universität wien

DISSERTATION | DOCTORAL THESIS

Titel | Title

Modeling Kinetics of RNA-RNA Interaction Formation

verfasst von | submitted by

Maria Waldl BSc MSc

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

Doktorin der Naturwissenschaften (Dr.rer.nat.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 796 605 419
Dissertationsgebiet lt. Studienblatt | Field of study as it appears on the student record sheet:	Chemie
Betreut von | Supervisor:	Univ.-Prof. Dipl.-Phys. Dr. Ivo Hofacker

Acknowledgements

To:

Ivo Hofacker who supervised this work and made this research possible. I especially appreciate our countless discussions that inspired my thought progress and shaped my approach to research. His ability to point out areas for further improvement or inquiry was much valued and substantially improved the quality of this work.

My main collaborators at the TBI in the endeavour of understanding RNA-RNA interactions: Irene Beckmann and Sebastian Will who provided constant feedback, valuable input and moral support.

My current and former colleagues at the TBI who provided an engaging working environment, as collaborators on the papers presented in this thesis and for countless discussions providing constant input on new developments in RNA research.

The Freiburg Bioinformatics Group, especially Rolf Backofen and Martin Raden, for interesting meetings and providing the IntaRNA library.

The Bioinformatics group in Leipzig, for welcoming me in the new city and helping me get up to speed in new fields of research. They also helped free some headspace and provided much-appreciated feedback, which allowed me to finalize this thesis. Thanks in particular to Peter and Sven for giving me the time to wrap things up.

My colleagues in the doc:mov mentoring program – especially Verena, Ivanka and Tadeja; Vlad and Hua-Ting for the regular calls as well as Loki for discussions on managing academic endeavours.

Thanks to my Family, especially my mother Susanne as well as my siblings Teresa and Simon, for the unconditional support in all my endeavours, including my academic ambitions and interests.

My partner Fin for his support, patience, regular check-ins on my progress and critical reading of the thesis.

Thank you!

Abstract

This thesis explores the formation of RNA-RNA interactions, which play a crucial role in gene regulation and other essential cellular processes. Despite continuous advancements in experimental techniques for RNA structure determination, these methods are limited in scope, resolution, or scalability. Computational approaches provide essential complementary tools for predicting RNA interactions and understanding their functional roles.

Most existing computational models for predicting RNA-RNA interactions rely on thermodynamic stability, identifying interactions with minimal free energy.While these models are effective in predicting interaction structures, they are often generate a high number of false positives that require experimental validation. A critical element absent from these models is the fact that functional RNA interactions must not only be stable but also form sufficiently fast to function within cellular timescales. Although this principle is well-known for intramolecular RNA folding, the dynamics of RNA-RNA interaction formation remain relatively unexplored.

This work introduces a novel kinetic model for RNA-RNA interaction prediction, focusing on direct pathways from initial contact to full interaction. This approach computes mechanistically meaningful features enables a fast assessment of interaction candidates. The model was tested using a dataset of small RNA (sRNA) and messenger RNA (mRNA) interactions, revealing that biologically functional interactions exhibit accessible initiation regions and faster extension from a seed interaction to a stable complex compared to random decoy interactions.

The kinetic model was applied to identify and evaluate a long-range functional interaction within the Japanese encephalitis virus case. Additionally, methods for computing RNA accessibility in long RNAs were extended, creating the basis for genome-wide kinetic interaction screening. The thesis further explores the embedding of 2D kinetic folding paths into 3D structures, identifying steric constraints that influence interaction formation. Moreover, the effective (open source) implementation of the model facilitates the integration of kinetic properties into computational RNA-RNA interaction design strategies.

In summary, this work contributes novel insights into RNA-RNA interaction mechanisms, provides a basis for improved computational screening for functional interactions and supports applications in synthetic RNA design.

Kurzfassung

Diese Dissertation untersucht die kinetische Faltung von RNA-RNA-Interaktionen, die eine zentrale Rolle in zellulären Prozessen wie der Genregulation spielen. Trotz kontinuierlicher Fortschritte in experimentellen Methoden zur Bestimmung von RNA-Strukturen sind diese Methoden in ihrer Präzision, strukturellen Auflösung oder Skalierbarkeit begrenzt. Computergestützte Ansätze sind daher unverzichtbare komplementäre Werkzeuge für die Vorhersage von RNA-Interaktionen und das Verständnis ihrer Funktion.

Die meisten bestehenden Vorhersagemodelle für RNA-RNA Interaktionen beruhen auf thermodynamischer Stabilität, wobei Interaktionen mit minimaler freier Energie identifiziert werden. Obwohl diese Ansätze Strukturen effektiv voerhersagen, sagen sie auch viele falsch-positive regulatorische Paare voraus. Ein wesentliches Element, das in diesen Modellen fehlt, ist die Tatsache, dass funktionelle RNA-Interaktionen nicht nur stabil sein, sondern sich auch schnell genug bilden müssen, um innerhalb zellulärer Zeitskalen zu funktionieren. Obwohl dieses Prinzip für die intramolekulare RNA-Faltung bekannt ist, ist der kinetische Faltungsprozess von RNA-RNA-Interaktionen weitgehend unerforscht.

In dieser Arbeit wird ein neues kinetisches Modell zur Vorhersage von RNA-RNA-Interaktionen vorgestellt, das auf direkten Faltungspfaden vom ersten Kontakt bis zur vollständigen Interaktion basiert. Damit ermöglicht das Modell eine schnelle und physikalisch fundierte Bewertung der kinetischen Eigenschaften von Interaktionskandidaten. Eine Evaluation mit bekannten Interaktionen zwischen kleinen bakteriellen RNAs und Boten-RNAs zeigen, dass biologisch funktionelle Interaktionen weitgehend ungepaarte und damit zugängliche Startregionen besitzen und eine schnellere Expansion von einer 3-5 Basenpaare langen Startinteraktion zu einer stabilen Struktur zeigen als zufällig generierte Vergleichsinteraktionen.

Das kinetische Modell wurde in einer Fallstudie zur Beschreibung einer funktionelle Interaktion im Japanischen Enzephalitis-Virus angewandt. Zusätzlich wurden Methoden zur Berechnung der Ungepaartheit von Regionen in langen RNAs weiterentwickelt, um skalierbare (genomweite) Interaktionsanalysen zu ermöglichen. Darüber hinaus wurden kinetische Faltungspfade in komplexere 3D-Modelle eingebettet, um sterische Randbedingungen der Interaktionsfaltung zu identifizieren. Die effiziente Implementierung des Modells ermöglichte eine einfache Integration kinetischer Eigenschaften in das computergestützte Design von künstlichen RNA-RNA-Interaktionen.

Diese Arbeit liefert neue Einblicke in die Mechanismen der RNA-RNA Interaktionsfaltung, bietet eine Grundlage für bessere computergestützte Screeningverfahren nach funktionellen Interaktionen und unterstützt Anwendungen im Bereich des Designs synthetischer RNAs.

Contents

1 Introduction

Despite its apparent simplicity as a linear biopolymer composed of only four nucleotide building blocks, RNA (ribonucleic acid) plays remarkably diverse roles in cellular processes. Such roles include transcription and translation regulation, degradation, splicing as well as organisation of viral genomes. Beyond its biological significance, RNA's chemical versatility also supports applications in medicine and synthetic biology.

Many of these functions and applications rely on RNA-RNA interactions as part of a dynamic process. RNA is uniquely suited to this due to its ability to form stable structures at room temperature while maintaining the flexibility necessary for structural rearrangements in response to environmental changes. This balance between stability and flexibility is crucial for ensuring specific interactions through well-defined structures while enabling the formation of new interactions and structural refolding that regulate downstream processes. Through these properties, RNA is a core contributor to the complexity and adaptability of living systems.

Understanding the dynamics of RNA structure and RNA-RNA interactions is key to understanding the roles of RNAs and, by extension, cellular functions.

Background: Established methods and their key limitations Advances in experimental techniques have promoted the understanding of RNA-RNA interactions. Both improved methods for detecting base pairs or even 3D conformations of RNA structures, as well as recent high-throughput cross-linking experiments that identify interaction partners contribute to a more comprehensive model.

However, these approaches have limitations: Methods that resolve structure, such as NMR, X-ray crystallography, electron microscopy, or mutation rescue experiments are expensive and time-consuming, making them unsuitable for large-scale studies. Conversely, large-scale experiments can provide information about interaction partners but do not resolve structure details. Furthermore, most experimental methods struggle to capture dynamic and transient structures.

Computational tools play a crucial complementary role, enabling large-scale data processing, hypothesis testing, comparative genomics, and prediction of RNA structures and interactions from more readily available sequence data. These tools support a broader perspective for studying interactions and investigating underlying principles. RNA-RNA interactions rely on complementary intermolecular base pairs, analogous to intramolecular RNA structures. Simple computational approaches search for stretches of sequence complementarity between two candidate RNAs, while more advanced models predict

complex interaction structures that include loops and bulges. Typically, these models use algorithms that minimize the free energy of the predicted RNA-RNA complex, following the same principles as those employed in RNA secondary structure prediction. Although machine learning and 3D structure modeling approaches are emerging, they are limited by the availability of training data and computational cost. As a result, 2D structure prediction remains an essential approach that effectively captures key interaction mechanisms. These methods have successfully identified previously unknown interactions and helped to reduce the search space for potential RNA interaction targets, making experimental validation more feasible. However, false positive predictions remain a challenge: a large fraction of predicted interactions are not detected experimentally. This suggests that a plausible secondary structure is a critical component of a functional interaction while also highlighting the need to better understand other factors that influence RNA-RNA interactions.

Potential areas for improvement include accounting for steric constraints, protein-mediated interactions, and the accessibility of interaction sites within individual RNAs, as well as the influence of metabolite and RNA concentrations. While substantial progress has been made in predicting RNA structures and identifying interaction partners, most existing studies have focused primarily on thermodynamic stability. However, biological systems do not operate at equilibrium, and RNA folding often occurs under time constraints imposed by cellular processes. It is already well known from intramolecular RNA folding that not only the thermodynamic stability but also the kinetics are essential in understanding how functional structures emerge within relevant timescales. Despite its importance, this dynamic aspect remains unexplored in the context of RNA-RNA interactions.

Aim: Exploring the RNA-RNA interaction formation process This thesis aims to explore the role of kinetics in RNA-RNA interaction formation and to develop models that capture such dynamics. By investigating the interaction process over time, this work seeks to provide insights into the fundamental principles governing RNA interactions and their biological functions.

