNA (`FARNA`) [53], uses a MC algorithm guided by a low-resolution energy function. In a second step, the assembled models are refined using an all-atom potential to improve hydrogen bond geometry and reduce collisions. This refinement, called `FARFAR`, incorporates energy terms specific to RNA molecules. The `FARFAR` protocol involves generating and selecting models through 10,000 to 100,000 independent simulations, which are then clustered. For short RNA fragments, the Rosetta Online Server That Includes Everyone (`ROSIE`) [124] can be used. The `FARNA/FARFAR` method predicts 3D RNA structures using three-nucleotide fragments and the knowledge-based energy function. Later, `FARFAR2` [206] was introduced with improvements in prediction accuracy and efficiency, including an updated fragment library, a fractional filter, a special MC movement and a new all-atom scoring function. `FARFAR2` demonstrates improved accuracy in recovering near-native structures. `FARNA/FARFAR/FARFAR2` is applicable to approaches with very small motifs as fragments and has a very good prediction accuracy, but is computationally expensive.

`MC-Fold/MC-Sym`
: The two computational tools `MC-Fold` and `Mc-Sym` are included in the pipeline of Parisien and Major [134]. Both use the Nucleotide Cyclic Motif (NCM) to derive effective knowledge-based scoring functions. The NCM database contains single-pair loops and double-stranded NCMs and a library of corresponding 3D motifs. `MC-Fold` predicts secondary RNA structures using the NCM database, while `MC-Sym` assembles 3D structures based on the predicted secondary structures. This allows the pipeline to accurately predict both secondary and tertiary structures without the need for additional scoring in `MC-Sym`. Starting from an RNA sequence, the output of the pipeline is a 3D PDB structure.

`Vfold`

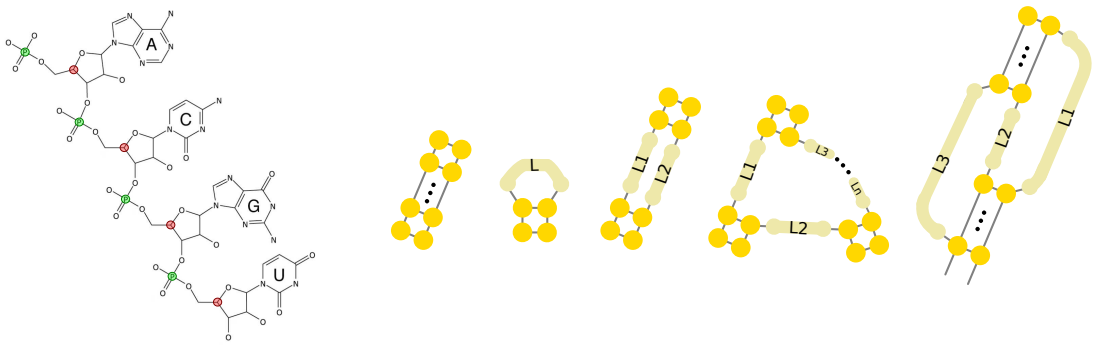
The virtual bond model pipeline consists of the 2D structure prediction tool `Vfold` [31] and the 3D structure prediction tool `Vfold3D` [34, 217] / `VfoldLA` [215, 216]. Experimental `SHAPE` data can also be used as a constraint for 2D structure prediction, and with a given 2D structure, 3D structure prediction can also be started directly. The 2D virtual bond model [31] is a statistical-mechanical calculation, which distinguishes it from the `MC-Sym` approach, with the aim of including the range of realistic three-dimensional RNA structures in its 2D prediction. The nucleotide backbone is defined by the bonds between $P - C4$ and their angles to each other, see figure I.10. In loop regions, these virtual bonds are more flexible. A diamond lattice is used to simulate the variation of loop structures, this is done by self-avoiding random walks of the virtual bonds on the diamond lattice. The off-lattice helix conformations are then connected to the nearest loop lattice conformations, measured by the minimum root-mean-square deviation (RMSD). This simplified 2D lattice model thus contains atomic level detail and accurately captures the different rotameric states of virtual bonds. The algorithm iteratively finds the lowest free energy conformation for segments of increasing length starting from the 3' end of an RNA.

This approach was then extended to predict pseudoknots of type H (see section I.1.2 and [32, 33]), and was also used to compute entropy parameters for kissing hairpin conformations (see section I.2.1 and [35]). A drawback of this model, in contrast to `SimRNA`, is that it does not include non-canonical interactions, which can play an important role in the formation of interactions and complex pseudoknots.

From the resulting ensemble of 2D structures, the low free energy structures are evaluated and assembled into a full 3D coarse-grained structure using 3D fragments, similar to but not as detailed as in `RNAComposer`, and then translated into an all-atom 3D structure refined in the `AMBER` force field. [34]

The fragment templates are categorised into A-shaped helices, hairpin, internal and bulge loops, n-way junctions, open junctions between two helices and H-type pseudoknots. When selecting a 3D fragment for sampling, the length of the fragment is taken into account and then the fragment with the highest sequence match is ranked higher. An extension of this approach, `VfoldLA`, searches for individual strands within a fragment rather than the entire fragment, e.g. the entire two-way junction. This increases both the chance of successfully sampling a structure in the first place and the structural diversity of this knowledge-based approach. However, it has the disadvantage that loop/fragment specific features are lost, resulting in

a less accurate prediction. The `Vfold2D` diamond lattice model used also has the disadvantage of being a generic model and is only suitable for short and simple loops due to its computational cost. `Vfold2D-MC` [44] is a follow-up approach in which both helices and loops represent off-lattice conformations and the angles for generating the loop conformations are generated using MC sampling. Weights based on experimental data are used for each angle. At the time of writing, the ability to predict pseudoknots in this 2D approach of Cheng et al. [44] was not yet implemented.



(a) **(b)**

Figure I.10.
`Vfold`: (a) The nucleotide backbone is determined by a connection between the phosphate (P) (green) and $C4$ (red) (b) Overview of the fragment templates used: helices built with A-form helices, hairpin and internal/bulge loops with different loop lengths, n-way junctions (3–7 and open junctions) and a few h-type pseudoknot templates [217]. Depending on the availability within the isolated structures for the fragment library, there are varying numbers of samples per fragment.

`RNAComposer`

`RNAComposer`[141] takes an user supplied RNA secondary structure and builds a 3D RNA structure from it. The RNA `FRABASE` [140, 142] database plays an essential role in this process. `FRABASE` contains 2D and 3D structural motifs selected and isolated from the PDB [17] (see section I.4). The resulting 3D structural elements contain information on the respective atoms, energy, stereochemistry and structural properties. The given secondary structure is decomposed into its motifs (stems, different loop types, single stranded) and the matching 3D elements are selected according to secondary structure topology, sequence similarity, base compatibility, resolution of the structure in `FRABASE` and energy. Finally, the assembled structure is refined and the energy is also improved using the `CHARMM` force field. A common problem with knowledge-based methods is the lack of basic data, which is handled differently from approach to approach. For example, `RNAComposer` selects primarily on the basis of sequence compatibility, unlike `Ernwin`, which selects structural elements from its database completely independently of sequence, thereby increasing the number of motifs. If this is not possible due to a lack of data, bases can be replaced in the motif search.

`Ernwin`

`Ernwin` [94, 187] is a helix-centred coarse-grained model for predicting 3D RNA structures from their secondary structures. It represents secondary structure elements, stems and loops, as cylinder segments, see figure I.11, and considers elements connecting helices as degrees of freedom. The `Ernwin` energy function consists of five separate terms: two constrained energies for collision and junction closure, and three unconstrained energies for radius of gyration, A-minor energy and loop-loop interaction energy, derived from known crystal structures in the PDB database. An important component of `Ernwin` is `forgi`, a library for RNA secondary structure analysis. With `forgi`, `Ernwin` has the ability to interpret most standard RNA secondary structure formats, see section II.2. An initial 3D model is then constructed from the secondary structure. The model uses MCMC simulation to efficiently predict RNA 3D structures, and the structure with the lowest energy in the conformational ensemble is selected as the predicted final structure.

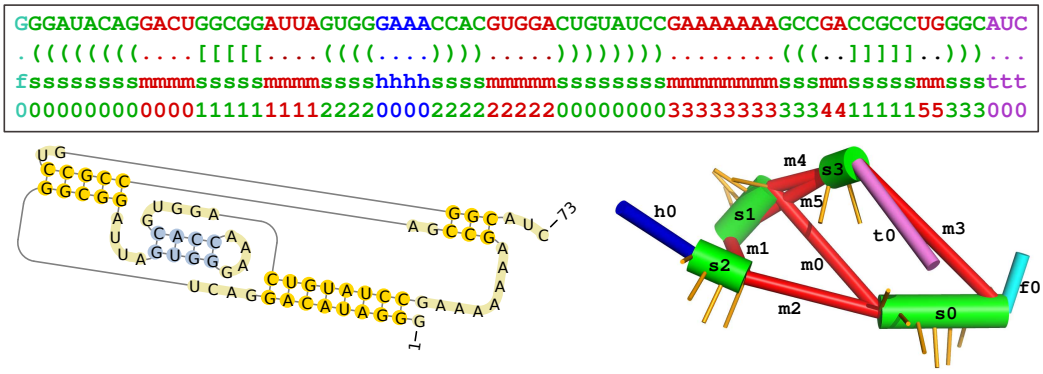


Figure I.11.
`Ernwin`: An RNA molecule is viewed as a collection of secondary structure elements, with the stems represented as cylinders connected by internal links and loops. This is shown here using the example of the ZMP (5-amino-4-imidazole carboxamide ribose-5'-monophosphate) riboswitch (PDB 4ZNP [150]). The upper box shows the sequence and the dot-bracket secondary structure. The secondary fragments are divided into the respective fragments and numbered consecutively; 5'-end (turquoise), stems (green), multiloops (red), hairpin loops (blue), 3'-end (magenta). Bottom left the secondary structure [82] and in comparison the `Ernwin` representation with the cylinders in the respective color including labelling (visualised using `PyMol` [163]).

In addition, `Ernwin` allows the integration of small-angle X-ray scattering (SAXS) data. These provide a good overview of the global shape and size of an RNA structure with their 1-dimensional profile. When using SAXS data in the `Ernwin` calculation, the radius of gyration energy is replaced by a pair distance distribution energy. The helix-centred method allows `Ernwin` to quickly predict plausible global shapes of large RNAs using RNA sequence and secondary structure as essential inputs. As with other approaches, `Ernwin` also provides an in-house tool for translating the resulting coarse-grained structure into an all-atom model.

YUP

In Yammp molecular mechanics package — Yammp Under Python (YUP) [113, 181] each nucleotide is represented by a bead located on the P atom, see figure I.12 (a). The beads are connected by pseudobonds, which form a kind of RNA backbone. The one bead per nucleotide model represents the highest possible resolution within YUP. In addition, there is a model in which one bead represents a complete helix (regardless of the length of the helix), and one in which a helix is represented by five or more beads, see figure I.12 (b). The five-bead model makes it possible to extrapolate the representation of the single-bead model, especially with the correct orientation of the helices. Here each helix is represented by four beads at the four ends of the helix and, depending on the length of a helix, one to three additional 'space filling' beads are added to fill the helix. YUP is designed to improve the quality of low resolution structures by incorporating additional experimental constraints. Harmonic functions are used to enforce constraints such as bonds, angles and improper torsions. A value of zero would therefore indicate that all experimental constraints have been successfully satisfied. Deviating values can consequently be used as a rough indicator of the quality of the model. In this sense, it is a refinement approach rather than an approach for predicting de novo structures. Nevertheless, it allows larger RNA models to be modelled within the given data and small adjustments to be made.

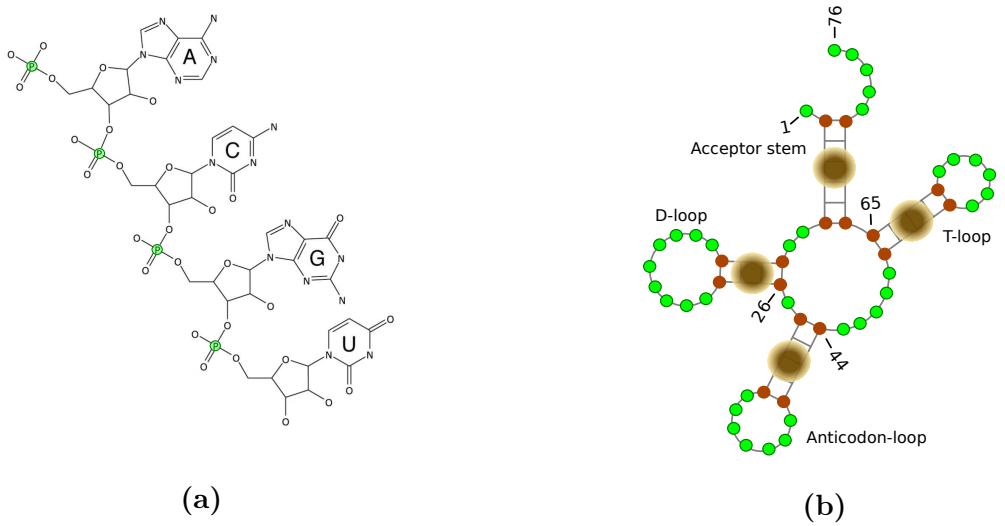


Figure I.12.
YUP: (a) Contrary to NAST, here the one bead is on the P atom (green). (b) A helix can also be reduced to five beads: 4 beads at the helix ends (brown) and, depending on the length, at least one bead in the helix centre (shaded brown), shown here as a tRNA example. A further reduction is to use only the helix centre.

NAST

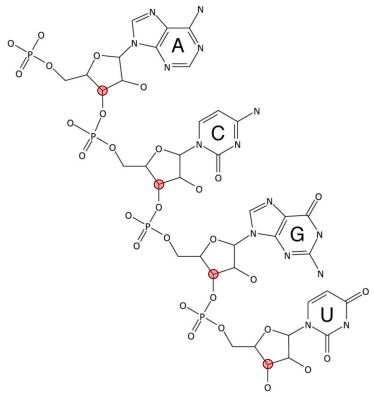


Figure I.13.
The NAST approach is a coarse-grained representation with the one bead on the $C3'$ atom (red).

The nucleic acid simulation tool (NAST) [88] is an one bead per nucleotide model with the bead at the $C3'$ atom, see figure I.13. It is based on a MD algorithm combined with knowledge-based potential energy. Like YUP, NAST is limited by the fact that it cannot make de novo predictions. It requires information on secondary structure and tertiary contacts. The NAST energy function integrates four types of statistical information: geometries from three high-resolution solved ribosome structures (distances, angles and dihedrals between $C3'$ atoms of consecutive nucleotides), a term to prevent steric overlap between unbonded residues separated by more than three residues in sequence, ideal helical geometry for nucleotides in the secondary structure, and long-range interactions between nucleotides in tertiary contacts. From the sampled structures, 1,500 randomly selected structures are clustered by similarity and ranked according to their original and experimental data. The additional supporting software Coarse to Atomic (C2A) [87] allows the structures predicted in NAST to be translated from coarse-grained to all-atom.

iFoldRNA/iFoldV2

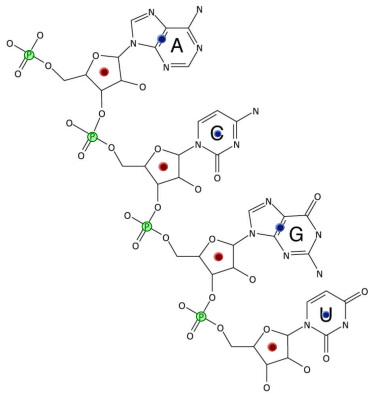


Figure I.14.
iFoldRNA: Bead representation with one bead at the P atom (green) and one each at the centre of the sugar (red) and the base (blue).

iFoldRNA [56, 168, 99] uses a 3-bead system (see figure I.14): at the centre of each mass, one is placed on the P and one on the sugar. The bases are all treated equally and have one bead in the centre of the six atom ring. Each of these beads, and the beads of the preceding and succeeding bases, are linked together by constraint parameters such as bonds, angles and dihedrals. These parameters are based on values obtained from high-resolution RNA structures. The system is highly simplified, but takes into account canonical base pairing, stacking (distinguishing between purines and pyrimidines), electrostatic effects such as short-range phosphate-phosphate repulsion, and hydrophobic effects to prevent overpacking between bases. The non-bonded parameters used are composed of values from Mathews et al. [119], which are known from 2D structure prediction, see chapter 1.2. This also allows a thermodynamic analysis of the respective simulation. In addition, there is an explicit loop entropy using experimentally validated free energies for different loop types and sizes. The replica exchange DMD simulations can be started de novo from a linear structure, but can also be supported by a secondary structure or hydroxyl radical probing (HRP) data. In general, each of the parallel replica exchange simulations runs at a different temperature, which differs from e.g. SimRNA where each replica exchange simulation starts at the same given start temperature and simulates to the end temperature. From the resulting ensemble, a set of structures with the best energy is clustered based on their RMSD, and the centroids of these clusters are reconstructed into a full all-atom structure.

SimRNA

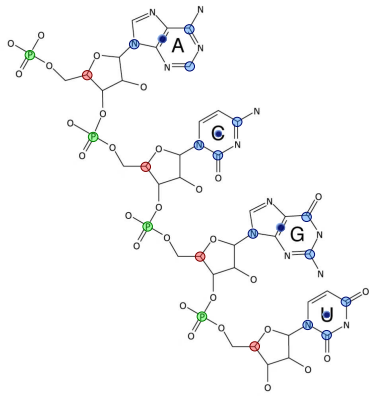


Figure I.15.
SimRNA is based on a five-bead system: P (green), $C4'$ (red) and for the bases with $N1/N9$, $C2$, $C4/C9$ (blue) and the centre of each base (blurred blue).

SimRNA [26, 112] reduces a nuclodide to a five-bead model, see figure I.15. One bead is located at P and one at $C4'$, and a distinction is made between pyrimidines ($C2$, $N1$, $C4$) and purines ($N9$, $C2$, $C6$). In addition, the centre of each base, between $N1$ and $C4$ for pyrimidienes and $N9$ and $C6$ for purines, and a central 3D layer around the base describe the excluded volume and show possible attractive interaction regions. SimRNA is a knowledge-based tool that works with REMC. The energy function is based on statistical potentials derived from observed structural patterns in known 3D RNA structures. It consists of two classes of terms: (a) sequence-independent local terms related to RNA backbone geometry, based on bond lengths, angles and torsion angles, and (b) sequence-dependent long-range terms representing interactions between nucleotide residues, such as base-base, base-backbone and backbone-backbone interactions. The interaction preferences and distances are stored in 3D layers, and backbone-backbone interactions are not sequence or orientation dependent. SimRNA allows de novo structure prediction and starts with an open ring structure. Alternatively, it is possible to start directly with a 3D structure and manipulate it with constraints in the form of distances. These can be integrated in different ways, e.g. by specifying a secondary structure or by specifying individual distances between nucleotides, the weight of which can be set individually. SimRNA provides an ensemble of possible movement sets. It is worth noting that the user can easily adjust the possible frequency of each move set used in the prediction process. An additional feature is that SimRNA, like HireRNA, allows non-canonical base pairs and multiples in its structure predictions. The resulting ensemble can be clustered based on the RMSD and translated into all-atom structures.

HiRe-RNA :

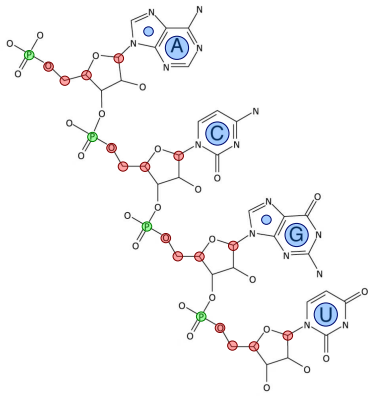


Figure I.16.
HiRE-RNA uses a system of six to seven beads; for the backbone one at the P atom (green), for the sugars those at $O5'$, $C5'$, $C4$, $C1'$ (red) and for the bases — in contrast to other approaches — not at single atoms but one to two in the respective ring centres (blue).

The High Resolution RNA (HiRE-RNA) coarse-grained force field of [46, 47, 135] represents nucleotides by using one bead at the phosphate group, four at the sugar and beads centred in the respective bases for purine (two beads) and for pyrimidine bases (one bead), see figure I.16. In HiRE-RNA, the interaction potential is composed of local interactions and non-local interactions, which in turn are divided into short-range, such as base-pairing/hydrogen bonding and stacking, and long-range, such as electrostatic potentials. There is also an excluded volume potential, as in SimRNA. For local interactions, harmonic potentials are used for the covalent bond parameters such as length, angle or torsion. Optimisation of the energy parameters is achieved using a genetic algorithm and is applied to the respective energy potentials by a weighting coefficient. The model uses the REMD conformational sampling algorithm. Structure prediction is possible from the fully extended state with no information other than the sequence or with minimal information such as the 2D structure.

3.2. Deep-Learning-Based Approaches

The application of deep learning-based approaches is bringing significant advances in many areas of research. A well-known example in both molecular biology and structure prediction is AlphaFold [90, 193], which provides accurate 3D protein structures. RNA structure prediction is also trying to benefit from the deep learning approach. End-to-end approaches such as E2Efold-3d/RhoFold[43, 169] and DRfold [104] or geometry-based approaches such as trRosettaRNA [202] and DeepFoldRNA [136] apply this approach in 3D. In addition, tools such as ARES (see section I.3.3), which are employed for the evaluation of the quality of a predicted RNA structure, make use of deep learning.

In principle, these approaches can be divided into 3 components: (i) feature extraction, (ii) structure prediction and (iii) refinement. (i) It attempts to compensate for the lack of data using a combination of a deep score network and a post-processing network. The

output of the former is a scoring matrix with pairing probabilities. The data for these approaches are mostly filtered from the known databases (see section I.4 for an overview of RNA 3D structure databases) and for the extraction of the pairwise residual features `Infernal` [128] and `rMSA` [222] have often been used. `Infernal` uses structurally annotated multiple sequence alignments (MSA) to generate probabilistic profiles of the sequence and secondary structure of the RNA family. These profiles are known as covariance models. They are used to identify new RNA family members in sequence databases and to generate comprehensive MSA. `rMSA` is a pipeline designed for sensitive search and construction of RNA MSA from RNA homologs for a target RNA sequence. It uses component methods such as `blastn` [4], `Infernal`, and `RNAfold` [107], but introduces a novel five-step hierarchical sequence search strategy to attempt to prevent the inclusion of unrelated sequences.

As an example `E2Efold-3D/Rohfold`: (i) The data for this model come from `Rfam` [91] and `RNAcentral` [176] and for the extraction of the pairwise residue features `Infernal` and `rMSA` were used. The post-processing is based on the RNA foundation model (`RNA-FM`) [41]. The pre-trained language model is based on selected and pre-processed non-coding RNA (ncRNA) structures from the `RNAcentral` database and is extended with additional data points, especially for unlabelled sequences. These two networks are combined and trained together in an end-to-end approach. (ii) The main element for structure prediction is the base pairing from which the 3D structure is reconstructed. For the 3D structure, two beads are placed on the sugars ($C1'$, $C4'$) and one each on the base ($N1/N2$) along the backbone. Four torsion angels are also added to reconstruct and subsequently optimise the backbone. For training, the dataset `bp-RNA-1m` [50] and `RNAStralign`, an RNA alignment and structure dataset aggregated by Tan et al. [182], are used for the ground truth secondary structure. In (iii), remaining structure clashes and steric violations are resolved by the `Amber` force field. As `E2Efold-3D/Rohfold` corresponds more to a 2D deep learning structure prediction with subsequent 3D embedding, the structure constraints are more similar to those of 2D structure prediction, also due to the external models used, such as `rMSA`. Furthermore, only $A-U$, $G-C$, $G-U$ base pairs are allowed, overlapping base pairs are not allowed, the one with the best covariance is taken, and sharp loops with a length of < 4 bases are not permitted.

`trRosetta` also consists of three steps. (i) Starting from a given RNA sequence, one MSA is extracted with `rMSA` and one with `Infernal` using the sequence databases of `NCBI` [161], `Rfam` and `RNAcentral` databases. Base pairing probabilities are predicted using `SPOT-RNA` [170], a deep learning network for RNA secondary structure prediction.

Each base is described by 5 beads: P, for the sugar $C3'$, $C1'$, for purines $N9$, $C2$ and for pyrimidines $N1$, $C4$. The backbone orientation, as well as the distances and orientation of the beads from one nucleotide to the next, are predicted using deep learning (ii). The structure of the transform network used for prediction is based on that of `AlphaFold2`. (iii) A full atom structure is then generated by energy minimisation using the deep learning potential and the physically based energy terms in `Rosetta`.

`DRFold` [104] obtains its learning and test data points from a filtered individual set. This is based on the `PDB` database [17]. From the given nucleotide sequence, the base pair probability is predicted using `RNAfold` and `PETfold` [164]. It uses a coarser grained representation with one bead on the phosphate backbone, one for the sugar and one for the base (P, $C4'$, $N1/N9$). This 3-bead system is used to build the full atom model. Instead of direct base pairing potentials, it relies on relative frame positions and geometric potentials. The full atom reconstruction is done in two steps — first `Arena` [138] is used for the standard conformation, where the 3-beads model remains stable and only refinement steps such as incorrect bond lengths and angles or atom clashes are corrected. In the second step the local structure geometry is refined using the `Amber` force field. `DRFold` is a standalone pipeline, but has also been tested as a hybrid pipeline with `DeepFoldRNA`. `DeepFoldRNA` uses a self-organising neural network that receives the input features generated by `rMSA` and `PETfold`. It predicts two main geometric constraints: the distances between the nitrogen atoms $N1/N9$ of the base attached to the ribose sugar $C4'$ and the backbone P atoms, and the orientations between the residues. These constraints are converted into composite potentials that guide the folding simulations. The energy function used here is based on a force field consisting of 7 weighted parameters determined from the RNA structures in the training set. The combination of an end-to-end approach with a geometry approach allows to optimise the default geometry potentials in `DRfold` and to improve the prediction accuracy according to their tests.

However, the success in the protein field cannot be directly transferred to the RNA field. A very important issue is the availability of RNA structures. Even in the 2D field, where more data is available, it is not versatile enough to allow efficient learning on a larger scale of different RNA types. The review by Flamm et al. [67], in which several 2D deep learning approaches were tested, clearly shows where the current major drawbacks of deep learning lie with respect to RNA structure prediction. Ideally, test and training sets should consist of different RNA families for optimal construction. This lack of data is particularly critical in the 3D domain, given the limited quantity and inherent bias of the data, as reviewed in Schneider et al. [162]. It is also important to consider that a 3D structure comprises

a far greater variety of structural motifs than its 2D counterpart. Furthermore, factors such as non-Watson-Crick base pairs, tertiary structure motifs, including pseudoknots, or RNA-RNA interactions, which play an important role in stabilising structures, are only partially available and insufficiently classified to allow for the precise determination and further follow-up deep learning prediction of sequence and structural properties.

Nevertheless, it must be acknowledged that the strategy of using existing experimental RNA data as reference points for prediction tools, as in knowledge-based approaches or tools such as `Ernwin` that use helix and loop elements based on previously identified structures, can be quite successful in predicting accurate and valid structures. Analysis of previously isolated structures can also help to extend the definition of limiting structural arrangements, such as the length of loops in H-type pseudoknots and the resulting changes in energy parameters in the prediction (see publication in section II.1 and II.2).

3.3. RNA 3D Structure Refinement and Analyse

Some of the tools presented above (section I.3) have an in-house tool, or integrated parts of existing tools, for fine-graining, building a full atomic structure, or tools that provide detailed information about the predicted 3D structures. This information provides additional information about the spatial arrangement and interactions within the RNA molecule, such as the type of base pairing, angles, etc., in addition to the full 3D structure, and shows whether a predicted structure can be built at all. Such analyses can also help to identify indicators of possible interactions with other molecules and to gain a more detailed understanding of structure formation based on experimentally or computationally determined structures, helping to improve 3D approaches, but also 2D approaches in the case of translational rules.

Detailed Structure Analysis

Common stand-alone tools that provide a detailed view within a PDB structure are for example Find RNA 3D (`FR3D`) [160], `MC-Annotate`[70], Dissecting the Spatial Structure of RNA (DSSR) [108], `RNAView` [219], `CompAnnotate` [84] and ClaRNA [198].

`FR3D` is a database and software tool for analysing 3D RNA structures and interactions. It provides another comprehensive resource for studying RNA tertiary structures, base interactions, base-phosphate interactions, stacking and other structural motifs, and RRIs in various biological contexts.

MC-Annotate combines geometric and symbolic methods to analyse 3D RNA structures. The approach transforms RNA structures into graphs in which nucleotides are nodes connected by edges labelled with the pairwise distance measure. Using a graph isomorphism algorithm, different structural motifs such as base pairs, base triples and backbone interactions are identified and characterised within a PDB.

DSSR is a software package that analyses and annotates RNA structures. It identifies the unmodified bases and their base pairs, but also recognises features such as modified nucleotides, base stacking interactions of multiple stems within a helix, and distinguishes between different types of loops. The non-redundant RNA set [103] (see section I.4) is used as a data source for the motif search. A feature that can be useful for 2D prediction tools is the ability to switch the "existence" of pseudoknots on and off when determining the secondary structure of a 3D structure.The generated images and diagrams can be displayed by including PyMol [163] or Visualization Applet for RNA (VARNA) [52].

Either MC-Annotate or DSSR can be used as a package in the 3D coarse-grained helix-centred structure prediction tool Ernwin.

RNAView is a program for the annotation, visualisation and analysis of RNA secondary and tertiary structures. RNAView provides detailed annotation of nucleic acid structures, including helical regions, base pairs (including the various pairing orientation combinations of Watson-Crick, Hoogsteen and sugar edge, in total 12 families of edge-to-edge base pairs), loops and modifications. It can generate 2D plots in a variety of formats, and with RNAMLview, a graphical editor, it is possible to customise 2D representations of RNA structures. RNAView is also part of the simulation tool SimRNA.

In summary, each of these tools serves a specific purpose in the analysis and annotation of RNA structures, offering unique features and functionalities tailored to different aspects of RNA structure analysis. The basic difficulty with all of these tools is that for a given PDB structure — especially from an experimental setting — it is necessary to determine exactly which particular bases interact with each other and in what way. The success of this task is highly dependent on the resolution of the structure. One approach to refining a structure is the comparative tool CompAnnotate. Geometric information that is not available with the required accuracy for a given PDB structure may be available at high resolution in structurally identical PDBs, especially for conserved local structures. Therefore, the tool uses input PDB files and existing annotation tools to improve the accuracy. Another approach is offered by classification of contacts in RNA tertiary structures (ClaRNA). Using a selected RNA dataset from the PDB, existing tools

(MC-Annotate, RNAView, FR3D and the ModeRNA stacking algorithm [155]) were applied to these structures. This classification was used to identify interactions that could be unambiguously identified, and a reference dataset is built based on the results. This allows the strengths and weaknesses of each method to be balanced and allows analysis on coarser grained structures.

Obviously, the capabilities of these tools depend heavily on the quality of the 3D structure given to them. If the resolution is too low, whether in the course of an experiment or a structure prediction, the structure information will be very inaccurate or, in the worst case, incorrect. This is where 3D structure refinement tools such as QRNAS [175], Backbone Rotameric and Quantum-mechanical-energy-scaled base-base knowledge-based potential (BRiQ refinement) [214], RNAfitme [9] and 3dRNA optimisation [200] can partially support, also to relax possible collisions or excessive stretching in the course of a coarse-grained structure prediction. However, the refinement of the local or global structure is a crucial process. For example, RNAfitme refines 3D RNA structures by reconstructing them with fixed backbones and simulating them guided by the CHARMM force field. Although this approach is effective in reducing local spatial conflicts, it struggles to improve the overall quality of the structure. In contrast, QRNAS uses a modified version of the Amber force field to improve local RNA 3D structures, including modified residues. The software provides the option to enforce backbone regularisation, which corrects outlying parts and improves base pair planarity compared to RNAfitme [203]. However, both tools have weaknesses in global RNA structure refinement.

Structure Prediction Evaluation

Each prediction tool generates an ensemble of structures. An important question after any structure prediction is whether the final predicted structure is a feasible and accurate structure. Many approaches already evaluate their ensemble, either by applying clustering algorithms and selecting the respective centres, or by evaluating the structures directly by their energy, or both. In addition, there are other stand-alone programs that specialise in the evaluation of RNA 3D structures, distinguishing between knowledge-based scoring functions/statistical potentials and deep-learning based scoring functions.

Examples of the former are Ribonucleic Acids Statistical Potential (RASP) [36], the all-atom knowledge-based potential approach of Bernauer et al. [19], distance-scaled finite ideal-gas reference state (DFIRE-RNA) [223], and the two tools residue separation based statistical potential (rsRNASP) [179] and coarse-grained statistical potential (cgRNASP) [180].

In general, any parameter information that provides insight into the general RNA geometry can help to evaluate a predicted RNA structure. Ideally, a diverse and extensive non-redundant set of structures is used to infer the potential. In the case of RASP, 85 selected experimentally determined PDB structures are used to compute and optimise the average pairwise distance-dependent energy scoring functions and subsequently to form a decoy set. The approach provides four knowledge-based potentials that differ substantially in the number and type of atoms representing the RNA nucleotides (similar to the beads of coarse-grained prediction methods), including an all-atom potential. For the decoy set, 500 decoy/non-native structures were generated from each of the 85 structures. Parts of the constraints were randomly removed or added to the native structure. These non-native structures can then be compared with the native structure to assess the accuracy and robustness of the RNA structure. Given a native PDB structure, the output is an overall RASP score, which is the sum of the individual scores and the contributing total number of interactions.

Another approach that uses the distance between atom pairs as a geometric parameter is the all-atom knowledge-based potential approach of Bernauer et al. [19]. It also allows a distinction to be made between coarse-grained (five beads per base) and all-atom knowledge-based potentials. As with the generation of the RASP dataset, only crystallographic RNA structures with a resolution of at least 3.5 Å were selected. After the final selection process, their dataset consists of 77 structures. The resulting potentials are differentiable, which means that they can be used for energy minimisation or MD simulations. Two decoy sets are generated; one generated by MD simulations to obtain near-native structures with a RMSD of <2 Å and the second further away from the native structure using REMD for a set with RMSD <10 Å. Scoring of the decoys was done with the `Rosetta` scoring function [53, 54].

DFIRE-RNA is also based on distance potentials (with 85 atom types) to model the atomic pair distribution of RNAs in their reference state they employ a function in ideal gas [224].

In general, an all-atom model with different atom types is not automatically better than a coarse-grained model, as the comparison of rsRNASP and cgRNASP shows [204, 180]. In particular, rsRNASP attempts to integrate the often neglected distinction between short and long range interactions into its potentials. For this purpose, the reference states of the training set (X-ray RNA structures <3.5 Å) from the non-redundant RNA set are divided into short-range (built by averaging) and long-range interactions (built by random-walk-chain), which are additionally weighted. Especially for coarse-grained 3D

structure prediction, a coarse-grained potential is essential. cgRNASP is based on the same principle as rsRNASP, with the difference that the finite ideal gas reference state is used for the long-range reference states, and that it provides three coarse-grained statistical potentials: one-bead ($C4'$), two-bead (P, $C4'$) and three-bead (P, $C4$', $N1/N9$), which are also the common beads in coarse-grained prediction tools. The disadvantage of all these methods is that they are exclusively distance dependent.

The same RNA training set used by RASP, but for an all-atom four-body statistical potential approach, is used by ribonucleic acids multibody potential (RAMP) [116]. This approach is an attempt to cover higher order interactions. In RAMP, each RNA molecule is represented by the four atom types C, N, O, P. The multibody potential energy function considers the positions and orientations of all four molecules simultaneously, generating quadruplet types based on this four-letter alphabet to accurately model their interactions. The stability of an RNA structure, which relies on base pairing via hydrogen bonding and stacking, is in turn strongly dependent on orientation. The tool 3dRNAscore [201] combines the averaged all-heavy atom distance statistical potential with a torsion angle dependent statistical potential.

As with structure prediction, attempts are being made to determine the accuracy of structural models using deep learning approaches: Li et al. [105] used an RNA 3D convolutional neural network (CNN) (RNA3DCNN) to create two scoring functions for evaluating RNA decoys. One was trained by MD simulation and one by MC structure prediction on 414 RNAs ≤ 3.5 Å selected from the Nucleic Acid Knowledgebase (NAKB) [18, 101]. The Atomic Rotationally Equivariant Scorer (ARES) [189] is a neural network that considers the rotational symmetry of RNA molecules and their atomic interactions to assess the quality of RNA structures. It's initial layers use 3D coordinates and element types of atoms as input to recognise structural motifs through training, allowing local information gathering to identify finer-scale motifs such as base pairs and coarser-scale motifs such as helices. ARES is trained on data from FARFAR2 and performs well on test data generated from it, but may be biased towards this dataset. Deep learning based scoring functions have advantages over traditional statistical potential but, like deep learning prediction approaches, are limited by a lack of structural diversity which reduces their performance outside of their training diversity.