Approach: New models for interaction formation, systematic evaluation and use cases This thesis proposes a novel kinetic model designed to facilitate a physically meaningful and computationally efficient assessment of potential RNA-RNA interactions. Rather than exhaustively exploring all possible paths to every potential interaction between two RNAs, the approach targets specific interaction candidates, such as those proposed by thermodynamic models, and focuses on direct folding paths from an initial contact to a fully formed interaction. This strategy assumes that natural systems favor efficient and direct folding into functional states rather than exhaustive exploration of all possible conformations. The model enables the computation of kinetic features, such as folding barriers and folding times.

Using a dataset of interactions between small RNAs (sRNAs) and messenger RNAs (mRNAs), the kinetic features of functional interactions are examined and compared

to those of randomized interactions. The objective is to test whether kinetic features provide additional information beyond what is captured by thermodynamic stability. This evaluation also offers a framework to assess how the model's expression capability changes with variations in model complexity.

Furthermore, several practical applications of this kinetic model are explored: In collaboration with biophysicists, the model was employed to identify and confirm a functional long-range interaction within the Japanese encephalitis virus. Additionally, methods for computing interaction site accessibility in long RNAs were revisited, aiming to enable more scalable screening for potential targets. Another key aspect of the work involved testing whether 2D kinetic folding paths could be embedded in 3D structures and identifying additional factors, such as steric constraints, that influence interaction formation. Finally, the viability of the kinetic tool was demonstrated by integrating kinetic features into sequence design approaches for RNA-RNA interactions.

Structure of the thesis Following this introduction, Chapter 2 expands upon the background presented in the introduction chapter and provides the corresponding references. It includes information on chemical structure, biological functions, and established experimental as well as computational methods for studying RNA structures. Chapter 3 summarizes the research questions, followed by the methods section in Chapter 4. Chapter 5 presents reprints of the six papers that form the main body of this thesis, starting with the kinetic model and accessibility computations for long RNAs. It proceeds through an case study in virology and a study of a pseudoknotted telomerase RNA structure, introduces the 3D embedding of kinetic folding paths and concludes with a strategy for incorporating kinetics into RNA design approaches. The final Chapter 6 summarizes and discusses the findings and offers an outlook on potential future research directions.

A note on the use of large language models (LLMs): Tools such as DeepL and ChatGPT were only used in the editing process for this manuscript, mainly to generate suggestions for textual improvements. The author remains fully responsible for the content.

2 Background

This chapter provides an introduction to the biochemical background of RNA structure, the biological function of RNA-RNA interactions and experimental as well as computational methods for studying RNA-RNA interactions. It sets the stage for discussing current challenges in the field of RNA-RNA interactions and provides context for the contributions presented in this thesis.

2.1 Biochemistry of RNAs and RNA-RNA interactions

An RNA-RNA interaction refers to the physical association or contact between two RNA molecules, typically mediated by complementary base pairing. The specificity of these base pairing interactions enables the critical role of RNA in various biological processes. The following section provide background on the biophysical and chemical properties of RNAs and their interactions and introduces formal notations for describing RNA structure and interactions can be found in Chapter 4.

RNA (ribonucleic acid) is a biopolymer composed of nucleotide monomers, each consisting of three chemical components: a ribose sugar, a phosphate group, and a nitrogenous base (adenine [A], guanine [G], cytosine [C], or uracil [U]), see Figure 2.1a. The nucleotides are connected by a phosphodiester bond as shown in Figure 2.1b.

One defining feature of RNA is the presence of a hydroxyl (-OH) group at the 2' carbon of the ribose sugar, distinguishing it from DNA, which lacks this group. This hydroxyl group increases RNA's chemical reactivity and susceptibility to hydrolysis (degradation) but also contributes to its structural flexibility, enabling it to form dynamic and intricate secondary and tertiary structures that are essential for its biological roles. The phosphate backbone, shared with DNA, imparts a negative charge, which influences RNA folding and interactions through electrostatic repulsion.

The structural organization of RNA can be described in **three hierarchical levels**:

- The primary structure describes the linear sequence of nucleotides linked by phosphodiester bonds, forming a sugar-phosphate backbone. The sequence has directionality defined by the 5' to 3' linkages in the sugar–phosphate backbone.
- The secondary structure is determined by intramolecular base pairs formed by hydrogen bonds between complementary bases (most commonly A-U, G-C, and G-U pairs, see Figure 2.1c [RW54, VM00]) that result in double helical regions. Key motifs include hairpins, internal loops, bulges, and pseudoknots. These structures

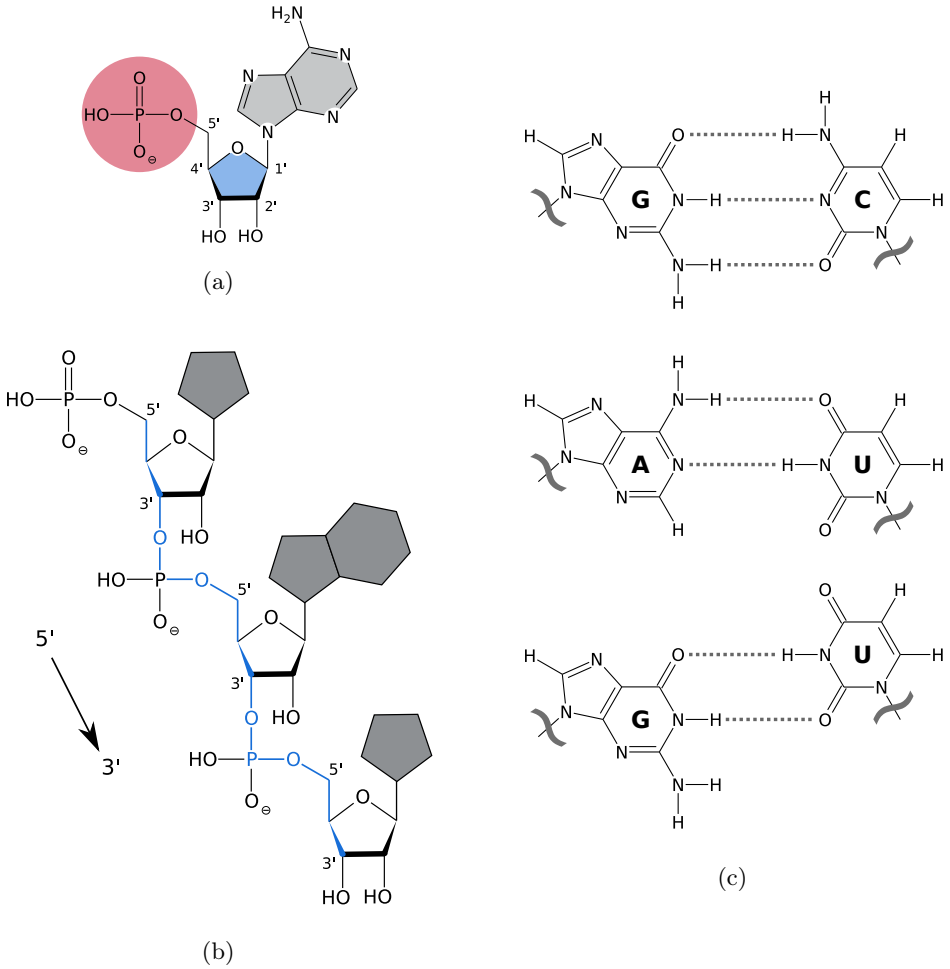


Figure 2.1: Chemical Structure of RNA: Nucleotides (a) consists of a ribose sugar (blue), a phosphate group (red) and a nitrogenous base (grey). They are connected by a sugar phosphate backbone (b). The nitrogenous bases can form complementary base pairs through H-bonds, commonly between guanine(G) and cytosine(C), adenine(A) and uracil(U) as well as guanin(G) and uracil(U). (Figures adapted from [Wal17])

are stabilized by hydrogen bonding and base stacking interactions, with the latter arising from π-π interactions between adjacent bases [YPFK06].

- The tertiary structure represents the three-dimensional conformation of the RNA, a.i. the spatial coordinates of atoms, nucleotides or (coarse grained) structure elements. The elements defined by the secondary structure are folded in 3D space and stabilized by long-range interactions, including tertiary base pairs, base-phosphate contacts, and coordination with metal ions (e.g., Mg^{2+}). This results in tertiary motifs like coaxial stacking, kissing loops, and A-minor interactions that enable the formation

of complex structured molecules like ribozymes and ribosomes.

These chemical and structural properties of RNA directly influence its capacity for **RNA-RNA interactions**. Analogously to RNA structures in general, RNA-RNA interactions occur via specific nucleotide pairing (Watson-Crick base pairs and beyond [SZWL09]) between regions of two RNA molecules. These regions can include small stretches of complementary sequences or more complex structural motifs like loops and bulges, the main difference being that interactions occur between two molecules. Furthermore, they are commonly thought to form after the RNA already formed some intramolecular structure. The interactions can be transient or stable, depending on their biological purpose.

One can differentiate between intra- and intermolecular interactions:

- Intermolecular interactions involve two separate RNA molecules, such as Small RNAs (sRNAs) binding to mRNAs to regulate translation (e.g., bacterial sRNAs targeting mRNAs); microRNAs (miRNAs) interacting with target mRNAs to induce degradation or inhibit translation; or lncRNA-mRNA interactions in which long noncoding RNAs modulate mRNA stability or splicing [DC20, JPK20].
- Intramolecular interactions, on the other hand, occur within the same RNA molecule and contribute to the formation of secondary and tertiary structures, e.g. as pseudoknots or long-range interactions. Locally, they are indistinguishable from interactions between two RNA molecules, but are not affected by concentration effects. Functions include the stabilization of complex 3D structures, e.g. in ribosome assembly, bringing together distant exons for splicing or genome packaging, and frame shifts in virus genomes [GC09].

The chemical properties of RNA allow for its versatile roles in biology and enable its emerging applications in synthetic biology and therapeutic development. Extending the understanding of the chemical basis of RNA-RNA interactions (including modified nucleotides, non-canonical nucleotides, backbone flexibility, contributions to the conformational free energy and the speed of conformation changes) thus provides a critical foundation for advancing both understanding of fundamental biology and applied research.