Besides the possibility to test predicted RNA structures for their feasibility and correctness, RNApuzzles [49, 123, 121, 122] offers the possibility to evaluate RNA 3D prediction tools and, in a broader sense, to report on the progress in prediction accuracy in the 3D field, their strengths and weaknesses. The concept behind RNAPuzzle is very similar to

its counterpart in protein prediction; critical assessment of protein structure prediction (CASP). RNAPuzzle is a blind experiment in which a target sequence for an as yet unpublished experimentally determined crystal structure can be predicted by interested research groups and/or their computational tools. These results are then compared with the experimental structure for correct stereochemical structure, topology and geometric similarity. The RMSD is a commonly used metric for ranking predicted models against a crystal structure and provides a measure for global topology comparisons, but may not always accurately reflect the ranking due to errors distributed throughout the structure. Therefore, several metrics are used to evaluate different structural properties of predictions [111]. These include the deformation index (DI) [133], interaction network fidelity (INF) [133], clash score [211] and mean of circular quantities (MCQ) [225, 210]. Each metric focuses on a particular aspect of prediction. For example, DI standardises the RMSD relative to the sequence length, INF evaluates different interaction types including stacking interactions, the clash score evaluates atomic harmony and MCQ compares the torsion angle space. These metrics have advantages and disadvantages, but when considered collectively, they offer valuable insights into various aspects of the predicted structures. Structure prediction tools that have participated in the RNAPuzzle experiment, such as FARFAR, SimRNA, iFoldRNA or RNAComposer, have proven to be quite powerful and have achieved results in terms of stereochemical accuracy similar to those of human experts. [38]

4. 3D Datasets

RNA-related data sets are an important key element in RNA research. They contain a wide range of information such as sequence, structure, functional annotations, etc. It is important to have well-founded benchmarks, especially for deep learning approaches, but also for traditional structure prediction. Altogether, databases play a key role in enabling the development of computational tools and algorithms that advance RNA research by providing valuable resources for the study of RNA biology, evolution and function.

A central hub for ncRNA sequence information is the `RNAcentral` [176] database, which serves as an access point from several separate established RNA databases, such as the RNA families database (`RFAM`) [91]. Newer versions also include additional information such as genomic coordinates or RNA family classifications based on the linked databases. The more recent versions also provide the 2D structure of a wide range of RNA structures. In the `Rfam` mentioned above, each RNA family is represented by a MSA by `Infernal`, the consensus secondary structure and the covariance model. From version 14, search and 2D visualisation including pseudoknots is possible.

Databases such as the Protein Data Bank (`PDB`) [17] provide access to 3D structures determined by NMR, X-ray crystallography or cryoelectron microscopy. The Non-Redundant/Reduced Redundancy Data Set of RNA (`http://rna.bgsu.edu/rna3dhub/nrlist/`) [103] has been developed to offer an alternative set of RNA 3D structures, which are selected to avoid redundancy inherent in the PDB database, due to the existence of multiple 3D structure representations of the same molecule from the same organism. The selection is based on a sequence comparison utilising the length of each chain and the number of identical bases in an alignment, followed by a structure comparison using `FR3D` (see section I.3.3). The dataset is updated weekly and includes structures with a 4Å resolution or better. It allows a closer look at the composition, e.g. the abundance of certain RNA species, to counteract any bias in further studies such as Beckmann [15] or the publication in section II.1. Other large databases containing 3D RNA structures include the Cambridge Structural Database (`CSD`) [77], the Nucleic Acid Database (`NDB`) [127] and its successor the `NAKB` [18, 101]. The `CSD` also contains a collection of RNA structures, but is not primarily focused on nucleic acids. The `NAKB` database integrates information on nucleic acids, including RNA, from a variety of sources to provide a comprehensive knowledge of their structures and functions. The database uses both experimentally determined structural information and various tools, such as

DSSR, to analyse it.

Especially for specific structural conformations such as pseudoknot, it is challenging to obtain a well-founded parameter dataset that can be used for valuable prediction, as overall there is limited and partially biased data available. For example, there are more experimentally determined tRNA or rRNA that can be used to analyse pseudoknot conformations than other ncRNAs [15]. Many computational approaches focus on selected individual structures to determine the properties of pseudoknots, such as Gultyaev et al. [79]. This is not fundamentally wrong, as it allows the parameter set to be extended by pseudoknots, but it only covers a very limited and one-sided view. Approaches to counteract this include PseudoBase++ (https://rnavlab.utep.edu/database) [183], an extension of the PseudoBase database (https://www.ekevanbatenburg.nl/PKBASE/) [13]. Both use the same data set of pseudoknot secondary structures, which are regularly synchronised with each other, but differ in the amount of data prepared for the user. The disaggregation of certain values around the pseudoknot, such as loop lengths, stem lengths, bulges within the stems, base pair composition, allows the targeted study of patterns within the pseudoknot [1], which may refute or support common theses, such as the asymmetry caused by the deep/shallow groove [79].

Another database focussing on specific structural information is the 'Genus for biomolecules' database (http://genus.fuw.edu.pl) [158]. It records the genus of RNA and protein chains [221], for a further description of the genus see chapter I.1. For RNAs the BGSU RNA Site database is used, for base pairs the classical Leontis and Zirbel [103] classification, but the user can also have an individual PDB structure determined. Although the database recognises elements such as pseudoknots based on the base pair structure, these are not classified further.

However, structure evaluation is based on 2D aspects and neglects essential 3D aspects, such as the steric possibility of forming a stable pseudoknot at all. A tool for systematic analysis of RNA PDB structures is the `pseudoknot_analyzer` distributed with the `forgi` library, see publication in section II.1. In addition to identifying pseudoknots within a 3D RNA structure, the tool allows classification according to Reidys et al. [149] and provides further information about the 3D nature of a pseudoknot, such as the angels between helixes or whether coaxial stacking is present or not.

II. Results and Published Work

1. 3D based on 2D: Calculating helix angles and stacking patterns using forgi 2.0, an RNA Python library centered on secondary structure elements.

Thiel BC, Beckmann IK, Kerpedjiev P and Hofacker IL. 3D based on 2D: Calculating helix angles and stacking patterns using forgi 2.0, an RNA Python library centered on secondary structure elements. *F1000Research* 8 (2019), p. 287. DOI: 10.12688/f1000research.18458.2

Summary

The accuracy of RNA structure prediction has improved considerably in recent decades, except when tertiary structural motifs such as pseudoknots are involved. But the assembly of tertiary structure motifs including pseudoknots, their rotation and angles to each other play a crucial role in stabilising and shaping. In this paper, I show that already existing high-resolution 3D RNA structures can provide reliable additional information and geometries about tertiary structure motifs, in particular the structure-giving and functionally important class of pseudoknots. For this purpose, `forgi 2.0`, a Python library for RNA structure analysis, has been further developed and enhanced.The newly developed extension allows the identification, classification and analysis of a wide range of different pseudoknots in RNA structures and within RRI, and as an application perform statistics on the stems involved in multiloops and pseudoknots. These offer new insights and the possibility to integrate valid structural principles into existing RNA structure prediction approaches, thus extending and ensuring their predictive accuracy to a widely neglected important structural and high-functioning structural motif.

Authors contribution

I (IKB) was significantly involved in the further development of the `forgi` algorithm. An essential part of this work was the design and implementation of a concept for the identi-

fication and classification of pseudoknots in 3D RNA structures, as well as the evaluation of class-specific features and their statistical analysis, including the visualisation of the results.

Software availability

https://pypi.org/project/forgi/

License

SOFTWARE TOOL ARTICLE

REVISED 3D based on 2D: Calculating helix angles and stacking patterns using *forgi 2.0*, an RNA Python library centered on secondary structure elements. [version 2; peer review: 2 approved]

Bernhard C. Thiel[1], Irene K. Beckmann[1], Peter Kerpedjiev[2], Ivo L. Hofacker[1,3]

[1]Department of Theoretical Chemistry, Faculty of Chemistry, University of Vienna, Vienna, 1090, Austria
[2]Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, 02115, USA
[3]Research Group Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, 1090, Austria

v2 **First published:** 14 Mar 2019, **8**:287
https://doi.org/10.12688/f1000research.18458.1
Latest published: 23 Apr 2019, **8**:287
https://doi.org/10.12688/f1000research.18458.2

Abstract

We present *forgi*, a Python library to analyze the tertiary structure of RNA secondary structure elements. Our representation of an RNA molecule is centered on secondary structure elements (stems, bulges and loops). By fitting a cylinder to the helix axis, these elements are carried over into a coarse-grained 3D structure representation. Integration with Biopython allows for handling of all-atom 3D information. *forgi* can deal with a variety of file formats including dotbracket strings, PDB and MMCIF files. We can handle modified residues, missing residues, cofold and multifold structures as well as nucleotide numbers starting at arbitrary positions. We apply this library to the study of stacking helices in junctions and pseudoknots and investigate how far stacking helices in solved experimental structures can divert from coaxial geometries.

Keywords

RNA, Python, RNA tertiary structure, RNA secondary structure, coaxial stacking, pseudo knots

This article is included in the Bioinformatics gateway.

This article is included in the Python collection.

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 23 Apr 2019	✓ view	
version 1 14 Mar 2019	? view	✓ view

1. **Marta Szachniuk**, Poznan University of Technology, Poznań, Poland

2. **Jerome Waldispühl**, McGill University, Montreal, Canada
Roman Sarrazin-Gendron, McGill University, Montreal, Canada

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Ivo L. Hofacker (ivo@tbi.univie.ac.at)

Author roles: Thiel BC: Data Curation, Investigation, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Beckmann IK**: Data Curation, Investigation, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Kerpedjiev P**: Methodology, Software, Writing – Review & Editing; **Hofacker IL**: Conceptualization, Funding Acquisition, Methodology, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was partly funded by the Austrian science fund (FWF) projects F 43 "RNA regulation of the transcriptome" and I 2874 "Prediction of RNA-RNA interactions" and Doctoral College W 1207 "RNA Biology".
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Thiel BC, Beckmann IK, Kerpedjiev P and Hofacker IL. **3D based on 2D: Calculating helix angles and stacking patterns using *forgi 2.0*, an RNA Python library centered on secondary structure elements. [version 2; peer review: 2 approved]** F1000Research 2019, **8**:287 https://doi.org/10.12688/f1000research.18458.2

First published: 14 Mar 2019, **8**:287 https://doi.org/10.12688/f1000research.18458.1

REVISED **Amendments from Version 1**

As requested by reviewer 1 we include the exact formulas and complete code used to compute angles between helices.

See referee reports

Introduction

In RNA 3D structure prediction, knowledge-based potentials are commonly used, especially for coarse-grained approaches that are suitable for larger RNA molecules[1]. The creation of such potentials requires knowledge extraction from solved RNA structures, usually taken from the Protein Data Bank (PDB)[2]. PDB files and their newer replacement, MMCIF files, contain atomic coordinates and additional information in the header fields, but do not contain any base-pairing annotations. To extract information about base pairs and their types[3], dedicated software like MC-Annotate[4], RNAView[5], FR3D[6], DSSR[7] or the RNApdbee web server[8] is required. Due to the hierarchical organization of the RNA energy landscape[9], it is often most convenient to treat secondary and tertiary structure of RNA molecules separately and predict tertiary structures given a secondary structure[10].

For knowledge extraction from RNA-containing PDB files, it is highly desirable to have a software library at hand, which understands the semantics of RNA secondary structures and makes tasks like iterating over loops of a certain type straightforward. Ideally, such a library should be written in an easily accessible scripting language, be well documented and tested and should be available under an open-source license.

Libraries other than the `forgi` library presented here only partly fill these needs. The Vienna RNA package[11] can be used to predict secondary structures from sequence, including advanced features like G-Quadruplex prediction and incorporation of SHAPE data. It provides Python and Perl bindings, but deals exclusively with secondary structure. Biopython[12,13] is useful for dealing with RNA sequence data and can be used to load RNA 3D structures, but has no dedicated support for RNA secondary structure. PyCogent[14], a library for genomic biology, has extensive support for nucleic acid sequences, but contains only a lightweight class for RNA secondary structures and code for pseudoknot removal[15]. A stand-alone version of the latter was included into the `forgi` library under the terms of the GNU General Public License 3.0. More specialized libraries include `modeRNA`[16] (homology modeling, Python) and the FR3D suite[6] (RNA motif search, Matlab). Here, we present the `forgi` library[17], which is centered on RNA secondary structure elements (such as stems, bulges and loops) and makes them usable for 3D structure analysis. `forgi` aims at providing a high level API for many common operations, but can be easily extended with new functionality. The flexibility of providing an open source library in a scripting language is a clear benefit over other programs that are only distributed as binaries. While not restricted to work with PDB or MMCIF files, `forgi` shines especially where 3D information is analyzed in the context of its secondary structure environment.

Methods

Implementation

The `forgi` library[17] is strongly object-oriented, but takes advantage of module-level Python functions where appropriate. The core object representing the secondary structure is the `BulgeGraph` object, which holds a `Sequence` instance for the primary sequence. To include secondary structure based 3D coordinates, the `BulgeGraph`'s subclass, `CoarseGrainRNA` is available. For all-atom 3D analysis, the `forgi` library has a built-in integration with Biopython.

We will briefly describe the three main data structures that hold the sequence, secondary structure and the tertiary structure representation of an RNA.

Primary structure: The `Sequence` class. Each `BulgeGraph` holds a `Sequence` object. Since `forgi` supports loading of data from PDB files (see below), this `Sequence` object has to account for many special cases arising from experimental considerations, which will be detailed in the following paragraphs. There are two numbering schemes commonly used for sequences: 1-based indexing and indexing based on an external reference. In particular, many sequences in structural experiments use "1" to dedicate the first residue of a biological macromolecule, whereas the first residue actually used in the experiment can be upstream of the functional RNA (leading to negative indices) or after the start of the biological unit (leading to an index above 1). The latter is especially common if fragments of larger RNA molecules like the ribosome are studied. Finally, experimenters might decide to insert nucleotides in the middle of a molecule. In order not to affect the numbering of subsequent residues, these inserted residues get the same number as the previous one, followed by a letter (called insertion code).

To handle both kinds of indexing, the `Sequence` class distinguishes indices by type. Integer indices always refer to 1-based indexing, while tuples compatible to the indices used in Biopython's PDB module are interpreted as the second kind of indices.

For many applications, it is necessary to restrict the RNA alphabet to 4 letters (i.e. 4 residue types), "G", "C", "A" and "U". However, in the cell many RNA molecules are post-transcriptionally modified at certain positions. Many modifications, including the methylation of OH or NH_2 groups and A to I editing, have been implicated with a variety of biological functions[18]. During parsing of PDB files, we automatically convert such modified residues to the unmodified parent, but in addition store the modification as an annotation in the `Sequence` object. The corresponding unmodified parents for 3-letter codes of common modified residues were obtained from PDBeChem[19] (http://www.ebi.ac.uk/pdbe-srv/pdbechem/) and the `forgi` library has the ability to query this database on the fly if it encounters a new 3-letter code.

Finally, many experimental 3D structures do not contain coordinates for all residues present in the experiment. The `forgi Sequence` class can store two version of the sequence, with and without missing residues, and the file parser can extract this information from PDB and MMCIF files.

The secondary structure of an RNA is internally represented as a graph, the Bulge Graph, where secondary structure elements (stems, single-stranded regions, interior loops and hairpin loops) form the nodes. Whenever these elements are adjacent along the backbone, they are connected by an edge in this graph. During the Bulge Graph creation, each node gets a unique name such as "s0" for the first stem or "h0" for the first hairpin. The concept of the Bulge Graph, illustrated in Figure 1, has been described previously in more detail[20] and is related to the independently developed RAG (RNA as Graph) approach[21].

`forgi` supports element-based transformations of the Bulge Graph, such as condensing the secondary structure to an representation similar to RNAshapes[22], which we use to classify pseudoknots, for example (see below).

The `BulgeGraph` object allows for easy identification, selection and classification of structural domains, such as multi loops, helices consisting of multiple stems and bulges (termed "rods" in `forgi`), and pseudoknots.

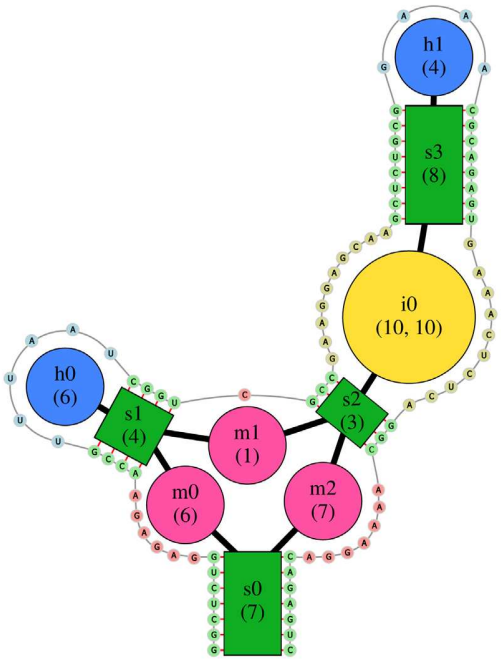


Figure 1. Illustration of the Bulge Graph representation underlying the `forgi` library. Secondary structure elements (stems, bulges and single-stranded regions) are nodes connected by edges. The sequence is shown around the Bulge Graph.

The `CoarseGrainRNA` class holds 3D structures. 3D structures are loaded from PDB or MMCIF files using Biopython's PDB module[13]. In order to assign a secondary structure to the RNA, `forgi` can call either MC-Annotate[4] (Linux only, no MMCIF-support, binary available at http://major.iric.ca/MajorLabEn/MC-Tools.html) or DSSR[7] (binary available after free registration at http://forum.x3dna.org). As an alternative, `forgi` has a built-in heuristic for the detection of canonical base pairs and GU wobble pairs. This heuristic is based on distances along the hydrogen bonds and the coplanarity of the bases, and is not intended to compete with the power of more specialized tools since it will fail in some edge cases, e.g. involving modified residues or residues with missing atoms. This heuristic is a useful fallback, if the above mentioned programs are not available.

During loading of the RNA, a helix axis is assigned to stems as described previously[20]. For each stem, we store start and end coordinates of the helix axis as well as two twist vectors that point towards the minor groove at the beginning and the end of the stem. Similarly, start and end coordinates for bulges, loops and single-stranded regions are stored and can be accessed using the element's name.

`forgi` 2.0 now fully supports co- and multi-fold structures and can load multiple chains that are connected by base pairs into a single `CoarseGrainRNA` object, while chains not connected by any base pair are loaded into separate objects.

Using Biopython's KD-Tree implementation (`Bio.PDB.NeighborSearch`), a list of residues within 6 angstrom from a non-RNA C or N atom is obtained. These residues are considered protein-/ligand interacting. Knowledge of interacting residues is particularly useful to avoid biases in statistics about structural features of bare RNA.

A cleaned version of the PDB - with modified residues converted to their canonical parent and non-RNA molecules removed is stored as Biopython chains in the `CoarseGrainRNA`'s `chains` attribute.

Operation

`forgi`[17] is compatible with Python 2.7, 3.5 and 3.6, should run on all operating systems where its dependencies are available and has been tested on Linux, Mac and Windows. It makes heavy use of the NumPy[23] library to speed-up array-based calculations and also depends on SciPy[24], NetworkX[25], Biopython[12,13], pandas[26,27] and `appdirs`, all available via the Python Package Index (PyPi) or Anaconda.

Helpful utility scripts. `forgi` comes with the two very useful scripts: `rnaConvert.py` can be used to convert between many common file formats for RNA structures, including the Vienna format and fasta variants, the bpseq format, the connectivity table (ct) format, MMCIF format and PDB files. `visualize_rna.py` can be used to display a coarse-grained representation of an RNA's secondary structure in PyMol alongside the all atom structure from PDB files, producing visualizations like those in Figure 3 and Figure 5.

Analysis of stacking geometries. To illustrate how the `forgi` library makes secondary structure elements usable for 3D structure analysis, we used it to analyze the stacking of adjacent helices in multi loops and pseudoknots in a representative set of RNA 3D structures[28] (version 3.36, available at http://rna.bgsu.edu/rna3dhub/nrlist/).

While loading these 3D structures into `forgi`, DSSR[7] (Version v1.7.1-2017nov01) was called to obtain the secondary structure and nucleotide level reference stacking annotations. We count a pair of connected helices as stacking, if at least one nucleotide of the first helix's closing/opening base pair is in a continuous stack with at least one nucleotide of the second helix's opening/closing base pair. This allows for any number of stacking nucleotides between the stems (not necessarily connected via the backbone) and for bulged out nucleotides which do not contribute to the stack. Our definition of stacking is more relaxed than stricter criteria used elsewhere[29].

For each pair of adjacent stems within a junction, we used `forgi` to calculate a number of properties, such as the angle between the stem vectors, the separation vector between the stems' ends and an offset value between the stems.

These properties are based on the axes that were fitted to the helices (see above) and the multi loop segment that connects the two stems. We define the stem vectors $\boldsymbol{v}_h$ as pointing away from the multi loop segment along the helix axis, where $\boldsymbol{s}_h$ and $\boldsymbol{e}_h$ are the helix's start and end coordinates.

$$\boldsymbol{v}_h = \begin{cases} \boldsymbol{e}_h - \boldsymbol{s}_h & \text{if } \boldsymbol{s}_h \text{ is at the side of the multiloop} \\ \boldsymbol{s}_h - \boldsymbol{e}_h & \text{if } \boldsymbol{e}_h \text{ is at the side of the multiloop} \end{cases}$$

The angle between the adjacent stems i and h is then defined as the angle between their stem vectors $\boldsymbol{v}_i$ and $\boldsymbol{v}_h$:

$$\cos(\alpha) = \frac{\boldsymbol{v}_h \cdot \boldsymbol{v}_i}{\|\boldsymbol{v}_h\| \cdot \|\boldsymbol{v}_i\|}$$

The offset between stems is calculated as distance between two rays R_h that start at the helix's end closer to the multi loop segment, $\boldsymbol{c}_h$, and extend the helix axis:

$$\boldsymbol{c}_h = \begin{cases} \boldsymbol{s}_h & \text{if } \boldsymbol{s}_h \text{ is at the side of the multiloop} \\ \boldsymbol{e}_h & \text{if } \boldsymbol{e}_h \text{ is at the side of the multiloop} \end{cases}$$

$$R_h = \left\{ \boldsymbol{X} \mid \boldsymbol{X} = \boldsymbol{c}_h + \lambda \boldsymbol{v}_h, \lambda \geq 0 \right\}$$

Depending on the helix orientation, the distance between these rays is either the distance between two lines or the distance between two points or between one point and one line, all of which can be calculated with standard formulas.

The following code shows how `forgi` was used to calculate these properties for multi loops:

```
from forgi import load_rna
from forgi.threedee.utilities.vector import vec_angle, vec_distance
from forgi.threedee.utilities.vector import line_segment_distance

def main():
    # Load the PDB into CoarseGrainRNA instances
    # rnas is a list, because a PDB can contain multiple connected components
    # This also loads the DSSR json annotations.
    rnas = load_rna("path/to/file.cif", pdb_remove_pk=False,
                    pdb_annotation_tool="DSSR")
    for rna in rnas:
        for junction in rna.junctions:
            print_junction_parameters(rna, junction)

def print_junction_parameters(rna, junction):
    """
    Prints the angles and offsets of a junction.
    :param rna: A BulgeGraph Object
    :param junction: A list of element names (strings)
    """
    for ml in junction:
        # rna.edges holds adjacent nodes in the BulgeGraph
        stem1, stem2 = rna.edges[ml]
        # rna.coords holds the 3D coordinates of structure elements
        # rna.coords["s0"] returns a tuple start-, end-coordinates of stem "s0"
        # Furthermore rna.coords has the get_direction function to give the
        # vector pointing from the start to the end coordinate.
        direction1 = rna.coords.get_direction(stem1)
        direction2 = rna.coords.get_direction(stem2)
        # Get the indices into rna.coords for the stem sides closer to
        # and further away from the multiloop segment ml
        c1, f1 = rna.get_sides(stem1, ml)
        c2, f2 = rna.get_sides(stem2, ml)
        # Make sure the stem vector points away from the multiloop
        if c1 == 1:
            direction1 = - direction1
        if c2 == 1:
            direction2 = - direction2
        angle = vec_angle(direction1, direction2) # imported above
        is_stacking_dssr = (ml in rna.dssr.stacking_loops())
        closer1 = rna.coords[stem1][c1]
```

```
closer2 = rna.coords[stem2][c2]
# Calculate the offset as distance between rays.
offset = vec_distance(*line_segment_distance(
                            closer1, closer1+100000*direction1,
                            closer2, closer2+100000*direction2))
print("{}\t{}\t{}".format(angle, is_stacking_dssr, offset))
```

We then used pandas[26,27], Matplotlib[30] and a custom library (https://github.com/Bernhard10/filterAndView) to analyze and visualize the collected data. The results were collected for different classes of RNA independently. This was necessary, because the representative sets of RNA structures contain, by design, homologs of the same molecule in multiple species.

For the classification of pseudoknots, Reidys' concept[31] based on the definition of the mathematical genus is used. Classes of pseudoknots are defined based on their shadow representations, which contain only crossing base pairs and only one base pair each. On the level of these shadows, only four distinct classes exhibit genus 1, two of which are well known: the H-type pseudoknot and the kissing hairpin. pseudoknots with higher genus contain, among others, the case where a genus 1 pseudoknot is nested within another pseudoknot.

In combination with the `forgi` library, we wrote a tool that is able to convert structures to their shadow representation, identify and classify the pseudoknots within the structure and describe their helix arrangement in the 3D structure. This tool, called `pseudoknot_analyzer.py`, is distributed with the `forgi` library in the folder "examples". We used it to gather statistics about simple H-type and kissing hairpin pseudoknots. Figure 4a, b illustrates how we measured the angle between stems in pseudoknots via vector directions. In kissing hairpins the angles α and β restrict the possible values of the angle γ. We also include the representative structure of an intermolecular kissing hairpin interaction (lacking the green connection in Figure 4b) in our analysis. In this special case α and β are indistinguishable and were assigned arbitrarily.

Results

We used the helix-centered representation of the 3D structure to analyze the geometry of coaxially stacking helices. The relative geometry of two cylinders in 3D space can be described by five parameters: A separation vector (three parameters) and two angles for the relative orientation in 3D space. In Figure 2 and Figure 4, we show the single angle calculated between the vectors along the helix axes, as a proxy for these parameters.

The distribution of angles between adjacent stems in tRNAs multi loops (see Figure 2a) shows the expected bimodal distribution with one mode slightly below 90° and another mode between 160° and 170°, which fits to the known helix arrangement in the L-shaped tRNA. As confirmed by comparison to annotations with DSSR, the second peak is almost exclusively due to stacking helices, even at angles as small as 140°. In Figure 3 we show an example of such a coaxial stack that has been strongly bent (possibly by the tRNA synthetase) without completely breaking the stack or affecting the canonical tRNA secondary structure.

The second family of RNA molecules with lots of data available is ribosomal RNA. Here we find that the angle of adjacent stems in multi loops is almost uniformly distributed above 20° until a peak from 135° to 170°, where DSSR annotates roughly half of the geometries as stacking (see Figure 2c). 7% of the stacking geometries and

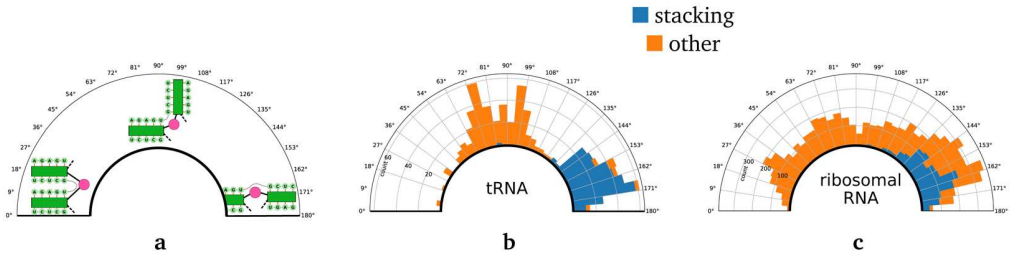


Figure 2. Distribution of angles between adjacent stems in multi loops. (**a**) Angles close to 0° mean parallel stems whereas angles close to 180° correspond to potentially stacking stems. Angle distributions in (**b**) tRNAs and (**c**) ribosomal RNA are shown as a histogram mapped onto a circular plot illustrating the angles. The (inner) blue bars are instances where DSSR detects stacking on the atom-level scale. The orange bars start at the top of the blue bars and indicate geometries where DSSR does not detect stacking.

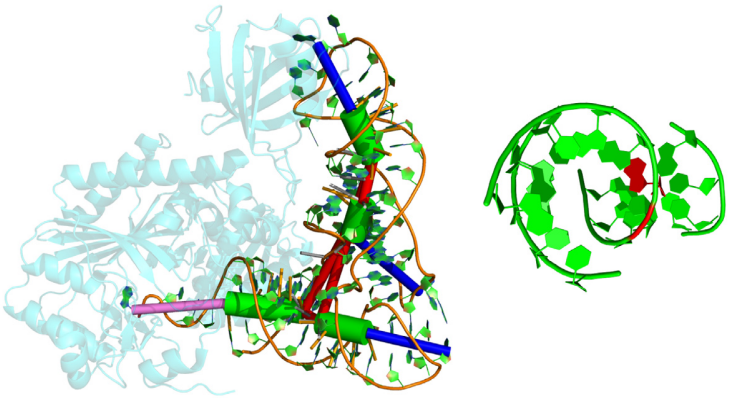


Figure 3. Example of a tRNA where stacking between two non-collinear stems occurs (PDB id 4WJ4[32]). Green cylinders were fitted to stems, blue cylinders represent hairpins and the pink cylinder the unpaired nucleotides at the 3' end. Red cylinders connect stems and indicate single-stranded connections between these stems (multi loop segments). Left: The stems of the anticodon arm (top) and the D-arm stack (according to DSSR) despite being at an angle of 143°. Right: View along the axis of the anticodon arm's stem. The red nucleotide is the unpaired multi loop segment which mediates stacking to the stem of the D-arm. This illustration was generated using PyMol[33] via the `visualize_rna.py` wrapper in `forgi`.

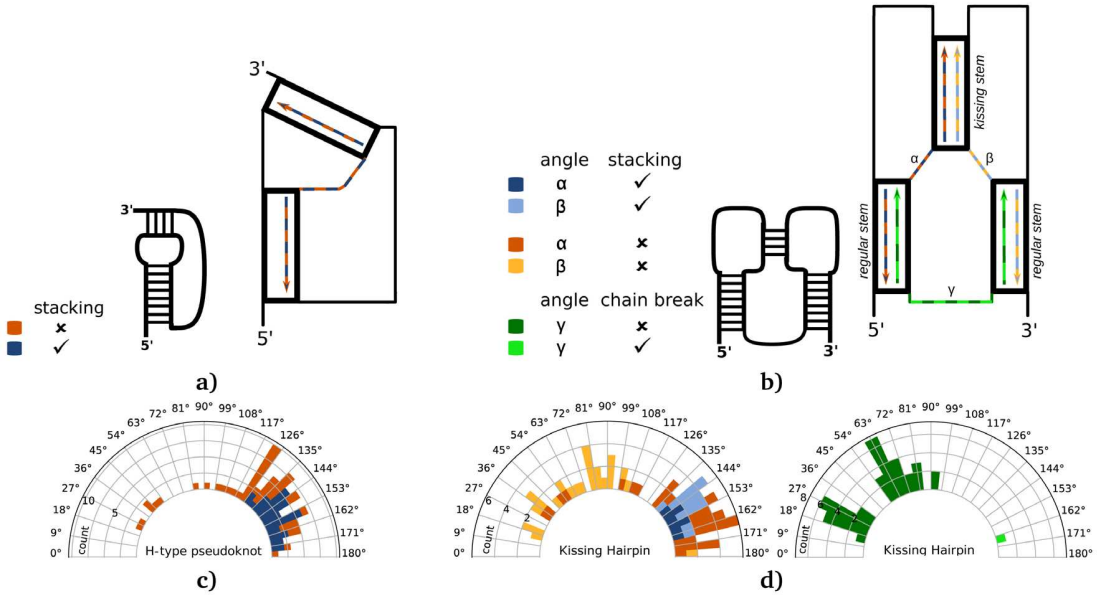


Figure 4. Angles between the stems in pseudoknots. (**a**) and (**b**) show the vector directions used for the angle measurement. The distributions of the angles between adjacent stems building **c**) an H-type pseudoknot or (**d**) a kissing hairpin are shown as histograms mapped to a circle. Furthermore, the distribution of angles between the regular stems of kissing hairpin pseudoknots is shown in the second panel of (**d**). Like mentioned in Figure 2, the angles in **c**) measured between stacking stems (detected via DSSR) are colored in blue. Geometries without stacking are colored in orange. The color scheme in (**d**) (shades of orange, blue and green) refer to the respective vectors/measured angles (α, β, γ) in the same color scheme as in (**b**). Blue bars stand for stacking between the two related stems, whereas orange ones for non-stacking stems. Dark green bars in the second kissing hairpin associated histogram show angles measured between a intramolecular interaction (kissing hairpin pseudoknot), whereas light green bars stand for angles measured between two RNA chains (kissing hairpin interaction).

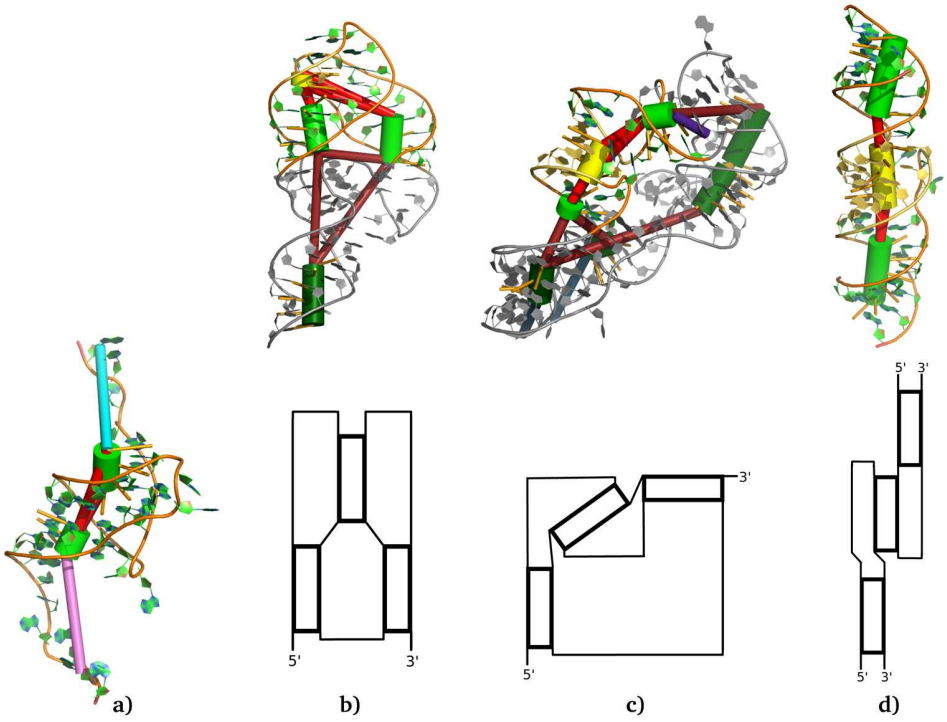


Figure 5. Examples of coaxial stacking within (**a**) an H-type pseudoknot (PDB id 2XD0) and (**b**-**d**) the three major structural families of kissing hairpins (PDB ids 5KPY, 4FRN and 1ZCI, from left to right). In all four representations the green cylinders were fitted to stems, the turquoise and pink cylinder represent the unpaired nucleotides at the 5' and 3' end respectively. Red cylinders connect stems and indicate single-stranded connections between these stems (pseudoknot or multi loop segments). The light-colored stems in (**b**–**d**) represent the regular stems of kissing hairpins, whereas the kissing stem is colored yellow. Note that panel (**d**) (PDB id 1ZCI) shows a kissing hairpin interaction between two RNA chains. Below the 3-dimensional representation of the kissing hairpins, we show a schematic sketch of the structural family's helix arrangement. These illustrations were generated using PyMol[33] via the `visualize_rna.py` wrapper in `forgi`.

67% of the non-stacking geometries with an angle above 140° also have an offset above 10Å between the extended stem axes.

Additionally we analyzed the angle distribution between the stems forming simple H-type pseudoknots or kissing hairpins. Most angles measured within an H-type pseudoknot are between 120° and 180° (see Figure 4c). In this range, 35 out of 67 instances correspond to stacking. One example within this range is shown in Figure 5a, which represents a processed non-coding RNA, which regulates a bacterial antiviral system (PDB id 2XD0[34]).

The distribution illustrated in Figure 4d shows mainly values between 130° and 180° for the angles β and especially α. One additional angle measurement between the two regular stems (see Figure 4d, angle γ) shows one peak at about 20° and one at about 65°. Within the class of kissing hairpins we often find coaxial stacking, but most of the time only one of the two regular stems stacks with the kissing stem in the middle. With the help of the coarse grained representation we were able to divide the class of kissing hairpins into three different structural families (see Figure 5b-d).

The first family is especially common among (A-)riboswitches. Here the two regular stems are oriented almost parallel (γ below to 32°) with the second one (counting from the 5' end) stacking onto the kissing stem. The angles α and β are above 130°. One example is the structure with PDB id 5KPY[35] shown in Figure 5b. Here, the link from the first regular stem stacks onto it and interacts with the kissing stem via base multiplets and A-Minor interactions. This way, the roughly parallel orientation of all 3 stems is stabilized by stacking and base-pairing interactions. Interestingly, the kissing stem is more parallel to the first stem than to the second stem, onto which it stacks directly ($\alpha > \beta$).

The second structural family is shown in Figure 5c. Here the two regular stems are close to 90°, whereas the kissing stem is along the arc between them. This conformation is found in several groups of non coding RNA including some 23S ribosomal RNA and the cobalamin riboswitch regulatory element (PDB id 4FRN[36]) shown in Figure 5c.

The two families as well as some conformations in between are observable within intramolecular pseudoknots. As mentioned, `forgi` is also able to model multiple RNA chains that interact with each other forming an intermolecular complex. One example is the dimerization initiation site of the HIV type 1 (PDB id 1ZCI[37]), illustrated in Figure 5d and being the only example of the third family of kissing hairpin pseudoknots. Here the kissing hairpin interaction shows a nearly perfect coaxial stacking between all three involved stems. In this case β is close to 0°, because stacking takes place at the other side of the kissing stem and thus the complementary angle is close to 180°.

Discussion

Using the `forgi` library[17] we have analyzed the relative orientation of helices in junctions and pseudoknots. While stacking between adjacent helices is common, we found that their orientation often deviates significantly from co-linearity, suggesting that junctions introduce flexibility into the RNA structure while still maintaining the benefit of stacking interactions. Previous studies that assign coaxial geometries based on manual inspection[38] miss this deviation from coaxiality that becomes apparent once you actually calculate the helix axis. Our approach allowed us to perform this analysis on the scale of stems, as opposed to the smaller all-atom scale used by the program DSSR and in previous surveys[39] or the level of networks of (noncanonical) base pairs[38].

Conversely, we found that out of 1257 pairs of adjacent stems in ribosomal RNA structures with an angle above 140°, only roughly half (608) are annotated as stacking by DSSR. This becomes clear if we consider that the angle between stems is only a proxy for a 5-dimensional orientation parameter. In particular, a large offset means that stem axes with an angle close to 180° can be parallel without being coaxially stacked. Indeed, it is not uncommon that all three stems in a 3-way junction are roughly parallel, with two stems forming a coaxial stack and the third having a higher offset[40].

There is growing interest in predicting pseudoknots in RNA structures as they are involved in a variety of biological functions[41]. The `forgi` library allows us to easily identify pseudoknots in RNA 3D structures and gather statistics on the frequency of pseudoknot types, sizes, and composition. Kissing hairpins formed between two RNA molecules are known to often form continuous stacks between all three helices, such as the pseudoknot in Figure 5d. In contrast to these intermolecular kissing hairpins, we find that stacking between all three helices is seldom possible in intramolecular pseudoknots. Instead we find that the limited length of all loop segments makes it nearly impossible to form coaxial stacking of all three stems in one line in a single RNA chain. Thus, in the main families of kissing hairpin conformations, the regular (outer) stems are oriented parallelly or almost perpendicular with the kissing helix stacking onto at most one of the two regular stems. These structures are often further stabilized by base triplets like those found in Figure 5b.

Similar to stacking conformations in multi loops, many stacking helices in pseudoknots form angles below 180°. A deviation of the coaxiality between stems can have multiple reasons such as a higher flexibility of short helices within a pseudoknot as well as strain from short stem connections. But there are also structures that show more classical coaxial stacking like PDB id 2XD0, see Figure 5a.

Results and challenges of the application of `forgi` to the analysis of pseudoknots are described in more detail in the thesis by Beckmann[42].

Varying additional parameters like the minimal number of base pairs required to count an interaction as a helix allow for a better understanding about the principles of the formation of three-dimensional RNA structures. The insight in the world of pseudoknots and multi loops presented here can support the improvement of the prediction of RNA structures and help identify unrealistic multi-loop conformations predicted by current RNA structures prediction tools. We are now in the process of implementing additional features for the RNA 3D structure prediction program Ernwin[20] based on our findings about stacking and pseudoknots.

Conclusions

`forgi`[17] is a multi-purpose Python library for dealing with RNA on the levels of sequence, secondary structure and tertiary structure. By providing our code as a Python library, we give users of our tool more flexibility than a single executable program could provide. Furthermore, our code is fully open source, giving researchers the possibility to comprehend and where needed extend the inner workings of the code.

`forgi` is well documented, with a tutorial available at https://ViennaRNA.github.io/forgi/ and a complete API documentation following the Python standard of docstrings parsable by Sphinx. It contains an extensive automatic test suite (unit and integration tests) and has been optimized for speed, maintainability and usefulness.

Data availability

Underlying data

The PDB ids analyzed were taken from version 3.36 of the representative sets of RNA 3D structures[28] at a cutoff of 4 Å, http://rna.bgsu.edu/rna3dhub/nrlist/release/3.36/4.0A.

Extended data

Open Science Framework: 3D based on 2D: `forgi` 2.0 Extended Data. https://doi.org/10.17605/OSF.IO/HDJRU[43]. This project contains the following extended data files:

- **assignment_of_classes_to_pdbids.csv**: Assignment of PDB ids to RNA type. These annotations were generated in a semi-manual way.
- **multiloop_angles.csv**: Angles and offsets measured in PDB structures for multi loops, generated with the script `describe_rna.py`, which is available in the examples folder of `forgi`.
- **pseudoknot_angles.tsv**: Angles measured in pseudoknots, generated with `pseudoknot_analyzer.py`, which is available in the examples folder of `forgi`.

Software availability

Software available from: https://pypi.org/project/forgi/ (`pip install forgi`) and Bioconda.

Archived source code at time of publication: https://doi.org/10.5281/zenodo.2582870[17].

License: GNU General Public License 3.0.

Grant information

This work was partly funded by the Austrian science fund (FWF) projects F 43 "RNA regulation of the transcriptome" and I 2874 "Prediction of RNA-RNA interactions" and Doctoral College W 1207 "RNA Biology".

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

We thank all contributors, who submitted a pull-request or filed an issue in the `forgi` repository on github. We thank Roman Ochsenreiter for proofreading, Gregor Entzian for useful inputs about many quirks in the PDB format and for proofreading and Lorena Pantano for creating the first version of a Bioconda recipe for `forgi`.

References

1. Dawson WK, Maciejczyk M, Jankowska EJ, *et al.*: **Coarse-grained modeling of RNA 3D structure.** *Methods.* 2016; **103**: 138–156.
PubMed Abstract | Publisher Full Text
2. Berman H, Henrick K, Nakamura H, *et al.*: **The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data.** *Nucleic Acids Res.* 2007; **35**(Database issue): D301–D303.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Leontis NB, Westhof E: **Geometric nomenclature and classification of RNA base pairs.** *RNA.* 2001; **7**(4): 499–512.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Gendron P, Lemieux S, Major F: **Quantitative analysis of nucleic acid three-dimensional structures.** *J Mol Biol.* 2001; **308**(5): 919–936.
PubMed Abstract | Publisher Full Text
5. Yang H, Jossinet F, Leontis N, *et al.*: **Tools for the automatic identification and classification of RNA base pairs.** *Nucleic Acids Res.* 2003; **31**(13): 3450–3460.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Sarver M, Zirbel CL, Stombaugh J, *et al.*: **FR3D: finding local and composite recurrent structural motifs in RNA 3D structures.**

J Math Biol. 2008; **56**(1–2): 215–252.
PubMed Abstract | Publisher Full Text | Free Full Text

7. Lu XJ, Bussemake HJ, Olson WK: **DSSR: an integrated software tool for dissecting the spatial structure of RNA.** *Nucleic Acids Res.* 2015; **43**(21): e142.
PubMed Abstract | Publisher Full Text | Free Full Text

8. Zok T, Antczak M, Zurkowski M, *et al.*: **RNApdbee 2.0: multifunctional tool for RNA structure annotation.** *Nucleic Acids Res.* 2018; **46**(W1): W30–W35.
PubMed Abstract | Publisher Full Text | Free Full Text

9. Mustoe AM, Brooks CL, Al-Hashimi HM: **Hierarchy of RNA functional dynamics.** *Annu Rev Biochem.* 2014; **83**(1): 441–466.
PubMed Abstract | Publisher Full Text | Free Full Text

10. Thiel BC, Flamm C, Hofacker IL: **RNA structure prediction: from 2D to 3D.** *Emerg Top Life Sci.* 2017; **1**(3): 275–285.
Publisher Full Text

11. Lorenz R, Bernhart SH, Höner Zu Siederdissen C, *et al.*: **ViennaRNA Package 2.0.** *Algorithms Mol Biol.* 2011; **6**(1): 26.
PubMed Abstract | Publisher Full Text | Free Full Text

12. Cock PJ, Antao T, Chang JT, *et al.*: **Biopython: freely available Python tools for computational molecular biology and bioinformatics.** *Bioinformatics.* 2009; **25**(11): 1422–1423.
PubMed Abstract | Publisher Full Text | Free Full Text

13. Hamelryck T, Manderick B: **PDB file parser and structure class implemented in Python.** *Bioinformatics.* 2003; **19**(17): 2308–2310.
PubMed Abstract | Publisher Full Text

14. Knight R, Maxwell P, Birmingham A, *et al.*: **PyCogent: a toolkit for making sense from sequence.** *Genome Biol.* 2007; **8**(8): R171.
PubMed Abstract | Publisher Full Text | Free Full Text

15. Smit S, Rother K, Heringa J, *et al.*: **From knotted to nested RNA structures: a variety of computational methods for pseudoknot removal.** *RNA.* 2008; **14**(3): 410–416.
PubMed Abstract | Publisher Full Text | Free Full Text

16. Rother M, Rother K, Puton T, *et al.*: **ModeRNA: a tool for comparative modeling of RNA 3D structure.** *Nucleic Acids Res.* 2011; **39**(10): 4007–4022.
PubMed Abstract | Publisher Full Text | Free Full Text

17. Thiel B, Kerpedjiev P, tcarlile, *et al.*: **Viennarna/forgi: Forgi version 2.0**. 2019.
http://www.doi.org/10.5281/zenodo.2582870

18. Licht K, Jantsch MF: **Rapid and dynamic transcriptome regulation by RNA editing and RNA modifications.** *J Cell Biol.* 2016; **213**(1): 15–22.
PubMed Abstract | Publisher Full Text | Free Full Text

19. Dimitropoulos D, Ionides J, Henrick K: **Using MSDchem to search the PDB ligand dictionary.** *Curr Protoc Bioinformatics.* 2006; Chapter 14: Unit14.3.
PubMed Abstract | Publisher Full Text

20. Kerpedjiev P, Höner Zu Siederdissen C, Hofacker IL: **Predicting RNA 3D structure using a coarse-grain helix-centered model.** *RNA.* 2015; **21**(6): 1110–1121.
PubMed Abstract | Publisher Full Text | Free Full Text

21. Schlick T: **Adventures with RNA graphs.** *Methods.* 2018; **143**: 16–33.
PubMed Abstract | Publisher Full Text | Free Full Text

22. Steffen P, Voss B, Rehmsmeier M, *et al.*: **RNAshapes: an integrated RNA analysis package based on abstract shapes.** *Bioinformatics.* 2006; **22**(4): 500–503.
PubMed Abstract | Publisher Full Text

23. Oliphant TE: **Guide to NumPy**. 2nd Edition. CreateSpace Independent Publishing Platform, 2015.
Reference Source

24. Jones E, Oliphant T, Peterson T, *et al.*: **SciPy: Open source scientific tools for Python, since 2001**. [online, accessed 8 Oct 2018].
Reference Source

25. Hagberg AA, Schult DA, Swart PJ: **Exploring network structure, dynamics, and function using networkx**. In Gaël Varoquaux, Travis Vaught, and Jarrod Millman, editors, *Proceedings of the 7th Python in Science Conference*. Pasadena, CA USA. 2008; 11–15.
Reference Source

26. McKinney W: **Data structures for statistical computing in python**. In Stéfan van der Walt and Jarrod Millman, editors, *Proceedings of the 9th Python in Science Conference.* 2010; 51–56.
Reference Source

27. McKinney W: **pandas: a foundational python library for data analysis and statistics**. *Python for High Performance and Scientific Computing.* 2011; 1–9.
Reference Source

28. Leontis NB, Zirbel CL: **Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking**. In *Nucleic Acids and Molecular Biology*. Springer Berlin Heidelberg. 2012; 281–298.
Publisher Full Text

29. Tyagi R, Mathews DH: **Predicting helical coaxial stacking in RNA multibranch loops.** *RNA.* 2007; **13**(7): 939–951.
PubMed Abstract | Publisher Full Text | Free Full Text

30. Hunter JD: **Matplotlib: A 2D graphics environment.** *Comput Sci Eng.* 2007; **9**(3): 90–95.
Publisher Full Text

31. Reidys CM, Huang FW, Andersen JE, *et al.*: **Topology and prediction of RNA pseudoknots.** *Bioinformatics.* 2011; **27**(8): 1076–1085.
PubMed Abstract | Publisher Full Text

32. Suzuki T, Nakamura A, Kato K, *et al.*: **Structure of the *Pseudomonas aeruginosa* transamidosome reveals unique aspects of bacterial tRNA-dependent asparagine biosynthesis.** *Proc Natl Acad Sci U S A.* 2015; **112**(2): 382–387.
PubMed Abstract | Publisher Full Text | Free Full Text

33. Schrödinger, LLC: **The PyMOL molecular graphics system.** version 1.8. 2015.

34. Blower TR, Pei XY, Short FL, *et al.*: **A processed noncoding RNA regulates an altruistic bacterial antiviral system.** *Nat Struct Mol Biol.* 2011; **18**(2): 185–191.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Porter EB, Polaski JT, Morck MM, *et al.*: **Recurrent RNA motifs as scaffolds for genetically encodable small-molecule biosensors.** *Nat Chem Biol.* 2017; **13**(3): 295–301.
PubMed Abstract | Publisher Full Text | Free Full Text

36. Johnson JE Jr, Reyes FE, Polaski JT, *et al.*: **B_{12} cofactors directly stabilize an mRNA regulatory switch** *Nature.* 2012; **492**(7427): 133–137.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Ennifar E, Dumas P: **Polymorphism of bulged-out residues in HIV-1 RNA DIS kissing complex and structure comparison with solution studies.** *J Mol Biol.* 2006; **356**(3): 771–782.
PubMed Abstract | Publisher Full Text

38. Laing C, Schlick T: **Analysis of four-way junctions in RNA structures.** *J Mol Biol.* 2009; **390**(3): 547–559.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Holbrook SR: **Structural principles from large RNAs.** *Annu Rev Biophys.* 2008; **37**(1): 445–464.
PubMed Abstract | Publisher Full Text

40. Lescoute A, Westhof E: **Topology of three-way junctions in folded RNAs.** *RNA.* 2006; **12**(1): 83–93.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Staple DW, Butcher SE: **Pseudoknots: RNA structures with diverse functions.** *PLoS Biol.* 2005; **3**(6): e213.
PubMed Abstract | Publisher Full Text | Free Full Text

42. Beckmann IK: **Identification and Classification of Pseudoknots and their Impact on RNA 3D Structure Prediction**. Master's thesis, University of Vienna, 2018.
Reference Source

43. Thiel BC, Beckmann IK, Kerpedjiev P, *et al.*: **3D based on 2D: Forgi 2.0 Extended Data**. 2019.
https://www.doi.org/10.17605/OSF.IO/HDJRU

2. 3D feasibility of 2D RNA–RNA interaction paths by stepwise folding simulations

Beckmann IK, Waldl M, Will S and Hofacker IL. 3D feasibility of 2D RNA–RNA interaction paths by stepwise folding simulations. *RNA* 30.2 (2023), pp. 113–123. DOI: 10.1261/rna.079756.123

Summary

Intra- and intermolecular interactions are essential for RNA functionality. 2D structure prediction tools often overpredict interactions because they do not take into account the steric hindrances and kinetic inaccessibility underlying the 3D structure. In addition, many interaction approaches do not, or only to a limited extent, allow kissing hairpin interactions in their models. This is also due to the fact that there are only a limited number of high-resolution kissing hairpin interactions (pathways) available, which in turn means that there is only a limited and biased 2D rule set. The developed RNA-RNA interaction pipeline allows 2D interaction pathways to be evaluated for steric feasibility and kinetic accessibility by stepwise 3D structure simulation. These small conformational changes allow to identify how long interactions can be formed, how steric hindrance can obstruct the folding path, and finally to ensure that the final state is kinetically accessible. Furthermore, the pipeline allows the construction of a set of 3D proven constraints for 2D structure prediction methods.

Authors contribution

I (IKB), as the main developer, was significantly involved in the design and implementation of the pipeline as well as in the experimental design and the subsequent analysis and interpretation of the data. I was also responsible for drafting the manuscript. Overall, as first author, I managed the project with the utmost care and expertise, ensured the quality of the final publication and thus made a significant and formative contribution to this study.

Software availability

www.github.com/irenekb/RRI-3D

License

3D feasibility of 2D RNA–RNA interaction paths by stepwise folding simulations

IRENE K. BECKMANN,[1,2] MARIA WALDL,[1,3,4] SEBASTIAN WILL,[5] and IVO L. HOFACKER[1,6]

[1]Department of Theoretical Chemistry, Faculty of Chemistry, University of Vienna, 1090 Wien, Austria
[2]Vienna BioCenter PhD Program, Doctoral School of the University of Vienna and Medical University of Vienna, A-1030 Vienna, Austria
[3]Vienna Doctoral School in Chemistry (DoSChem), University of Vienna, 1090 Vienna, Austria
[4]Center for Anatomy and Cell Biology, Medical University of Vienna, 1090 Vienna, Austria
[5]LIX - Batiment Turing, Ecole Polytechnique, 91120 Palaiseau, France
[6]Faculty of Computer Science, Research Group Bioinformatics and Computational Biology, University of Vienna, 1090 Vienna, Austria

ABSTRACT

The structure of an RNA, and even more so its interactions with other RNAs, provide valuable information about its function. Secondary structure-based tools for RNA–RNA interaction predictions provide a quick way to identify possible interaction targets and structures. However, these tools ignore the effect of steric hindrance on the tertiary (3D) structure level, and do not consider whether a suitable folding pathway exists to form the interaction. As a consequence, these tools often predict interactions that are unrealistically long and could be formed (in three dimensions) only by going through highly entangled intermediates. Here, we present a computational pipeline to assess whether a proposed secondary (2D) structure interaction is sterically feasible and reachable along a plausible folding pathway. To this end, we simulate the folding of a series of 3D structures along a given 2D folding path. To avoid the complexity of large-scale atomic resolution simulations, our pipeline uses coarse-grained 3D modeling and breaks up the folding path into small steps, each corresponding to the extension of the interaction by 1 or 2 bp. We apply our pipeline to analyze RNA–RNA interaction formation for three selected RNA–RNA complexes. We find that kissing hairpins, in contrast to interactions in the exterior loop, are difficult to extend and tend to get stuck at an interaction length of 6 bp. Our tool, including source code, documentation, and sample data, is available at www.github.com/irenekb/RRI-3D.

Keywords: RNA–RNA interaction; folding pathways; steric feasibility; coarse-grained folding simulation

INTRODUCTION

A large part of the transcripts in organisms are noncoding RNAs (ncRNAs). They perform diverse functions as regulatory elements for essential cellular processes in all domains of life (Shabalina and Koonin 2008; Waters and Storz 2009). These RNA functions crucially depend on their structure as well as their interaction with other RNAs. Experimental studies of kinetics and folding pathways for RNA–RNA interactions are challenging and time consuming. Therefore, computational tools that model the formation of RNA–RNA interactions offer an appealing alternative for studying these interactions.

In many cases, the level of secondary (2D) structure is sufficient to understand the function of an RNA. Computationally efficient tools for predicting reasonably accurate RNA 2D structures, such as are available in the ViennaRNA Package (Lorenz et al. 2011), are widely used. Further, there are extensions to predict RNA–RNA interactions like RNAup (Mückstein et al. 2006), RNAcofold (Bernhart et al. 2006), pairfold (Andronescu et al. 2005), NUPACK (Fornace et al. 2020), AccessFold (DiChiacchio et al. 2015), or IntaRNA (Mann et al. 2017). These prediction tools silently assume that any 2D structure can also be realized in three dimensions (3D). While uncritical for noncrossing single molecule structures, this assumption is no longer valid for RNA–RNA interactions. Consequently, 2D interaction prediction tools tend to "over-predict" interactions, disregarding that long interactions are often sterically infeasible. Even worse, they are often kinetically inaccessible, since the pathway toward the full interaction will be obstructed by steric effects. This lack of steric and kinetic

Corresponding author: ivo@tbi.univie.ac.at
Handling editor: Eric Westhof
Article is online at http://www.rnajournal.org/cgi/doi/10.1261/rna.079756.123. Freely available online through the *RNA* Open Access option.

considerations compromises the accuracy of conventional tools.

RNA 3D structure prediction remains a challenging and computationally demanding problem. However, available tools are well able to model smaller structural changes and to test whether 2D RNA–RNA interactions structure motifs are sterically feasible in 3D. We therefore designed a pipeline based on 3D prediction tools that breaks up the simulation into small steps to shorten the simulation time. In the stepwise extension setup, the required conformational change is small, making computations more efficient and less sensitive to the limitations of 3D modeling tools. The approach mimics the formation pathway of RNA–RNA interactions and therefore tests whether the final state as well as intermediaries are sterically feasible, thus assuring that the final state is kinetically reachable. If the extension process stops due to steric hindrance, we can identify how long interactions can be formed (which is often shorter than predicted by 2D tools). Applying the pipeline to several interaction examples leads to insights into the types of feasible interaction structures, which in turn can be incorporated as constraints in 2D structure prediction methods or used as a postprocessing filter. Ideally, this will facilitate efficient larger scale screens on a 2D level, without sacrificing steric feasibility. In addition, the developed pipeline with all its features allows for the investigation of specific interactions and different scenarios of interaction formation in depth.

Concretely, we selected three known RNA–RNA interaction systems, where at least some experimental data are available for further study: CopA–CopT, a well-studied antisense interaction (Kolb et al.2000a,b); the HIV-1 dimerization initiation site (DIS) (Ennifar and Dumas 2006), for which a 3D structure is available; and the DsrA–*rpoS* small RNA–mRNA interaction (Wu et al. 2017).

CopA–CopT. CopA is an antisense RNA that binds to its target CopT, which is part of the leader region of the *repA* mRNA, for plasmid replication control. The CopA–CopT interaction indirectly regulates the plasmid R1 replication by inhibition of the *RepA* translation. The CopA and CopT fragments used for our simulations and their initial binding at their complementary hairpins are shown in Figure 1A. On the CopT side, the hairpin contains a YUNR motif that induces a U-turn structure that promotes the initial binding (Franch et al. 1999). This system consists of two almost perfectly complementary chains (Supplemental Fig. S1). Although 2D prediction tools, without hesitation, predict the full-length duplex as the thermodynamically most stable state, it is not at all obvious (see Kolb et al. 2000b) that this state is very accessible starting from an initial (kissing hairpin) seed contact of the RNAs.

HIV-1 DIS. The HIV-1 dimerization initialization site is a strongly conserved feature in the 5′-untranslated region of the viral genome. Here, a slightly different experimental setup was used, because X-ray 3D structures are available. The 1ZCI (subtype F of the HIV-1 DIS) PDB structure shows a homodimer with each chain forming a 7 bp helix enclosing a 9 nt hairpin (Fig. 1D). The two hairpins interact to form a 6 bp intermolecular helix. Even though the monomers are not perfectly self-complementary, 2D interaction prediction tools will predict an extended interaction, interrupted by two 2 × 1 interior loops, spanning the full length of the 23 nt fragment. For our simulation, we chose the 1ZCI structure as the starting point and tested whether a refolding into the extended interaction is possible.

DsrA–*rpoS*. As a more complex example, we model the DsrA (downstream of rcsA)–*rpoS* interaction. DsrA is an Hfq-dependent small regulatory RNA in *Escherichia coli*. One of its targets is the *rpoS* mRNA, whose translation is up-regulated upon interaction with DsrA. Different structures have been proposed for DsrA, but all include stable stem–loop structures SL1 and SL3 at the 5′ and 3′ end. For the linker region (LR) and stem–loop SL2, we consider three different structures corresponding to Fold-A and Fold-B from Wu et al. (2017) (Fig. 1B), as well as the consensus structure from Rfam (Fig. 1C; Kalvari et al. 2020). The ViennaRNA package (Lorenz et al. 2011) predicts the Rfam structure as minimum free energy (MFE), followed by Fold-B as a near optimal alternative structure. Since Wu et al. (2017) suggest that DsrA needs to refold from Fold-A to Fold-B in order to interact with *rpoS*, we use this model to analyze how interaction formation depends on the starting structure and start site. To simplify the model and since little is known about the structure of *rpoS*, we model it as a 41 nt unstructured RNA.

3D modeling tools

Our approach is characterized by the use of coarse-grained 3D prediction and stepwise 3D simulation. To enable the 3D embedding of 2D interaction paths in reasonable computation time, we make use of two (comparatively fast) 3D modeling tools. One of them, Ernwin (Kerpedjiev et al. 2015b), is used to derive initial 3D structure sketches, while the other, SimRNA (Boniecki et al. 2015), is used to refine 3D structures and simulate folding paths. Crucial for the feasibility of our entire approach, both of these tools feature structure abstraction and coarse-graining (at different levels). Furthermore, we perform folding simulations only in iteration-limited steps to explore local structural neighborhoods. This keeps execution times low while improving control and reducing dependence on the exact details of SimRNA's coarse-grained simulation.

Ernwin generates 3D structures for a given secondary structure, applying a fragment assembly strategy on the level of loops and helices. It thus achieves very fast sampling, making it well suited to quickly sample candidate 3D structures in our approach.

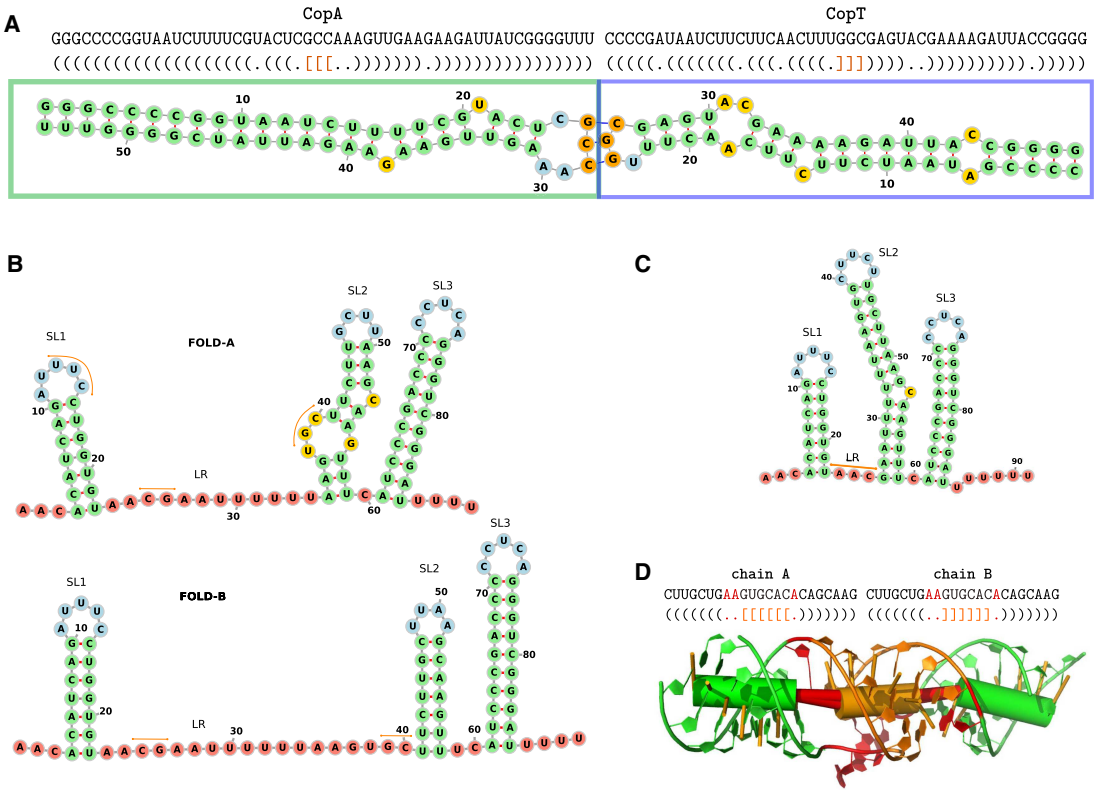


FIGURE 1. (*A*) Sequence and secondary structure representation of the 54 nt part of CopA (green box) and the corresponding 47 nt CopT (blue box) counterpart used in our simulation study. The three central residues of the hairpins (CCG/GGC) are the initial interaction site (marked in orange). (*B*) Fold-A and Fold-B of DsrA from Wu et al. (2017) in 2D structure representation. We label the stem–loops SL1–SL3 and the LR between SL1 and SL2. Fold-B opens the entire region from SL1 up to C40. (*C*) 2D DsrA consensus structure representation based on Rfam and predicted by the ViennaRNA package. The orange lines in (*B*) and (*C*) mark potential initial sites used later. (*D*) Sequence and dot-bracket representation of the HIV-1 DIS stem–loop interaction with the 6 nt kissing hairpin interaction marked in orange, and the conserved, unbounded, noncomplimentary nucleotides marked in red. The 3D representation below (visualized with PyMOL [Schrödinger 2022]) represents both: the individual nucleotides and the associated secondary structure in Ernwin-style, represented as cylinders. Thus, the green cylinders stand for the respective intramolecular stem in each chain, the orange one for the interaction, and the small red ones for the unbounded nucleotides. All 2D representations in this figure are generated with forna (Kerpedjiev et al. 2015a).

SimRNA combines a knowledge-based potential with Monte-Carlo simulation to sample slightly coarse-grained structures. In contrast to Ernwin, it can predict completely novel structures, since it is not restricted to a fragment library. Both tools allow for translating their coarse-grained models back to atomic resolution. In the case of the Ernwin model, these back-translated models often contain local gaps and clashes that can be improved by a refinement step using SimRNA. SimRNA does not require a fixed secondary structure as input; however—essential for our stepwise approach—one can add (soft) constraints that steer the simulation toward a secondary structure. In particular, we use these soft constraints to specify the desired interaction base pairs. Soft constraints are implemented as an energy penalty in SimRNA, making all conformations that violate the constraint less likely. If a desired 2D structure is sterically impossible, SimRNA should not be able to fulfill the constraints. Thus, failure to fulfill the constraints is at least an indicator that a desired 2D structure cannot be embedded in 3D.

RESULTS

Pipeline

Our pipeline starts from a given path for RNA–RNA interaction formation in 2D. Such a path is a sequence of secondary structures, starting with an initial interaction, where each structure introduces a small step toward a target interaction (see, e.g., Supplemental Fig. S1). Then, it systematically attempts to find possible 3D explanations of the given path. Recall that not all 2D paths are expected to have (energetically and kinetically favorable) 3D support, such that the pipeline essentially tests 2D hypotheses. Operationally, the pipeline breaks up the computation into steps, each corresponding to one step of the 2D

pathway. This avoids performing simulation of large structural changes which are computationally expensive and unreliable. We describe the general working mechanism along with various options that allow the pipeline to cover diverse concrete application scenarios.

To define the 2D path, one can either specify every 2D structure explicitly or let the path be generated systematically by extending an initial interaction toward its maximal bulge-free extensions or toward a given target structure; the latter possibly including bulges. The automatic extension can follow different schemes, alternatingly adding base pairs to the left and right or simultaneously adding a base pair to the left and the right of the interaction. To allow opening and refolding of the intramolecular structure, automatic path generation can moreover keep a spacer around the growing interaction site free from intramolecular base pair constraints. The structures along the 2D pathways are later used as soft constrains in the SimRNA simulations.

As shown in Figure 2, our pipeline is divided into the generation of a start structure (*s1*)–(*s4*) and the stepwise extension of the interaction site (*e1*)–(*e3*):

1. Generation of a start structure

If a 3D start structure is provided (in PDB format), this step is omitted. Otherwise, 3D start structures are generated based on the sequence and an initial secondary structure.

(*s1*) A pool of coarse-grained start conformations is modeled using Ernwin.

(*s2*) The sampled conformations are clustered into n_{cluster} clusters based on their Ernwin fragments used. From each cluster, a representative is selected as the structure with the best energy.

(*s3*) For each of the n_{cluster} representatives, n_{run} short SimRNA simulations are performed to refine the structure, ensuring that gaps are closed and clashes are resolved. Constraints are used to ensure that the secondary structure is conserved during this refinement.

(*s4*) From each of the $n_{\text{cluster}} \times n_{\text{run}}$ SimRNA simulations, we select one structure that must (i) exactly match the specified initial 2D structure, and (ii) has the lowest SimRNA energy. These selected structures serve as start points for the stepwise extension.

2. Stepwise extension

We now perform a series of stepwise extensions toward the full-length interaction, as described below. This procedure is performed independently for each of the start structures generated above.

(*e1*) n_{sim} short parallel SimRNA simulations (each of length n_{step}) are performed with a secondary structure

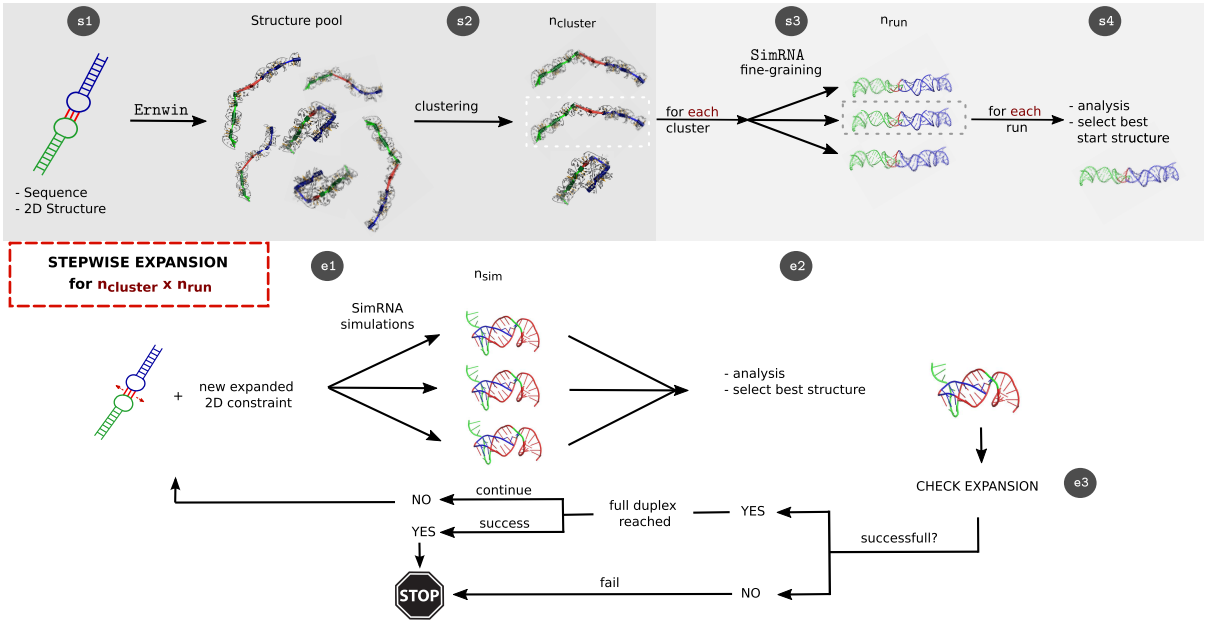


FIGURE 2. Graphical representation of the developed pipeline: An initial interaction is extended until the full duplex, a predefined target interaction, is reached or an extension of the interaction side is no longer possible due to steric and kinetic effects. (*s1*) Starting from a sequence and the corresponding secondary structure, 3D models are sampled using Ernwin and subsequently clustered (*s2*), yielding one representative per cluster. (*s3*) Each of these cluster representatives is relaxed in n_{run}-independent SimRNA simulations. (*s4*) Each simulation is analyzed and a best structure is selected. Each of these $n_{\text{cluster}} \times n_{\text{run}}$ structures forms the start point for a stepwise expansion. In the first step, (*e1*) n_{sim} parallel, constrained SimRNA runs are performed. (*e2*) From the pooled structures the start structure for the next extension step is elected based on hierarchic 2D and 3D criteria. At the checkpoint (*e3*), we test whether the interaction region could indeed be expanded and terminate unsuccessful runs.

constraint designed to expand the interacting region. Structures from all n_{sim} simulations are pooled together.

(*e2*) From this pool of structures, the start structure for the next iteration is selected in a hierarchical fashion. To this end, we compare the secondary structure of the SimRNA model with the desired structure as defined by the constraints used in (*e1*) via the base pair distance. The selection is done by (i) comparing only the interaction region, (ii) considering the structure of the whole complex, and (iii) using the SimRNA energy to break ties.

(*e3*) If for a simulation the interaction does not expand despite the forcing constraint potential, we conclude that the maximum sterically possible interaction has been reached for this specific run and it stops. If the desired target interaction has been reached after an extension step, the simulation terminates with success. Otherwise, we continue with the next step of our 2D pathway at (*e1*).