2.2 RNA-RNA interactions in biology and biotechnology

RNAs participate in diverse biological processes by forming base-pairing interactions with other RNAs, contributing, for example, to gene regulation. The essential role of these interactions in cellular life becomes particularly evident when considering how life on a fundamental level does happen in a non-equilibrium state and depends on dynamic processes: a living organism must sense environmental changes–such as temperature fluctuations, nutrient availability, or cellular damage–and initiate appropriate molecular responses. A substantial part of this regulatory capacity involves RNA molecules interacting with each other to coordinate changes in gene expression, translation, and cellular metabolism.

The importance and complexity of these regulatory systems is also illustrated by the abundance of non-coding RNAs (ncRNAs), which for example constitute larger portions of the genome than protein-coding genes in mammals [Edd01, MM06, SS04, LMT13]. They often serve regulatory functions through interactions that are not only conserved across evolutionary domains of life but are also central to core cellular processes.

A classical example of RNA-RNA interaction is the interaction among messenger RNAs (mRNAs), transfer RNAs (tRNAs), and ribosomal RNAs (rRNAs) during translation as part of the "central dogma of molecular biology", which describes the decoding of genetic information into proteins [CRI70]. The regulation of this multistep process also relies on additional RNA interactions. It is the best-known example of a process catalyzed by a ribozyme, in this case the ribosome, whose function depends on RNA-RNA interactions [DC02].

The following section highlights examples of the biological functions of RNA-RNA interactions:

RNA-RNA interactions in gene regulation Beyond interactions that directly involve tRNAs and the ribosome, many regulatory RNA-RNA interactions modulate gene expression:

- Small RNAs (sRNAs) in bacteria base pair with mRNAs to control their translation and stability, often in response to environmental changes. These interactions typically occur in untranslated regions and can be assisted by RNA chaperones such as Hfq [Bar04, VL11, YNG13, SLSL23, UZS16].
- Micro RNAs (miRNAs) and small interfering RNAs (siRNAs) in eukaryotes base pair with target mRNAs to repress translation or promote degradation via the RNA-induced silencing complex (RISC) [Bar04, FXM$^+$98, HZW$^+$20].
- Antisense RNAs transcribed from the opposite strand of a gene locus interact with their corresponding mRNAs to, for example, modulate splicing, stability, or translation [FW09].
- Spliceosomal RNAs, such as U1, U2, U4, U5, and U6 snRNAs, engage in complex RNA-RNA interactions with pre-mRNA during splicing to ensure correct intron removal [WWL09].
- Small nucleolar RNAs (snoRNAs) guide site-specific modifications (methylation or pseudouridylation) of ribosomal RNAs and other RNAs through specific base-pairing interactions [Kis01].
- Long non-coding RNAs (lncRNAs) often act as scaffolds or guides for protein recruitment by forming interactions with RNA targets. For example, lncRNA-mediated RNA-RNA interactions contribute to the spatial regulation of gene expression and chromatin structure [RC12, GR12].

In all these examples, RNA-RNA base pairing serves as a specific recognition mechanism

for the temporal and spatial regulation of gene expression.

Stress response and RNA condensates RNA-RNA interactions are also essential in the formation of membraneless organelles such as stress granules and P-bodies. In response to cellular stress, untranslated mRNAs and non-coding RNAs aggregate via RNA-RNA and RNA-protein interactions into condensates that dynamically regulate translation and RNA decay [VTPM$^+$18]. These assemblies are thought to arise from RNA-RNA interactions that provide the molecular framework or "glue" required for phase separation.

RNA-RNA interactions in viral genomes RNA viruses rely extensively on RNA-RNA interactions for genome organization, replication, and translation. Structural elements such as long-range interactions, pseudoknots, and kissing-loop interactions regulate essential steps in the viral life cycle. For instance, in flaviviruses, a long-range RNA-RNA interaction between the 5' and 3' ends cyclizes the genome into a "panhandle" structure, which is essential for efficient replication and translation [VG09, NSAB$^+$17]. Retroviruses, such as HIV, package two genomic RNAs as a dimer stabilized by a kissing-hairpin interaction [CT97, NDF$^+$13]. Understanding such interactions provides potential targets for antiviral strategies.

Applications in biotechnology and synthetic biology

RNA-RNA interactions offer unique advantages for biotechnology and synthetic biology due to their programmability, reversibility, and structural diversity. Synthetic RNAs can be engineered to form specific base-pairing interactions, enabling the design of gene circuits, molecular switches, biosensors, and therapeutic agents.

In synthetic biology, RNA-based molecular switches—such as riboregulators and toehold switches—exploit engineered RNA-RNA interactions to regulate gene expression in a programmable manner. These systems can respond to endogenous or exogenous RNA inputs, allowing context-specific activation or repression of gene targets [GSCY14, CCC12].

RNA-RNA interactions also underpin next-generation diagnostics. CRISPR-based detection systems, such as those using Cas12a or Cas13, leverage RNA-guided interactions for sequence-specific cleavage, which can be coupled to fluorescent or lateral flow readouts. Such tools enable rapid, multiplexed detection of viral or cellular RNA targets with high sensitivity [GAL$^+$17, HLD$^+$25].

In therapeutics, siRNAs and antisense oligonucleotides act through RNA-RNA base pairing to silence or modulate gene expression. The success of mRNA vaccines has further demonstrated the feasibility of RNA-based medicines, where careful control over RNA structure and interaction potential is critical for stability and translational efficiency [UP90, CLW06, VBC$^+$25, PHPW18, GBK$^+$23].

Understanding the chemistry and biophysics of RNA-RNA interactions is central to advancing these technologies. Future developments may lead to increasingly complex,

adaptive RNA-based systems for diagnostics, nucleic acid nanotechnology, programmable cell behavior, and personalized therapies.

Conclusion

RNA-RNA interactions are indispensable to fundamental biological processes, including gene regulation, RNA processing, translation, cellular organization and viral replication. Their diverse roles in fundamental biology, combined with their emerging applications in synthetic biology and medicine, underscore the relevance of continued research on RNA-RNA interactions.

2.3 Experimental methods for investigating RNA–RNA interactions and available datasets

RNA-RNA interactions play many critical roles, incuding in gene regulation, enzymatic activities and viral replication. Methods for investigating these interactions experimentally are essential for understanding their structure and function. Over recent decades, a wide range of experimental methods has been developed to identify interaction partners, resolve interaction structures, and characterize their dynamics. Despite advancements, experimental approaches face significant challenges in capturing the transient, complex, and context-dependent nature of RNA interactions. This section reviews some of the primary techniques used in the field, their advantages, and their limitations, emphasizing the need to integrate these data with computational models.

Biophysical structural methods

To resolve the three-dimensional architecture of RNA complexes, several high-resolution structural techniques have been instrumental [DFH$^+$23]:

X-Ray Crystallography has resolved atomic structures of stable RNA complexes (e.g., riboswitch–ligand, ribosomal A-site RNA) at high resolution, but requires crystallizable constructs and cannot capture dynamics [RGB09, JSR23].

NMR Spectroscopy provides residue-level insights into RNA flexibility and transient contacts, but is limited to RNAs $\leq$50 nt and yields ensemble-averaged data [FRWS03, MSP19, MTP23].

Cryo-Electron Microscopy (cryo-EM) has advanced to near-atomic resolution for large RNP assemblies (e.g., spliceosome, ribosome), yet flexible or transient duplexes often lack sufficient contrast for reliable modeling [BK22, MJZS22].

Small angle X-ray scattering (SAXS) provides low-resolution 3D structures on a topological level. It needs to be integrated with 3D modeling approaches to place the given RNA sequences withing the spatial constraints provided by SAXS and to refine the structure [TS23].

Strengths of this methods include the atomic or near-atomic resolution and direct visualization of tertiary contacts. They are limited by low throughput, challenging sample preparation and expensive equipment, and often poor capture of dynamic or weakly interacting regions.

Classic biochemical assays

These established methods offer direct and often accessible ways to probe RNA-RNA pairing and structure:

EMSA (Electrophoretic Mobility Shift Assay) detects duplex formation by reduced gel mobility of RNA complexes, offering a simple, qualitative readout of binding affinity but no information on the interaction site [HF07].

Toeprinting / Primer Extension maps where an interaction or ribosome blocks reverse transcriptase, identifying interaction sites on mRNAs with single–nucleotide precision. However, it is labor-intensive and low-throughput [Nil13, YB22].

Chemical and Enzymatic Probing (SHAPE, DMS, RNases) reports local nucleotide flexibility. Comparing reactivity profiles of free versus bound RNA reveals protected (paired) sites upon interaction, with single–nucleotide resolution—but typically reflects equilibrium states and requires extensive controls due to the inherent noise [Wee10, SCW15].

Mutation–Rescue Experiments validate direct interactions by introducing compensatory mutations that restore duplex formation, confirming selected base pairs [TCKD14].

Co-expression Assays can reveal interactions when RNAs are expressed together in a model system, often coupled with pull-downs or probing to assess complex formation [LXL$^+$19, YMMS$^+$21].

Overall, they provide an essential contribution to understanding RNA-RNA interactions as they can provide insights in interaction structures while being less expensive than the higher resolution biophysical methods discussed in the previous section.

High-throughput crosslinking methods

Modern transcriptome-wide methods employ in vivo crosslinking and proximity ligation to capture RNA contacts at scale. A brief overview:

CLASH (Crosslinking, Ligation, And Sequencing of Hybrids): This method uses UV crosslinking in live cells to UV-crosslink RNA-RNA-protein complexes–most often Argonaute-bound miRNAs and their targets. After immunoprecipitation, the bound RNAs are fragmented, ligated, and sequenced, producing chimeric reads that directly report both canonical and non-canonical RNA–RNA duplexes mediated by the bait protein [HT14].

Psoralen-Based Methods (PARIS, SPLASH, RIC-seq) use psoralen (derivatives)

to crosslinking RNA duplexes indpendent of RNA binding proteins (RBPs) in living cells on transcriptome-wide scale, followed by proximity ligation and deep sequencing to detect chimeric reads. PARIS achieves near–base-pair resolution of intra- and intermolecular contacts, especially illuminating long-range interactions [LZL+16]. SPLASH enhances the recovery of low-abundance duplexes by selective enrichment of crosslinked strands [ASW+16]. RIC-seq captures broader three-dimensional proximity, revealing both base-paired and non-canonical contacts, with reduced resolution on the interaction region [CCJ+20].