The raw output of our pipeline is a large number of SimRNA trajectories that are available for further analysis. In order to facilitate the analysis of the large number of sampled 3D structures, each 3D structure is translated back into a 2D structure, i.e., a list of base pairs, including noncanonical base pairs recognized by the SimRNA_trafl2pdb tool. The main output of the RRI-3D pipeline are tab-separated files with statistics on these 2D structures. A more detailed description of pipeline output is available in the supplemental List 1 and the GitHub project page at www.github.com/irenekb/RRI-3D.

Simulation results for model systems

CopA–CopT

We used the start conformation proposed in Kolb et al. (2000b), consisting of the two stem structures with three defined interaction base pairs as interaction start. From the initial Ernwin simulations, we derived $n_{cluster} = 10$ clusters and started the extension simulation with n_{run} and $n_{sim} = 5$. The stepwise extension part of the pipeline (*e1–e3*) was repeated three times with different n_{step} (5000, 10,000, 100,000). In the course of the stepwise extension, both intra- and intermolecular pairs from the corresponding structure in the 2D path (Fig. 3A; Supplemental Fig.

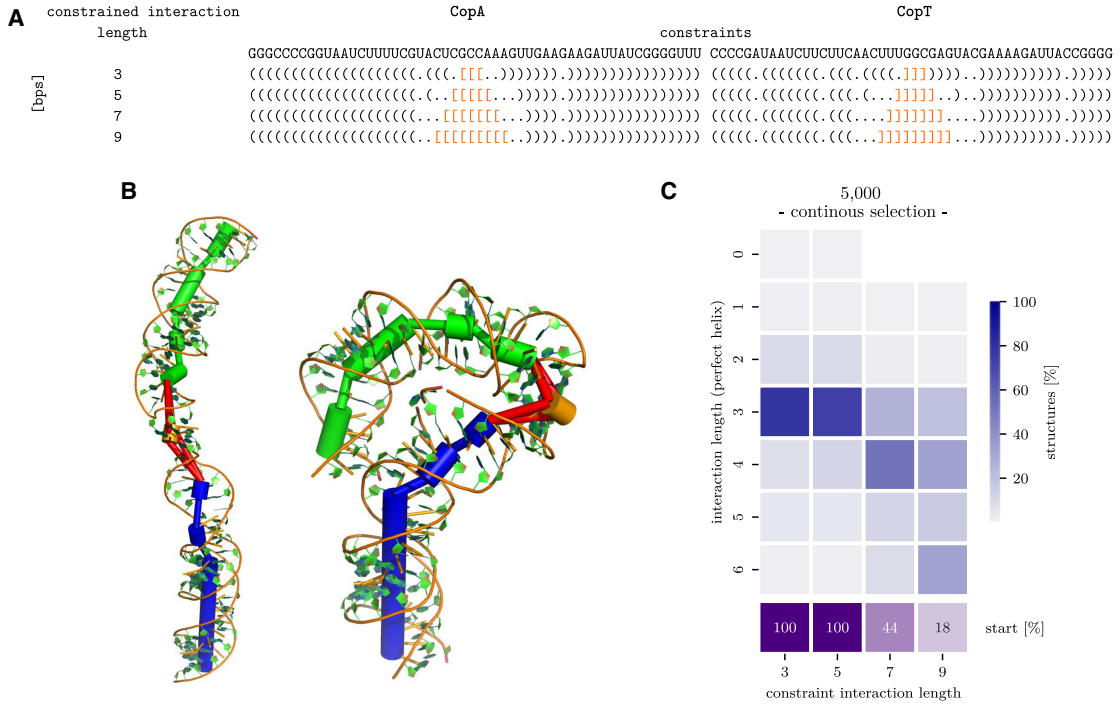


FIGURE 3. (*A*) CopA–CopT sequence and the dot-bracket representation (interaction in orange) of the constrained 2D path (full 2D path in Supplemental Fig. S1). (*B*) Two representative starting clusters with very different 3D conformations. CopA is shown in green, CopT in blue, and the interaction region in orange. (*C*) Summary of simulation results for $n_{step} = 5000$: Each column of the heatmap represents one extension step (defined by the length of the interaction constraint). Each column shows the histogram of observed interaction lengths (blue color gradient). In the first two extension steps, most structures retain the initial interaction length of 3 bp; 6 bp interactions become prevalent in the last extension step. The last row (purple) reports, for every extension step, the percentage of the (originally) 50 runs that successfully extended in all previous steps (i.e., passed checkpoint [*e3*]) and are still active. Thus, in the second extension step (constraint length 5), only 44% of runs manage to extend the interaction and proceed to the next extension step (7 bp). Only 18% of the runs reach the final extension step (9 bp), and none of these extends even further.

S1) were used as SimRNA soft constraints as well as in the selection process. Adding intramolecular constraints is helpful to avoid helix ends opening during the simulation and focuses on exploring the dynamics of the interaction region. Additionally, two different stop criteria (*e3*) were compared. In the first setting (I), the pipeline stops as soon as the simulation was not able to further extend the interaction as a perfect helix. In the second setting (II), the pipeline continues if the interaction region can be extended, even at the expense of interruptions such as bulges or small interior loops. The latter procedure allows flexibility within the interacting region and thus can proceed to longer interactions. Detailed descriptions of the simulation parameters, including the settings for the examples presented here, are available on the RRI-3D website. All starting structures contained an initial interaction of 3 bp; however, the structures of the 10 clusters generated by Ernwin differed considerably (see, e.g., Fig. 3B). Especially for short SimRNA simulations (n_{step} = 5000, 10,000), the starting cluster strongly affected the ability to extend the interaction site. While for some start clusters a 6 bp interaction formed spontaneously within the very first SimRNA simulation, others did not reach the 6 bp stage at all unless n_{step} was increased. Results for shortest (n_{step} = 5000) SimRNA simulations are shown in Figure 3C and Supplemental Figure S2A.

Pipeline runs with n_{step} = 5000 or 10,000 SimRNA steps hardly differ (Supplemental Fig. S2B). Long SimRNA simulations with n_{step} = 100,000 allow nearly every cluster to extend to 6 bp continuous interaction (Fig. 4A). In setting I, some runs even reach 7–8 bp interaction length; however, this only happens when the SimRNA constraint has already been extended to 11–13 intermolecular base pairs. Indeed, all structures with more than 6 bp interaction exhibit very high constraint penalties; see Supplemental Figure S2C. This suggests that an interaction length of 6 bp presents an optimum that can only be overcome by very strong constraints and indicates that quite some effort is needed for further folding. Setting II, i.e., with loops and bulges within the interaction region, allows the formation of longer interactions. While the interacting region could contain up to 13 bp (with a peak at 12 bp), no uninterrupted helices longer than 8 bp were observed (Fig. 4B). The results of our simulations are consistent with the CopA–CopT pathway proposed by Kolb et al. (2000b), in which an interior loop is formed in the interaction site in order to allow longer interaction. Even in this setting, most runs terminate at an interaction length of 6 bp. Again, longer interaction regions require more base pair constraints (13–23 bp) and thus stronger constraints (Supplemental Fig. S3). Moreover, the relative orientation of the CopA and CopT stems can influence how easily interactions can be extended; see Supplemental Figure S4. An analysis of the SimRNA energies (Supplemental Fig. S3) shows that 6 bp conformations are energetically favorable with small constraint violation energies, while longer interactions exhibit higher constraint energies. Moreover, conformations with interacting regions longer than 6 bp have consistently worse energies, indicating that these structures could be the result of strong constraints, rather than naturally forming conformations.

HIV-1 DIS

Since the full extended interaction target of the HIV-1 DIS segment includes two loops, we specify the start and end conformation for our pipeline. The enclosing intramolecular helices were left unconstrained (see Supplemental Fig.

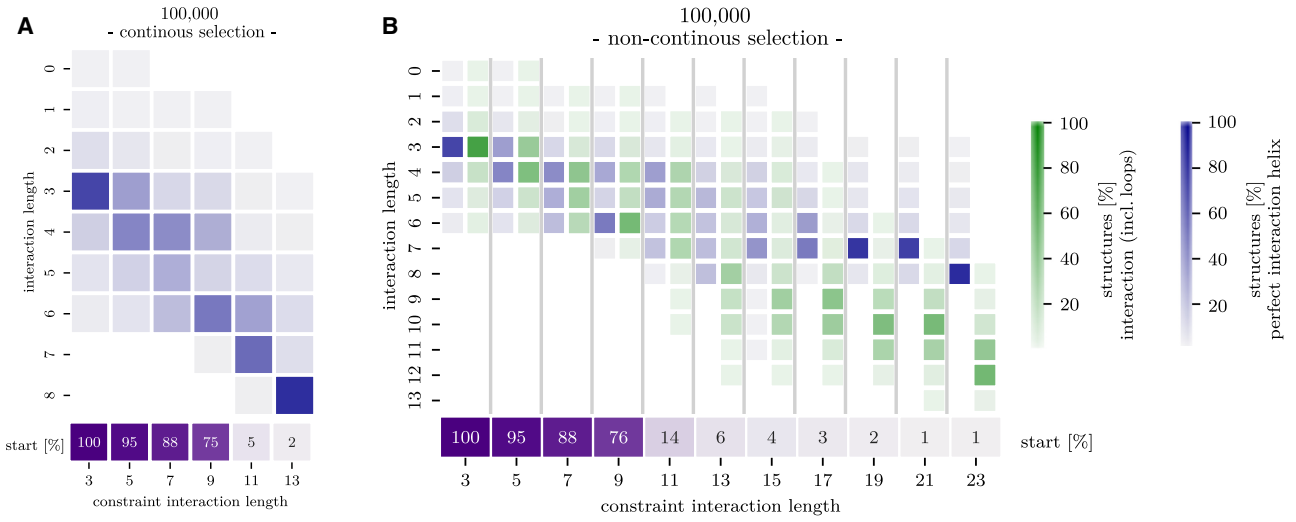


FIGURE 4. Distribution of CopA–CopT interaction lengths for two different pipeline settings and long SimRNA runs (n_{step} = 100,000); see Figure 3 for description. (*A*) Setting I, extensions form perfect helices. (*B*) Setting II, with loops in the interaction region allowed. Histograms are shown for perfect interaction helices (blue), as well as interactions with loops (green).

S5 for the full 2D path). Instead of using Ernwin, we started from the 1ZCI PDB structure (and thus have $n_{cluster} = 1$), but increased n_{run} and n_{sim} to 10 in order to increase 3D structure diversity. With the previously used settings for SimRNA simulations ($n_{step} = 10{,}000$ or $100{,}000$), we were unable to extend the initial 6 bp kissing interaction. We conjecture three causes: the interaction length of 6 bp appears to be especially stable for kissing hairpins; the interior loop needs to be bridged in order to reach longer interactions (Fig. 1D); and the PDB structure provides a particularly good initial 3D fold. We therefore performed additional pipeline runs with extremely long SimRNA simulations of $n_{step} = 1$ million, and two different expansion modes were compared. In setting I, each extension step elongates the interaction region by 1 bp in both directions (symmetric); see Supplemental Figure S5A for the given constraints and Figure 5A for the results. In setting II, we extend only in one direction until the maximum extent is reached (asymmetric); see Supplemental Figure S5B and Figure 5B. With these extremely long SimRNA simulations, two out of 10 runs were able to form an extended duplex for both settings. In both settings, two intermediates with an interaction length of 8 bp (especially in setting I) and around 12 bp can be observed. After the second intermediate state, the enclosing helices unfold completely and the full-length interaction is formed spontaneously. A notable difference between the two settings is that for setting I, the simulation passes through intermediates with very high constraint energies (i.e., many unsatisfied base-pairing constraints); see Supplemental Figure S5.

At interaction length 12, the enclosing stem has shrunk to 3–4 bp in length, and further extension leads to the complete unfolding of the enclosing stem. This unfolding of the remaining enclosing stem corresponds to the final energy barrier, after which the extended duplex can be reached. Looking at the SimRNA energies (Supplemental Fig. S5) and stem lengths, we observe that the second intermediate state has its energy minimum at an interaction length of 12 and a remaining enclosing stem of 3–4 bp. This seems to be the minimum length at which the enclosing stem remains stable. With the next extension, it unfolds completely, allowing it to reach the full extended duplex by a downhill walk on the energy landscape.

Note, however, that even with these settings eight out of 10 runs remain trapped close to the 6 bp state. Moreover, the PDB structure contains only a 23 nt long fragment of the HIV genome and therefore a shortened enclosing helix, which would be even more difficult to unfold in the context of the full genome.

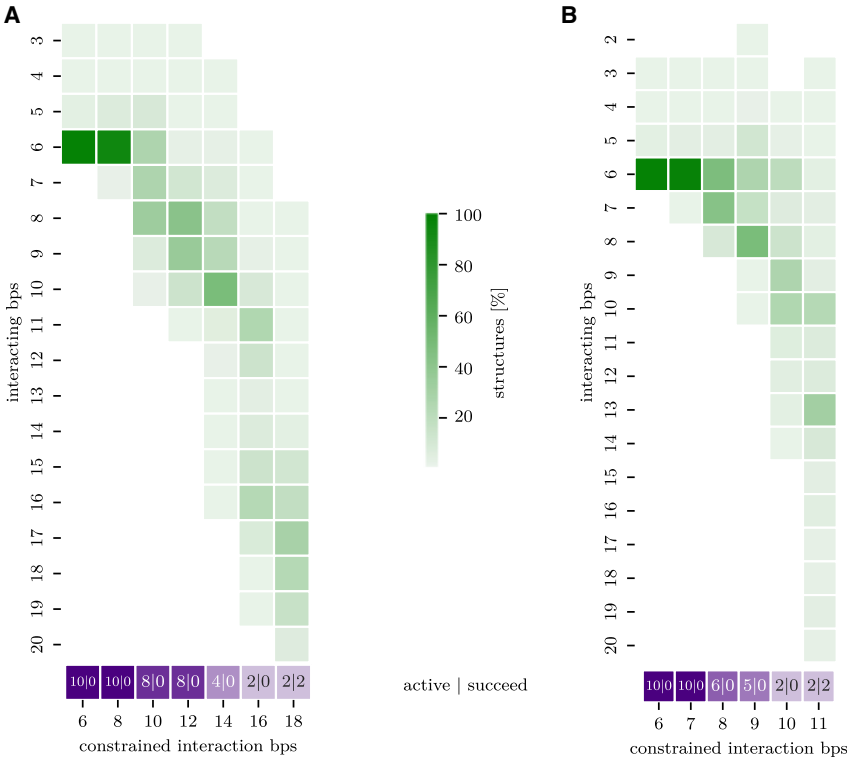


FIGURE 5. (*A*) Setting I. Symmetric interaction extension of the HIV-1 DIS interaction site with $n_{step} = 1$ million. Each column shows the histogram (green) of interaction lengths including loops (*y*-axis) at the corresponding constrained interaction length in bp (*x*-axis). For every elongation step (corresponding to a number of constrained interaction base pairs), the purple box reports firstly the number of still active runs (from a total of 10); and secondly, the number of runs that successfully reach the full duplex. (*B*) Setting II with an asymmetric interaction extension.

DsrA–rpoS

We selected the interaction between DsrA and *rpoS* as an example with different interaction formation scenarios that can be investigated with our pipeline. IntaRNA, a thermodynamics-based 2D interaction prediction tool, predicted an extended interaction, depicted in the lower half of Figure 6 (see also Wu et al. 2017). Specifically, we compared three different possible intramolecular start structures of DsrA (Fig. 1) as well as the influence of the initial contact point. Based on RRIkinDP (Waldl et al. 2023), the energy landscape (see heatmap in Fig. 6) of all possible intermediate interactions was computed in 2D. This provided three favorable start interaction sites with high accessibility which we selected for modeling using our 3D pipeline: (i) in the SL1 loop (Fold-A only), (ii) at the start of LR (all three conformations), and (iii) in the bulge of SL2 in Fold-A (respectively the end of LR in Fold-B). As for CopA–CopT, we ran our pipeline with $n_{cluster} = 10$, $n_{run} = 5$, and

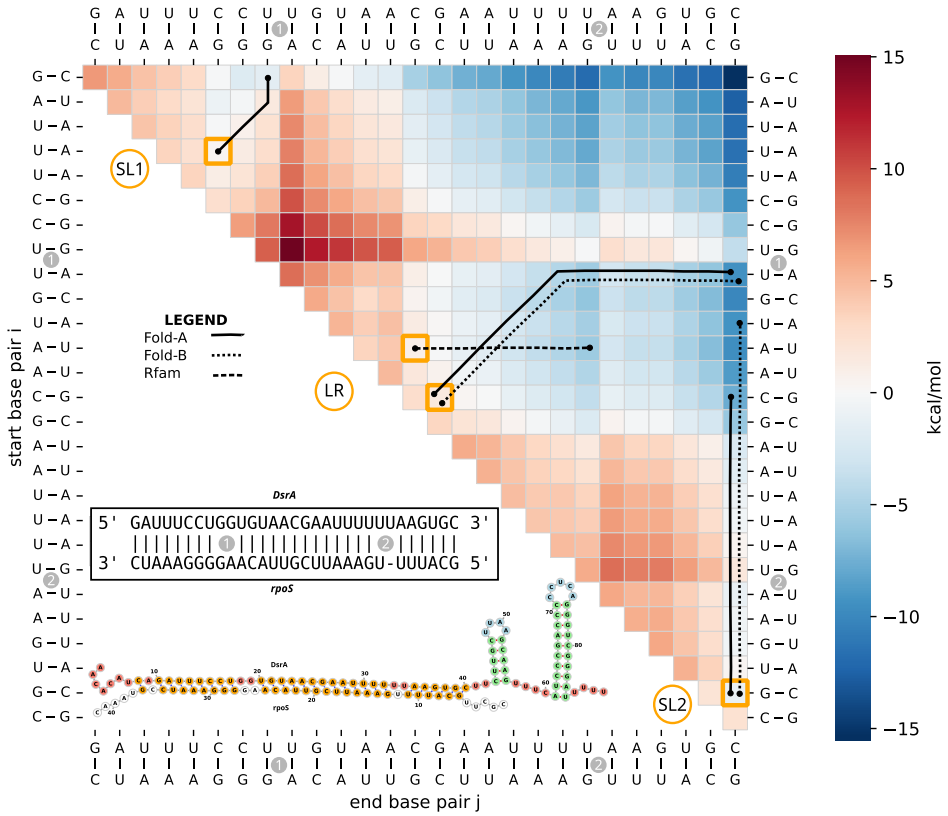


FIGURE 6. Interaction of DsrA–*rpoS* in *E. coli* as a direct path energy landscape following RRIkinDP (Waldl et al. 2023) and comparison of different formation scenarios. RRIkinDP, a novel 2D structure-based tool for studying RNA–RNA interaction formation on direct paths, was used to generate the energy landscape of structures along folding pathways from a first interaction base pair to the full DsrA–*rpoS* interaction. To study interaction formation, RRIkinDP considers intermediate interactions formed by consecutive interaction base pairs. Each intermediate interaction is therefore defined by the first and last interaction pair. For visualization, intermediate states can therefore be arranged in an upper triangular matrix, indexed by two enclosing pairs on the *x*- and *y*-axis. Each cell of the matrix represents an intermediate interaction, with minimal interactions consisting of a single base pair on the diagonal and the full, maximal length, interaction in the *upper right* corner. Cell colors depict the free energy ΔG on a red-to-blue scale. For details, see Waldl et al. (2023). The predicted 2D interaction for DsrA–*rpoS* appears in the *lower left* triangle, presented in two ways: the interaction site *above*, with interior loops marked as (1) and (2), and a 2D representation of the complete interaction complex *below*. Using the energy landscape, we identified three energetically favorable start sites (highlighted in orange): within the loop of the first stem–loop (SL1), within the LR, and within the bulge of the second stem–loop (SL2), as shown in Figure 1. From these sites, we generated direct folding paths in 2D, which were subsequently embedded in 3D using the RRI-3D pipeline. The resulting 3D pathways were projected back into 2D and are depicted as black lines within the heatmap. In the path, moving diagonally corresponds to extending the interaction on both sides, while moving up or to the right extends the interaction by 1 bp on the left or right, respectively. Different linestyles represent different initial folds of the DsrA SL2 structure (see legend and Fig. 1).

three different simulation lengths n_{step} = 5000, 10,000 and 100,000. In total, we tested six combinations of initial contact and starting structure; see Figure 1B,C. The corresponding 2D pathways used for SimRNA constraints are shown in Supplemental Figures S7–S10. Two-dimensional projections of representative folding paths for each scenario are also marked in the energy landscape (Fig. 6).

Out of the three start points, the hairpin of SL1 is clearly the worst. The 3 bp initial interaction spontaneously extends to 6 bp in length. However, all simulations terminated without fully unfolding the SL1 stem. Even with an increasing number of constraints, almost no interaction length >8 is sampled; see Supplemental Figure S11. This was somewhat expected from the energy barrier seen in the 2D landscape.

Simulations starting in LR quickly extend in the 3′ direction over the whole LR, regardless of whether Fold-A (Supplemental Fig. S13) or Fold-B (Supplemental Fig. S14) is assumed. When starting in Fold-A the interaction unwinds the lower part of SL2, suggesting that a refolding into Fold-B would eventually happen. Even though the constraints are extended symmetrically in both directions, the observed interaction region grows much more quickly in the 3′ direction, while <10% of runs manage to extend into SL1. For the Rfam structure, our initial contact is wedged in between SL1 and SL2. Since SL2 is much weaker, we used a 2D path that extends toward SL2 (Supplemental Fig. S10). The simulations manage to partially unfold SL2 (Supplemental Fig. S15). Even though

this version of SL2 consists of weak AU and GU pairs, it slows down interaction formation due to its length. Nineteen out of initially 50 simulations manage to at least dissolve the 6 bp helix until the bulge.

Starting with SL2 as the start site, Fold-B (Supplemental Fig. S12B) has a slight advantage compared to Fold-A (Supplemental Fig. S12A), since it avoids opening the additional 3 bp of SL2. However, the fraction of simulations that extend all the way to SL1 is almost the same. Only a single simulation manages to extend slightly into SL1.

In comparison, our simulations suggest that optimal interaction formation should start within the LR of Fold-A or Fold-B. In contrast to Wu et al. (2017), we find that starting in Fold-A works almost equally well as starting in Fold-B and may even trigger refolding into Fold-B. The Rfam structure offers fewer highly accessible start points in LR and additionally, interaction formation is slower. In all cases, SL1 was too stable to be dissolved during our simulations. However, once a stable interaction has been formed within the LR region, the DsrA–*rpoS* complex will not dissolve and could eventually extend beyond SL1 on a timescale beyond what is covered in our simulations.

DISCUSSION

Our presented approach enables detailed studies of potential 2D folding pathways for the formation of RNA–RNA interactions in 3D. To overcome the typically extreme computational cost of 3D structure prediction, we use coarse-grained 3D simulations and, moreover, construct interactions step-by-step in a series of length-limited simulation runs. Following a predefined 2D path of secondary structures, each step corresponds to an elementary extension of the interaction by 1 or 2 bp. This allows us to simulate large structural changes that are inaccessible for atom-level simulations such as molecular dynamics. Moreover, it allows us to perform a sufficiently large number ($n_{\text{cluster}} \times n_{\text{run}} \times n_{\text{sim}}$) of independent simulations in order to sample diverse starting conformations and alternative pathways. As an example, a typical CopA–CopT simulation with $n_{\text{cluster}} = 10$, $n_{\text{run}} = 5$, $n_{\text{sim}} = 5$, $n_{\text{step}} = 10{,}000$ took about 19 h, using up to 10 cores.

In contrast to models that only consider secondary structure, this allows us to identify likely obstacles in the folding path due to steric hindrance. A paradigmatical example is the interaction ("kissing") between two hairpins as in the case of CopA–CopT. In this case, due to the perfect complementarity, secondary structure tools would typically extend any initial contact up to the full duplex, completely neglecting steric effects. Radically changing the picture, our pipeline reveals that extending the interaction beyond the 6 bp stage becomes increasingly difficult; or even impossible while maintaining perfect stacking: rarely, we observed kissing hairpin interactions of more than 8 bp; all of them contained bulges or small interior loops, presumably in order to relieve strain. This is consistent with the structures proposed by Kolb et al. (2000b) on the basis of mutation and probing experiments. We find similar behavior for the HIV-DIS interaction, which also forms a kissing hairpin. In this case, a few simulations reach a full duplex, but this is only possible because we simulate a short fragment of the HIV genome, and thus the enclosing helix is short enough to completely unfold. A more complex example is represented by the DsrA–*rpoS* interaction, where we show that our approach can be used to compare different scenarios of interaction formation, such as starting from different monomer structures or different initial contact points. Our observation that kissing hairpins are difficult to extend into long interactions is consistent with the fact that there are almost no known natural kissing hairpin structures (including those in pseudoknots) with more than six intermolecular pairs. Using a modified Forgi script (Thiel et al. 2019) and the nonredundant 3D structure set of Leontis and Zirbel (2012), we found that among all kissing hairpin pseudoknots with genus 1 (as defined by Reidys et al. 2011), out of 44 pseudoknots only the group II intron (PDB 3IGI by Toor et al. 2009) with an interaction length of 7 bp is longer, but this occurs between a hairpin loop and an interior loop. We note that the pipeline can also be used to test the 3D feasibility of pseudoknots predicted on the basis of secondary structure, as done by Gosavi et al. (2022).

An attractive future application of our pipeline would be to postprocess and evaluate conventional secondary structure-based interaction predictions. Here, an RNA targeting screen as performed by fast 2D tools like RISearch2 (Alkan et al. 2017) could be scrutinized for 3D steric feasibility by our pipeline. Adding to this idea, insights from example systems can be used to flag 2D-based predictions that call for further investigation. In particular, predictions relying on long (>6 bp) intermolecular helices within stable intramolecular structures should be contested. Conversely, long interactions are uncritical, even favorable in exterior loops.

Finally, our findings motivate the future improvement of 2D-based tools by incorporating 3D features (e.g., context-dependent length restrictions) or even tighter integration with 3D modeling.

DATA DEPOSITION

Python source code for our pipeline, including documentation as well as data and scripts to analyze the three example systems, is available at https://github.com/irenekb/RRI-3D.

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

This work was supported by the Austrian FWF: I-2874-N28 "Prediction of RNA–RNA interactions," W-1207 "DK RNA Biology," I-4520 "Deciphering Complex RNA structure by probing and predictions," and F-80 "RNAdeco"; and also by the French ANR: ANR-21-CE45-0034-01 "INSSANE."

Received July 28, 2023; accepted November 16, 2023.

REFERENCES

Alkan F, Wenzel A, Palasca O, Kerpedjiev P, Rudebeck AF, Stadler PF, Hofacker IL, Gorodkin J. 2017. RIsearch2: suffix array-based large-scale prediction of RNA-RNA interactions and siRNA off-targets. *Nucleic Acids Res* **45:** e60. doi:10.1093/nar/gkw1325

Andronescu M, Zhang ZC, Condon A. 2005. Secondary structure prediction of interacting RNA molecules. *J Mol Biol* **345:** 987–1001. doi:10.1016/j.jmb.2004.10.082

Bernhart SH, Tafer H, Mückstein U, Flamm C, Stadler PF, Hofacker IL. 2006. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol* **1:** 3. doi:10.1186/1748-7188-1-3

Boniecki MJ, Lach G, Dawson WK, Tomala K, Lukasz P, Soltysinski T, Rother KM, Bujnicki JM. 2015. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Res* **44:** e63. doi:10.1093/nar/gkv1479

DiChiacchio L, Sloma MF, Mathews DH. 2015. AccessFold: predicting RNA–RNA interactions with consideration for competing self-structure. *Bioinformatics* **32:** 1033–1039. doi:10.1093/bioinformatics/btv682

Ennifar E, Dumas P. 2006. Polymorphism of bulged-out residues in HIV-1 RNA DIS kissing complex and structure comparison with solution studies. *J Mol Biol* **356:** 771–782. doi:10.1016/j.jmb.2005.12.022

Fornace ME, Porubsky NJ, Pierce NA. 2020. A unified dynamic programming framework for the analysis of interacting nucleic acid strands: enhanced models, scalability, and speed. *ACS Synth Biol* **9:** 2665–2678. doi:10.1021/acssynbio.9b00523

Franch T, Petersen M, Wagner EH, Jacobsen JP, Gerdes K. 1999. Antisense RNA regulation in prokaryotes: rapid RNA/RNA interaction facilitated by a general U-turn loop structure. *J Mol Biol* **294:** 1115–1125. doi:10.1006/jmbi.1999.3306

Gosavi D, Wower I, Beckmann IK, Hofacker IL, Wower J, Wolfinger MT, Sztuba-Solinska J. 2022. Insights into the secondary and tertiary structure of the Bovine Viral Diarrhea Virus Internal Ribosome Entry Site. *RNA Biol* **19:** 496–506. doi:10.1080/15476286.2022.2058818

Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, et al. 2020. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Res* **49:** D192–D200. doi:10.1093/nar/gkaa1047

Kerpedjiev P, Hammer S, Hofacker IL. 2015a. Forna (force-directed RNA): simple and effective online RNA secondary structure diagrams. *Bioinformatics* **31:** 3377–3379. doi:10.1093/bioinformatics/btv372

Kerpedjiev P, Siederdissen Cz, Hofacker IL. 2015b. Predicting RNA 3D structure using a coarse-grain helix-centered model. *RNA* **21:** 1110–1121. doi:10.1261/rna.047522.114

Kolb FA, Engdahl HM, Slagter-Jäger JG, Ehresmann B, Ehresmann C, Westhof E, Wagner EGH, Romby P. 2000a. Progression of a loop-loop complex to a four-way junction is crucial for the activity of a regulatory antisense RNA. *EMBO J* **19:** 5905–5915. doi:10.1093/emboj/19.21.5905

Kolb FA, Malmgren C, Westhof E, Ehresmann C, Eheresman B, Wagner EGH, Romby P. 2000b. An unusual structure formed by antisense-target RNA binding involves an extended kissing complex with a four-way junction and a side-by-side helical alignment. *RNA* **6:** 311–324. doi:10.1017/s135583820099215x

Leontis NB and Zirbel CL. 2012. Nonredundant 3D structure datasets for RNA knowledge extraction and benchmarking. In *Nucleic acids and molecular biology*, pp. 281–298. Springer, Berlin/Heidelberg

Lorenz R, Bernhart SH, Siederdissen CH, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6:** 1. doi:10.1186/1748-7188-6-26

Mann M, Wright PR, Backofen R. 2017. IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions. *Nucleic Acids Res* **45:** W435–W439. doi:10.1093/nar/gkx279

Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF, Hofacker IL. 2006. Thermodynamics of RNA–RNA binding. *Bioinformatics* **22:** 1177–1182. doi:10.1093/bioinformatics/btl024

Reidys CM, Huang FWD, Andersen JE, Penner RC, Stadler PF, Nebel ME. 2011. Addendum: topology and prediction of RNA pseudoknots. *Bioinformatics* **28:** 300. doi:10.1093/bioinformatics/btr643

Schrödinger LLC. 2022. *The PyMOL Molecular Graphics System*, Version 2.5.

Shabalina SA, Koonin EV. 2008. Origins and evolution of eukaryotic RNA interference. *Trends Ecol Evol* **23:** 578–587.

Thiel BC, Beckmann IK, Kerpedjiev P, Hofacker IL. 2019. 3D based on 2D: calculating helix angles and stacking patterns using *forgi 2.0*, an RNA Python library centered on secondary structure elements. *F1000Res* **8:** 287. doi:10.12688/f1000research.18458.2

Toor N, Keating KS, Fedorova O, Rajashankar K, Wang J, Pyle AM. 2009. Tertiary architecture of the *Oceanobacillus iheyensis* group II intron. *RNA* **16:** 57–69. doi:10.1261/rna.1844010

Waldl M, Beckmann IK, Will S, Hofacker IL. 2023. Modeling kinetics of RNA-RNA interactions on direct paths. bioRxiv. doi:10.1101/2023.07.28.548983

Waters LS, Storz G. 2009. Regulatory RNAs in bacteria. *Cell* **136:** 615–628. doi:10.1016/j.cell.2009.01.043

Wu P, Liu X, Yang L, Sun Y, Gong Q, Wu J, Shi Y. 2017. The important conformational plasticity of DsrA sRNA for adapting multiple target regulation. *Nucleic Acids Res* **45:** 9625–9639. doi:10.1093/nar/gkx570

See the following page for ***Meet the First Author***

3. Insights into the secondary and tertiary structure of the Bovine Viral Diarrhea Virus Internal Ribosome Entry Site

Gosavi D, Wower I, Beckmann IK, Hofacker IL, Wower J, Wolfinger MT and Sztuba-Solinska J. Insights into the secondary and tertiary structure of the Bovine Viral Diarrhea Virus Internal Ribosome Entry Site. *RNA Biology* 19.1 (2022), pp. 496–506. DOI: 10.1080/15476286.2022.2058818

Summary

Probing data such as from SHAPE-MaP can be used as a guide for secondary structure prediction and subsequent computational approaches can derive further structural information. However, these approaches are based on 2D structure prediction and although the experimental data can be used to generate the structure, the 2D results are highly dependent on the tool used. The integration of 3D structure prediction, as shown here, is a valuable addition and demonstrates that it provides a comprehensive understanding of RNA structure. In particular, it allows to verify whether the proposed 2D structure is physically feasible in 3D, especially in the case of pseudoknots. The relationship between flexibility and stability is often a critical factor in the functionality of an RNA. The integration of 3D structure prediction through massively parallel simulations and clustering techniques has enabled the identification of low-energy conformations and their dynamics, thus expanding the possible structural landscape. In conclusion, the integration of 3D structure prediction in this study highlights a powerful and novel approach to identify further details of RNA architecture and helps to learn more about the biological functions of RNAs.

Authors contribution

DG, IW, JW and JSS were responsible for the planning and execution of the SHAPE-MaP experiments, while MTW, ILH and I (IKB) were responsible for the bioinformatics part. My role was to develop a concept for embedding the 2D structure in 3D to further investigate its feasibility, dynamics and steric properties. I implemented the corresponding pipeline and performed statistical analyses for the 3D structure evaluation, in particular for the pseudoknot detected by SHAPE-MaP.

RNA BIOLOGY
2022, VOL. 19, NO. 1, 496–506
https://doi.org/10.1080/15476286.2022.2058818

Taylor & Francis
Taylor & Francis Group

RESEARCH PAPER

OPEN ACCESS Check for updates

Insights into the secondary and tertiary structure of the Bovine Viral Diarrhea Virus Internal Ribosome Entry Site

Devadatta Gosavi[a], Iwona Wower[b], Irene K. Beckmann[c], Ivo L. Hofacker [c,d], Jacek Wower[b], Michael T. Wolfinger [c,d], and Joanna Sztuba-Solinska [a,e,*]

[a]Department of Biological Sciences, Auburn University, 120 W. Samford Ave, Rouse Life Sciences Building, Auburn, AL, United States; [b]Department of Animal and Dairy Sciences, Auburn University, Auburn, AL, United States; [c]Department of Theoretical Chemistry, University of Vienna, Vienna, Austria; [d]Research Group Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Vienna, Austria; [e]Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznan, Poland

ABSTRACT

The internal ribosome entry site (IRES) RNA of bovine viral diarrhoea virus (BVDV), an economically significant *Pestivirus*, is required for the cap-independent translation of viral genomic RNA. Thus, it is essential for viral replication and pathogenesis. We applied a combination of high-throughput biochemical RNA structure probing (SHAPE-MaP) and *in silico* modelling approaches to gain insight into the secondary and tertiary structures of BVDV IRES RNA. Our study demonstrated that BVDV IRES RNA in solution forms a modular architecture composed of three distinct structural domains (I–III). Two regions within domain III are represented in tertiary interactions to form an H-type pseudoknot. Computational modelling of the pseudoknot motif provided a fine-grained picture of the tertiary structure and local arrangement of helices in the BVDV IRES. Furthermore, comparative genomics and consensus structure predictions revealed that the pseudoknot is evolutionarily conserved among many *Pestivirus* species. These studies provide detailed insight into the structural arrangement of BVDV IRES RNA H-type pseudoknot and encompassing motifs that likely contribute to the optimal functionality of viral cap-independent translation element.

ARTICLE HISTORY
Received 22 November 2021
Revised 1 March 2022
Accepted 23 March 2022

KEYWORDS
Internal ribosome entry site (IRES); bovine viral diarrhoea virus (BVDV); selective 2′ -hydroxyl acylation analysed by primer extension and mutational profiling (SHAPE-MaP); H-type pseudoknot; covariance analysis

Introduction

Bovine viral diarrhoea virus (BVDV) is a member of the genus *Pestivirus*, family Flaviviridae, that includes the causative agents of economically significant diseases of cattle, pigs, and sheep. The BVDV infections in cattle lead to decreased fertility and milk production, slow foetal growth, diarrhoea, respiratory symptoms, reproductive and immunological dysfunction [1]. According to the ICTV virus taxonomy profile [2], two main genotypes of BVDV, namely type 1 (*Pestivirus* A) of low-virulence and type 2 (*Pestivirus* B) of high-virulence, have been recognized and are estimated to cause considerable economic loss [3]. Besides the two BVDV genotypes, several genetic isolates have been distinguished within BVDV type 1 [4,5].