Crosslinking Immunoprecipitation (CLIP) Variants identify RNA binding sites of specific RNA binding proteins (RBPs), indirectly revealing RNA–RNA contacts mediated by those proteins. While indirect, sites of protein–RNA binding often mark regions involved in RNA–RNA interactions (e.g., Hfq–sRNA–mRNA) [JD08].

Crosslinking methods offer powerful ways to study RNA interactions in vivo. RBP-dependent techniques (e.g., CLASH, CLIP) reveal protein-mediated contacts, while psoralen-based approaches (e.g., PARIS, SPLASH, RIC-seq) capture direct RNA duplexes independently of proteins. Methods like MARIO and LIGR-Seq further highight the diversity of available protocols [NCY+16, SSWOB16]. Each method has trade-offs in resolution, specificity, and interaction type, underscoring the value of ntegrate multiple experimental and computational techniques (e.g. providing base pairing information) for a complete view of the RNA interactome.

Kinetic approaches

The following methods focus on the dynamic and temporal aspects of RNA folding and interaction formation:

smFRET (Single-Molecule Förster Resonance Energy Transfer) monitors real-time conformational changes during RNA folding or hybridization by labeling two sites with donor and acceptor fluorophores. This yields kinetic rates and intermediate states for duplex formation, albeit requiring specialized microscopy and site-specific labeling [SDH+14].

Optical Tweezers apply mechanical force to unfold RNA hairpins or duplexes, measuring extension changes that report on folding/unfolding kinetics and thermodynamic stability. These assays provide single-molecule force–extension curves but demand complex instrumentation and data analysis [WML+07, BCLW21].

Time-Resolved SHAPE combines rapid mixing with chemical probing to capture folding pathways and interaction kinetics on the second to minute timescale, linking structural snapshots to dynamic processes [MW08, SRB+15].

Despite their power to shed light on kinetic properties, these methods are technically demanding and often require significant expertise and specialized equipment, which limits their accessibility for routine studies.

Available data repositories

Multiple public repositories aggregate interactions from diverse experimental methods–structural, biochemical, and high-throughput. In addition to generalist databases, several resources focus on specific RNA classes:

- **NPInter**: ncRNA–RNA interactions (miRNA–mRNA, lncRNA–protein–RNA, snoRNA–rRNA) from CLIP/CLASH and literature mining, with functional context and expression profiles [ZLT+23].
- **RNAInter**: Coverage of RNA–RNA, RNA–protein, and RNA–DNA interactions; incorporates literature mining, crosslinking evidence, co-expression data, and computational predictions; annotates canonical and non-canonical base pairs [KTH+22a].
- Currently unavailable data collections that contained RNA-RNA interactiond include **RAID** and **RISE** [YZL+17, GSX+18].
- **miRTarBase**: A comprehensive repository of experimentally validated microRNA–target interactions, curated from reporter assays, western blot, qPCR and high-throughput sequencing studies with comprehendsive annotations [HLC+22, KAA+23].
- **sRNATarBase**: Provides a currated set of bacterial sRNA–mRNA interactions validated by Northern blot, EMSA, and other biochemical assays, as well as predicted interactions with detailed mapping of seed regions and binding affinities [WLZ+16]. Other databases focused on sRNAs include sRNAdb and sInterBase [PKB+12, CMR+23].
- **snoDB**: Specializes in small nucleolar RNA (snoRNA) interactions and modification guide relationships, integrating CLIP data, curated pseudouridylation and methylation site annotations as well as interaction and structure predictions [BPFC+23].

These resources collectively span regulatory sRNA–mRNA pairs, structural long-range and tertiary contacts (pseudoknots, kissing loops), viral genome cyclization elements, and RNP-mediated interactions. However, biases toward well-studied species, abundant transcripts, and stabile complexes persist, underscoring the need for continued method development and data expansion.

Summary

Experimental methods for RNA–RNA interaction analysis range from high-resolution structural techniques to transcriptome-wide mapping and single-molecule kinetics. Each technique addresses different scales—atomic detail, interaction sites, or dynamic pathways—but also faces trade-offs in throughput, resolution, and the ability to capture transient or low-abundance interactions. Integrative pipelines that combine these data with computational modeling (e.g., thermodynamic prediction, kinetic simulations) are essential for a comprehensive understanding of RNA–RNA interactomes.

2.4 Computational RNA-RNA interaction prediction methods

Advancements in experimental techniques, both high-throughput as well as more resolution focused, have greatly expanded our understanding of RNA–RNA interactions, but they still face limitations. Challenges for example include detecting transient or weak interactions, scalability issues for genome-wide analyses, and difficulties in capturing structure details like base pairs or three-dimensional (3D) structure. Computational methods complement these techniques by offering scalable, prediction models for exploring explore RNA interactions and uncover structural and dynamic insights beyond what is accessible though experiments. This chapter introduces computational prediction methods, covering thermodynamic, kinetic, and data-driven approaches.

Thermodynamic models

Thermodynamic models for predicting RNA-RNA interactions on a 2D level are among the earliest and most fundamental approaches. These models seek the configuration with the lowest Gibbs free energy (ΔG), assuming that RNA molecules naturally adopt the most energetically favorable structure. The foundational works by [NPGK78a, WS78b, ZS81a, McC90a] introduced dynamic programming algorithms to efficiently predict RNA secondary structures based on free energy minimization, forming the basis of many current prediction tools. The formal gammer that underlies these algorithms restricts the predictions structures 2D structures that do not have any crossing base pairs (pseudoknots, see Chapter 4 for details).

Extensions of this single RNA algorithms to interactions followed different approaches: Tools like RNAcofold link interacting sequences and treated them like a single RNA molecule, thus only allowing interactions that do not introduce crossing base pairs the overall structure [BTM$^+$06]. Other tools including RNAduplex, RNAhybrid and DuplexFold, focus on optimizing interaction base pairs but do not consider competition with intramolecular base pairs [HFS$^+$94a, RSHG04, RM10]. Some software packages, e.g. NUPACK, RNAstructure and ViennaRNA, can also model concentration effects and interactions between more than two RNAs [DBS$^+$07, LBHzS$^+$11, RM10].

All these methods rely on the neighbour energy model with Turner parameters to score the free energy of structures [TUL71, TBD$^+$73, MSZT99]. These parameters quantify the stability contributions of individual local structure components,loops (e.g. hairpins and stacking base pairs) enabling accurate free energy calculations as sum of the component contributions.

Despite their robustness, thermodynamic models face limitations: they typically assume equilibrium conditions and struggle to account for dynamic folding pathways or kinetic barriers.

Accessibility computation

The prediction performance of thermodynamic interaction models was significantly improved by integrating intramolecular structure as accessibility [UG17]: The probability of a given interaction site to be unpaired within the ensemble of all possible intramolecular structures of the individual RNAs can be computed from established 2D structure prediction algorithms introduced in the previous section. It directly corresponds to the energy required to make the interaction site unpaired (accessible) in the intramolecular structure such that interaction base pairs can form. Combining this energy therms with the energy contribution of the interaction base pairs serves as an energy model for predicting interaction structures.

In the context of this thesis the RNAup model as well as IntaRNA are used for such predictions [MTH+06, BRB08a, MWB17]. On top of the RNAup model, IntaRNA provides heuristics for faster interaction screens.

On a conceptional level, these tools better reflect biologically relevant interactions by penalizing interactions that invade intramolecular strongly paired regions and including interactions that introduce crossing base pairs if evaluated as a single RNA.

Machine learning approaches

Recent works on machine learning (ML) tools for RNA structure prediction also cover RNA-RNA interactions. ML based tools for interaction prediction mostly exist in the form of specialized tools for micro RNA (miRNA) target prediction. Some general RNA structure prediction tools can take multiple sequences as input or can run on chained sequences [PMYK19, WCZ+18, MLY22]. However the performance of these tools usually is not evaluated on interactions separately (or even if so only on small data sets that do not allow for conclusive results on performance on interactions)[AAD+24].

This highlights the main challenge in these data-driven approaches: They are the limited availability of reliable RNA-RNA interaction data and possible over-training on known structures [FWW+22, SWD+22, SSB+23]. Approaches to counter this issue could include the integration with traditional methods, extensions towards interpretable methods and a focus on learning underlying physical principles. The release of larger data sets on RNA-RNA interactions and further development of ML methods holds promise for more accurate and scalable RNA-RNA interaction predictions in the future.

Hybrid and specialized models

Hybrid models can incorporate experimental data with computational models to improve prediction accuracy. An example are tools that process cross-linking data and integrate them with thermodynamic models to identify interactions [SV21]. In addition, software package like ViennaRNA and RNAstructure can integrate experimental data as constraint on a range of folding algorithms, including into thermodynamic interaction prediction.

Especially the integration of chemical probing data such as SHAPE reactivity profiles is well established [LLH+16].

Specialized models for example exist for specific RNA classes like microRNA, piRNA, lncRNAs and sRNAs [WHB+18, LMZ+15]. For example, CopraRNA identifies sRNA-mRNA interactions across species [WRP+13]. Other tools like, Risearch2 are optimized for fast heuristics interaction predictions to enable genome wide screens [AWP+17].

Kinetic models

RNA folding and interaction formation are inherently dynamic processes. Unlike thermodynamic models that assume equilibrium states, kinetic models capture transient folding intermediates and metastable states that influence RNA-RNA interactions.

One of the earliest kinetic modeling approaches is Kinfold, which uses stochastic simulations based on Monte Carlo methods to explore folding pathways[FFHS00]. It incorporates structure changes on a base pair level with the 2D energy model implemented in ViennaRNA.

Kinetic models are particularly valuable for studying co-transcriptional folding, where RNA begins folding as it is synthesized. They also provide insights into regulatory mechanisms such as riboswitch activation, where structural transitions e.g. triggered by metabolite binding play a crucial role.

The computational complexity of kinetic modeling poses challenges, especially for long sequences. Nevertheless, these models offer a more realistic depiction of RNA behavior, capturing biologically significant states that purely thermodynamic methods miss.

Some of the existing approaches for kinetic RNA folding extend to multiple RNAs and some can model pseudoknots [XBTI03, GFW+08, BLH23]. In practice however their runtime limits applications to interaction formation and the faster models are to coarse grained to model interaction formation, eg relying introducing or deleting whole helices at a time. Existing kinetic models that directly target interaction formation are limited to low complexity intra- and intermolecular structures, e.g. strand replacement systems [SOS+13, DCR+22].