The BVDV genome is a positive-sense single-stranded RNA (12.5 kb) that codes for one open reading frame (ORF). The coding region is preceded by the distinctly structured 5′ untranslated region (5′ UTR), which folds into the Internal Ribosomal Entry Site (IRES) [6]. By definition, IRES is the RNA domain that recruits ribosomes to the internal region of mRNAs to initiate translation through a cap-independent pathway. IRES domains can be found in genomic RNAs of many pathogenic viruses. In Hepatitis C virus (HCV) genomic RNA, it mediates the initiation of translation by recruiting the subset of canonical translation initiation factors (eIF2, eIF3), methionine tRNA (Met-tRNAi), and by orienting the 40S ribosomal subunit at the translational initiation codon [7,8]. Cricket paralysis virus (CrPV) IRES RNA directs ribosomes without the requirement of additional translational factors [9], while IRES RNAs of poliovirus (PV) [10] and foot and mouth disease virus (FMDV) [11] rely on binding of eIFs, Met-tRNAi, and additional proteins referred to as the IRES trans-acting factors (ITAFs) [5,12–14]

According to previous structural studies, the viral IRES RNAs fold into intricate secondary and tertiary arrangements that provide a structural scaffold for cellular initiation factors and trigger conformational changes in the 40S ribosomal that drive translation initiation by an active mechanism [15–18]. For example, the HCV IRES consists of four structurally defined domains (I–IV), with domain II inducing the open conformation of the 40S subunit and III and IV forming a functionally essential pseudoknot [19]. Pseudoknots are formed upon base-pairing of a single-stranded region of RNA in the loop of a hairpin or a bulge to a stretch of complementary nucleotides elsewhere in the RNA chain. The flexibility and transient formation of pseudoknots might

CONTACT Michael T. Wolfinger michael.wolfinger@univie.ac.at Department of Theoretical Chemistry, University of Vienna, Vienna, Austria
*To whom correspondence should be addressed J.S.S. Tel: +1-334-844-4830, Email: jzs0165@auburn.edu

Supplemental data for this article can be accessed here

be essential for recognition and favourable interaction with translating ribosomes, or the translation factors associated with the ribosomes [20]. Thus, the folding strategy can generate a vast number of three-dimensional folds, which exhibit a diverse range of highly specific functions [21]. CrPV IRES also displays a multi-domain composition with the residues of one domain supporting a pseudoknot. In the case of BVDV IRES (SD-1 isolate, genotype 1), thermodynamic and phylogenetic studies suggested that its translational functionality is also controlled by the formation of a pseudoknot [21,22]. Burks et al., using comparative structural studies, predicted that the BVDV IRES pseudoknot involves interaction between the loops of hairpin 3 and 4 within domain III [5,22,23]. Toeprinting *in vitro* analysis involving BVDV IRES, 40S rRNA, and eIFs suggested that the ribosome makes multiple contacts with the BVDV IRES. Similar observations have been made for HCV IRES RNA [24–27], outlining particularly interesting contacts existing between 18S rRNA and the pseudoknot formed within IRES domain III [18,23]. Additionally, the presence of an RNA pseudoknot in BVDV IRES RNA was supported by compensatory mutations, which restored translation [22].

In this study, we investigated the secondary structure of BVDV IRES RNA using selective 2′-hydroxyl acylation analysed by primer extension and mutational profiling (SHAPE-MaP) [28,29]. Our analysis, which has been performed on two *in vitro* synthesized constructs that contain all essential elements supporting the cap-independent translation, revealed that BVDV IRES RNA folds into three distinct structural domains (I–III). Two single-stranded regions of domain III were shown to support the formation of an H-type pseudoknot, thus corroborating earlier studies that predicted the existence of a pseudoknot in BVDV IRES RNA and in related pestiviruses [30,31]. To obtain a better picture of pseudoknot formation, we performed RNA 3D simulations with ernwin [32] and SimRNA [33] on a trimmed construct of the BVDV-1 strain NADL IRES. Tertiary structure simulations confirmed that an H-type pseudoknot can be formed under the condition that two adjacent helices are oriented in a well-defined range of angles against each other. On a broader scale, we performed a comparative genomics screen of the 5′ UTR in the genus *Pestivirus*, which confirmed pervasive evolutionary conservation of the IRES region, and provided evidence for the presence of a pseudoknot in all studied viruses.

Results

The secondary structure of BVDV IRES RNA

SHAPE-MaP is a high-throughput biochemical probing technique that can be used to characterize RNA secondary structure at a single-nucleotide resolution (Fig. S1). In SHAPE-MaP, unpaired nucleotides (nts) are acylated more readily at the 2′-hydroxyl ribose moieties by an electrophilic reagent, e.g. 1-methyl-7-nitroisatoic anhydride (1M7) [34] than those that are paired, and the reactivity biases are employed to address RNA secondary structure [35]. The modified RNA is directed to reverse transcription reaction in the presence of Mn^{2+} ions that induce reverse transcriptase to read through the bulky 2′-O-adducts and insert mutations [36]. The resulting cDNA products carrying the mutations are directed to stepwise amplification to generate DNA libraries, which are directed to next-generation sequencing.

We performed structural probing of BVDV IRES RNA on two *in vitro* synthesized transcripts; the short transcript (560 nts) included the BVDV IRES RNA sequence only; the long transcript (1296 nts) additionally contained the EGFP cassette (nts 561–1296), and the MS2 hairpin (nts 523–543) at its 3′ end. The IRES RNA constructs used in this study were derived from BVDV-1 strain NADL genomic RNA and contain RNA sequences that were outlined as necessary for supporting the expression of EGFP protein by the cap-independent translation [37,38]. BVDV-1 strain NADL is a cytopathic strain, originally isolated from an animal infected with fatal mucosal disease [39] (Fig. S2). We confirmed the homogeneous conformation of both folded RNA constructs by native gel electrophoresis prior to structural probing (Fig. S3). The comparison of SHAPE reactivity profiles for short and long IRES RNA constructs, performed by custom R scripts, showed a positive correlation established by an average Pearson correlation coefficient of 0.8398 (p-value 0.022) (Fig. S4). Since we obtained a more comprehensive mapping output for the long transcript, we used these data for modelling the BVDV IRES RNA.

Our analysis revealed that the BVDV IRES can be divided into three domains labelled I–III, with individual stem-loops (SL) within each domain being assigned letter codes (Figure. 1). Domain I contain SL Ia, which includes an apical loop with an equal number of purines and pyrimidines, SL Ib, which has a purine-rich apical loop, and SL Ic, which has an adenine-rich apical loop. Domain II forms a Y-shaped hairpin that includes two SLs interconnected by a three-way junction. The 5′ SL folds into two short stems interrupted by an asymmetrical internal loop with three nts at the 5′ side and mispaired adenine at the 3′ side. The apical loop includes predominantly adenines, which are largely not reactive to SHAPE reagent, likely due to extensive base stacking interactions characteristic for A-rich sequences [40]. The 3′ SL contains a 3′ four nts bulge and a UNRN tetraloop (UAGU), where N stands for any nucleotide and R is a purine base. These types of tetra-loops are frequently present in apical loops involved in the U-turn formation, where the backbone changes direction immediately following the initiating U, and the Watson–Crick faces of the bases immediately 3′ of the U are exposed to the solvent [41,42]. Domain III includes the main stem that is formed by SL IIIf and IIIg, interrupted by a 2 nt bulge, followed by a three-way junction that connects SL IIIe and the apical IRES portion. Successively, a four-way junction connecting SL IIId1 and IIId2 is present, followed by a four-way junction connecting SL IIIa, b, and c. The stem of SL IIIa includes an AGUA tetra-loop with less-reactive UA residues [41]. The SL IIIb forms the longest hairpin of this region, closed with a purine-rich apical loop. It includes an asymmetrical internal loop at its 3′ side, a downstream 3 nts bulge, and a cytosine mismatch. The SL IIIc includes three canonical base pairs and a CAUG tetra-loop belonging to the YNMG tetra-loop family, where Y stands for U/C, N, and M is for A/C/U. Previous NMR, circular dichroism, and functional group substitution studies in 16S rRNA indicated the unusual thermodynamic stability of YNMG tetra-loop and showed its involvement in

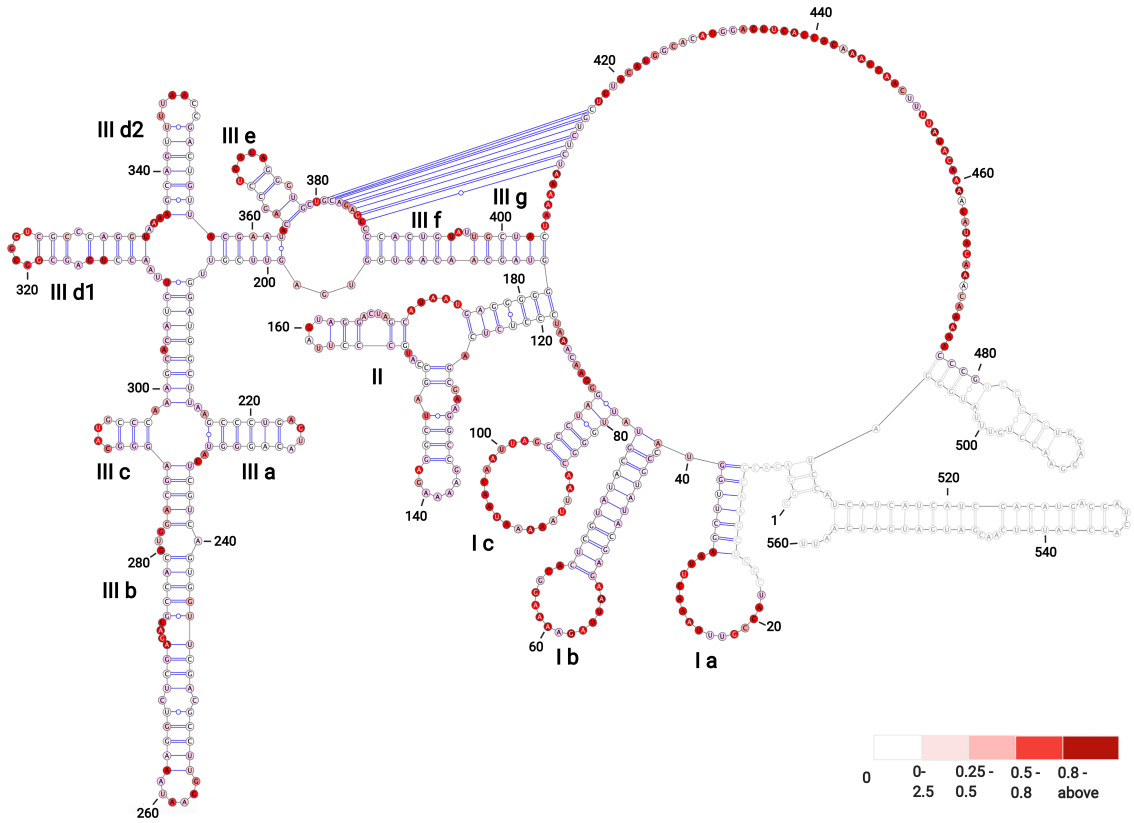


Figure 1. The secondary structure model of BVDV IRES RNA derived from SHAPE-MaP probing. The three main domains are labelled (I–III) with outlined individual RNA motifs: SL Ia, SL Ib, SL Ic, SL II, SL IIIa, SL IIIb, SL IIIc, SL IIId1, SL IIId2, SL IIIe, SL IIIf, SL IIIg. The H-type pseudoknot has been manually inserted within the single-stranded residues at positions 381–387 and 410–416 of domain III. The residues lacking reactivity values correspond to the sites of primer annealing (light grey). The secondary structure is colour-coded according to the SHAPE reactivity values represented by low (< 0.25), medium (0.25–0.80), high reactivity (0.8–2), and hyperreactivity (>2).

tertiary interactions [43]. SL IIId1 is closed with an apical penta-loop that consists of highly reactive G residues (median SHAPE reactivity of 1.020). This type of G-rich apical loop was found in the 5′ UTR SL2 and SL3 of the HIV-1 genome involved in binding the nucleocapsid protein [44]. The SL IIId2 involves a stem with two GU wobble base pairs and a 6 nt apical pyrimidine-rich loop. The SL IIIe is closed with a UGAUA penta-loop.

The residues immediately downstream of SL IIIe and the residues near the AUG codon (nts 429–431) were proposed previously to be involved in a H-type pseudoknot [21,22]. In the structure predicted from SuperFold, however, part of these residues was involved in the formation of two hairpins held by two reactive GC pairs (G381, reactivity value of 0.219 pairing with C388, reactivity value of 0.408; G387, reactivity value of 0.714 pairing with C382, reactivity value of 0.175 for hairpin 1; and G415, reactivity value of 0.121 pairing with C426, reactivity value of 0.57; G425, reactivity value of 0.346 pairing with C416, having no reactivity value for hairpin 2). Considering that cytosines are known to be the least reactive towards SHAPE electrophilic reagents and are generally involved in single-stranded regions [45,46], we resolved these hairpins. Subsequently, the pseudoknot in domain III has been manually inserted within the single-stranded regions at positions 381–387 (5′-GCAGAGG-3′, median SHAPE reactivity of 0.617) and 410–416 (5′-UCUCUGC-3′, median SHAPE reactivity of 0.115).

The single-stranded region immediately downstream of the pseudoknot harbours the AUG codon. We noted that numerous residues within BVDV IRES RNA display hyperreactivity towards SHAPE reagents (Table S1, Fig. S4). These are found within the loops (nts 26, 22, 54, 58, 60, 91, 99), bulges (nts 271, 273), single-stranded regions (nts 116, 438, 444, 455, 461, 465, 469), and near the pseudoknot (nts 409, 417) (Fig. S5). These residues were previously reported as accommodating a C2-endo ribose conformation [47]. This conformation of ribose shows slow dynamics and favours intra-residue H bonding between the 2′-hydroxyl and the 3′-oxygen, enhancing the reactivity of the 2′-hydroxyl towards SHAPE reagent [48].

We used SuperFold [29] to compute a SHAPE-MaP assisted minimum free energy structure, and base pair probabilities of the BVDV IRES RNA by partition function folding. SHAPE reactivities are incorporated into the folding recursions as pseudo energies, following the approach proposed by Deigan et al. [49]. The obtained model was visualized as a colour-coded arch plot (Figure. 2). Well-defined pairings outlined by high pairing probability between 0.8 and 1.0 (green arcs) are visible in regions folding into SL Ia – c and SL IIIa – e. Regions with low pairing probabilities <0.3 (yellow and grey arcs) are scarce and dispersed among residues that, per SHAPE modelling, are predicted to be single-stranded. Overall, the arc model displayed strong agreement

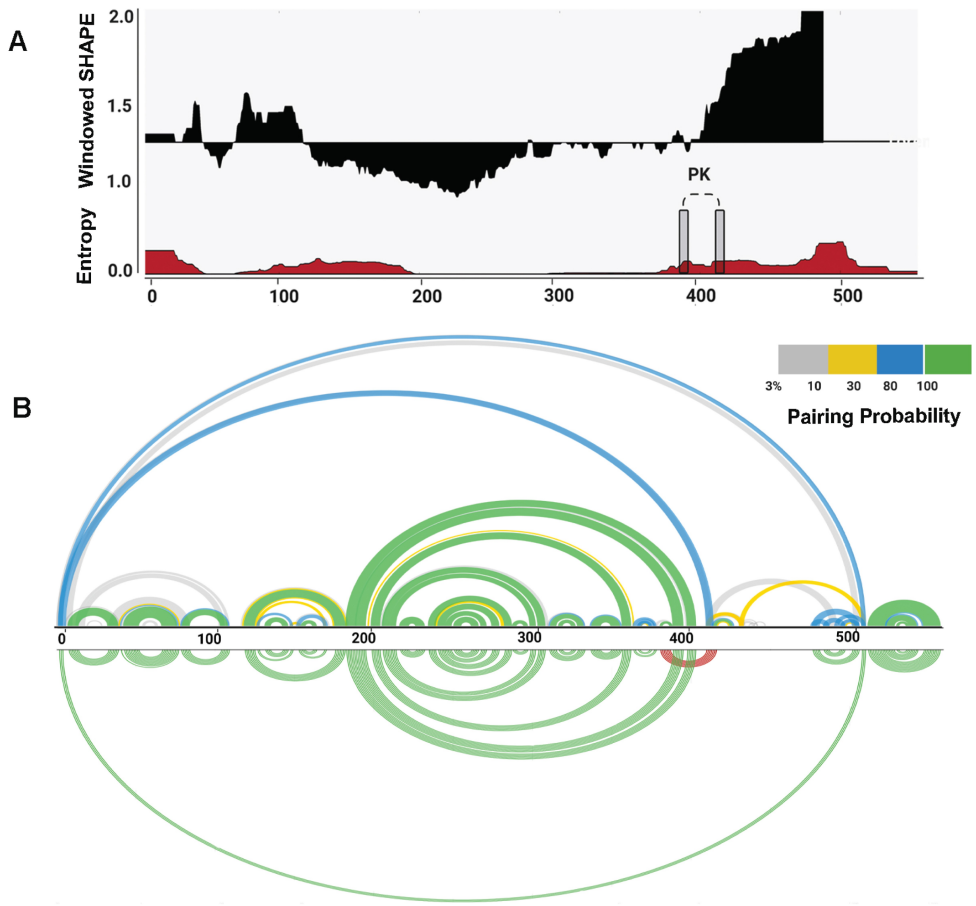


Figure 2. Structural features of BVDV IRES RNA. (A) Windowed SHAPE reactivities (black) and Shannon entropies (brown) are indicated. The X-axis represents the nucleotide position, and Y-axis corresponds to windowed SHAPE reactivities (top) and entropy values (bottom). Single-stranded regions involved in pseudoknot formation are indicated with grey rectangles. (B) Arc plots represent the base-pairing probabilities for BVDV IRES RNA. The top arc plot is colour-coded according to the pairing probability scale. The bottom arc plot includes the predicted H-type pseudoknot (red).

between nucleotide-resolution experimental SHAPE reactivity data and modelled base-pairing. We also assessed how accurately the secondary structure motifs are defined by their sequence and the probing data by calculating the Shannon entropies at nucleotide resolution [36,50]. We noted that the region corresponding to the Domain III (nts 200–380) displays low SHAPE reactivity values, low Shannon entropy values, and high pairing probability, which likely reflects its structural and functional significance. On the other hand, the regions that display low (nts 130–200) and high (nts 420–490) SHAPE reactivity values and high Shannon entropy correspond to Domain II and single-stranded residues of Domain III that represent the formation of a pseudoknot (nts 381–387 and 410–416). These parameters were shown previously to be characteristic of flexible regions that can fold into interconverting RNA structures [51,52].

Evolutionary conservation of pseudoknot in the genus Pestivirus

To obtain deeper insight into the evolutionary conservation of the BVDV IRES region, and in particular conservation of the pseudoknot, we performed a comparative genomics screen within the genus *Pestivirus*. To this end, we selected 13 representative viral genomes, comprising the ICTV-approved *Pestivirus* species A-K, as well as Linda virus and Norway Rat *Pestivirus*, and searched for hits of the Rfam *Pestivirus* IRES (RF00209) covariance model. Only high-quality (high bit score) hits for each input sequence were considered. Multiple sequence alignment (Fig. S6), and consensus structure prediction of these structurally homologous elements revealed rich covariation patterns throughout the entire IRES region (Figure. 3). A test for statistically significantly covarying base pairs with R-scape reported 23 hits, one of them in the pseudoknot.

Comparison of probing-assisted structure model of BVDV IRES RNA with previously proposed models

We performed a comparative analysis of our probing-assisted BVDV IRES RNA structure model and previously proposed models, which revealed both commonalities and differences (Fig. S7). In the BVDV IRES model that was proposed by Deng and Brock, a largely equivalent architecture of domain III computed by early versions of modern RNA structure prediction algorithms highlights its functional importance [53]. Likewise, prediction of pseudoknot formation in domain III on a shorter construct by Le et al., is in agreement with our experimental data [30]. Contrary, domain II exhibits a lower degree of similarity. While our prediction is in favour of

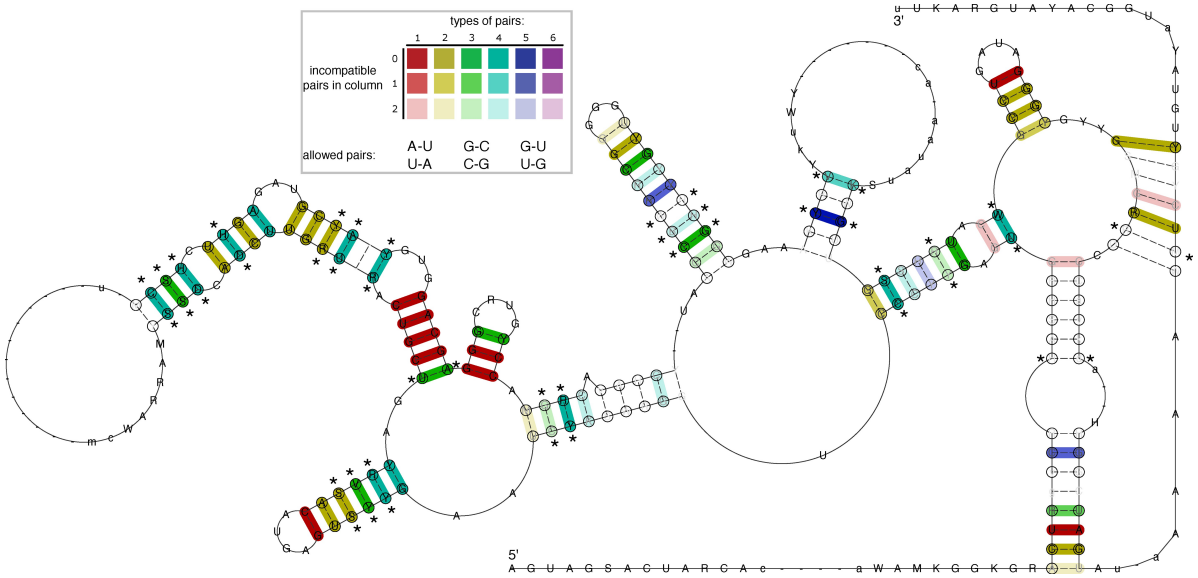


Figure 3. The consensus of IRES Domain III secondary structure prediction generated for 13 *Pestivirus* genomes highlights rich covariation patterns by the different colouring of base pairs, following the RNAalifold colour scheme. Red indicates nucleotide-level sequence conservation, while other colours highlight increasing covariation levels by structure-conserving nucleotide substitutes, as depicted in the insert. The model includes nucleotides with a frequency greater than the average in the underlying alignment in IUPAC notation. Dashes along the consensus sequence indicate positions where the majority of sequences had gaps. Circled nucleotides highlight compensatory mutations, i.e. cases where both nucleotides of a base pair are mutated, such as UA -> GC. Statistical significantly covarying base pairs are marked with an asterisk.

a Y-shaped three-way junction element (Figure. 1), earlier studies reported an extended stem-loop structure that is interspersed with bulges. We used homologous sequences from the 5′ UTR of ICTV-approved *Pestivirus* species A-K to construct a consensus structure of domain II. Interestingly, an extended stem-loop structure is supported by multiple covariations in the consensus model (Fig. S8). Partition function folding and re-evaluation of the corresponding element in our model revealed that domain II encompasses two entities, i.e. a structurally constrained closing stem, as well as a conformationally flexible apical portion that can fold into either an extended stem-loop or a Y-shaped structure (Fig. S9). As these conformations co-exist within a narrow energy range of 1 kcal/mol, we consider the apical portion of domain II is multi-stable. Taken together, earlier models of the BVDV-1 NADL IRES RNA, in particular domain III stay, to a large extent, in agreement with our SHAPE-MaP assisted structural analysis (Supplemental Data: https://github.com/jzs0165/BVDV-Files)

The modelling of BVDV IRES RNA 3D structure

SHAPE-assisted secondary structure prediction reported the existence of an H-type pseudoknot near the basal stem of Domain III. By analysing the predicted secondary structure, one might expect a co-linear arrangement of SL IIIf and IIIg, resulting in coaxial stacking of the two helices. However, under such a scenario, spatial separation of the complementary pseudoknot interaction regions induced by the putatively coaxially stacked helical stems SL IIIf and SL IIIg might be too large to form the pseudoknot. We were intrigued by the question of whether this pseudoknot interaction is sterically feasible in 3D.

To this end, we performed coarse-grained RNA 3D simulations with a purposefully trimmed construct that encompasses relevant parts of Domain III that are involved in pseudoknot formation. Specifically, 5′ and 3′ portions of the original construct were truncated after positions 179 and 425, respectively. Likewise, the four-way junction that contains SL IIId1, SL IIId2, and the substructure comprising SL IIIa, b, and c, were retained as an apical loop element with a connected backbone. This construct, denoted BVDVsegment_180_425 (Fig. S10), features the same structural traits as the original full-length construct. In particular, pKiss single sequence folding predictions suggest the formation of an H-type pseudoknot as expected. Likewise, this truncated construct represents a reasonably small system that allows for detailed tertiary structure analysis *in silico*.

We used ernwin to test whether the proposed secondary structure can be embedded in 3D. Operating on secondary structure elements, ernwin employs a coarse-grained sampling approach and distinguishes between stems, bulges, and various loop types, each represented by cylinders that fit the helix axis. 3D structures are built from fragments of already known RNAs, and tertiary structures are guaranteed to always match the input secondary structure. Building on this concept, we constructed initial 3D models of BVDVsegment_180_425, thereby confirming that the proposed H-type pseudoknot is sterically feasible.

In parallel, we set out to predict the tertiary structure of BVDVsegment_180_425 with SimRNA. We performed massively parallel replica-exchange Monte Carlo simulations to sample the low-energy portion of the underlying energy landscape, using secondary structure restraints that ensure the formation of all helices predicted for BVDVsegment_180_425, including the pseudoknot. A snapshot of a low-energy conformation is shown in (Figure. 4). Clustering of predicted low-energy structures allowed us to characterize two sets of highly compact target

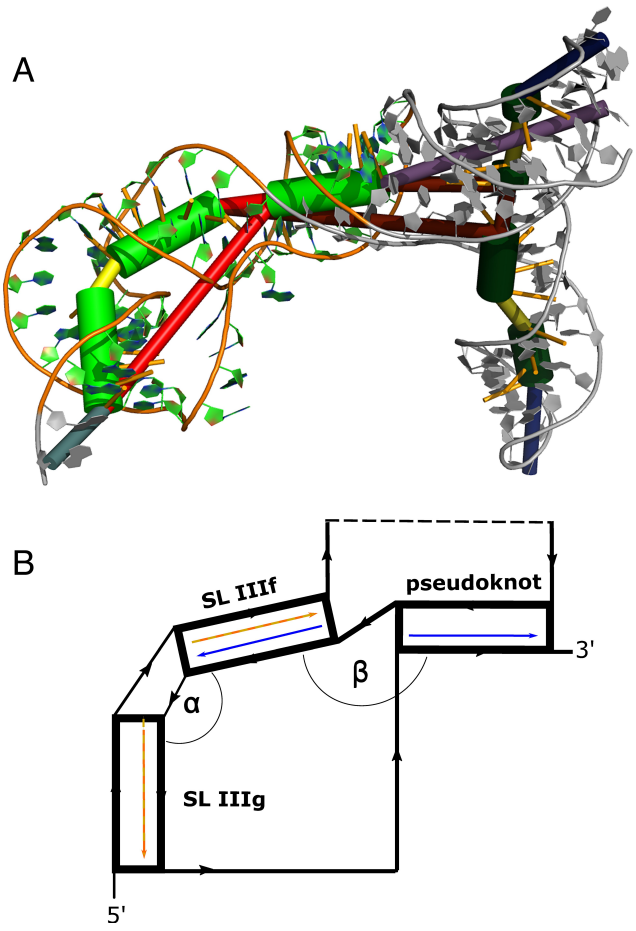


Figure 4. (A) Rendering of a low-energy tertiary structure snapshot of the truncated construct BVDVsegment_180_425 overlaid with coarse-grained 3D elements. Green cylinders highlight helical regions, with the rightmost representing H-type pseudoknot, and the middle and leftmost green cylinders portraying stem-loops IIIf and IIIg, respectively. The bulge between SL IIIf and IIIg is depicted in yellow, and the multiloop enclosed by the formation of the pseudoknot is plotted as a long red connection. The short red connection between the pseudoknot helix and SL IIIf represents a single unpaired C between these helices. Other structural elements are greyed out. The quasi-colinear arrangement of the pseudoknot and SL IIIf is noticeable, while the bulge induces a marked twist between SL IIIf and IIIg. (B) Schematic representation of the three helices encompassing the pseudoknot, SL IIIf, and IIIg, sketching their approximate spatial arrangement. Angles between adjacent helical segments are defined by stem vectors, i.e. angle α between SL IIIf and SL IIIg (orange and red vectors), and β between SL IIIf and the pseudoknot helix (blue vectors). The directionality of the backbone is indicated by black arrows. Regions of BVDVsegment_180_425, which are not directly involved in the spatial arrangement of these helices, are depicted as a dashed segment along the backbone.

structures, 38 in total, that differ in the size of the bulge between helical segments IIIf and IIIg, i.e. 2 nt in set 1 (17 candidate structures) and 3 nt in set 2 (21 candidate structures). The sets of candidate structures were then subjected to follow-up analysis with forgi [54], focusing on the angles between the pseudoknot helix and SL IIIf (angle β) and SL IIIf and SL IIIg (angle α). As shown in (Figure. 5), both angles are distributed in well-defined intervals for all 38 low-energy structures. While α falls between 37 and 149 degrees, β is dispersed between 129 and 174 degrees. This suggests that all 3D candidate structures are well-defined and that there is nearly quasi-coaxial stacking between the pseudoknot helix and SL IIIf. Likewise, our data indicate that there is a marked twist between SL IIIf and IIIg. Notably, structures with a 2 nt bulge are less arched than those with a 3 nt bulge, which is in good agreement with their expected behaviour. An important finding of this analysis is that the pseudoknot can only be formed when SL IIIf and SL IIIg are oriented at an acute angle, facilitated by the central bulge.

Discussion

RNA viruses are known to use exceptionally thrifty molecular strategies to optimize their replication cycle. They often contain functional RNA motifs within non-coding regions that control crucial steps of their infectivity cycle and structural switches that aid their regulation. The IRES RNA within the 5′ UTRs of BVDV is one of such complex structural domains known to govern the efficiency of cap-independent translation [23]. In this study, we applied RNA structural probing method and computational 3D modelling to gain insight into the secondary and tertiary structure of BVDV IRES RNA.

Overall, BVDV IRES RNA folds into three major structural domains (I–III), each containing well-defined motifs, i.e. hairpins and stems, interspersed among less structurally constrained regions, such as multi-way junctions and loops. This model stays, to a large extent, in agreement with the previously proposed secondary structure models for the 5′ and 3′ UTRs of BVDV strains, i.e. NADL, OSLOSS and SD-1 [5,30,53]. In particular, a good agreement for domain III folding has been found through the comparison of previous BVDV IRES structural models with our probing-assisted model. However, we have found an alternative folding of domain II into a Y-shaped hairpin instead of an extended hairpin, which can be justified by thermodynamic parameters, i.e. higher Shannon entropy and a small minimum free energy difference of 1 kcal/mol between both structures, which support the structural fluidity of domain II.

The key finding relates to the most structurally well-defined domain III (low SHAPE reactivities, low Shannon entropy values, high pairing probability), which includes two single-stranded regions and manually represented H-type pseudoknot. The 3D modelling of the BVDV IRES RNA provides further insight into the structural arrangements of individual motifs, predicting quasi-coaxial stacking between SL IIIf and the pseudoknot helix. These coaxial stackings can be critical for stabilizing the IRES pseudoknot and can be formed only if the angle between SL IIIg and SL IIIf is acute. Similar stabilization processes have been shown for T2 bacteriophage mRNA of gene 32 [55], turnip yellow mosaic virus (TYMV) [56], and simian retrovirus-1 (SRV-1) RNA pseudoknots [57]. The large junction connecting the individual RNA motifs of domains I–III likely provides a suitable movement and intrinsic flexibility to the single-stranded regions within domain III, which support the formation of a pseudoknot.

Previously, the formation of a pseudoknot in BVDV IRES RNA has been proposed based on thermodynamic predictions and Monte Carlo simulations [30,31]. The comparative phylogenetic analysis performed on BVDV IRES RNA sequences from genotypes 1 and 2 provided further evidence supporting the formation of a pseudoknot [5]. These studies showed that the pseudoknot engages between four and seven nucleotides,

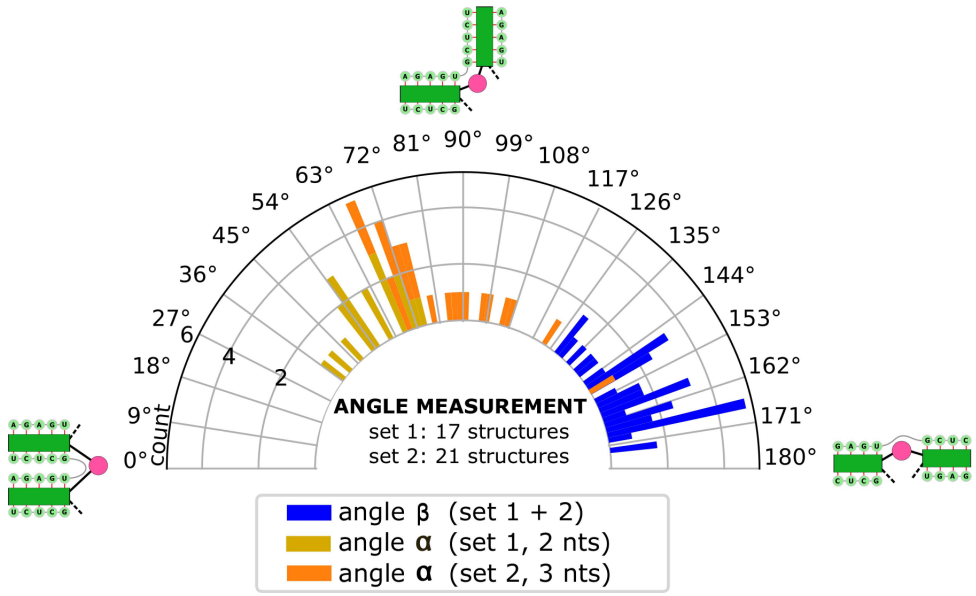


Figure 5. Distribution of angles α and β in the low-energy 3D structure of BVDVsegment_180_425. The relative orientation of helices was assessed by a helix-centred representation of RNA 3D structure. As stretched by the relative geometry of green helical segments, angles close to 0° account for parallel stems, whereas angles close to 180° indicate collinear arrangement and can be indicative of coaxial stacking. Analysis of candidate structures from SimRNA simulations of BVDVsegment_180_425 shows that angles between helices are distributed in well-defined intervals. Set 1 and set 2 comprise low-energy structures predicted by SimRNA, whose helices SL IIIf and IIIg are separated by a bulge of 2 and 3 nts, respectively. For set 1, α lies between 37 and 75 degrees, while for set 2, values are distributed between 66 and 149 degrees. This is expected, as a larger bulge allows for more conformational flexibility of the two adjacent helices. Contrary, β falls in a range between 130 and 174 degrees for set 1 and 129 and 173 degrees for set 2. Values for both sets were combined and depicted in blue.

depending on the BVDV strain, with residues 5′-CAGA-3′ and 5′-UCUG-3′ being shared by all isolates. Here, we performed a computational screen of the IRES domain III in more than a dozen phylogenetically related viruses. Consensus structure prediction revealed rich covariation support in this evolutionarily conserved region of the *Pestivirus* IRES. Importantly, we could identify statistically significant covariation in 22 base pairs, including the pseudoknot, thus providing computational evidence for the crucial role of the pseudoknot in IRES functionality.

Site-directed mutagenesis analysis indicated that BVDV IRES RNA constructs with impaired base-pairing within these residues lost their translational activity, while compensatory mutations reversed that effect [22]. Furthermore, mutational and structural studies suggested the formation of structurally similar pseudoknots in the IRES of CSFV and HCV RNAs [58,59]. Additionally, it has been shown that the BVDV and HCV 5′ UTRs share a considerable extent of sequence and structure similarity [30,57,60]. The shared features include the formation of multiple SLs, which are globally organized into three domains surrounding the centrally positioned pseudoknot [61], with HCV IRES folding into additional domain IV, positioned externally [62]. Interestingly, in both HCV and BVDV IRES RNAs, domain III includes hairpins IIIa, IIId1, and IIIe, which apical loops, i.e. tetra-loop, G-rich loop, and penta-loop have highly conserved nucleotide sequence [19].

Although the relationship between the IRES pseudoknot and initiation codon in BVDV RNA has not been addressed, previous studies performed on HCV IRES RNA indicated that the formation of a pseudoknot leads to the correct positioning of mRNA open reading frame (ORF) in the 40S ribosomal binding cleft [63]. In HCV IRES, it was shown by SHAPE probing that upon binding of the 40S subunit P-site to the AUG codon, the residues at and beyond the start codon show a substantial increase in the flexibility of the translational initiation facilitation [64]. Similar structural probing experiments in the presence and absence of ribosomes can be performed to understand the effects of the translational machinery assembly on BVDV IRES RNA. Additionally, various therapeutic strategies involving the identification of small molecule inhibitors, antisense oligonucleotides, ribozymes, which target and disrupt the HCV IRES RNA and prevent the binding of translational factors, abolishing the viral replication, have also been developed [65]. Considering the similarity between HCV and BVDV IRES RNA, analogous strategies can be applied for interfering with BVDV replication.