Recent work kinetic folding contributes an approaches based on Fast Fourier Transform (FFT) that significantly reduces the computational complexity [OMMS22]. Current results suggest further fine tuning is needed to get similar prediction performance as established models but is promising on the side of making kinetic folding realizable for longer RNAs and even interaction complexes.

Overall, tools for predicting and analyzing the kinetics of RNA-RNA interaction formation of biological relevant RNAs with complex intramolecular structure are not yet available.

3 Research Question

While RNA-RNA interactions have been widely studied, most research has focused on thermodynamic stability, assuming that interactions occur at equilibrium. However, given that biological systems are inherently dynamic, RNA folding and interaction processes occur under kinetic constraints. The formation of RNA-RNA interactions is influenced not only by sequence complementarity but also by factors such as intermediate folding states, steric constraints, and competition between alternative structures.

At the level of secondary structures, the number of structures that an RNA sequence can form grows exponentially in sequence length [HSS98]. Introducing interactions further increases the number of potential conformations. On a computational level, it is infeasible to exhaustively explore a dynamic system of this size on a fine-grained level. Moreover, it is also implausible that an RNA in vivo explores the entire structure space to find its functional conformation[Lev69].

Following these two ideas, this work investigates the following questions:

- (RQ1) How do kinetics influence RNA-RNA interaction formation and is this observable in characteristic kinetic and steric properties of biologically functional RNAs?
- (RQ2) Can a reduced (coarse-grained) structure space, based on plausible assumptions on underlying biological processes, accurately capture the dynamics of RNA-RNA folding while reducing the complexity of the interaction formation problem?

By addressing these questions, this thesis aims to provide a deeper understanding of how kinetic constraints shape inter- and intra-molecular RNA-RNA interactions. It seeks to integrate dynamic processes that govern RNA function in living systems with the predominantly static, thermodynamic-based models used in interaction predictions so far.

4 Bioinformatics Methods

This chapter provides background on existing bioinformatics methods on top of which this thesis builds and presents some of the new definitions and algorithms in this context. Thereby it provides a more systematic background than what can be fitted into the individual papers. It covers formal notations, energy models, dynamic programming algorithms for secondary structure prediction and finding folding barriers, modeling folding as a Markov process, ML classifiers, covariance analysis, as well as a short introduction to RNA design. Some background sections and notations were adapted from the paper contributions presented in this thesis as well as from [Wal17].

4.1 Formal notation for RNA 2D structure and RNA-RNA interactions

A formal representation of RNA structure is essential for computational approaches. This section introduces notations for RNA primary and secondary structures based on [Wat78, WS78a] and presents the model used for describing RNA-RNA interactions in the newly developed kinetic framework.

Primary structure As discussed in the background chapter on the Biochemistry of RNAs and their interactions, the most basic structure level is the primary sequence.

The primary structure is the linear sequence of nucleotides and is represented as a string S over the alphabet $\Sigma = \{A, U, G, C\}$:

$$S = s_1 s_2 \dots s_n, \quad s_i \in \Sigma,\, i = 1, 2, \dots, n$$

Here, n is the length of the RNA sequence, and s_i denotes the nucleotide at position i. An example is: $S =$ "AUGCGAUGGUCAGU".

Secondary structure The secondary structure is described by the set of base pairs $R \subseteq (i, j)$, where $1 \leq i < j \leq n$, such that:

1. s_i and s_j form a valid base pair ($(s_i s_j) \in \{A-U, U-A, G-C, C-G, G-U, U-G\}$).
2. No two base pairs (i, j) and (k, l) overlap or cross (If $(i < k < j)$, then $k < l < j$).
3. each nucleotide can be involved in at most one pase pair ($i = k$ if and only if $j = l$).

4. base pairs that are less than three positions apart can not pair for steric reasons (for any base pair (i, j) $j - i - 1 > m$ with minimum hairpin loop length $m = 3$).

This defines a nested structure with no pseudoknots.

These restrictions on the base pairing cover the most common structures and allow efficient DP algorithms as presented below. Relaxing some of these constraints can help with covering a wider range of known biological structures, as for example non-canonical base pairs (eg used in the 3D embedding), modified nucleotides (strategies for detection of such are presented in [RWC+24]), crossing structures like pseudoknots (as e.g. investigated in [WTO+18]).

RNA-RNA interactions RNA-RNA interactions can be described by concatenating the sequences of the two interacting RNAs and applying the same formalism as for single RNA molecules. This approach assumes that interactions occur exclusively in exterior loops. An example for such a tools is RNAcofold [BTM+06]. This model can not predict interactions that would correspond to pseudoknots, e.g. H-type pseudoknots or kissing hairpins.

A more general representation can be achieved by separating intramolecular and intermolecular base pairs. This approach is for example used in methods like RNAup and IntaRNA [MTH+06, BRB08a, MWB17]. The intramolecular structures of the two RNAs still fulfill all the constraints introduced above. The base pairs between the two RNAs also fulfill the constraints with respect to all other interaction base pairs. However, interaction base pairs can cross intramolecular base pairs.

Interactions in the RRIkinDP model We study interactions of two RNAs with positions $1, \ldots, N$ and $N + 1, \ldots, N + M$. *RNA interaction structures* are defined at the level of secondary structure as sets of base pairs (i, j), $1 \leq i < j \leq N+M$ between the nucleotides of the RNA(s). Common non-crossing restrictions are imposed on both intramolecular structures P_R and Q_R and the intermolecular base pairs I_R. Observe that interaction base pairs I_R are naturally ordered by their position in the first RNA: $(i, j) \prec (i', j')$ if and only if $i < i'$.

In our kinetic model, we abstract from the intramolecular base pairs, treating them implicitly in the energy model as described below. This makes it possible to consider a state space $\mathcal{X}_C$ of states $S\langle k, l\rangle$ for all $1 \leq k \leq l \leq n$, which correspond to these "k, l"-equivalence classes. In this way, each state represents a set of interaction structures, which have identical interaction base pairs.

4.2 Nearest neighbor energy model

The probability of an interaction structure in thermodynamic equilibrium is determined by the free energy of the interaction. Furthermore, transitions between structures depend

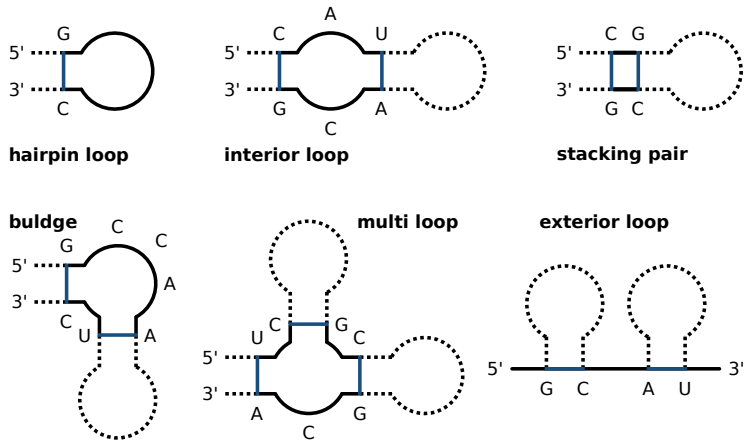


Figure 4.1: Loop types according to the nearest neighbor model. (Figure adapted from [Wal17]).

on the difference in their Gibbs free energies. Therefore, methods for predicting the free energy of a given structure are needed. A thermodynamic framework for such energy evaluations is provided by the nearest neighbor energy model [TUL71, TBD+73]. It decomposes a structure into structural motifs, called loops, and computes the free energy as the sum of the free energies of these loops:

Loop decomposition A loop is a part of an RNA structure that is enclosed by base pairs (and backbone) but does not contain any base pair within. Such loops can capture the important influence of the stacking between base pairs and entropic effects much better than energy models that use base pairs as smallest building blocks of structure. Loops differ in how many helices branch from them and how many unpaired bases they enclose. Based on the number of branching helices, three main types of loops can be identified: hairpin loops as the origin of a single helix, interior loops where two helices branch off, and multiloops that have more than two helices branching from them. Two special cases of interior loops can be identified based on the unpaired bases they enclose: If only one side of the interior loop includes unpaired bases a bulge is formed. If two base pairs are directly adjacent, they form a stacking pair. Figure 4.1 provides a graphical representation of these loop types.

Energy parameters Based on the decomposition of a structure R into its loops, its energy $E(R)$ can be computed from the sum over its loops:

$$E(R) = \sum_{L \in R} E_L$$

While the energy contribution of some small hairpin and interior loops have been determined in UV-melting experiments, most energies are derived from mathematical models by interpolation between experimental data points and a parameterized multi loop model. These parameters can be obtained from the Nearest Neighbor Data Base [MSZT99, MTM24].

4.3 Thermodynamic 2D structure prediction

Thermodynamic RNA structure prediction aims to find an RNA's structure by a minimization process, assuming that RNA adopts the most thermodynamically stable configuration under cellular conditions.

Decomposition of nested secondary structures

To come up with a description for all structures that can be formed by a sequence S, one can start by analyzing any structure that can be formed by a subsequence $S[i:j]$, where i is the first nucleotide of the subsequence and j is the last nucleotide of the subsequence.

For any structure formed by a subsequence $S[i:j]$, there are two ways in which the nucleotide in position i can contribute to the structure of the subsequence. Either its is unpaired or it is paired with another nucleotide k with $i < k \leq j$. In the first case the structure, defined as a set of base pairs, is the same as the structure for the sequence $S[i+1:j]$. In the other case the base pair (i,k) splits the sequence in two independent parts $S[i+1:k-1]$ and $S[k+1:j]$. This subsequences can be treated like the initial subsequence $S[i:j]$. Thus, a recursive decomposition of nested secondary RNA structures can be reached. Figure 4.2 is a visualization of this recursive decomposition.

It is possible to write this decomposition as a grammar:

$S \rightarrow X$

$X \rightarrow xX|uXaX|aXuX|gXcX|cXgX|gXuX|uXgX$

$X \rightarrow \epsilon$

Where X represents any nested structure, x represents any unpaired nucleotide, and a,c,g and u represents base pairing nucleotides.

Based on this grammar an algorithm for finding the structure with most base pairs, the number of possible structures or the structure with the minimum free energy can be derived.