Materials and methods

Cloning IRES from BVDV-NADL into pIW-IRES (eMS2hp)-EGFP expression vector

Oligonucleotide primers used in the study were purchased (IDT) and are listed in (Table S2). All DNA products were purified from agarose gels and eluted using the elution kit (Qiagen, 28,706). Quick Ligase (NEB, M2200L) was used in all ligation reactions. PCR reactions were carried using Q5 Hot Start Polymerase (NEB, M0493S).

Cloning IRES under CMV promoter control

Inactivated BVDV-NADL virus was obtained from the laboratory of Paul Walz (Veterinary School, Auburn University). Viral RNA was extracted from 10^6 viral particles following

Ribozol extraction protocol (VWR Life Science, 97,064–950). Two hundred and fifty microlitres of virus suspension was mixed with three volumes of RiboZol and incubated for 5 min at room temperature. With an addition of 200 µl chloroform, the sample was shaken for 15 s and incubated for 15 min at room temperature. Phases were separated by centrifugation at 12,000 × *g* for 15 min in a cold microfuge. Six hundred microlitres aqueous phase was purified using Zymo Clean & Concentrator-25 (Zymo, R1018). Purified RNA (~1.2 µg) was used as a template for the synthesis of 461 bp DNA fragment coding for the 5′ UTR sequence and 25 codons of the first viral gene Npro. The oligonucleotide primers were designed using BVDV-NADL. cDNA was synthesized using 40 ng of viral DNA and 1 µM reverse primer IRES-NADL-rev in 20 µl reaction with SuperScript IV reverse transcriptase (Thermo Fisher Scientific, 18090010). The 461 bp DNA fragment was amplified using IRES-NADL forward and reverse primers and Q5 Hot Start polymerase (NEB, M0493S). The second round of PCR amplification extended the 461 bp fragment to include the KpnI restriction site at the 5′ end and the BamHI site at the 3′ end. Amplification with primer pair T7-IRES-for-3 and T7-NADL-rev yielded 509 bp DNA fragment, which was then digested with KpnI-HF (NEB, R3142S) and BamHI-HF (NEB, R3136S). The final 463 bp fragment was ligated with the pcDNA3.1(+)circ RNA Mini Vector (Addgene, 60,648) digested with KpnI-HF and BamHI-HF. Clones were selected and maintained in *E. coli* strain NEB5-alpha (NEB, C2987I). All constructs were verified by Sanger sequencing (Eurofins). New plasmid was designated pIW-IRES.

Cloning gene for EGFP protein downstream from IRES

Plasmid pIRES-EGFP-puro (Addgene, 45567-DNA.cg) was digested with BstXI (NEB, R0113S) and BsrGI (NEB, R0575S), and 710 bp DNA fragment coding for EGFP protein was prepared for ligation. Plasmid pIW-IRES was digested with BamHI-HF (NEB, R3136S) and ApaI (NEB, R0114S). Insertion of the BstXI-BsrGI fragment with EGFP between BamHI and ApaI sites on the core plasmid was performed using the following adapters: A-5′EGFP-adapt-top and B-5′ EGFP-adapt-bott for the 5′ end of the insert plus C-3′EGFP-top and D-3′EGFP-bott for the 3′ end of the insert. The 5′ end adapter pair added an EcoRI restriction site, while the 3′ end pair added an EcoRV site. Double digestion with EcoRI (NEB, R0101S) and EcoRV (NEB, R0195S) was used for the initial screening of clones. All putative correct clones were verified by sequencing. New plasmid was designated pIW-IRES-EGFP.

Insertion of eMS2 hp at 3′ end of IRES

The design of extended hairpin binding bacteriophage MS2 protein was described previously. Plasmid pIW-IRES-EGFP was linearized with BamHI-HF (the 3′ end of IRES) and BstXI (the 5′ end of EGFP). Gblock spanning the 3′ end of IRES, extended protein MS2-binding hairpin (eMS2hp), and the 5′ portion of EGFP were purchased (IDT) and amplified with primers R17L-for and R17L-rev gblock was digested with BamHI and BstXI and ligated to linearized pIW-IRES-EGFP. The sequence of plasmid pIW-IRES(eMS2hp)-EGFP was verified by sequencing.

RNA isolation and purification

RNA was extracted with TRIzol reagent (Thermo Fisher Scientific, 15596018) followed by the RNA Clean and Concentrator kit (Zymo Research, 76020–604) and DNase treatment provided with the kit according to manufacturer's protocols. The RNA was eluted in 15 µl RNAse free water and stored at −80°C.

In vitro transcription

Plasmid pIW-IRES/eMS2hp-EGFP was linearized either with EcoRI for synthesis of IRES/eMS2hp, or with EcoRV for synthesis of IRES/eMS2-EGFP-mRNA. 100 µl transcription reaction consisted of 2.5 µg linearized plasmid, 200 mM HEPES-KOH, 7.5, 30 mM $MgCl_2$, 40 mM dithiothreitol (DTT), 1 mM spermidine, 7.5 mM NTP, SUPERNase-In (Ambion) – 20 U/100 µl reaction, inorganic pyrophosphatase (NEB, M0361S) – 0.1 U/100 µl reaction and 100 µg/ml T7 RNA polymerase. The reactions were incubated for 2 h at 37°C with gentle rotation.

BVDV IRES RNA SHAPE-MaP

The sequence corresponding to the BVDV IRES was divided into two overlapping zones (nts 1–283 and 217–504) that were reverse transcribed and amplified with zone-specific RT and PCR primers (Table S3). Five picomoles of BVDV IRES *in vitro* transcript (either short, 560 nts or long, 1296 nts) was heated at 95°C for 3 min and slowly cooled to 4°C (0.1°C/s). The *in vitro* transcripts were folded in a final volume of 36 µl in 1X folding buffer containing 50 mM Tris–HCl pH 8.0, 100 mM NaCl, 5 mM $MgCl_2$, and incubated at 37°C for 15 min. The folded transcripts were divided into two reactions: positive and negative (18 µl each) with one sample treated with 2 µl of 100 mM 1-methyl-7-nitroisatoic anhydride (1M7) (DC Chemicals, 73043–80-8) for positive (+) reaction (final 1M7 concentration of 10 mM), and the second sample treated with 2 µl dimethyl sulphoxide (DMSO) for negative (-) reaction. Both tubes were incubated at 37°C for 5 min to facilitate RNA modification. Additionally, the denaturing control (DC) was prepared and included 5 pmol of transcript in 3 µl TE-like buffer (0.1 mM EDTA, 10 mM Tris-HCl, pH 7.5), 5 µl formamide, and 1 µl denaturing buffer (1 M HEPES, pH 8, 0.5 M EDTA), that were mixed and incubated at 95°C. To 9 µl of the denatured transcript, 1 µl of 100 mM 1M7 was added and incubated at 95°C for 1 min. All samples were purified using ethanol precipitation with 5 mg/ml glycogen (Thermo Fisher Scientific, AM9510). After modification, RNA was reverse transcribed at 42°C for 3 h using zone-specific RT primers in the SHAPE-MaP buffer, including 50 mM Tris pH 8, 75 mM KCl, 10 mM dithiothreitol (DTT), 6 mM $MnCl_2$ (Sigma-Aldrich, 7773–01-5), 0.7 mM dNTPs with 10 U SuperScript II reverse transcriptase (Invitrogen, 18,064,022). The obtained cDNA products were purified using Mag-Bind

Total Pure NGS beads (Omega Bio-Tek, M1378-01). Each sample was amplified separately by two PCRs using Q5 high-fidelity polymerase (NEB, M0149S) and zone-specific PCR primers. PCR 1 included 15 cycles: 98°C, 30 s; 98°C, 10 s; 55°C, 30 s; 72°C, 30 s; 72°C, 2 min. The obtained products were checked on 0.8% agarose gel, purified using Mag-Bind Total Pure NGS beads, and subjected to the second round of PCR amplification using 15 cycles: 98°C, 30 s; 98°C, 10 s; 55°C, 30 s; 72°C, 30 s; 72°C, 2 min. The final libraries were purified on 2% preparative agarose gel by Monarch DNA Gel Extraction kit (NEB, T1020L). The concentration of each library was measured by performing quantitative PCR (qPCR) using PerfeCTa® NGS Quantification Kit (Quanta Biosciences 10029–558) and following the manufacturer's protocol. Purified amplicon libraries were pooled and sequenced on Illumina MiSeq Instrument.

SHAPE-assisted structure predictions

The dataset obtained from Illumina MiSeq was saved into separate FASTQ files for each of the respective zones and reactions. A custom Python script 'trimmer.py' was used to trim raw sequences and remove sequencing artefacts. The FASTQ files corresponding to the (+), (-) and DC reactions were concatenated and processed by ShapeMapper v1.2 [66] to obtain mutation counts, sequence read depth and normalized reactivity values in graphical and CSV format, which help detect potential experimental issues. SuperFold v1.1 [29] was utilized to determine the partition function, minimum free energy and structural analysis. Shannon entropy was computed from base-pairing probabilities analysed during SuperFold partition function calculation with a partition Window size of 1200 nts and step size of 100 nts. The local median SHAPE reactivity and Shannon Entropy were calculated over a 55 nt sliding window with a fold step size of 300. RNAstructure v6.0 [67] was used for generating the secondary structure models by incorporating SHAPE reactivities as pseudo-energy constraints using slope and intercept values of 2.6 and −0.8, respectively. To determine the presence of a pseudoknot in the IRES RNA, we incorporated the reactivity data into the SHAPEknot v5.8.1 algorithm with the default parameters of SHAPEintercept = −0.8 and SHAPEslope = 2.6. VARNA applet v3-93, ViennaRNA's RNAplot v2.4.17 and R-chie webserver [68] were used to visualize RNA secondary structures and arc plots.

Comparative genomics screen within the genus Pestivirus

Complete genomes of all ICTV-listed *Pestivirus* species (A-K), as well as two unclassified Pestiviruses, i.e. Linda virus and Norway rat *Pestivirus* were downloaded from the NCBI GenBank database. The cmsearch tool from the Infernal suite v1.1.4 [69] was used to screen the Rfam covariance model RF00209 (*Pestivirus* internal ribosome entry site) against the library of 13 *Pestivirus* genomes, yielding structurally homologous elements within the 5′ UTR in each genome (Table S4). These were aligned with mafft v7.475 [70] and subjected to consensus structure prediction with RNAalifold v2.4.17 [71] from the ViennaRNA Package v2.4.17 [72]. As RNAalifold cannot directly predict structures that contain pseudoknots, a consensus structure was predicted in two steps. First, a nested, i.e. pseudoknot-free consensus structure was predicted, thereby constraining the regions involved in pseudoknot formation to being unpaired. In a second step, only the pseudoknot was allowed to form, with all other base pairs constrained to being unpaired. A combination of the two predictions yielded the anticipated consensus structure that involves an H-type pseudoknot. Statistical significance of covariation was assessed with R-scape v1.5.10 [73].

Three-dimensional (3D) structure predictions of BVDV IRES RNA

RNA tertiary structure predictions were performed on a trimmed 106 nt construct that encompasses the portion of the BVDV1-NADL IRES that is involved in pseudoknot formation. The construct was verified to form the expected H-type pseudoknot with pKiss v2.2.13 [74]. To test whether the predicted H-type pseudoknot is sterically feasible in 3D, we used ernwin v1.0.1, a tool for sampling coarse-grained 3D structures from fragments of known RNA tertiary structures. Moreover, we performed 3D structure predictions with SimRNA v3.20. This Monte Carlo sampling approach uses a five-bead system per nucleotide to model a coarse-grained representation of RNA structure and employs an empirically derived knowledge-based potential. SimRNA was run for 16 million iterations in replica-exchange mode. The resulting trajectories were subjected to a clustering procedure, thereby filtering sets of similar structures in terms of root mean square deviation (RMSD) within 1% of all trajectories with the lowest energy. Two sets of low-energy candidate structures, comprising 17 and 21 structures, were then evaluated for specific helix angles with forgi v2.0 [54], a library for analysing RNA tertiary structures. The coarse-grained representations from ernwin and SimRNA, together with the structure probing data, allowed us to assess the feasibility and the structural characteristics of the detected pseudoknot. RNA 3D structures were visualized in PyMol v2.4.2 [75].

Accession numbers

The high-throughput sequencing data generated in this study are deposited into SRA databases with bio-sample accession number: PRJNA725417, BVDV-NADL accession number: NC_011461.

Acknowledgments

We would like to thank the Alabama Agricultural Experiment Station, Hatch Funding Program for the funding. J.S.S and D.G. are also supported by start-up funds from the Department of Biological Sciences, College of Science and Mathematics, and Office of the Vice President for Research, Auburn University. This work was partly funded by the Austrian Science Fund (FWF) Doctoral College W 1207 "RNA Biology". We would also like to thank University of Vienna for the Open Access funding.

Funding

This work was funded by the Alabama Agriculture Experiment Station, Hatch Funding program and partly funded by the Austrian Science Fund (FWF) Doctoral College W 1207 "RNA Biology".

Data availability

Custom R and Python scripts for data analysis are available at https://github.com/jzs0165/BVDV-Files

Disclosure statement

No potential conflict of interest was reported by the author(s).
Supplemental data for this article can be accessed online at https://doi.org/10.1080/15476286.2022.2058818

ORCID

Ivo L. Hofacker http://orcid.org/0000-0001-7132-0800
Michael T. Wolfinger http://orcid.org/0000-0003-0925-5205
Joanna Sztuba-Solinska http://orcid.org/0000-0002-5227-6629

References

[1] Khodakaram-Tafti A, Farjanikish GH. Persistent bovine viral diarrhea virus (BVDV) infection in cattle herds. Iran J Vet Res. 2017;18:154–163.
[2] Simmonds P, Becher P, Bukh J, et al. ICTV virus taxonomy profile: flaviviridae. J Gen Virol. 2017;98:2–3.
[3] Houe H. Epidemiological features and economical importance of bovine virus diarrhoea virus (BVDV) infections. Vet Microbiol. 1999;64:89–107.
[4] Vilček Š, Paton DJ, Durkovic B, et al. Bovine viral diarrhoea virus genotype 1 can be separated into at least eleven genetic groups. Arch Virol. 2001;146:99–115.
[5] Burks JM, Zwieb C, Müller F, et al. Comparative structural studies of bovine viral diarrhea virus IRES RNA. Virus Res. 2011;160:136–142.
[6] Eterhans EP, Achofen CB, Talder HS, et al. Cytopathic bovine viral diarrhea viruses (BVDV): emerging pestiviruses doomed to extinction. Vet Res. 2010;41:44.
[7] Tsukiyama-Kohara K, Iizuka N, Kohara M, et al. Internal ribosome entry site within hepatitis C virus RNA. J Virol. 1992;66:1476–1483.
[8] Rijnbrand R, Bredenbeek P, van der Straaten T, et al. Almost the entire 5′ non-translated region of hepatitis C virus is required for cap-independent translation. FEBS Lett. 1995;365:115–119.
[9] Wilson JE, Powell MJ, Hoover SE, et al. Naturally occurring dicistronic cricket Paralysis virus RNA is regulated by two Internal Ribosome Entry Sites. Mol Cell Biol. 2000;20:4990–4999.
[10] Kühn R, Luz N, Beck E. Functional analysis of the internal translation initiation site of foot-and-mouth disease virus. J Virol. 1990;64:4625–4631.
[11] Pelletier J, Sonenberg N. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. Nature. 1988;334:320–325.
[12] Thakor N, Holcik M. IRES-mediated translation of cellular messenger RNA operates in eIF2 a -independent manner during stress. Nucleic Acids Res. 2011;40:541–552.
[13] Martinez-salas E, Francisco-Velilla R, Fernandez-Chamorro J. Insights into structural and mechanistic features of viral IRES elements. Front Microbiol. 2018;8:1–15.
[14] Burks JM, Zwieb C, Müller F, et al. In silico analysis of IRES RNAs of foot-and-mouth disease virus and related picornaviruses. Arch Virol. 2011;156:1737–1747.
[15] Komar AA, Hatzoglou M. Exploring internal ribosome entry sites as therapeutic targets. Front Oncol. 2015;5:1–10.
[16] Ghassemi F, Madadgar O, Roohvand F, et al. Translational efficiency of BVDV IRES and EMCV IRES for T7 RNA polymerase driven cytoplasmic expression in mammalian cell lines. Mol Biol. 2017;51:324–333.
[17] Yu Y, Abaeva IS, Marintchev A, et al. Common conformational changes induced in type 2 picornavirus IRESs by cognate trans-acting factors. Nucleic Acids Res. 2011;39:4851–4865.
[18] Yamamoto H, Collier M, Loerke J, et al. Molecular architecture of the ribosome-bound hepatitis C virus internal ribosomal entry site RNA. EMBO J. 2015;34:3042–3058.
[19] Lukavsky PJ. Structure and function of HCV IRES domains. Virus Res. 2009;139:166–171.
[20] Shen LX, Tinoco I. The structure of an RNA pseudoknot that causes efficient frameshifting in mouse mammary tumor virus. J Mol Biol. 1995;247:963–978.
[21] Brierley I, Pennell S, Gilbert RJC. Viral RNA pseudoknots: versatile motifs in gene expression and replication. Nat Rev Microbiol. 2007;5:598–610.
[22] Moes L, Wirth M. The internal initiation of translation in bovine viral diarrhea virus RNA depends on the presence of an RNA pseudoknot upstream of the initiation codon. Virol J. 2007;4:124.
[23] Pestova TV, Hellen CUT. Internal initiation of translation of bovine viral diarrhea virus rna. Virology. 1999;258:249–256.
[24] Fukushi S, Okada M, Stahl J, et al. Ribosomal protein S5 interacts with the internal ribosomal entry site of hepatitis C virus. J Biol Chem. 2001;276:20824–20826.
[25] Otto GA, Lukavsky PJ, Lancaster AM, et al. Ribosomal proteins mediate the hepatitis C virus IRES-HeLa 40S interaction. RNA. 2002;8:913–923.
[26] Laletina E, Graifer D, Malygin A, et al. Proteins surrounding hairpin IIIe of the hepatitis C virus internal ribosome entry site on the human 40S ribosomal subunit. Nucleic Acids Res. 2006;34:2027–2036.
[27] Babaylova E, Graifer D, Malygin A, et al. Positioning of subdomain IIId and apical loop of domain II of the hepatitis C IRES on the human 40S ribosome. Nucleic Acids Res. 2009;37:1141–1151.
[28] Smola MJ, Calabrese JM, Weeks KM. Detection of RNA-protein interactions in living cells with SHAPE. Biochemistry. 2015;54:6867–6875.
[29] Smola MJ, Rice GM, Busan S, et al. Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. Nat Protoc. 2015;10:1643–1669.
[30] Le SY, Siddiqui A, Maizel JV. A common structural core in the internal ribosome entry sites of picornavirus, hepatitis C virus, and pestivirus. Virus Genes. 1996;12:135–147.
[31] Le SY, Sonenberg N, Maizel JV. Unusual folding regions and ribosome landing pad within hepatitis C virus and pestivirus RNAs. Gene. 1995;154:137–143.
[32] Kerpedjiev P, Höner Zu Siederdissen C, Hofacker IL. Predicting RNA 3D structure using a coarse-grain helix-centered model. RNA. 2015;21:1110–1121.
[33] Boniecki MJ, Lach G, Dawson WK, et al. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. Nucleic Acids Res. 2015;44:1–20.
[34] Mortimer SA, Weeks KM. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. J Am Chem Soc. 2007;129:4144–4145.
[35] Sztuba-Solinska J, Rausch JW, Smith R, et al. Kaposi's sarcoma-associated herpesvirus polyadenylated nuclear RNA: a structural scaffold for nuclear, cytoplasmic and viral proteins. Nucleic Acids Res. 2017;45:6805–6821.
[36] Siegfried NA, Busan S, Rice GM, et al. RNA motif discovery by SHAPE and mutational profiling (SHAPE-MaP). Nat Methods. 2014;11:959–965.
[37] Ghassemi F, Madadgar O, Roohvand F, et al. Translational efficiency of BVDV IRES and EMCV IRES for T7 RNA polymerase

driven cytoplasmic expression in mammalian cell lines. Mol Biol. 2017;51:283–292.
[38] Ghaderi M, Sabahi F, Sadeghi-Zadeh M, et al. Construction of an eGFP expression plasmid under control of T7 promoter and IRES sequence for assay of T7 RNA polymerase activity in mammalian cell lines. Iran J Cancer Prev. 2014;7:137–141.
[39] Mendez E, Ruggli N, Collett MS, et al. Infectious Bovine viral Diarrhea virus (Strain NADL) RNA from stable cDNA clones: a cellular insert determines ns3 production and viral cytopathogenicity. J Virol. 1998;72:4737–4745.
[40] Hurst T, Xu X, Zhao P, et al. Quantitative understanding of SHAPE mechanism from RNA structure and dynamics analysis. J Phys Chem B. 2018;122:4771–4783.
[41] Stallings SC, Moore PB. The structure of an essential splicing element: stem loop IIa from yeast U2 snRNA. Structure. 1997;5:1173–1185.
[42] Moore PB. Structural motifs in RNA. Annu Rev Biochem. 1999;68:287–300.
[43] Proctor DJ, Schaak JE, Bevilacqua JM, et al. Isolation and characterization of a family of stable RNA tetraloops with the motif YNMG that participate in tertiary interactions. Biochemistry. 2002;41:12062–12075.
[44] Paoletti AC, Shubsda MF, Hudson BS, et al. Affinities of the nucleocapsid protein for variants of SL3 RNA in HIV-1. Biochemistry. 2002;41:15423–15428.
[45] Wilkinson KA, Vasa SM, Deigan KE, et al. Influence of nucleotide identity on ribose 2′-hydroxyl reactivity in RNA. RNA. 2009;15:1314–1321.
[46] Mlýnský V, Bussi G. Molecular dynamics simulations reveal an interplay between SHAPE reagent binding and RNA flexibility. J Phys Chem Lett. 2018;9:313–318.
[47] McGinnis JL, Dunkle JA, Cate JHD, et al. The mechanisms of RNA SHAPE chemistry. J Am Chem Soc. 2012;134:6617–6624.
[48] Vicens Q, Gooding AR, Laederach A, et al. Local RNA structural changes induced by crystallization are revealed by SHAPE. RNA. 2007;13:536–548.
[49] Deigan KE, Li TW, Mathews DH, et al. Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci. U S A. 2009;106:97–102.
[50] Huynen M, Gutell R, Konings D. Assessing the reliability of RNA folding using statistical mechanics. J Mol Biol. 1997;267:1104–1112.
[51] Huber RG, Lim XN, Ng WC, et al. Structure mapping of dengue and Zika viruses reveals functional long-range interactions. Nat Commun. 2019;10:1408.
[52] Mauger DM, Golden M, Yamane D, et al. Functionally conserved architecture of hepatitis C virus RNA genomes. Proc Natl Acad Sci U S A. 2015;112:3692–3697.
[53] Deng R, Brock KV. 5′ and 3′ untranslated regions of pestivirus genome: primary and secondary structure analyses. Nucleic Acids Res. 1993;21:1949–1957.
[54] Thiel BC, Beckmann IK, Kerpedjiev P, et al. 3D based on 2D: calculating helix angles and stacking patterns using forgi 2.0, an RNA Python library centered on secondary structure elements. F1000Res. 2019;8:1–16.
[55] Holland JA, Hansen MR, Du Z, et al. An examination of coaxial stacking of helical stems in a pseudoknot motif: the gene 32 messenger RNA pseudoknot of bacteriophage T2. RNA. 1999;5:257–271.
[56] Dumas P, Moras D, Florentz C, et al. 3-d graphics modelling of the tRNA-like 3'-end of turnip yellow mosaic virus RNA: structural and functional implications. J Biomol Struct Dyn. 1987;4:707–728.
[57] Sung D, Kang H. Mutational analysis of the RNA pseudoknot involved in efficient ribosomal frameshifting in simian retrovirus-1. Nucleic Acids Res. 1998;26:1369–1372.
[58] Fletcher SP, Jackson RJ. Pestivirus Internal Ribosome Entry Site (IRES) structure and function: elements in the 5′ untranslated region important for IRES function. J Virol. 2002;76:5024–5033.
[59] Wang C, Le SY, Ali N, et al. An RNA pseudoknot is an essential structural element of the internal ribosome entry site located within the hepatitis C virus 5′ noncoding region. RNA. 1995;1:526–537.
[60] Spahn CMT, Kieft JS, Grassucci RA, et al. Hepatitis C virus IRES RNA-induced changes in the conformation of the 40S ribosomal subunit. Science. 2001;291:1959–1962.
[61] Grassmann CW, Yu H, Isken O, et al. Hepatitis C virus and the related bovine viral diarrhea virus considerably differ in the functional organization of the 5′ non-translated region: implications for the viral life cycle. Virology. 2005;333(2):349–366.
[62] Lemon SM, Honda M. Internal ribosome entry sites within the RNA genomes of hepatitis C virus and other flaviviruses. Semin Virol. 1997;8:274–288.
[63] Berry KE, Waghray S, Doudna JA. The HCV IRES pseudoknot positions the initiation codon on the 40S ribosomal subunit. RNA. 2010;16(8):1559–1569.
[64] Berry KE, Waghray S, Mortimer SA, et al. Crystal structure of the HCV IRES central domain reveals strategy for start-codon positioning. Structure. 2011;19(10):1456–1466.
[65] Pawlotsky JM, Chevaliez S, McHutchison JG. The hepatitis C virus life cycle as a target for new antiviral therapies. Gastroenterology. 2007;132(5):1979–1998.
[66] Busan S, Weeks KM. Accurate detection of chemical modifications in RNA by mutational profiling (MaP) with ShapeMapper 2. RNA. 2018;24(2):143–148.
[67] Bellaousov S, Reuter JS, Seetin MG, et al. RNAstructure: web servers for RNA secondary structure prediction and analysis. Nucleic Acids Res. 2013;41(W1):W471–W474.
[68] Tsybulskyi V, Mounir M, Meyer IM. R-chie: a web server and R package for visualizing cis and trans RNA–RNA, RNA–DNA and DNA–DNA interactions. Nucleic Acids Res. 2020;48(18):E105–E105.
[69] Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–2935.
[70] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772–780.
[71] Bernhart SH, Hofacker IL, Will S, et al. RNAalifold: improved consensus structure prediction for RNA alignments. In: BMC bioinformatics. BioMed Central Ltd. 2008. p. 9.
[72] Lorenz R, Bernhart SH, Höner Zu Siederdissen C, et al. ViennaRNA package 2.0. Algorithms Mol Biol. 2011;6(1). DOI:10.1186/1748-7188-6-26
[73] Rivas E, Clements J, Eddy SR. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. Nat Methods. 2017;14(1):45–48.
[74] Janssen S, Giegerich R. The RNA shapes studio. Bioinformatics. 2015;31(3):423–425.
[75] Yuan S, Chan HCS, Hu Z. Using PyMOL as a platform for computational drug design. Wiley Interdiscip Rev Comput Mol Sci. 2017;7(2):1–17.

III. Summary and Outlook

The function of an RNA is critically dependent on its structural characteristics and its interactions with other RNAs. The tertiary structure elements of RNA increase the complexity of structure prediction, but they are an elementary part of an extremely well-considered and fine-tuned system. The neglect of specific interactions or the improper calibration of parameters may result in the generation of infeasible conformations, and subsequently, incorrect conclusions regarding the structure and function of a biomolecule.

A restricted number of tools include functionally significant tertiary structure elements, such as pseudoknots, in their structure prediction. Moreover, even if such tools do include these elements, the underlying parameters are constrained or favour only a subset of pseudoknots. Consequently, only those pseudoknots that are particularly stable within the given scheme are likely to be predicted. This is problematic because different metastable intermediates may also be functionally essential. A further aspect to consider is that the helices of a kissing hairpin pseudoknot may differ from those of a classical A-form. Given the non-classical shape and behaviour, it is essential to implement bespoke parameters rather than the standard ones to predict kissing hairpin structures.

The process of learning from three-dimensional structures can be obtained by identifying which structures are feasible and which are not, and by determining specific parameter sets that can subsequently be introduced into structure prediction tools. One strategy for achieving this is de novo modelling of three-dimensional structures, whereby the potential structural landscape and its folding pathways are investigated. Another strategy is to perform statistical analysis on experimentally determined RNA structures, assuming that energetically favourable motifs will be observed frequently, while unobserved motifs are either sterically impossible or at least energetically unfavourable. In the course of my thesis, I have employed both strategies and developed tools and pipelines to facilitate the extraction of two-dimensional parameters based on three-dimensional data, as well as the evaluation of the feasibility of predicted 2D structures in 3D, including the valuable folding pathways from an initial contact to the final interaction.

I developed a `forgi` extension (section II.1) that allows to identify and classify pseudoknots and kissing hairpin interactions according to the topological genus concept of Reidys et al. [149] by analysing secondary structure elements present in the tertiary structure. Classification by genus allows complex RNA structures to be broken down into a simplified representation. The extension also analyses the structural elements such as helix orientation and frequency of coaxial stacking. This allows the collection of patterns and statistics on helix arrangement per pseudoknot type based on 3D. My research identified the H-type pseudoknot and the kissing hairpin as the most frequent pseudoknots observed at the genus level. Particular attention was paid to the frequency of coaxial stacking in these conformations. The data indicate that intramolecular kissing hairpin structures were rarely stacked due to steric constraints. However, stacking between all three helices of an intermolecular kissing hairpin is possible. These results also highlight the need to make a steric distinction between intramolecular (pseudoknot) and intermolecular interactions when predicting kissing hairpin interactions.

The computational time and memory requirements for 3D structure prediction are still a handicap. To achieve an optimal trade-off between computational efficiency and realism, I used two different modelling tools, `Ernwin` and `SimRNA`, to develop the published pipeline `RRI-3D` (section II.2). In contrast to other 3D structure prediction approaches, both tools also permit the simulation of RNA duplexes. This in turn allows the study of pseudoknots as well as the dynamics and constraints of RRIs formation, including kissing hairpin interactions. Because `Ernwin` operates at the level of secondary structure elements, it scans a large structural landscape for potential 3D embedded starting structures in an extremely fast manner, allowing prediction of large structures as well as long-range interactions. However, being fragment-based, it can only model structures that can be built from fragments already present in known structures. A subsequent `SimRNA` simulation not only refines the structure, but also tests whether the structure is stable. Since the secondary structure is not fixed in the `SimRNA` simulation, energetically unstable structures will quickly relax by (partially) unfolding helices. At the same time, the combination of the two methods allows rapid and comprehensive step-by-step analysis of the formation of very large RNA complexes. The continuous translation of the coarse-grained models of both approaches into all-atom structures, which is anchored in the pipeline, allows to check whether these are sterically possible even during the ongoing process of interaction formation. The developed pipeline offers the unique opportunity to provide information about folding pathways and local intermediate structures based on the trajectories generated at the tertiary structure level. This is particularly important

for the folding of RNA structures, including the formation of pseudoknots and RRIs, as these states are often critical for the functionality of an RNA [132]. In addition to the investigation of the interaction site, the `RRI-3D` approach also allows an analysis of the folding behaviour of the remaining RNA structure.

Computational tools play an important role in supporting experimental results in RNA structure studies, complementing and enhancing the understanding gained from experimental techniques. At the experimental level, 2D RNA structure probing techniques are much faster and cheaper than 3D techniques, but evaluation of the 3D structure provides a wealth of additional information. Here, 3D-based computational approaches can sufficiently complement experimental techniques by providing structural insights, pseudoknot detection and visualisation capabilities, as shown in the publication "Insights into the secondary and tertiary structure of the Bovine Viral Diarrhoea Virus. Internal Ribosome Entry Site" in section II.3. Based on the experimental data obtained by SHAPE reactivity, the secondary structure was predicted using 2D tools, which supports the interpretation of the experimental results. The observed H-type pseudoknot plays a central role in shaping the overall structure and are proposed to be essential for the functionality of viral replication and pathogenesis. 3D modelling and clustering of low energy candidate structures provided insight into the feasibility and structural properties of the tertiary structure and allowed a deeper understanding of the folding patterns. Analysis and visualisation of the structural features of the pseudoknot, including the orientation of helices, loops and single-stranded regions, helped to determine how specific interactions contribute to its stability and function. The quasi-coaxial stacking between the pseudoknot, stem loops and other motifs revealed the intricate folding patterns that contribute to the overall structure of the RNA. The novel combination of computational 3D approaches with experimental methods verifies the feasibility of 3D embedding and can provide a comprehensive understanding of RNA structure and function.

In the context of this thesis, I have identified and critically examined the shortcomings of current pseudoknot prediction approaches and their parametrization, including kissing hairpin interactions. In particular, predicted pseudoknot interactions tend to be sterically impossible, primarily due to the fact that they are predominantly based on energy minimisation, thereby ignoring the aspect of folding kinetics. Approaches such as the `RRI-3D` pipeline provide insights into the formation of structural dynamics and demonstrates the extent to which 2D structure predictions are compromised by the neglect of 3D structure modelling. The developed tools and pipelines, along with the resulting information, can now be used to more effectively investigate experimental-approved pseudoknots and to

efficiently test the feasibility of potential folding pathways. Furthermore, these methodologies allow the formulation of rules, such as maximal helix lengths of an interaction, which can be employed to filter 2D structure predictions or to determine and analyse potential folding pathways.

A limitation of the present study (section II.1) is that although the observed data set was larger than in most other studies, the sample size for pseudoknot diversity is still small, which may affect the degree of generalisability of the findings. Given the limitations of the available data, three-dimensional approaches can be a valuable addition to support experimental two-dimensional methods and to gain new data. The success of this approach is best illustrated by cooperative work on the BVDV IRES RNA and its pseudoknot.

An further possibility is offered by the `RRI-3D` pipeline, which includes an `RNAblueprint` [81] based extension to design RRI sequences for specific secondary structures. This enables the generation of multiple sequence models for a specific secondary structure interaction setup. Subsequently, each of these sequence models can be translated and clustered fully automatically in the `RRI-3D` pipeline in parallel runs into a set of 3D interaction start structures. Thereafter, a folding path simulation can be initiated from each of these clusters. This procedure offers the possibility of analysing an extremely large structure landscape with many samples along a specific folding path. Furthermore, it permits the examination of sequence dependency. In this context, the addition of a full-atom model which addresses the various base pairing types in more detail would be advantageous. This is a computationally expensive process, particularly for larger structures, where a potential solution would be to restrict sequence dependency analysis to the interaction site. However, future research into RNA functionality will benefit from investigating the relationship between RNA sequences, secondary structures and tertiary structural elements, including pseudoknots and kissing hairpin interactions. Computational 3D approaches play a key role here by providing valuable insights into RNA pseudoknots and kissing hairpins to enhance our understanding of these structures.

Bibliography

[1] Aalberts DP and Hodas NO. Asymmetry in RNA pseudoknots: observation and theory. *Nucleic Acids Research* 33.7 (2005), pp. 2210–2214. DOI: 10.1093/nar/gki508 (cit. on p. 42).

[2] Abrahams JP, den Berg MV, van Batenburg E and Pleij C. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Research* 18.10 (1990), pp. 3035–3035. DOI: 10.1093/nar/18.10.3035 (cit. on pp. 7, 9, 104).

[3] Alkan C, Karakoç E, Nadeau JH, Sahinalp SC and Zhang K. RNA–RNA Interaction Prediction and Antisense RNA Target Search. *Journal of Computational Biology* 13.2 (2006), pp. 267–282. DOI: 10.1089/cmb.2006.13.267 (cit. on p. 19).

[4] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25.17 (1997), pp. 3389–3402. DOI: 10.1093/nar/25.17.3389 (cit. on p. 33).

[5] Andersen AA and Collins RA. Intramolecular secondary structure rearrangement by the kissing interaction of the Neurospora VS ribozyme. *Proceedings of the National Academy of Sciences* 98.14 (2001), pp. 7730–7735. DOI: 10.1073/pnas.141039198 (cit. on p. 18).

[6] Andronescu M, Condon A, Hoos HH, Mathews DH and Murphy KP. Computational approaches for RNA energy parameter estimation. *RNA* 16.12 (2010), pp. 2304–2318. DOI: 10.1261/rna.1950510 (cit. on p. 14).

[7] Andronescu M, Pop C and Condon AE. Improved free energy parameters for RNA pseudoknotted secondary structure prediction. *RNA* 16.1 (2009), pp. 26–42. DOI: 10.1261/rna.1689910 (cit. on pp. 12, 13, 107).

[8] Andronescu M, Zhang ZC and Condon A. Secondary Structure Prediction of Interacting RNA Molecules. *Journal of Molecular Biology* 345.5 (2005), pp. 987–1001. DOI: 10.1016/j.jmb.2004.10.082 (cit. on p. 17).

[9] Antczak M, Zok T, Osowiecki M, Popenda M, Adamiak RW and Szachniuk M. RNAfitme: a webserver for modeling nucleobase and nucleoside residue conformation in fixed-backbone RNA structures. *BMC Bioinformatics* 19.1 (2018). DOI: 10.1186/s12859-018-2317-9 (cit. on p. 37).