As an example, $N_{i,j}$ can be defined as the number of possible structures of a given sequence S.

$$N_{i,j} = N_{i+1,j} + \sum_{k \in \text{ possible } (i,k) \text{ pairs}} (N_{i+1,k-1} \cdot N_{k+1,j})$$

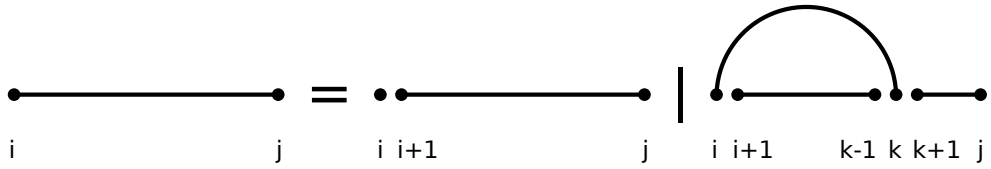


Figure 4.2: Decomposition of nested RNA secondary structures: Any secondary (sub-)structure $s[i:j]$ can be decomposed into smaller substructures by considering the possible contributions of the i^{th} nucleotide. Either the nucleotide is unpaired and the structure $s[i:j]$ is equivalent to the structure $s[i+1:j]$ or the nucleotide is paired with some other nucleotide in position k. Thus, the structure $s[i:j]$ is the base pair (i,k) plus the structures $s[i+1:k-1]$ and $s[k+1:j]$.

Nussinov algorithm

In 1980, Ruth Nussinov published an algorithm for predicting optimal secondary structures based on this decomposition. It is a dynamic programming algorithm that maximizes the number of base pairs [NPGK78b].

It uses following decomposition:

$$E_{i,j} = \max \begin{cases} E_{i+1,j} \\ \max\limits_{\substack{i<k\leq j \\ (S[i]S[k])\in B}} \Big\{ E_{i+1,k-1} + E_{k+1,j} + W(i,k) \end{cases}$$

with $W(i,k) = 1$, where $E_{i,j}$ is the maximum number of non-crossing base pairs in the interval from i to j and $W(i,k)$ represents the score added for every base pair that is introduced into the structure. As the algorithm is supposed to find the maximum number of base pairs, it should increase the base pair count by one for every base pair found. Therefore $W(i,k)$ is set to 1. Implicitly, the algorithm checks all possible nested RNA secondary structures and returns highest possible number of base pairs that can occur within one structure.

$E_{i,j}$ can also be interpreted as an energy that is maximized by maximizing the number of base pairs. The model can be extended by introducing a base type specific scoring [NJ80]:

$$W(i,k) = \begin{cases} 3 \text{ if } S[i]S[k] \in (GC, CG) \\ 2 \text{ if } S[i]S[k] \in (AU, UA) \\ 1 \text{ if } S[i]S[k] \in (GU, UG) \end{cases}$$

However, these maximum matching algorithms lead to poor results, mainly because the main contribution to free energy is not by the formation of base pairs through hydrogen bonds but by the stacking of π-electron systems. A major improvement to the algorithm was achieved by defining the energy function with respect to the loops that are enclosed by base pairs instead of the base pairs themselves [WS78c]. Using the nearest neighbor parameters introduced in the previous section, this enables a more reliable computation of the free energy of a given structure.

Zuker's algorithm

In 1981 Michael Zuker and Patrick Stiegler published a dynamic programming algorithm based on Nussinov's maximum matching algorithm that distinguished between hairpin, interior and multi loops. Thus, it is capable of computing the minimum free energy structure based on loop energies [ZS81b].

The original Zuker decomposition is not unique. Therefore, some structures are checked multiple times. While this does not influence the result of minimum free energy predictions, it cannot be used if properties of the structure ensemble, like the partition function [McC90b], are of interest. Furthermore, while it takes multi loops into account in the decomposition, it lacks a model for computing their energy contribution.

This was solved by adding the multi loop energy model introduced in the Nearest Neighbor parameters section and by extending the recursion by another dynamic programming matrix (M^1) [ZS84].

Nowadays, a unique recursive decomposition is widely used, amongst others in the RNAfold program of the ViennaRNA package [HFS$^+$94b, LBH$^+$11]. This is essential when the ensemble of structures that can be formed is of interest rather than the most stable structure(s).

4.4 Accessibility

In the previously-introduced definition of secondary structures, each nucleotide can be part of at most one base pair. Therefore, a nucleotide needs to be unpaired in the intermolecular structure to be able to form an interaction base pair.

RNA accessibility quantifies the likelihood that specific regions of a RNA molecule remain unpaired, making them available for interactions with other molecules. Computing accessibility involves evaluating the thermodynamic stability of all possible secondary structures and estimating the probability that a given nucleotide or segment remains unpaired.

To achieve this, partition functions are used to sum over all possible secondary structures

(partitions), weighted by their Boltzmann factors:

$$Z = \sum_{S} e^{-\frac{\Delta G(S)}{RT}}$$

where Z is the partition function, S denotes a possible secondary structure, $\Delta G(S)$ is its free energy, R is the gas constant, and T is the temperature.

The probability that a nucleotide i remains unpaired is given by:

$$P_{\text{unpaired}}(i) = \frac{Z_{\text{unpaired}}(i)}{Z}$$

where $Z_{\text{unpaired}}(i)$ is the partition function for structures in which nucleotide i remains unpaired.

For a contiguous segment of length l spanning from position i to $i + l - 1$, the probability that the entire region remains unpaired is:

$$P_{\text{unpaired}}(i, i + l - 1) = \frac{Z_{\text{unpaired}}(i, i + l - 1)}{Z}$$

where $Z_{\text{unpaired}}(i, i + l - 1)$ sums the Boltzmann weights of all configurations with these nucleotides unpaired.

This unpaired probabilities can be computed with the ViennaRNA package [LBH+11] and e.g. the included constraints frame work. Efficient computation of accessibility is a key component for thermodynamic interaction prediction tools introduced in the next section.

4.5 Direct interaction formation paths and folding landscape

This section describes how direct interaction folding paths can be summarized into an energy landscape and how the barriers on such direct folding paths can be computed.

4.5.1 Targeted interaction energy landscapes

Summarizing, RNA interaction landscapes can be described as energy landscapes [FFHS00], which consists of a state space, a neighborhood and an energy function. The targeted energy landscape for a pair of RNAs, given the targeted candidate interaction C, is defined by

- the state space $\mathcal{X}_C$,
- a neighborhood $\mathcal{N}$, characterized by changes of single interaction base pairs as defined below, and
- the energy function $\mathcal{E} : S\langle k, l\rangle \mapsto E(k, l)$.

The energy function E is determined according to the concrete model (equilibrated or frozen, as described previously). The neighborhood $\mathcal{N}$specifies the elementary steps, also called moves, that change a state by adding or removing a single base pair of I_C. Formally, $\mathcal{N}$ is a function from states $S\langle k,l\rangle$ to the set of its neighbor states, defined by $\mathcal{N}(S\langle k,l\rangle) = \{S\langle k-1,l\rangle, S\langle k,l+1\rangle, S\langle k+1,l\rangle, S\langle k,l-1\rangle\} \cap \mathcal{X}_C$.

Since any interaction structure along direct paths can be uniquely described by two indices, the entire space $\mathcal{X}_C$ of states $S\langle k,l\rangle$ can be arranged in two dimensions, with k on the x-axis and l on the y-axis.

4.5.2 Efficiently computing energy barriers

Following Arrhenius laws, the folding rate along a particular path p is dominated by the highest energy difference between its start and any other state on the path; let's call this difference *path saddle energy*. This suggests the study of "energy barriers" as quality scores for evaluating transitions.

In particular, we are interested in seed-specific energy barriers on transitions from seeds $s = (s_\mathrm{l}, s_\mathrm{r})$ to the full interaction. Here, we have to consider all possible paths between the two states. Since we study interaction growth between two states $x = S\langle s_\mathrm{l}, s_\mathrm{r}\rangle$ and $y = S\langle 1, n\rangle$, we define the *saddle energy* between x and y as the lowest path saddle energy among all *direct* paths from x to y. Then, the *energy barrier* between the states is the difference between their saddle energy and the energy of the starting state x.

Recall that successive states in direct paths are connected by moves that grow the interaction site, i.e. each state $S\langle k,l\rangle$, $i > 1$, is preceded by either $S\langle k+1,l\rangle$ or $S\langle k,l-1\rangle$.

To efficiently minimize the energy barrier from the seed to the full candidate interaction, let's more generally consider the saddle energy from seed s to $S\langle k,l\rangle$ and denote it as $B_s(k,l)$. Then, we are ultimately interested in the energy barrier $B_s(1,n) - E(s_\mathrm{l}, s_\mathrm{r})$, whereas $B_s(k,l)$ can be recursively obtained by

$$B_s(k,l) = \min \begin{cases} \max \begin{cases} E(k,l) \\ B_s(k+1,l) \end{cases} & \text{if } k < s_\mathrm{l} \\ \max \begin{cases} E(k,l) \\ B_s(k,l-1) \end{cases} & \text{if } l > s_\mathrm{r} \\ E(k,l) & \text{if } s_\mathrm{l} = k \text{ and } s_\mathrm{r} = l. \end{cases} \tag{4.1}$$

This recursion tailors a dynamic programming algorithm that allows us to compute energy barriers efficiently through memoization of $B_s(k,l)$.

The recursion minimizes over three cases: (i) an extension to the left, (ii) an extension to the right, and (iii) the initialization of the saddle energy as the free energy of the start interaction.

4.6 Interaction formation as Markov process

To understand the folding dynamics on a more fine grained level, we can model the interaction formation as a stochastic process within the state space defined by our direct paths model. Given the manageable size of the state space, it is feasible to solve the master equation of the transition system exactly. Starting with an initial structure or a probabilistic combination of initial structures, solving the master equation describes the time evolution of the system. This results in the probabilities of all structures/states, representing the interaction complex, over time.

$$\frac{P_u(t)}{dt} = \sum_{u \neq v} [P_v(t)k_{vu} - P_u(t)k_{uv}] \tag{4.2}$$

Here, $P_u(t)$ denotes the probability of the system being in state u at time t, a.e. the probability of the two RNAs forming interaction structure u at time t. k_{uv} is the transition rate from state u to state v. The term $P_v(t)k_{vu}$ represents the flux from state v to state u.

All possible interaction structures $S_{k,l}$ are considered for all combinations of k and l that satisfy $1 \leq k \leq l \leq n$, with n being the total number of interaction base pairs in the candidate structure. Neighboring states of $S_{a,b}$ are $\{S_{a-1,b}, S_{a+1,b}, S_{a,b-1}, S_{a,b+1}\}$.