[10] Asano K and Mizobuchi K. Structural Analysis of Late Intermediate Complex Formed between Plasmid ColIb-P9 Inc RNA and Its Target RNA. *Journal of Biological Chemistry* 275.2 (2000), pp. 1269–1274. DOI: 10.1074/jbc.275.2.1269 (cit. on p. 19).

[11] Backofen R and Hess WR. Computational prediction of sRNAs and their targets in bacteria. *RNA Biology* 7.1 (2010), pp. 33–42. DOI: 10.4161/rna.7.1.10655 (cit. on pp. 17, 18).

[12] Bartel DP. MicroRNAs: Target Recognition and Regulatory Functions. *Cell* 136.2 (2009), pp. 215–233. DOI: 10.1016/j.cell.2009.01.002 (cit. on p. 1).

[13] van Batenburg FHD, Gultyaev AP, Pleij CWA, Ng J and Oliehoek J. PseudoBase: a database with RNA pseudoknots. *Nucleic Acids Research* 28.1 (2000), pp. 201–204. DOI: 10.1093/nar/28.1.201 (cit. on p. 42).

[14] van Batenburg F, Gultyaev A and Pleij C. An APL-programmed genetic algorithm for the prediction of RNA secondary structure. *Journal of Theoretical Biology* 174.3 (1995), pp. 269–280. DOI: 10.1006/jtbi.1995.0098 (cit. on pp. 7, 9).

[15] Beckmann IK. Identification and classification of pseudoknots and their impact on RNA 3D structure prediction. en (2018). DOI: 10.25365/THESIS.52151 (cit. on pp. 41, 42).

[16] Beckmann IK, Waldl M, Will S and Hofacker IL. 3D feasibility of 2D RNA–RNA interaction paths by stepwise folding simulations. *RNA* 30.2 (2023), pp. 113–123. DOI: 10.1261/rna.079756.123 (cit. on p. 57).

[17] Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN and E. BP. The Protein Data Bank. *Nucleic Acids Research* 28.1 (2000), pp. 235–242. DOI: 10.1093/nar/28.1.235 (cit. on pp. 26, 34, 41).

[18] Berman HM, Lawson CL and Schneider B. Developing Community Resources for Nucleic Acid Structures. *Life* 12.4 (2022), p. 540. DOI: 10.3390/life12040540 (cit. on pp. 39, 41).

[19] Bernauer J, Huang X, Sim AY and Levitt M. Fully differentiable coarse-grained and all-atom knowledge-based potentials for RNA structure evaluation. *RNA* 17.6 (2011), pp. 1066–1075. DOI: 10.1261/rna.2543711 (cit. on pp. 37, 38).

[20] Bernhart SH, Mückstein U and Hofacker IL. RNA Accessibility in cubic time. *Algorithms for Molecular Biology* 6.1 (2011). DOI: 10.1186/1748-7188-6-3 (cit. on p. 17).

[21] Bernhart SH, Tafer H, Mückstein U, Flamm C, Stadler PF and Hofacker IL. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms for Molecular Biology* 1.1 (2006). DOI: 10.1186/1748-7188-1-3 (cit. on p. 17).

[22] Bernhart SH, Hofacker IL and Stadler PF. Local RNA base pairing probabilities in large sequences. *Bioinformatics* 22.5 (2005), pp. 614–615. DOI: 10.1093/bioinformatics/btk014 (cit. on p. 17).

[23] Biedenbänder T, de Jesus V, Schmidt-Dengler M, Helm M, Corzilius B and Fürtig B. RNA modifications stabilize the tertiary structure of tRNAfMet by locally increasing conformational dynamics. *Nucleic Acids Research* 50.4 (2022), pp. 2334–2349. DOI: 10.1093/nar/gkac040 (cit. on p. 6).

[24] Bindewald E, Kluth T and Shapiro BA. CyloFold: secondary structure prediction including pseudoknots. *Nucleic Acids Research* 38.Web Server (2010), W368–W372. DOI: 10.1093/nar/gkq432 (cit. on p. 13).

[25] Bonhoeffer S, McCaskill J, Stadler P and Schuster P. RNA multi-structure landscapes: A study based on temperature dependent partition functions. *European Biophysics Journal* 22.1 (1993). DOI: 10.1007/bf00205808 (cit. on p. 17).

[26] Boniecki MJ, Lach G, Dawson WK, Tomala K, Lukasz P, Soltysinski T, Rother KM and Bujnicki JM. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Research* 44.7 (2015), e63–e63. DOI: 10.1093/nar/gkv1479 (cit. on p. 31).

[27] Bouchard P and Legault P. A remarkably stable kissing-loop interaction defines substrate recognition by the Neurospora Varkud Satellite ribozyme. *RNA* 20.9 (2014), pp. 1451–1464. DOI: 10.1261/rna.046144.114 (cit. on pp. 18, 19).

[28] Brooks B, Brooks CI, Mackerell AJ, Nilsson L, Petrella R, Roux B, Won Y, Archontis G, Bartels C, Boresch S et al. CHARMM: The biomolecular simulation program. *Journal of computational chemistry* 30.10 (2009), pp. 1545–1614 (cit. on p. 22).

[29] Brunel C, Marquet R, Romby P and Ehresmann C. RNA loop–loop interactions as dynamic functional motifs. *Biochimie* 84.9 (2002), pp. 925–944. DOI: 10.1016/s0300-9084(02)01401-3 (cit. on pp. 18, 19).

[30] Busch A, Richter AS and Backofen R. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics* 24.24 (2008), pp. 2849–56. DOI: 10.1093/bioinformatics/btn544 (cit. on p. 17).

[31] Cao S and Chen SJ. Predicting RNA folding thermodynamics with a reduced chain representation model. *RNA* 11.12 (2005), pp. 1884–1897. DOI: 10.1261/rna.2109105 (cit. on p. 24).

[32] Cao S and Chen SJ. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Research* 34.9 (2006), pp. 2634–2652. DOI: 10.1093/nar/gkl346 (cit. on pp. 9, 10, 13, 14, 24, 106).

[33] Cao S and Chen SJ. Predicting structures and stabilities for H-type pseudoknots with interhelix loops. *RNA* 15.4 (2009), pp. 696–706. DOI: 10.1261/rna.1429009 (cit. on pp. 13, 24).

[34] Cao S and Chen SJ. Physics-Based De Novo Prediction of RNA 3D Structures. *The Journal of Physical Chemistry B* 115.14 (2011), pp. 4216–4226. DOI: 10.1021/jp112059y (cit. on p. 24).

[35] Cao S and Chen SJ. Structure and stability of RNA/RNA kissing complex: with application to HIV dimerization initiation signal. *RNA* 17.12 (2011), pp. 2130–2143. DOI: 10.1261/rna.026658.111 (cit. on pp. 19, 24, 108).

[36] Capriotti E, Norambuena T, Marti-Renom MA and Melo F. All-atom knowledge-based potential for RNA structure prediction and assessment. *Bioinformatics* 27.8 (2011), pp. 1086–1093. DOI: 10.1093/bioinformatics/btr093 (cit. on p. 37).

[37] Carr CE and Marky LA. Thermodynamic investigation of kissing-loop interactions. *Biochimie* 157 (2019), pp. 177–183. DOI: 10.1016/j.biochi.2018.11.012 (cit. on p. 19).

[38] Carrascoza F, Antczak M, Miao Z, Westhof E and Szachniuk M. Evaluation of the stereochemical quality of predicted RNA 3D models in the RNA-Puzzles submissions. *RNA* 28.2 (2021), pp. 250–262. DOI: 10.1261/rna.078685.121 (cit. on p. 40).

[39] Chang KY and Tinoco I. The structure of an RNA "kissing" hairpin complex of the HIV TAR hairpin loop and its complement. *Journal of Molecular Biology* 269.1 (1997), pp. 52–66. DOI: 10.1006/jmbi.1997.1021 (cit. on p. 19).

[40] Chastain M and Tinoco I. Structural Elements in RNA. *Progress in Nucleic Acid Research and Molecular Biology*. Elsevier, 1991, pp. 131–177. DOI: 10.1016/s0079-6603(08)60008-2 (cit. on p. 7).

[41] Chen J, Hu Z, Sun S, Tan Q, Wang Y, Yu Q, Zong L, Hong L, Xiao J, Shen T et al. Interpretable RNA Foundation Model from Unannotated Data for Highly Accurate RNA Structure and Function Predictions (2022). DOI: 10.48550/ARXIV.2204.00300 (cit. on p. 33).

[42] Chen X, He SM, Bu D, Zhang F, Wang Z, Chen R and Gao W. FlexStem: improving predictions of RNA secondary structures with pseudoknots by reducing the search space. *Bioinformatics* 24.18 (2008), pp. 1994–2001. DOI: 10.1093/bioinformatics/btn327 (cit. on pp. 12, 13, 107).

[43] Chen X, Li Y, Umarov R, Gao X and Song L. RNA Secondary Structure Prediction By Learning Unrolled Algorithms (2020). DOI: 10.48550/ARXIV.2002.05810 (cit. on p. 32).

[44] Cheng Y, Zhang S, Xu X and Chen SJ. Vfold2D-MC: A Physics-Based Hybrid Model for Predicting RNA Secondary Structure Folding. *The Journal of Physical Chemistry B* 125.36 (2021), pp. 10108–10118. DOI: 10.1021/acs.jpcb.1c04731 (cit. on p. 25).

[45] Chitsaz H, Salari R, Sahinalp SC and Backofen R. A partition function algorithm for interacting nucleic acid strands. *Bioinformatics* 25.12 (2009), pp. i365–i373. DOI: 10.1093/bioinformatics/btp212 (cit. on p. 19).

[46] Cragnolini T, Derreumaux P and Pasquali S. Coarse-Grained Simulations of RNA and DNA Duplexes. *The Journal of Physical Chemistry B* 117.27 (2013), pp. 8047–8060. DOI: 10.1021/jp400786b (cit. on pp. 21, 32).

[47] Cragnolini T, Laurin Y, Derreumaux P and Pasquali S. Coarse-Grained HiRE-RNA Model for ab Initio RNA Folding beyond Simple Molecules, Including Noncanonical and Multiple Base Pairings. *Journal of Chemical Theory and Computation* 11.7 (2015), pp. 3510–3522. DOI: 10.1021/acs.jctc.5b00200 (cit. on pp. 21, 32).

[48] Crothers D, Cole P, Hilbers C and Shulman R. The molecular mechanism of thermal unfolding of Escherichia coli formylmethionine transfer RNA. *Journal of Molecular Biology* 87.1 (1974), pp. 63–88. DOI: 10.1016/0022-2836(74)90560-9 (cit. on pp. 5, 6).

[49] Cruz JA, Blanchet MF, Boniecki M, Bujnicki JM, Chen SJ, Cao S, Das R, Ding F, Dokholyan NV, Flores SC et al. RNA-Puzzles: A CASP-like evaluation of RNA three-dimensional structure prediction. *RNA* 18.4 (2012), pp. 610–625. DOI: 10.1261/rna.031054.111 (cit. on p. 39).

[50] Danaee P, Rouches M, Wiley M, Deng D, Huang L and Hendrix D. bpRNA: large-scale automated annotation and analysis of RNA secondary structure. *Nucleic Acids Research* 46.11 (2018), pp. 5381–5394. DOI: 10.1093/nar/gky285 (cit. on p. 33).

[51] Dang Y, Yang Q, Xue Z and Liu Y. RNA Interference in Fungi: Pathways, Functions, and Applications. *Eukaryotic Cell* 10.9 (2011), pp. 1148–1155. DOI: 10.1128/EC.05109-11 (cit. on p. 1).

[52] Darty K, Denise A and Ponty Y. VARNA: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics* 25.15 (2009), pp. 1974–1975. DOI: 10.1093/bioinformatics/btp250 (cit. on p. 36).

[53] Das R and Baker D. Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences* 104.37 (2007), pp. 14664–14669. DOI: 10.1073/pnas.0703836104 (cit. on pp. 22, 23, 38).

[54] Das R, Karanicolas J and Baker D. Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature Methods* 7.4 (2010), pp. 291–294. DOI: 10.1038/nmeth.1433 (cit. on pp. 22, 23, 38).

[55] Dieterich C and Stadler PF. Computational biology of RNA interactions. *WIREs RNA* 4.1 (2012), pp. 107–120. DOI: 10.1002/wrna.1147 (cit. on p. 17).

[56] Ding F, Sharma S, Chalasani P, Demidov VV, Broude NE and Dokholyan NV. Ab initio RNA folding by discrete molecular dynamics: From structure prediction to folding mechanisms. *RNA* 14.6 (2008), pp. 1164–1173. DOI: 10.1261/rna.894608 (cit. on pp. 21, 30).

[57] Dirks RM and Pierce NA. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of computational chemistry* 24.13 (2003), pp. 1664–1677. DOI: 10.1002/jcc.10296 (cit. on pp. 11–14, 19, 107).

[58] Dirks RM and Pierce NA. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *Journal of computational chemistry* 25.10 (2004), pp. 1295–1304. DOI: 10.1002/jcc.20057 (cit. on p. 11).

[59] Do CB, Woods DA and Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* 22.14 (2006), e90–e98. DOI: 10.1093/bioinformatics/btl246 (cit. on p. 18).

[60] Draper DE. Strategies for RNA folding. *Trends in Biochemical Sciences* 21.4 (1996), pp. 145–149. DOI: 10.1016/s0968-0004(96)80169-1 (cit. on pp. 5–7).

[61] Duan S, Mathews DH and Turner DH. Interpreting Oligonucleotide Microarray Data To Determine RNA Secondary Structure: Application to the 3‘ End of Bombyx mori R2 RNA.

Biochemistry (Easton) 45.32 (2006), pp. 9819–9832. DOI: 10.1021/bi052618x (cit. on p. 21).

[62] Ducongé F, Di Primo C and Toulmé JJ. Is a Closing "GA Pair" a Rule for Stable Loop-Loop RNA Complexes? *Journal of Biological Chemistry* 275.28 (2000), pp. 21287–21294. DOI: 10.1074/jbc.m002694200 (cit. on p. 19).

[63] Durand G, Dausse E, Goux E, Fiore E, Peyrin E, Ravelet C and Toulmé JJ. A combinatorial approach to the repertoire of RNA kissing motifs; towards multiplex detection by switching hairpin aptamers. *Nucleic Acids Research* 44.9 (2016), pp. 4450–4459. DOI: 10.1093/nar/gkw206 (cit. on p. 19).

[64] Eggenhofer F, Tafer H, Stadler PF and Hofacker IL. RNApredator: fast accessibility-based prediction of sRNA targets. *Nucleic Acids Research* 39.suppl_2 (2011), W149–W154. DOI: 10.1093/nar/gkr467 (cit. on p. 16).

[65] Ennifar E, Walter P, Ehresmann B, Ehresmann C and Dumas P. *Nature Structural Biology* 8.12 (2001), pp. 1064–1068. DOI: 10.1038/nsb727 (cit. on p. 19).

[66] Fallmann J, Will S, Engelhardt J, Grüning B, Backofen R and Stadler PF. Recent advances in RNA folding. *Journal of Biotechnology* 261 (2017), pp. 97–104. DOI: 10.1016/j.jbiotec.2017.07.007 (cit. on pp. 17, 18).

[67] Flamm C, Wielach J, Wolfinger MT, Badelt S, Lorenz R and Hofacker IL. Caveats to Deep Learning Approaches to RNA Secondary Structure Prediction. *Frontiers in Bioinformatics* 2 (2022). DOI: 10.3389/fbinf.2022.835422 (cit. on p. 34).

[68] Franch T and Gerdes K. U-turns and regulatory RNAs. *Current Opinion in Microbiology* 3.2 (2000), pp. 159–164. DOI: 10.1016/s1369-5274(00)00069-2 (cit. on pp. 19, 20).

[69] Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, Neilson T and Turner DH. Improved free-energy parameters for predictions of RNA duplex stability. *Proceedings of the National Academy of Sciences* 83.24 (1986), pp. 9373–9377. DOI: 10.1073/pnas.83.24.9373 (cit. on pp. 10, 11, 105).

[70] Gendron P, Lemieux S and Major F. Quantitative analysis of nucleic acid three-dimensional structures. *Journal of Molecular Biology* 308.5 (2001), pp. 919–936. DOI: 10.1006/jmbi.2001.4626 (cit. on p. 35).

[71] Georg J and Hess WR. cis-Antisense RNA, Another Level of Gene Regulation in Bacteria. *Microbiology and Molecular Biology Reviews* 75.2 (2011), pp. 286–300. DOI: 10.1128/mmbr.00032-10 (cit. on p. 1).

[72] Gerlach W and Giegerich R. GUUGle: a utility for fast exact matching under RNA complementary rules including G–U base pairing. *Bioinformatics* 22.6 (2006), pp. 762–764. DOI: 10.1093/bioinformatics/btk041 (cit. on p. 17).

[73] Gluick TC and Draper DE. Thermodynamics of Folding a Pseudoknotted mRNA Fragment. *Journal of Molecular Biology* 241.2 (1994), pp. 246–262. DOI: 10.1006/jmbi.1994.1493 (cit. on pp. 6, 7).

[74] Gong J, Ju Y, Shao D and Zhang QC. Advances and challenges towards the study of RNA-RNA interactions in a transcriptome-wide scale. *Quantitative Biology* 6.3 (2018), pp. 239–252. DOI: 10.1007/s40484-018-0146-5 (cit. on p. 16).

[75] Gosavi D, Wower I, Beckmann IK, Hofacker IL, Wower J, Wolfinger MT and Sztuba-Solinska J. Insights into the secondary and tertiary structure of the Bovine Viral Diarrhea Virus Internal Ribosome Entry Site. *RNA Biology* 19.1 (2022), pp. 496–506. DOI: 10.1080/15476286.2022.2058818 (cit. on p. 68).

[76] Gregorian Jr RS and Crothers DM. Determinants of RNA Hairpin Loop-Loop Complex Stability. *Journal of Molecular Biology* 248.5 (1995), pp. 968–984. DOI: 10.1006/jmbi.1995.0275 (cit. on p. 19).

[77] Groom CR, Bruno IJ, Lightfoot MP and Ward SC. The Cambridge Structural Database. *Acta Crystallographica Section B Structural Science, Crystal Engineering and Materials* 72.2 (2016), pp. 171–179. DOI: 10.1107/s2052520616003954 (cit. on p. 41).

[78] Gultyaev AP. The computer simulation of RNA folding involving pseudoknot formation. *Nucleic Acids Research* 19.9 (1991), pp. 2489–2494. DOI: 10.1093/nar/19.9.2489 (cit. on pp. 7, 10).

[79] Gultyaev AP, van Badenburg F and Pleij CW. An approximation of loop free energy values of RNA H-pseudoknots. *RNA* 5.5 (1999), pp. 609–617. DOI: 10.1017/s135583829998189x (cit. on pp. 9, 42, 104).

[80] Gultyaev AP, van Batenburg F and Pleij CW. The Computer Simulation of RNA Folding Pathways Using a Genetic Algorithm. *Journal of Molecular Biology* 250.1 (1995), pp. 37–51. DOI: 10.1006/jmbi.1995.0356 (cit. on p. 10).

[81] Hammer S, Tschiatschek B, Flamm C, Hofacker IL and Findeiß S. RNAblueprint: flexible multiple target nucleic acid sequence design. *Bioinformatics* 33.18 (2017). Ed. by C Sahinalp, pp. 2850–2858. DOI: 10.1093/bioinformatics/btx263 (cit. on p. 84).

[82] Han K and Byun Y. PSEUDOVIEWER2: visualization of RNA pseudoknots of any type. *Nucleic Acids Research* 31.13 (2003), pp. 3432–3440. DOI: 10.1093/nar/gkg539 (cit. on pp. 3, 27, 99).

[83] Huang FWD, Qin J, Reidys CM and Stadler PF. Partition function and base pairing probabilities for RNA–RNA interaction prediction. *Bioinformatics* 25.20 (2009), pp. 2646–2654. DOI: 10.1093/bioinformatics/btp481 (cit. on p. 19).

[84] Islam S, Ge P and Zhang S. CompAnnotate: a comparative approach to annotate base-pairing interactions in RNA 3D structures. *Nucleic Acids Research* 45.14 (2017), e136–e136. DOI: 10.1093/nar/gkx538 (cit. on p. 35).

[85] Jabbari H, Condon A and Zhao S. Novel and Efficient RNA Secondary Structure Prediction Using Hierarchical Folding. *Journal of Computational Biology* 15.2 (2008), pp. 139–163. DOI: 10.1089/cmb.2007.0198 (cit. on p. 13).

[86] Jabbari H, Wark I, Montemagno C and Will S. Sparsification Enables Predicting Kissing Hairpin Pseudoknot Structures of Long RNAs in Practice. en. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2017. DOI: 10.4230/LIPICS.WABI.2017.12 (cit. on p. 13).

[87] Jonikas MA, Radmer RJ and Altman RB. Knowledge-based instantiation of full atomic detail into coarse-grain RNA 3D structural models. *Bioinformatics* 25.24 (2009), pp. 3259–3266. DOI: 10.1093/bioinformatics/btp576 (cit. on p. 29).

[88] Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D and Altman RB. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. *RNA* 15.2 (2009), pp. 189–199. DOI: 10.1261/rna.1270809 (cit. on pp. 21, 29).

[89] Jossinet F, Paillart JC, Westhof E, Hermann T, Skripkin E, Lodmell JS, Ehresmann C, Ehresmann B and Marquet R. Dimerization of HIV-1 genomic RNA of subtypes A and B: RNA loop structure and magnesium binding. *RNA* 5.9 (1999), pp. 1222–1234. DOI: 10.1017/s1355838299990982 (cit. on p. 18).

[90] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596.7873 (2021), pp. 583–589. DOI: 10.1038/s41586-021-03819-2 (cit. on p. 32).

[91] Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z et al. Rfam 14: expanded coverage of metagenomic, viral and microRNA families. *Nucleic Acids Research* 49.D1 (2020), pp. D192–D200. DOI: 10.1093/nar/gkaa1047 (cit. on pp. 33, 41).

[92] Kato Y, Sato K, Hamada M, Watanabe Y, Asai K and Akutsu T. RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics* 26.18 (2010), pp. i460–i466. DOI: 10.1093/bioinformatics/btq372 (cit. on p. 18).

[93] Kehr S, Bartschat S, Stadler PF and Tafer H. PLEXY: efficient target prediction for box C/D snoRNAs. *Bioinformatics* 27.2 (2010), pp. 279–280. DOI: 10.1093/bioinformatics/btq642 (cit. on p. 18).

[94] Kerpedjiev P, zu Siederdissen CH and Hofacker IL. Predicting RNA 3D structure using a coarse-grain helix-centered model. *RNA* 21.6 (2015), pp. 1110–1121. DOI: 10.1261/rna.047522.114 (cit. on pp. 21, 26).

[95] Kery MB, Feldman M, Livny J and Tjaden B. TargetRNA2: identifying targets of small regulatory RNAs in bacteria. *Nucleic Acids Research* 42.W1 (2014), W124–W129. DOI: 10.1093/nar/gku317 (cit. on p. 17).

[96] Kim CH and Tinoco I. A retroviral RNA kissing complex containing only two G-C base pairs. *Proceedings of the National Academy of Sciences* 97.17 (2000), pp. 9396–9401. DOI: 10.1073/pnas.170283697 (cit. on p. 19).

[97] Kolb FA, Engdahl HM, Slagter-Jäger JG, Ehresmann B, Ehresmann C, Westhof E, Wagner EGH and Romby P. Progression of a loop-loop complex to a four-way junction is crucial for the activity of a regulatory antisense RNA. *The EMBO Journal* 19.21 (2000), pp. 5905–5915. DOI: 10.1093/emboj/19.21.5905 (cit. on p. 19).

[98] Kolb FA, Malmgren C, Westhof E, Ehresmann C, Eheresman B, Wagner EGH and Romby P. An unusual structure formed by antisense-target RNA binding involves an extended kissing complex with a four-way junction and a side-by-side helical alignment. *RNA* 6.3 (2000), pp. 311–324. DOI: 10.1017/s135583820099215x (cit. on p. 19).

[99] Krokhotin A, Houlihan K and Dokholyan NV. iFoldRNA v2: folding RNA with constraints. *Bioinformatics* 31.17 (2015), pp. 2891–2893. DOI: 10.1093/bioinformatics/btv221 (cit. on pp. 21, 30).

[100] Lai D and Meyer IM. A comprehensive comparison of general RNA–RNA interaction prediction methods. *Nucleic Acids Research* 44.7 (2015), e61–e61. DOI: 10.1093/nar/gkv1477 (cit. on pp. 17, 18).

[101] Lawson CL, Berman HM, Chen L, Vallat B and Zirbel CL. The Nucleic Acid Knowledgebase: a new portal for 3D structural information about nucleic acids. *Nucleic Acids Research* 52.D1 (2023), pp. D245–D254. DOI: 10.1093/nar/gkad957 (cit. on pp. 39, 41).

[102] Lease RA and Belfort M. A trans-acting RNA as a control switch in Escherichia coli: DsrA modulates function by forming alternative structures. *Proceedings of the National Academy of Sciences* 97.18 (2000), pp. 9919–9924. DOI: 10.1073/pnas.170281497 (cit. on p. 18).

[103] Leontis NB and Zirbel CL. Nonredundant 3D Structure Datasets for RNA Knowledge Extraction and Benchmarking. *RNA 3D Structure Analysis and Prediction*. Nucleic acids and molecular biology ; 27. Heidelberg [u.a.]: Springer, 2012, pp. 281–298. DOI: 10.1007/978-3-642-25740-7{_}13 (cit. on pp. 36, 41, 42).

[104] Li J and Chen SJ. RNA 3D Structure Prediction Using Coarse-Grained Models. *Frontiers in Molecular Biosciences* 8 (2021). DOI: 10.3389/fmolb.2021.720937 (cit. on pp. 32, 34).

[105] Li J, Zhu W, Wang J, Li W, Gong S, Zhang J and Wang W. RNA3DCNN: Local and global quality assessments of RNA 3D structures using 3D deep convolutional neural networks. *PLOS Computational Biology* 14.11 (2018). Ed. by SJ Chen, e1006514. DOI: 10.1371/journal.pcbi.1006514 (cit. on p. 39).

[106] Lindgreen S, Umu SU, Lai ASW, Eldai H, Liu W, McGimpsey S, Wheeler NE, Biggs PJ, Thomson NR, Barquist L et al. Robust Identification of Noncoding RNA from Transcrip-

tomes Requires Phylogenetically-Informed Sampling. *PLoS Computational Biology* 10.10 (2014). Ed. by K Chen, e1003907. DOI: 10.1371/journal.pcbi.1003907 (cit. on p. 17).

[107] Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF and Hofacker IL. ViennaRNA Package 2.0. *Algorithms for Molecular Biology* 6.1 (2011). DOI: 10.1186/1748-7188-6-26 (cit. on pp. 1, 13, 14, 17, 33).

[108] Lu XJ, Bussemaker HJ and Olson WK. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Research* (2015), gkv716. DOI: 10.1093/nar/gkv716 (cit. on p. 35).

[109] MacKerell Jr AD, Banavali N and Foloppe N. Development and current status of the CHARMM force field for nucleic acids. *Biopolymers* 56.4 (2000), pp. 257–265. DOI: 10.1002/1097-0282(2000)56:4<257::aid-bip10029>3.0.co;2-w (cit. on p. 22).

[110] Madigan M, Aiyer J, Buckley DH, Sattley WM and Stahl D. Brock biology of microorganisms, global edition. en. 16th ed. London, England: Pearson Education, 2021 (cit. on p. 5).

[111] Magnus M, Antczak M, Zok T, Wiedemann J, Lukasiak P, Cao Y, Bujnicki JM, Westhof E, Szachniuk M and Miao Z. RNA-Puzzles toolkit: a computational resource of RNA 3D structure benchmark datasets, structure manipulation, and evaluation tools. *Nucleic Acids Research* (2019). DOI: 10.1093/nar/gkz1108 (cit. on p. 40).

[112] Magnus M, Boniecki MJ, Dawson W and Bujnicki JM. SimRNAweb: a web server for RNA 3D structure modeling with optional restraints. *Nucleic Acids Research* 44.W1 (2016), W315–W319. DOI: 10.1093/nar/gkw279 (cit. on p. 31).

[113] Malhotra A, Tan R and Harvey S. Modeling large RNAs and ribonucleoprotein particles using molecular mechanics techniques. *Biophysical Journal* 66.6 (1994), pp. 1777–1795. DOI: 10.1016/s0006-3495(94)80972-5 (cit. on pp. 21, 28).

[114] Mann M, Wright PR and Backofen R. IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. 45.W1 (2017), W435–W439. DOI: 10.1093/nar/gkx279 (cit. on p. 17).

[115] Marino JP, Gregorian RS, Csankovszki G and Crothers DM. Bent Helix Formation Between RNA Hairpins with Complementary Loops. *Science* 268.5216 (1995), pp. 1448–1454. DOI: 10.1126/science.7539549 (cit. on p. 19).

[116] Masso M. All-atom four-body knowledge-based statistical potential to distinguish native tertiary RNA structures from nonnative folds. *Journal of Theoretical Biology* 453 (2018), pp. 58–67. DOI: 10.1016/j.jtbi.2018.05.022 (cit. on p. 39).

[117] Mathews DH and Turner DH. Prediction of RNA secondary structure by free energy minimization. *Current Opinion in Structural Biology* 16.3 (2006), pp. 270–278. DOI: 10.1016/j.sbi.2006.05.010 (cit. on p. 1).

[118] Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M and Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences* 101.19 (2004), pp. 7287–7292. DOI: 10.1073/pnas.0401799101 (cit. on pp. 8, 18).

[119] Mathews DH, Sabina J, Zuker M and Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology* 288.5 (1999), pp. 911–940. DOI: 10.1006/jmbi.1999.2700 (cit. on pp. 8, 14, 19, 30).

[120] Merino EJ, Wilkinson KA, Coughlan JL and Weeks KM. RNA Structure Analysis at Single Nucleotide Resolution by Selective 2'-Hydroxyl Acylation and Primer Extension (SHAPE). *Journal of the American Chemical Society* 127.12 (2005), pp. 4223–4231. DOI: 10.1021/ja043822v (cit. on p. 21).

[121] Miao Z, Adamiak RW, Antczak M, Batey RT, Becka AJ, Biesiada M, Boniecki MJ, Bujnicki JM, Chen SJ, Cheng CY et al. RNA-Puzzles Round III: 3D RNA structure prediction of five riboswitches and one ribozyme. *RNA* 23.5 (2017), pp. 655–672. DOI: 10.1261/rna.060368.116 (cit. on p. 39).

[122] Miao Z, Adamiak RW, Antczak M, Boniecki MJ, Bujnicki J, Chen SJ, Cheng CY, Cheng Y, Chou FC, Das R et al. RNA-Puzzles Round IV: 3D structure predictions of four ribozymes and two aptamers. *RNA* 26.8 (2020), pp. 982–995. DOI: 10.1261/rna.075341.120 (cit. on p. 39).

[123] Miao Z, Adamiak RW, Blanchet MF, Boniecki M, Bujnicki JM, Chen SJ, Cheng C, Chojnowski G, Chou FC, Cordero P et al. RNA-Puzzles Round II: assessment of RNA structure prediction programs applied to three large RNA structures. *RNA* 21.6 (2015), pp. 1066–1084. DOI: 10.1261/rna.049502.114 (cit. on p. 39).

[124] Moretti R, Lyskov S, Das R, Meiler J and Gray JJ. Web-accessible molecular modeling with Rosetta: The Rosetta Online Server that Includes Everyone (ROSIE). *Protein Science* 27.1 (2017), pp. 259–268. DOI: 10.1002/pro.3313 (cit. on p. 23).

[125] Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF and Hofacker IL. Thermodynamics of RNA–RNA binding. *Bioinformatics* 22.10 (2006), pp. 1177–1182. DOI: 10.1093/bioinformatics/btl024 (cit. on p. 17).

[126] Mujeeb A, Parslow TG, Zarrinpar A, Das C and James TL. NMR structure of the mature dimer initiation complex of HIV-1 genomic RNA. *FEBS Letters* 458.3 (1999), pp. 387–392. DOI: 10.1016/s0014-5793(99)01183-7 (cit. on p. 19).

[127] Narayanan CB, Westbrook J, Ghosh S, Petrov AI, Sweeney B, Zirbel CL, Leontis NB and Berman HM. The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Research* 42.D1 (2013), pp. D114–D122. DOI: 10.1093/nar/gkt980 (cit. on p. 41).

[128] Nawrocki EP and Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* 29.22 (2013), pp. 2933–2935. DOI: 10.1093/bioinformatics/btt509 (cit. on p. 33).

[129] Nussinov R and Jacobson AB. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences* 77.11 (1980), pp. 6309–6313. DOI: 10.1073/pnas.77.11.6309 (cit. on p. 19).

[130] Paillart JC, Westhof E, Ehresmann C, Ehresmann B and Marquet R. Non-canonical interactions in a kissing loop complex: the dimerization initiation site of HIV-1 genomic RNA. *Journal of Molecular Biology* 270.1 (1997), pp. 36–49. DOI: 10.1006/jmbi.1997.1096 (cit. on p. 19).

[131] Pain A, Ott A, Amine H, Rochat T, Bouloc P and Gautheret D. An assessment of bacterial small RNA target prediction programs. *RNA Biology* 12.5 (2015), pp. 509–513. DOI: 10.1080/15476286.2015.1020269 (cit. on pp. 17, 18).

[132] Papenfort K and Vogel J. Multiple target regulation by small noncoding RNAs rewires gene expression at the post-transcriptional level. *Research in Microbiology* 160.4 (2009), pp. 278–287. DOI: 10.1016/j.resmic.2009.03.004 (cit. on pp. 20, 83).

[133] Parisien M, Cruz JA, Westhof É and Major F. New metrics for comparing and assessing discrepancies between RNA 3D structures and models. *RNA* 15.10 (2009), pp. 1875–1885. DOI: 10.1261/rna.1700409 (cit. on p. 40).

[134] Parisien M and Major F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature* 452.7183 (2008), pp. 51–55. DOI: 10.1038/nature06684 (cit. on p. 23).

[135] Pasquali S and Derreumaux P. HiRE-RNA: A High Resolution Coarse-Grained Energy Model for RNA. *The Journal of Physical Chemistry B* 114.37 (2010), pp. 11957–11966. DOI: 10.1021/jp102497y (cit. on pp. 21, 32).

[136] Pearce R, Omenn GS and Zhang Y. De NovoRNA Tertiary Structure Prediction at Atomic Resolution Using Geometric Potentials from Deep Learning (2022). DOI: 10.1101/2022.05.15.491755 (cit. on p. 32).

[137] Pérez A, Marchán I, Svozil D, Sponer J, Cheatham TE, Laughton CA and Orozco M. Refinement of the AMBER Force Field for Nucleic Acids: Improving the Description $\alpha\backslash\gamma$ Conformers. *Biophysical Journal* 92.11 (2007), pp. 3817–3829. DOI: 10.1529/biophysj.106.097782 (cit. on p. 22).

[138] Perry ZR, Pyle AM and Zhang C. Arena: Rapid and Accurate Reconstruction of Full Atomic RNA Structures From Coarse-grained Models. *Journal of Molecular Biology* 435.18 (2023), p. 168210. DOI: 10.1016/j.jmb.2023.168210 (cit. on p. 34).

[139] Pleij CW. Pseudoknots: a new motif in the RNA game. *Trends in Biochemical Sciences* 15.4 (1990), pp. 143–147. DOI: 10.1016/0968-0004(90)90214-v (cit. on pp. 4, 8).

[140] Popenda M, Błażewicz M, Szachniuk M and Adamiak RW. RNA FRABASE version 1.0: an engine with a database to search for the three-dimensional fragments within RNA structures. *Nucleic Acids Research* 36.suppl_1 (2007), pp. D386–D391. DOI: 10.1093/nar/gkm786 (cit. on p. 26).

[141] Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, Blazewicz J and Adamiak RW. Automated 3D structure composition for large RNAs. *Nucleic Acids Research* 40.14 (2012), e112–e112. DOI: 10.1093/nar/gks339 (cit. on p. 26).

[142] Popenda M, Szachniuk M, Blazewicz M, Wasik S, Burke EK, Blazewicz J and Adamiak RW. RNA FRABASE 2.0: an advanced web-accessible database with the capacity to search the three-dimensional fragments within RNA structures. *BMC Bioinformatics* 11.1 (2010). DOI: 10.1186/1471-2105-11-231 (cit. on p. 26).

[143] Puglisi JD, Wyatt JR and Tinoco I. A pseudoknotted RNA oligonucleotide. *Nature* 331.6153 (1988), pp. 283–286. DOI: 10.1038/331283a0 (cit. on p. 7).

[144] Qiu H, Kaluarachchi K, Du Z, Hoffman DW and Giedroc DP. Thermodynamics of Folding of the RNA Pseudoknot of the T4 Gene 32 Autoregulatory Messenger RNA. *Biochemistry* 35.13 (1996), pp. 4176–4186. DOI: 10.1021/bi9527348 (cit. on p. 6).

[145] Rastogi T, Beattie TL, Olive JE and Collins RA. A long-range pseudoknot is required for activity of the Neurospora VS ribozyme. *The EMBO Journal* 15.11 (1996), pp. 2820–2825. DOI: 10.1002/j.1460-2075.1996.tb00642.x (cit. on p. 19).

[146] Reeder J and Giegerich R. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC bioinformatics* 5.1 (2004), pp. 104–104. DOI: 10.1186/1471-2105-5-104 (cit. on pp. 11, 105).

[147] Reeder J, Steffen P and Giegerich R. pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows. *Nucleic Acids Research* 35.Web Server issue (2007), W320–W324. DOI: 10.1093/nar/gkm258 (cit. on pp. 11, 105).

[148] Rehmsmeier M, Steffen P, HÖchsmann M and Giegerich R. Fast and effective prediction of microRNA/target duplexes. *RNA* 10.10 (2004), pp. 1507–1517. DOI: 10.1261/rna.5248604 (cit. on p. 17).

[149] Reidys CM, Huang FWD, Andersen JE, Penner RC, Stadler PF and Nebel ME. Addendum: topology and prediction of RNA pseudoknots. *Bioinformatics* 28.2 (2011), pp. 300–300. DOI: 10.1093/bioinformatics/btr643 (cit. on pp. 4, 12, 42, 82).

[150] Ren A, Rajashankar KR and Patel DJ. Global RNA Fold and Molecular Recognition for a pfl Riboswitch Bound to ZMP, a Master Regulator of One-Carbon Metabolism. *Structure* 23.8 (2015), pp. 1375–1381. DOI: 10.1016/j.str.2015.05.016 (cit. on p. 27).

[151] Ren J, Rastegari B, Codon A and Hoos HH. HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. *RNA* 11.10 (2005), pp. 1494–1504. DOI: 10.1261/rna.7284905 (cit. on p. 13).

[152] Reuter JS and Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics* 11.1 (2010). DOI: 10.1186/1471-2105-11-129 (cit. on p. 17).

[153] Richter A and Backofen R. Accessibility and conservation: General features of bacterial small RNA–mRNA interactions? *RNA Biology* 9.7 (2012), pp. 954–965. DOI: 10.4161/rna.20294 (cit. on p. 17).

[154] Rivas E and Eddy SR. A dynamic programming algorithm for RNA structure prediction including pseudoknots 1 1Edited by I. Tinoco. *Journal of Molecular Biology* 285.5 (1999), pp. 2053–2068. DOI: 10.1006/jmbi.1998.2436 (cit. on pp. 10, 11, 19, 107).

[155] Rother M, Rother K, Puton T and Bujnicki JM. ModeRNA: a tool for comparative modeling of RNA 3D structure. *Nucleic Acids Research* 39.10 (2011), pp. 4007–4022. DOI: 10.1093/nar/gkq1320 (cit. on p. 37).

[156] Ruan J, Stormo GD and Zhang W. ILM: a web server for predicting RNA secondary structures with pseudoknots. *Nucleic Acids Research* 32.Web Server (2004), W146–W149. DOI: 10.1093/nar/gkh444 (cit. on p. 13).

[157] Ruan J, Stormo GD and Zhang W. An Iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. *Bioinformatics* 20.1 (2004), pp. 58–66. DOI: 10.1093/bioinformatics/btg373 (cit. on p. 13).

[158] Rubach P, Zajac S, Jastrzebski B, Sulkowska JI and Sułkowski P. Genus for biomolecules. *Nucleic Acids Research* 48.D1 (2019), pp. D1129–D1135. DOI: 10.1093/nar/gkz845 (cit. on p. 42).

[159] Salim N, Lamichhane R, Zhao R, Banerjee T, Philip J, Rueda D and Feig AL. Thermodynamic and Kinetic Analysis of an RNA Kissing Interaction and Its Resolution into an Extended Duplex. *Biophysical Journal* 102.5 (2012), pp. 1097–1107. DOI: 10.1016/j.bpj.2011.12.052 (cit. on p. 20).

[160] Sarver M, Zirbel CL, Stombaugh J, Mokdad A and Leontis NB. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *Journal of Mathematical Biology* 56.1–2 (2007), pp. 215–252. DOI: 10.1007/s00285-007-0110-x (cit. on p. 35).

[161] Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, Connor R, Funk K, Kelly C, Kim S et al. Database resources of the national center for biotechnology information. *Nucleic Acids Research* 50.D1 (2021), pp. D20–D26. DOI: 10.1093/nar/gkab1112 (cit. on p. 33).

[162] Schneider B, Sweeney BA, Bateman A, Cerny J, Zok T and Szachniuk M. When will RNA get its AlphaFold moment? *Nucleic Acids Research* 51.18 (2023), pp. 9522–9532. DOI: 10.1093/nar/gkad726 (cit. on p. 34).

[163] Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 2.5. 2022 (cit. on pp. 27, 36).

[164] Seemann SE, Gorodkin J and Backofen R. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Research* 36.20 (2008), pp. 6355–6362. DOI: 10.1093/nar/gkn544 (cit. on p. 34).

[165] Seemann SE, Richter AS, Gesell T, Backofen R and Gorodkin J. PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics* 27.2 (2010), pp. 211–219. DOI: 10.1093/bioinformatics/btq634 (cit. on p. 17).

[166] Serra MJ and Turner DH. [11] Predicting thermodynamic properties of RNA. *Energetics of Biological Macromolecules*. Elsevier, 1995, pp. 242–261. DOI: 10.1016/0076-6879(95)59047-1 (cit. on pp. 8, 11, 14, 105).

[167] Shabalina SA and Koonin EV. Origins and evolution of eukaryotic RNA interference. *Trends in Ecology & Evolution* 23.10 (2008), pp. 578–587. DOI: 10.1016/j.tree.2008.06.005 (cit. on p. 1).

[168] Sharma S, Ding F and Dokholyan NV. iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics* 24.17 (2008), pp. 1951–1952. DOI: 10.1093/bioinformatics/btn328 (cit. on pp. 21, 30).

[169] Shen T, Hu Z, Peng Z, Chen J, Xiong P, Hong L, Zheng L, Wang Y, King I, Wang S et al. E2Efold-3D: End-to-End Deep Learning Method for accurate de novo RNA 3D Structure Prediction. 2022. DOI: 10.48550/ARXIV.2207.01586 (cit. on p. 32).

[170] Singh J, Hanson J, Paliwal K and Zhou Y. RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning. *Nature Communications* 10.1 (2019). DOI: 10.1038/s41467-019-13395-9 (cit. on p. 33).

[171] Sleiman D, Goldschmidt V, Barraud P, Marquet R, Paillart JC and Tisné C. Initiation of HIV-1 reverse transcription and functional role of nucleocapsid-mediated tRNA/viral genome interactions. *Virus Research* 169.2 (2012), pp. 324–339. DOI: 10.1016/j.virusres.2012.06.006 (cit. on p. 1).

[172] Sperschneider J and Datta A. DotKnot: pseudoknot prediction using the probability dot plot under a refined energy model. *Nucleic Acids Research* 38.7 (2010), e103–e103. DOI: 10.1093/nar/gkq021 (cit. on p. 13).

[173] Sperschneider J, Datta A and Wise MJ. Heuristic RNA pseudoknot prediction including intramolecular kissing hairpins. *RNA* 17.1 (2010), pp. 27–38. DOI: 10.1261/rna.2394511 (cit. on p. 13).

[174] Staple DW and Butcher SE. Pseudoknots: RNA Structures with Diverse Functions. *PLoS Biology* 3.6 (2005), e213. DOI: 10.1371/journal.pbio.0030213 (cit. on p. 1).

[175] Stasiewicz J, Mukherjee S, Nithin C and Bujnicki JM. QRNAS: software tool for refinement of nucleic acid structures. *BMC Structural Biology* 19.1 (2019). DOI: 10.1186/s12900-019-0103-1 (cit. on p. 37).

[176] Sweeney BA, Petrov AI, Ribas CE, Finn RD, Bateman A, Szymanski M, Karlowski WM, Seemann SE, Gorodkin J, Cannone JJ et al. RNAcentral 2021: secondary structure integration, improved sequence search and new member databases. *Nucleic Acids Research* 49.D1 (2020), pp. D212–D220. DOI: 10.1093/nar/gkaa921 (cit. on pp. 33, 41).

[177] Tafer H and Hofacker IL. RNAplex: a fast tool for RNA–RNA interaction search. *Bioinformatics* 24.22 (2008), pp. 2657–2663. DOI: 10.1093/bioinformatics/btn193 (cit. on p. 17).

[178] Tafer H, Kehr S, Hertel J, Hofacker IL and Stadler PF. `RNAsnoop`: efficient target prediction for H/ACA snoRNAs. *Bioinformatics* 26.5 (2009), pp. 610–616. DOI: 10.1093/bioinformatics/btp680 (cit. on p. 18).

[179] Tan YL, Wang X, Shi YZ, Zhang W and Tan ZJ. rsRNASP: A residue-separation-based statistical potential for RNA 3D structure evaluation. *Biophysical Journal* 121.1 (2022), pp. 142–156. DOI: 10.1016/j.bpj.2021.11.016 (cit. on p. 37).

[180] Tan YL, Wang X, Yu S, Zhang B and Tan ZJ. cgRNASP: coarse-grained statistical potentials with residue separation for RNA structure evaluation. *NAR Genomics and Bioinformatics* 5.1 (2023). DOI: 10.1093/nargab/lqad016 (cit. on pp. 37, 38).

[181] Tan RKZ, Petrov AS and Harvey SC. YUP: A Molecular Simulation Program for Coarse-Grained and Multiscaled Models. *Journal of Chemical Theory and Computation* 2.3 (2006), pp. 529–540. DOI: 10.1021/ct050323r (cit. on pp. 21, 28).

[182] Tan Z, Fu Y, Sharma G and Mathews DH. TurboFold II: RNA structural alignment and secondary structure prediction informed by multiple homologs. *Nucleic Acids Research* 45.20 (2017), pp. 11570–11581. DOI: 10.1093/nar/gkx815 (cit. on p. 33).

[183] Taufer M, Licon A, Araiza R, Mireles D, van Batenburg FHD, Gultyaev AP and Leung MY. PseudoBase++: an extension of PseudoBase for easy searching, formatting and visualization of pseudoknots. *Nucleic Acids Research* 37.Database (2009), pp. D127–D135. DOI: 10.1093/nar/gkn806 (cit. on p. 42).

[184] Theimer CA, Wang Y, Hoffman DW, Krisch HM and Giedroc DP. Non-nearest neighbor effects on the thermodynamics of unfolding of a model mRNA pseudoknot. *Journal of Molecular Biology* 279.3 (1998), pp. 545–564. DOI: 10.1006/jmbi.1998.1812 (cit. on pp. 7, 10).

[185] Theis C, Janssen S and Giegerich R. Prediction of RNA Secondary Structure Including Kissing Hairpin Motifs. *Algorithms in Bioinformatics*. Vol. 6293. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 52–64. DOI: 10.1007/978-3-642-15294-8_5 (cit. on p. 11).

[186] Thiel BC, Beckmann IK, Kerpedjiev P and Hofacker IL. 3D based on 2D: Calculating helix angles and stacking patterns using forgi 2.0, an RNA Python library centered on secondary structure elements. *F1000Research* 8 (2019), p. 287. DOI: 10.12688/f1000research.18458.2 (cit. on p. 43).

[187] Thiel BC, Bussi G, Poblete S and Hofacker IL. Sampling globally and locally correct RNA 3D structures using ERNWIN, SPQR and experimental SAXS data. *bioRxiv* (2022). DOI: 10.1101/2022.07.02.498583 (cit. on pp. 21, 26).

[188] Tjaden B. TargetRNA: a tool for predicting targets of small RNA action in bacteria. *Nucleic Acids Research* 36.Web Server (2008), W109–W113. DOI: 10.1093/nar/gkn264 (cit. on p. 16).

[189] Townshend RJL, Eismann S, Watkins AM, Rangan R, Karelina M, Das R and Dror RO. Geometric deep learning of RNA structure. *Science* 373.6558 (2021), pp. 1047–1051. DOI: 10.1126/science.abe5650 (cit. on p. 39).

[190] Turner DH and Mathews DH. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research* 38.suppl_1 (2009), pp. D280–D282. DOI: 10.1093/nar/gkp892 (cit. on pp. 10, 18).

[191] Umu SU and Gardner PP. A comprehensive benchmark of RNA–RNA interaction prediction tools for all domains of life. *Bioinformatics* 33.7 (2016). Ed. by C Sahinalp, pp. 988–996. DOI: 10.1093/bioinformatics/btw728 (cit. on p. 17).

[192] Van Melckebeke H, Devany M, Di Primo C, Beaurain F, Toulme JJ, Bryce DL and Boisbouvier J. Liquid-crystal NMR structure of HIV TAR RNA bound to its SELEX RNA aptamer reveals the origins of the high stability of the complex. *Proceedings of the National Academy of Sciences* 105.27 (2008), pp. 9210–9215. DOI: 10.1073/pnas.0712121105 (cit. on p. 1).

[193] Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* 50.D1 (2021), pp. D439–D444. DOI: 10.1093/nar/gkab1061 (cit. on p. 32).

[194] Vasudevan S. Functional validation of microRNA-target RNA interactions. *Methods* 58.2 (2012), pp. 126–134. DOI: 10.1016/j.ymeth.2012.08.002 (cit. on p. 1).

[195] Vogel J. A rough guide to the non-coding RNA world of Salmonella. *Molecular Microbiology* 71.1 (2009), pp. 1–11. DOI: 10.1111/j.1365-2958.2008.06505.x (cit. on p. 1).

[196] Waldl M, Beckmann IK, Will S and Hofacker IL. RRIkinDP: Direct Paths RNA-RNA Interaction Kinetics Model (2023). DOI: 10.1101/2023.07.28.548983 (cit. on p. 20).

[197] Waldl M, Spicher T, Lorenz R, Beckmann IK, Hofacker IL, Löhneysen SV and Stadler PF. Local RNA folding revisited. *Journal of Bioinformatics and Computational Biology* 21.04 (2023). DOI: 10.1142/s0219720023500166 (cit. on p. 17).

[198] Waleń T, Chojnowski G, Gierski P and Bujnicki JM. ClaRNA: a classifier of contacts in RNA 3D structures based on a comparative analysis of various classification schemes. *Nucleic Acids Research* 42.19 (2014), e151–e151. DOI: 10.1093/nar/gku765 (cit. on p. 35).

[199] Walter AE and Turner DH. Sequence dependence of stability for coaxial stacking of RNA helixes with Watson-Crick base paired interfaces. *Biochemistry (Easton)* 33.42 (1994), pp. 12715–12719. DOI: 10.1021/bi00208a024 (cit. on pp. 8, 11, 105).

[200] Wang J, Mao K, Zhao Y, Zeng C, Xiang J, Zhang Y and Xiao Y. Optimization of RNA 3D structure prediction using evolutionary restraints of nucleotide–nucleotide interactions from direct coupling analysis. *Nucleic Acids Research* 45.11 (2017), pp. 6299–6309. DOI: 10.1093/nar/gkx386 (cit. on p. 37).

[201] Wang J, Zhao Y, Zhu C and Xiao Y. 3dRNAscore: a distance and torsion angle dependent evaluation function of 3D RNA structures. *Nucleic Acids Research* 43.10 (2015), e63–e63. DOI: 10.1093/nar/gkv141 (cit. on p. 39).

[202] Wang W, Feng C, Han R, Wang Z, Ye L, Du Z, Wei H, Zhang F, Peng Z and Yang J. trRosettaRNA: automated prediction of RNA 3D structure with transformer network. *Nature Communications* 14.1 (2023). DOI: 10.1038/s41467-023-42528-4 (cit. on p. 32).

[203] Wang X, Yu S, Lou E, Tan YL and Tan ZJ. RNA 3D Structure Prediction: Progress and Perspective. *Molecules* 28.14 (2023), p. 5532. DOI: 10.3390/molecules28145532 (cit. on p. 37).

[204] Wang X, Yu S, Lou E, Tan YL and Tan ZJ. RNA 3D Structure Prediction: Progress and Perspective. *Molecules* 28.14 (2023), p. 5532. DOI: 10.3390/molecules28145532 (cit. on p. 38).

[205] Waters LS and Storz G. Regulatory RNAs in Bacteria. *Cell* 136.4 (2009), pp. 615–628. DOI: 10.1016/j.cell.2009.01.043 (cit. on p. 1).

[206] Watkins AM, Rangan R and Das R. FARFAR2: Improved De Novo Rosetta Prediction of Complex Global RNA Folds. *Structure* 28.8 (2020), 963–976.e6. DOI: 10.1016/j.str.2020.05.011 (cit. on p. 23).

[207] Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S and Weiner P. A new force field for molecular mechanical simulation of nucleic acids and proteins. *Journal of the American Chemical Society* 106.3 (1984), pp. 765–784. DOI: 10.1021/ja00315a051 (cit. on p. 22).

[208] Wenzel A, Akbaşli E and Gorodkin J. RIsearch: fast RNA–RNA interaction search using a simplified nearest-neighbor energy model. *Bioinformatics* 28.21 (2012), pp. 2738–2746. DOI: 10.1093/bioinformatics/bts519 (cit. on p. 17).

[209] Westhof E and Jaeger L. RNA pseudoknots. *Current Opinion in Structural Biology* 2.3 (1992), pp. 327–333. DOI: 10.1016/0959-440x(92)90221-r (cit. on p. 4).

[210] Wiedemann J, Zok T, Milostan M and Szachniuk M. LCS-TA to identify similar fragments in RNA 3D structures. *BMC Bioinformatics* 18.1 (2017). DOI: 10.1186/s12859-017-1867-6 (cit. on p. 40).

[211] Williams CJ, Headd JJ, Moriarty NW, Prisant MG, Videau LL, Deis LN, Verma V, Keedy DA, Hintze BJ, Chen VB et al. MolProbity: More and better reference data for improved all-atom structure validation. *Protein Science* 27.1 (2017), pp. 293–315. DOI: 10.1002/pro.3330 (cit. on p. 40).

[212] Wright PR, Georg J, Mann M, Sorescu DA, Richter AS, Lott S, Kleinkauf R, Hess WR and Backofen R. CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains. *Nucleic Acids Research* 42.W1 (2014), W119–W123. DOI: 10.1093/nar/gku359 (cit. on p. 17).

[213] Wyatt JR, Puglisi JD and Tinoco I. RNA pseudoknots: Stability and loop size requirements. *Journal of Molecular Biology* 214.2 (1990), pp. 455–470. DOI: 10.1016/0022-2836(90)90193-p (cit. on pp. 7, 9).

[214] Xiong P, Wu R, Zhan J and Zhou Y. Pairing a high-resolution statistical potential with a nucleobase-centric sampling algorithm for improving RNA model refinement. *Nature Communications* 12.1 (2021). DOI: 10.1038/s41467-021-23100-4 (cit. on p. 37).

[215] Xu X and Chen SJ. Hierarchical Assembly of RNA Three-Dimensional Structures Based on Loop Templates. *The Journal of Physical Chemistry B* 122.21 (2017), pp. 5327–5335. DOI: 10.1021/acs.jpcb.7b10102 (cit. on p. 24).

[216] Xu X, Zhao C and Chen SJ. VfoldLA: A web server for loop assembly-based prediction of putative 3D RNA structures. *Journal of Structural Biology* 207.3 (2019), pp. 235–240. DOI: 10.1016/j.jsb.2019.06.002 (cit. on p. 24).

[217] Xu X, Zhao P and Chen SJ. Vfold: A Web Server for RNA Structure and Folding Thermodynamics Prediction. *PLoS ONE* 9.9 (2014). Ed. by Y Xu, e107504. DOI: 10.1371/journal.pone.0107504 (cit. on pp. 24, 25).

[218] Xue Y. Architecture of RNA–RNA interactions. *Current Opinion in Genetics & Development* 72 (2022), pp. 138–144. DOI: 10.1016/j.gde.2021.11.007 (cit. on p. 16).

[219] Yang H. Tools for the automatic identification and classification of RNA base pairs. *Nucleic Acids Research* 31.13 (2003), pp. 3450–3460. DOI: 10.1093/nar/gkg529 (cit. on p. 35).

[220] Yang Y, Wang YP and Li KB. MiRTif: a support vector machine-based microRNA target interaction filter. *BMC Bioinformatics* 9.S12 (2008). DOI: 10.1186/1471-2105-9-s12-s4 (cit. on p. 18).

[221] Zając S, Geary C, Andersen ES, Dabrowski-Tumanski P, Sulkowska JI and Sułkowski P. Genus trace reveals the topological complexity and domain structure of biomolecules. *Scientific Reports* 8.1 (2018). DOI: 10.1038/s41598-018-35557-3 (cit. on p. 42).

[222] Zhang C, Zhang Y and Pyle AM. rMSA: A Sequence Search and Alignment Algorithm to Improve RNA Structure Modeling. *Journal of Molecular Biology* 435.14 (2023), p. 167904. DOI: 10.1016/j.jmb.2022.167904 (cit. on p. 33).

[223] Zhang T, Hu G, Yang Y, Wang J and Zhou Y. All-Atom Knowledge-Based Potential for RNA Structure Discrimination Based on the Distance-Scaled Finite Ideal-Gas Reference State. *Journal of Computational Biology* 27.6 (2020), pp. 856–867. DOI: 10.1089/cmb.2019.0251 (cit. on p. 37).

[224] Zhou H and Zhou Y. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science* 11.11 (2002), pp. 2714–2726. DOI: 10.1110/ps.0217002 (cit. on p. 38).

[225] Zok T, Popenda M and Szachniuk M. MCQ4Structures to compute similarity of molecule structures. *Central European Journal of Operations Research* 22.3 (2013), pp. 457–473. DOI: 10.1007/s10100-013-0296-5 (cit. on p. 40).

[226] Zuker M and Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research* 9.1 (1981), pp. 133–148. DOI: 10.1093/nar/9.1.133 (cit. on p. 19).

The 2D representations in this work were primarily visualised using `pseudoviewer` [82] and then customised.

Acronyms

Mg^{2+}	Magnesium ions.
ARES	Atomic Rotationally Equivariant Scorer.
BRiQ refinement	Backbone Rotameric and Quantum-mechanical-energy-scaled base-base knowledge-based potential.
C2A	Coarse to Atomic.
CONTRAFOLD	CONditional TRAining for RNA Secondary Structure Prediction.
CSD	Cambridge Structural Database.
DFIRE-RNA	distance-scaled finite ideal-gas reference state.
FARFAR	Fragment Assembly of RNA with Full-Atom Refinement.
FARNA	Fragment Assembly of RNA.
FR3D	Find RNA 3D.
HiRE-RNA	High Resolution RNA.
ILM	Iterated Loop Matching approach.
IntaRNA	Interacting RNAs.
NAKB	Nucleic Acid Knowledgebase.
NAST	The nucleic acid simulation tool.
NDB	Nucleic Acid Database.
PDB	Protein Data Bank.
RAMP	ribonucleic acids multibody potential.
RASP	Ribonucleic Acids Statistical Potential.
RFAM	RNA families database.
RIP	RNA-RNA interaction problem.
RNA3DCNN	RNA 3D convolutional neural network.
ROSIE	Rosetta Online Server That Includes Everyone.
RactIP	RNA-RNA interACTion prediction.
VARNA	Visualization Applet for RNA.
YUP	Yammp molecular mechanics package — Yammp Under Python.
cgRNASP	coarse-grained statistical potential.
rsRNASP	residue separation based statistical potential.
bp	base pairs.

BVDV	bovine viral diarrhoea virus.
CASP	critical assessment of protein structure prediction.
ClaRNA	classification of contacts in RNA tertiary structures.
CLASH	crosslinking, ligation, and sequencing of hybrid.
CNN	convolutional neural network.
DI	deformation index.
DMD	discrete molecular dynamics.
DNA	deoxyribonucleic acid.
DSSR	Dissecting the Spatial Structure of RNA.
EMSA	electrophoretic mobility shift assay.
FRET	Forster resonance energy transfer.
HRP	hydroxyl radical probing.
INF	interaction network fidelity.
IRES	internal ribosome entry site.
L1	Loop-1.
L2	Loop-2.
L3	Loop-3.
LIGR-seq	ligation of interacting RNA followed by high-throughput sequencing.
MC	Monte Carlo.
MCMC	Markov chain Monte Carlo.
MCQ	mean of circular quantities.
MD	molecular dynamicis.
MEA	maximum expected accuracy.
MFE	minimum free energy.
miRNA	microRNA.
mRNA	messenger RNA.
MSA	multiple sequence alignments.

NCM	Nucleotide Cyclic Motif.
ncRNA	non-coding RNA.
NMR	nuclear magnetic resonance.
nt	nucleotides.
P	phosphate.
PARIS	psoralen analysis of RNA interactions and structures.
pk	pseudoknot.
RAP	RNA antisense purification.
REMC	replica exchange Monte Carlo.
REMD	replica exchange molecular dynamics.
RIA-seq	RNA interactome analysis followed by deep sequencing.
RMSD	root-mean-square deviation.
RNA	ribonucleic acid.
RRI	RNA-RNA interaction.
rRNA	ribosomal RNA.
S1	Stem-1.
S2	Stem-2.
SAXS	small-angle X-ray scattering.
snoRNAs	small nucleolar RNA.
SPLASH	sequencing of psoralen crosslinked, ligated, and selected hybrids.
SPR	surface plasmon resonance.
sRNA	small RNA.
tRNA	transfer RNA.

A. Appendix

Throughout the introduction to this dissertation, reference is made to the lack of or inadequate parameters for predicting pseudoknot and kissing hairpin interactions. Over the last few decades there have been repeated attempts to fill this data gap and to compile parameter sets for these tertiary structure elements. I believe that in the course of this dissertation it is important to provide a direct overview of these parameters, rather than simply referring to the respective papers and leaving the reader to create a basis for comparison. Therefore, this appendix provides an overview and, where possible, a comparison of the various parameter sets available to date.

,

Stem size [bp]	Loop size [nt]									
	1	2	3	4	5	6	7	8	9	10
(S2)	Deep groove (L1)									
1	99.9	99.9	99.9	4.2	99.9	99.9	99.9	99.9	4.2	99.9
2	99.9	99.9	4.2	99.99	99.9	99.9	99.9	99.9	4.2	99.9
15	99.9	99.9	4.2	4.2	99.9	99.9	99.9	99.9	4.2	99.9
(S1)	Shallow groove (L2)									
2	99.9	99.9	4.2	99.9	99.9	4.2	99.9	99.9	99.9	99.9
3	99.9	4.2	4.2	99.9	99.9	4.2	99.9	99.9	99.9	99.9
15	99.9	4.2	4.2	99.9	99.9	4.2	99.9	99.9	99.9	99.9

Table AA.1.
Free energy parameters (kcal/mol) for connecting H-Type pseudoknot loops proposed by Abrahams et al. [2]. All forbidden combinations of loop and stem length have been defined with an energy of 99.99 kcal/mol.

Stem size [bp]	Loop size [nt]										
	1	2	3	4	5	6	8	10	15	20	30
(S2)	Deep groove (L1)										
2	-	-	-	-	-	-	-	-	-	-	-
3	-	5.0	5.7	6.2	6.5	6.7	7.1	7.4	7.8	8.2	8.6
4	4.7	5.4	5.9	6.2	6.4	6.6	6.9	7.2	7.6	7.9	8.4
5	4.2	4.9	5.4	5.7	5.9	6.1	6.4	6.7	7.1	7.4	7.9
6	3.5	4.2	4.7	5.0	5.2	5.4	5.7	6.0	6.4	6.7	7.2
7	3.5	4.2	4.7	5.0	5.2	5.4	5.7	6.0	6.4	6.7	7.2
8	-	4.2	4.9	5.4	5.7	5.9	6.3	6.6	7	7.4	7.8
9	-	4.7	5.4	5.9	6.2	6.4	6.8	7.1	7.5	7.9	8.3
10	-	5.0	5.7	6.2	6.5	6.7	7.1	7.4	7.8	8.2	8.6
(S1)	Shallow groove (L2)										
2	-	-	-	-	-	-	-	-	-	-	-
3	-	3.5	4.2	4.7	5.0	5.2	5.6	5.9	6.3	6.7	7.1
4	-	-	4.2	4.9	5.4	5.7	6.1	6.4	7	7.3	7.8
5	-	-	-	4.7	5.4	5.9	6.4	6.8	7.4	7.8	8.3
6	-	-	-	5.0	5.7	6.2	6.7	7.1	7.7	8.1	8.6
7	-	-	-	-	5.2	5.9	6.7	7.1	7.8	8.2	8.7
8	-	-	-	-	5.4	6.1	6.9	7.3	8	8.4	8.9
9	-	-	-	-	-	5.6	6.8	7.3	8.1	8.4	9.0
10	-	-	-	-	-	5.7	6.9	7.4	8.2	8.6	9.2

Table AA.2.
Free energy parameters (kcal/mol) for pseudoknots proposed by Gultyaev et al. [79] based on the general theory of polymer loop thermodynamics and pseudoknots already experimentally verified at that time.

Parameter for		kcal/mol
Start a	pk	7.0
	pk in a multiloop	13.0
Overlapping pk		6.0
non-nested multiloop		8.43
Pair in a	pk	0.1
	non-nested multiloop	$0.1 \times g$
Unpaired base in a pk		0.2
Base dangling		$D \times g + 0.2$
Bulges, stems, internal loops in a gap matrix		$T \times g(0.83)$
Coaxial stacking in pk		$C \times g$

Table AA.3.
`pknot` pseudoknot specific parameters. The thermodynomic model is based on the standard parameters; for non-nested structures (e.g. bulges) within the pseudoknot, the Turner parameters (T) [166] and for the dangle energy (D) [69] were used and multiplied by a weighting factor $g < 1$. The same was done for the coaxial energies (C) [199]. In `pknotsRG` [146, 147] the free energy parameter for the start of a pseudoknot is set to 9.0 kcal/mol and for each unpaired base in the pk to 0.3 kcal/mol.

Stem size [bp]	Loop size [nt]											
	1	2	3	4	5	6	7	8	9	10	11	12
(S2)	Deep groove (L1)											
2	-	-	-	6.2	6.4	6.4	6.6	6.8	6.9	7.1	7.2	-
3	-	6.4	6.4	6.4	6.6	6.6	6.8	6.9	7.1	7.3	7.5	-
4	4.4	4.4	4.5	5.4	5.6	6	6.3	6.6	6.9	7.1	7.3	-
5	2.3	4.4	4.6	5.7	6	6.5	6.9	7.2	7.5	7.8	8	-
6	2.3	4.4	4.8	5.8	6	6.5	6.8	7.1	7.4	7.6	7.8	-
7	2.3	4.4	5	5.9	6.2	6.8	7	7.3	7.6	7.8	8	-
8	-	4.4	5.2	5.7	6.4	6.7	7.1	7.3	7.5	7.7	7.9	-
9	-	5.5	5.5	6.4	6.7	7.2	7.5	7.9	8.1	8.3	8.5	-
10	-	6.9	6.9	6.9	7.5	7.7	8.1	8.3	8.6	8.8	8.9	-
11	-	-	-	-	8.7	8.8	8.9	9.1	9.2	9.3	9.3	-
12	-	-	-	-	9.8	9.2	9.5	9.6	9.7	9.8	9.8	-
(S1)	Shallow groove (L2)											
2	-	-	-	7.6	7	7	7.1	7.2	7.3	7.4	7.5	7.7
3	-	6.5	6.5	6.5	6.6	6.7	6.9	7.1	7.2	7.4	7.6	7.7
4	-	-	9.2	9.2	9.2	8.9	8.9	8.9	9	9	9.1	9.2
5	-	-	-	9.8	9.8	9.8	9.1	8.9	8.8	8.8	8.8	8.8
6	-	-	-	11.9	11.9	11.9	11.9	11	10.4	10.1	9.9	9.8
7	-	-	-	-	12.4	12.4	12.4	12.4	11.4	11	10.7	10.5
8	-	-	-	-	12.1	12.1	12.1	12.1	11.6	11.4	11.2	11.1
9	-	-	-	-	-	13.7	13.7	13.7	13.7	12.6	12	11.5
10	-	-	-	-	-	13.7	13.7	13.7	12.7	12.2	11.8	11.5
11	-	-	-	-	-	-	-	-	15.9	14.1	13	12.4
12	-	-	-	-	-	-	-	-	18.7	15.8	14.2	13.2

Table AA.4.
Pseudoknot loop entropy parameters according to Cao and Chen [32]. The upper half of the table shows $-\Delta S_{L1}/k_B$ and the lower half $-\Delta S_{L2}/k_B$ as a function of the respective loop and its length with the respective stem length ($k_BT = 0.62kcal/mol$). The entries in grey are loop-stem combinations that cannot be calculated in the diamond lattice, but should still represent realistic parameters for a pseudoknot. An additional assemble free energy of $1.3kcal/mol$ at 37°C is introduced to account for the entropy change of the two subunits (L1 + S2 and L2 + S1) for the final calculation of a pseudoknot.

Penalty for		Parameters Dirks-Pierce	DP-A	Chen
Start	a external pk	9.60	-1.38	7.2[1]
	a pk in a multiloop	15.00	10.07	15
	a pk in a pk	15.00	15.00	15
	overlapping pk	–	–	6
	for each pk-stem	0.20	2.46	0.1[2]
Unpaired base in a pk		0.10	0.06	0.2
Nested substructure in a pk		0.10	0.96	–
Start		3.4	3.41	–
a branch in	a multiloop that spans a stem	0.4	0.56	–
an unpaired base in		0.0	0.12	–
Multiplicative penalty introduced by Andronescu et al. [7]				
for stacked pair that spans a stem			0.83	0.89
for an internal loop that spans a stem			0.83	074

Table AA.5.
Overview of the penalties introduced by Dirks and Pierce [57] for pseudoknot prediction. On the left are the original penalties, in the middle are the trained penalties DP-A from Andronescu et al. [7]. In addition, multiplicative penalties based on Rivas and Eddy [154] for stacked pairs and internal loops within a pseudoknot were introduced and trained in the base model. The last column shows the Chen et al. [42] penalties implemented in `Flexstem`. All free energy changes are given in kcal/mol. [1] for H types only [2] for each pair in a pseudoknot.

S2 = 3

L1 \ L3	2	3	4	5	6	7
2	0	0	1.8	2.6	4.2	5.8
3	0	-	1.6	1.1	1.4	2.5
4	1.8	1.6	3.8	4.2	5.8	7.4
5	2.6	1.1	4.2	4.1	5.4	7.0
6	4.2	1.4	5.8	5.4	6.3	7.8
7	5.8	2.5	7.4	7.0	7.8	9.3

S2 = 4

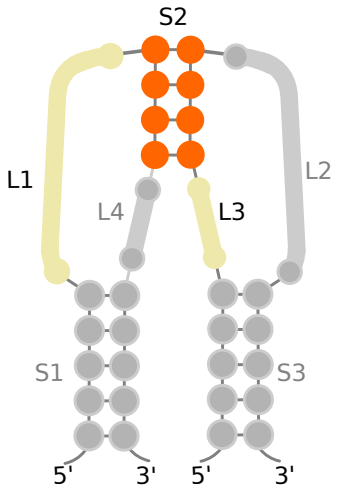
L1 \ L3	2	3	4	5	6	7
2	1.1	0.7	1.4	3.4	5.0	6.7
3	0.7	1.4	0.7	3.4	4.9	6.6
4	1.4	0.7	—	2.7	4.1	5.7
5	3.4	3.4	2.7	5.3	6.7	8.4
6	5.0	4.9	4.1	6.7	7.9	9.5
7	6.7	6.6	5.7	8.4	9.5	11.2

S2 = 5

L1 \ L3	1	2	3	4	5	6	7
1	0	1.4	1.4	2.8	3.7	5.2	6.7
2	1.4	2.8	2.4	4.1	4.8	6.3	7.8
3	1.4	2.4	2.1	3.7	4.4	5.8	7.3
4	2.8	4.1	3.7	5.4	6.1	7.6	9.0
5	3.7	4.8	4.4	6.1	6.8	8.3	9.7
6	5.2	6.3	5.8	7.6	8.3	9.7	11.2
7	6.7	7.8	7.3	9.0	9.7	11.2	12.6

S2 = 6

L1 \ L3	2	3	4	5	6	7
2	0	0.7	1.1	2.2	3.3	4.7
3	0.7	1.8	2.3	3.7	5.1	6.7
4	1.1	2.3	2.7	4.0	5.2	6.8
5	2.2	3.7	4.0	5.5	6.6	8.2
6	3.3	5.1	5.2	6.6	7.6	9.2
7	4.7	6.7	6.8	8.2	9.2	10.9



S2 = 7

L1 \ L3	4	5	6	7
4	2.2	3.3	5.0	6.7
5	3.3	4.2	6.0	7.6
6	5.0	6.0	7.8	9.5
7	6.7	7.6	9.5	11.2

S2 = 8

L1 \ L3	3	4	5	6	7
3	—	0	2.2	4.1	6.1
4	0	0.7	2.4	4.2	6.1
5	2.2	2.4	3.9	5.5	7.3
6	4.1	4.2	5.5	7.0	8.7
7	6.1	6.1	7.3	8.7	10.4

Table AA.6.
Model for the kissing loop entropy in the unit (k_B) by Cao and Chen [35]. Relevant for the selection of the parameters are the lengths of the S2 and the loops L1 and L3, see figure in the upper right quadrant. The individual tables refer to the respective S2 length (min 3 and max 8 bps), based on this other entropy values are used according to the loop length with a maximum loop length of 7 nts. For all loop lengths beyond this, a fitted formula with fitted entropy parameters is used.