Transition rates can be computed based on the energy landscape. We define a single base pair move in the interaction structure, along with the corresponding intramolecular rearrangement, as an elementary step. For neighboring states, transition rates follow a Metropolis-like criterion, from $\Delta G = E_v - E_u$:

$$k_{uv} = \begin{cases} k_0 \text{ if } \Delta G \leq 0, \\ k_0 e^{\frac{-\Delta G}{RT}} \text{ otherwise} \end{cases} \tag{4.3}$$

For non-neighboring states u and v, where the transition can not be achieved by adding or removing a single interaction base pair, the transition rate is $k_{uv} = 0$.

4.7 Evaluation of kinetic features

The kinetic featiures were evaluated on a data set of confirmed sRNA-mRNA interactions using a machine learning (ML) classifier. This section describes how the data set was curated and how the classifiers were applied:

4.7.1 Data set

A reliable evaluation requires high quality datasets with experimentally confirmed RNA–RNA interactions. Highly specialized interactions involving ribosomal RNA or tRNA were not selected, as they may not represent general RNA–RNA interactions. High-throughput

cross-linking datasets were also avoided due to their inherent noise. Instead, we focused on extending and manually curating a sRNA-mRNA interaction dataset centered on *E. coli*, starting with data from [RB12] and extending it with information from RNAinter [KTH+22b], RegulonDB [SGCL+23], and updates from recent publications.

The resulting dataset includes the names and sequences of sRNAs, their known mRNA regulatory targets represented by their 5' untranslated regions (UTRs) (from the transcription start site (TSS) to the translation start (TLS)+100; or TLS-200 to TLS+100 if TSS is unknown), experimentally confirmed base-pair interactions where available, and predicted minimum free energy (MFE) interactions obtained via IntaRNA [BRB08b] as lists of base pairs. This curated dataset is available in the supplementary materials, and all associated scripts are hosted on GitHub.

We also generated randomized datasets to serve as decoy interactions by performing dinucleotide shuffling on *E. coli* mRNA 5' UTRs, sRNAs, or both. Additionally, interactions were predicted between sRNAs and all extracted 5' UTRs of *E. coli* mRNAs to offer a baseline for off-target interactions and to highlight potential so far unknown functional interactions.

4.7.2 ML classifier

Thermodynamic and kinetic features were computed for each known interaction and decoy interactions to evaluate the classifier's ability to distinguish true interactions from decoy interactions. Because a given sRNA typically has far fewer target mRNAs than non-target mRNAs, we tested various unbalanced ratios of true-to-decoy interactions.

Three classifiers—a Support Vector Machine (SVM), linear discriminant analysis (LDA), and random forest— were trained to classify interactions as true or decoy based on various feature subsets. The baseline model used thermodynamic stability alone, establishing a benchmark accuracy without kinetic information. Subsequent models were trained with individual kinetic features and combinations of two, three, and four features to identify the most predictive feature sets. The classifiers were chosen for specific strengths: LDA for handling small datasets, SVM for managing non-linear feature relationships, and random forest for analyzing feature importance.

ML hyperparameters were tuned to small data sets and to prevent over fitting, with settings detailed in the supplementary materials. Feature importance was analyzed using established frameworks.

4.8 Interaction formation paths in 3D

Three different levels of coarse-graining were used to make the simulation of folding paths in 3D feasible. The folding paths were first computed in 2D, the starting point was embedded in 3D using Ernwin and SimRNA was used to refine the initial 3D structure and simulate the folding path in 3D guided by the 2D path [BLD+16].

Ernwin contributes a fast method for embedding a 2D structure in 3D by applying a fragment assembly strategy on the level of loops and helices.

SimRNA can refine the initial structure provided by Ernwin and represent the local folding space in more detail. It coarse-grains RNA structures to two pseudoatoms positioned at P and C4' to represent the phosphate and sugar moieties respectively and 2 to 3 pseudoatoms that represent the bases. It combines a knowledge-based potential with a Monte-Carlo simulation to sample 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 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.

4.9 Designing RNA-RNA interactions

Using the RRikinDP model, kinetic features were obtained that help to differentiate known sRNA-mRNA interactions from decoy interactions. Given the limited number of interactions that were identified with high confidence in experiments to date, an additional path to verifying the general validity of kinetic features is through designing interaction pairs with targeted kinetic features and verifying them in the lab. To this end an RNA sequence design approach was implemented that can generate RNAs with targeted kinetic properties. This implementation was done within the Infrared design framework because it provides an easy to use as well as flexible interface and an efficient sampling implementation [YMB$^+$24, YPW24].

At its core, Infrared employs a declarative framework powered by tree decomposition techniques to efficiently handle the complexities of RNA design. This approach allows for the systematic breakdown of complex RNA structures into simpler components, facilitating efficient computation and design. By leveraging these mathematical foundations, Infrared can perform weighted sampling algorithms that generate RNA sequences conforming to desired secondary structures and kinetic feature distributions.

In addition to enabling the verification of kinetic features in experiments, the integration of RRIkinDP with Infrared might be adapted and extended to designing regulators for specific mRNAs and further applications in synthetic biology or targeted medicine.

4.10 Evolution and covariance

The function of non-coding RNAs is strongly defined by their structure. During evolution pressure acts on RNAs to maintain their functional structure while the sequence can acquire mutations. If the sequences of multiple homologous RNAs (with a shared ancestor) are known conserved structural elements can be identified through the analysis of mutation patterns.

In more detail, the covariance analysis of sequence positions examines patterns of correlated mutations across multiple sequence alignments [ED94, BHW$^+$08]. The central assumption is: If two nucleotide positions in an RNA molecule consistently form base pairs, compensatory mutations (mutations that maintain base pairing) will often occur to preserve the structure. For example, if a guanine (G) is replaced by an adenine (A) at one position, a compensatory change from cytosine (C) to uracil (U) may occur at the paired position.

Mathematically, covariance analysis involves calculating a covariance score $C(i, j)$ for each pair of positions i and j in an aligned sequence set

$$C(i, j) = f_{i,j} - f_i f_j$$

where $f_{i,j}$ is the observed frequency of co-occurrence of specific nucleotide pairs at positions i and j, and f_i and f_j are the marginal frequencies of nucleotides at these positions.

This analysis provides insights into conserved base-pairing interactions, allowing researchers to infer secondary structures that are likely functionally important. Covariance analysis is a key component of structure prediction tools like R-scape and is frequently used in conjunction with comparative sequence analysis to improve the accuracy of RNA secondary structure models [RCE17].

Within this thesis this method was applied to detect interaction candidates:

- Telomerase RNAs in Saccharomycetes: The first example is the the pseudoknot in the telomerase RNA (TER) described in [WTO$^+$18]. The mutation rate and resulting sequence diversity in TERs made it necessary to follow multiple orthogonal approaches to identify TERs in new species. While interaction prediction tools contributed to the TER detection approach, we noticed that the interaction extent could not be fully explained by thermodynamic 2D structure models. This finding in part motivated the study of interaction kinetics.
- Combining computational and biophysical methods for RNA-RNA interaction detection: The second work during my PhD that makes use of covariance analysis was the investigation of a long range interaction in the RNA genome of the Japanese encephalitis virus (JEV) [MPW$^+$22]. The paper exemplarily shows how covariance analysis can be used to detect interaction candidates that were then further evaluated based on their kinetic features before being confirmed in experiments.

While the comparison of kinetics of homologous RNAs and their interactions were so far conducted manually, this could be formalized and automated in future.

5 Contributions

5.1 RRIkinDP: Targeted RNA–RNA interaction kinetics

Maria Waldl, Irene K. Beckmann, Sebastian Will and Ivo L. Hofacker (2025)
'RRIkinDP: Direct Paths RNA-RNA Interaction Kinetics Model',
bioRxiv, p. 2023.07.28.548983. doi: 10.1101/2023.07.28.548983v2 (preprint)

Summary: Understanding how kinetics influence RNA-RNA interaction formation is a key question of this thesis, and the RRIkinDP paper serves as a central contribution to addressing this challenge.

The paper introduces a novel model for RNA-RNA interaction kinetics within a restricted state space. Using this model, kinetic features are defined and evaluated on datasets of confirmed biological interactions and decoys. Meaningful features, such as the accessibility of interaction seeds and rapid folding into stable intermediate structures, are identified. The evaluation shows that the direct-paths-based model can differentiate functional RNA interactions from decoys, thus contributes to the understanding of what constitutes a functional interaction. To facilitate further research and applications, the model is made available as a free, open-source implementation. This tool enables case studies to explore specific interaction mechanisms and regulatory processes, as well as improve interaction prediction by serving as a filtering criterion for kinetically favorable interactions.

The RRIKinDP model serves as a foundation for the subsequent research presented in this thesis. It is utilized to generate 2D folding paths that guide 3D structure predictions, assess potential long-range interactions in the Japanese encephalitis virus—guiding the selection of candidates for experimental validation—and compute kinetic features in a RNA sequence design approach that optimizes both thermodynamic and kinetic properties of interactions.

Author's contributions: MW contributed to the framing of the research question and development of the direct paths approach; conceived essential steps in the methodology; organized the project and manuscript; performed the data collection, processing and analysis; implemented the software, wrote the paper draft and was involved in the manuscript revisions.

RRIkinDP: Targeted RNA–RNA interaction kinetics

Maria Waldl [0000-0001-7098-5712][1,2,3,4,*], Irene K. Beckmann [0000-0003-0929-9494][1,5], Sebastian Will [0000-0002-2376-9205][1,6], and Ivo L. Hofacker [0000-0001-7132-0800][1,7]

[1] *Institute for Theoretical Chemistry, University of Vienna, 1090 Vienna, Austria*
[2] *Vienna Doctoral School in Chemistry (DoSChem), University of Vienna, 1090 Vienna, Austria*
[3] *Center for Anatomy and Cell Biology, Medical University of Vienna, 1090 Vienna, Austria*
[4] *Department of Computer Science and Interdisciplinary Center for Bioinformatics, Leipzig University, 04109 Leipzig, Germany*
[5] *Vienna BioCenter PhD Program, Doctoral School of the University of Vienna and Medical University of Vienna, 1030 Vienna, Austria*
[6] *Laboratoire d'informatique de l'École polytechnique (LIX), Institut Polytechnique de Paris, 91120 Palaiseau, France*
[7] *Faculty of Computer Science, Research Group Bioinformatics and Computational Biology, University of Vienna, 1090 Vienna, Austria*

Abstract

Motivation: RNA–RNA interactions play essential roles in gene regulation and are controlled by both thermodynamics and kinetics. State-of-the-art tools predict thermodynamically optimal RNA–RNA interactions, but they neglect kinetic effects. While folding kinetics of single RNAs have been successfully modeled using transition systems between conformations, analogous approaches for RNA–RNA interactions quickly lead to infeasibly large systems. Novel models with controlled system size are required to overcome these limitations and enable computational analysis of RNA–RNA interaction kinetics. Such methods have the potential to improve our understanding of the governing principles of RNA–RNA interaction formation and improve target prediction tools.

Results: We propose a targeted interaction kinetics model that focuses on a given candidate interaction structure, and further limits the state space by considering only direct paths. This allows us to describe interaction formation as a Markov process on the limited state space and to study which properties are relevant for interaction formation. By comparing experimentally confirmed sRNA–mRNA interactions in *E. coli* with a randomized background, we show that native interactions are indeed kinetically favored and identify key features, such as seed accessibility and folding energy barrier. Using a machine learning classifier, we identified most-informative combinations of interaction features with respect to the kinetic behavior of native RNA–RNA interactions. Our kinetics model enables the efficient computation of various features that can be used to evaluate genome-wide target predictions by kinetic criteria. Beyond immediate practical improvements, our results contribute to long-standing general questions such as the influence of initial contact site accessibility.

Availability and implementation: RRIkinDP is available as free software on GitHub at `https://github.com/mwaldl/RRIkinDP`.

Contact: maria@tbi.univie.ac.at

1 Introduction

Non-coding RNAs (ncRNAs) act as principal agents in gene regulation. Typically, ncRNAs directly interact with other RNAs, such as messenger RNAs (mRNAs), via RNA–RNA interactions, which are influenced by thermodynamics and kinetics. Comprehensively understanding the role of ncRNAs in biological processes requires both identifying interaction

partners and, possibly even more, elucidating how stabilizing interaction structures form dynamically. Computational methods continue to be valuable in complementing wet-lab experiments, which remain costly, time-consuming, and challenging on a large scale. Moreover, although experiments often yield only partial information, e.g. approximate candidate target sites, complementary computation can provide detailed thermodynamic and kinetic interpretations.

Detailed structural data on RNA interactions are mostly available for a few very well-studied RNAs, such as tRNAs interacting with mRNAs in the ribosomal complex, or interactions involving microRNAs and small RNAs (sRNAs). Large scale interaction studies, e.g. based on crosslinking techniques (Aw et al., 2016; Cai et al., 2020; Ziv et al., 2018), despite general issues of accuracy and bias, first of all provide only interaction regions instead of detailed structure information. Computational methods are still indispensable to fill in the details by predicting the concrete RNA interactions and their features.

Thermodynamic interaction prediction and its limitations. Predicting RNA–RNA interaction structures with optimal, physically meaningful, free energy is NP-hard (Alkan et al., 2006). Consequently, several approaches have been developed to efficiently predict specific types of interactions. Some methods achieve high efficiency by imposing strict limitations, such as entirely forbidding intra-molecular base pairs (Rehmsmeier et al., 2004), or restricting inter-molecular base pairs to regions outside the intra-molecular structure (Bernhart et al., 2006). Other much less restricted methods remain computationally demanding (Alkan et al., 2006; Salari et al., 2010).

To resolve this dilemma, approaches like RNAup (Mückstein et al., 2006; Lorenz et al., 2011) and IntaRNA (Busch et al., 2008; Mann et al., 2017) predict complex interactions such as kissing-hairpins with high efficiency thanks to decomposing interactions into two steps: opening up of regions and their hybridization. These methods are suitable for large targeting-screens, especially when combined with index structures (Alkan et al., 2017).

While thermodynamic predictions already provide valuable information, previous studies have highlighted their limitations in identifying the true biological interactions. For instance, predicting interactions between an sRNA and the untranslated regions (UTRs) of all mRNAs in *E. coli* often ranks the biologically functional sRNA-mRNA interaction among the top 10-20% but generates numerous potential false positives that cannot be differentiated based purely on energy minimization (Busch et al., 2008; Mann et al., 2017). The high false positive rate indicates that stability alone cannot fully account for the functional relevance of interactions. Generally, while biologically functional interactions require a certain stability, it is finally decisive whether these stable interactions form sufficiently fast given the time scale of cellular processes.

Therefore, it is a key oversight and potentially critical limitation of existing methods to neglect the kinetic aspects of RNA–RNA interaction formation. RNA folding is a dynamic process, and kinetic factors such as seed stability, barriers along the folding path and the speed of folding play critical roles in determining whether an interaction can occur. Conventional models that focus solely on energy minimization overlook these dynamic pathways. Consequently, they are prone to predict stable interactions that do not form sufficiently fast; in turn, they are going to miss kinetically favorable interactions if they are energetically suboptimal.

The targeted interaction kinetics model. Resolving the dilemma between either neglecting crucial features or prohibitive computational complexity, we introduce a model of RNA–RNA interaction formation from an initial contact or *seed structure* to a given "targeted" *candidate structure*. This model allows studying interaction kinetics as a *stochastic process* on an interaction *energy landscape* in the sense of earlier work on single RNA folding (Flamm et al., 2000), reviewed by Flamm and Hofacker (2008). By focusing on the formation of the candidate interaction without detours through non-candidate interaction base pairs, we aim to capture important folding events while effectively restricting the state space and thereby enabling the efficient evaluation of kinetic interaction features.

Empirical evaluation and insights into kinetic features. The model is applied to a data set of sRNA–mRNA interaction in *E. coli*, using both native interactions and random-

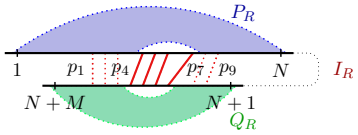


Figure 1: A state in the targeted interaction kinetics model that targets the interaction consisting of base pairs $p_1, \ldots, p_n$ ($n = 9$; solid and dotted, red). The shown state $S\langle 4, 7\rangle$ contains all candidate interaction base pairs from p_4 to p_7 (solid, red).

ized controls. This provides insights not only into the predictive capabilities of the computed kinetic features, but also into the broader principles governing RNA dynamics, such as the importance of an accessible interaction seed (start) and the interplay between intra- and intermolecular folding.

Algorithmic contributions, methods, and computational tools. We develop a methodological framework for the study of interaction kinetics. After precisely defining our targeted kinetics model, we provide methods that, based on this model, make it feasible to computationally study kinetics of the interaction between specific RNAs. First, we present an efficient dynamic programming (DP) algorithm to determine the folding barriers between seed interactions and full candidate interactions. Second, we establish techniques to solve the master equation of the kinetic interaction process, which allow us to calculate the probabilities of the single interaction states over time.

For deeper analysis based on our model, we introduce a machine learning approach to assess the suitability of features for differentiating native RNA–RNA interactions from a randomized background set of interactions; moreover, to quantify the information provided by these features.

Finally, we provide a pipeline for computing various interaction features, based on thermodynamics, putative first contact sites, and kinetics. Descriptive, non-trivial features are derived by applying our DP algorithm or by solving the master equation of the stochastic interaction formation process. Contributing intuitive visualization of the interaction landscape, we represent the state space as 2D energy landscape, enabling the visual assessment and study of interaction candidates.

2 Methods

2.1 Targeted interaction kinetics model

We study interactions of two RNAs with positions $1, \ldots, N$ and $N+1, \ldots, N+M$. *RNA interaction structures* are defined at the level of secondary structure as sets of base pairs (i, j), $1 \leq i < j \leq N + M$ between the nucleotides of the RNA(s). We impose the common non-crosssing restrictions (e.g. Lorenz et al., 2011) on both intra-molecular structures P_R and Q_R and the inter-molecular base pairs I_R. Observe that interaction base pairs I_R are naturally ordered by their position in the first RNA: $(i, j) \prec (i', j')$ if and only if $i < i'$.

Our primary goal is to evaluate the kinetic features and feasibility of a given *candidate interaction structure* C of the two RNAs. Such a candidate structure can be obtained, e.g. using thermodynamic prediction.

The modeled intermediary structures. We consider single-site interaction structures, where the interaction sites, defined by the minimal and maximal bases involved in interaction base pairs, are free of intramolecular structure. By targeting a candidate structure, we avoid considering the full space of interaction structures of two RNAs, but focus on direct formation of the target structure C starting from seeds.

Namely, we describe the step-by-step formation of the full interaction $I_C = \{bp_1, bp_2, ...bp_n\}$, $bp_1 \prec bp_2 \cdots \prec bp_n$ starting from *seed interactions* s, which could be single base pairs bp_k or short helices from bp_k to bp_l.

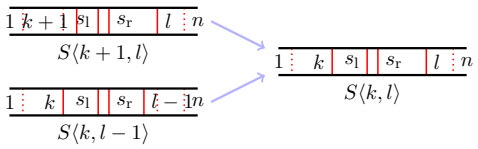


Figure 2: Transitions on the direct paths from a seed $s_\mathrm{l}, s_\mathrm{r}$ to the targeted interaction state $S\langle 1, n\rangle$. Each state $S\langle k, l\rangle$ ($k < s_\mathrm{l}$, $l > s_\mathrm{r}$) is accessed by an elementary move from either $S\langle k+1, l\rangle$ or $S\langle k, l-1\rangle$.

Generally, to keep our model well interpretable and computationally tractable, we model only *contiguous subsets of* I_C of the form $I_C^{k..l} := \{bp_k, bp_{k+1}, ...bp_l\}$ for $1 \leq k \leq l \leq n$. Any such interaction is characterized by the indices of its minimal and maximal candidate interaction base pair. We refer to the minimal and maximal base pair of the seed as s_l and s_r.

Partitioning of structures and states of the targeted kinetics model. Naturally, these interaction structures R with contiguous interaction subsets of I_C are partitioned into equivalence classes with identical $I_R = I_C^{k..l}$, $1 \leq k \leq\ l \leq n$.

In our kinetic model, we handle intra-molecular base pairs implicitly by the energy model. This makes it possible to consider a state space $\mathcal{X}_C$ of states $S\langle k, l\rangle$ for all $1 \leq k \leq l \leq n$, which correspond to these "k, l"-equivalence classes. In this way, each state represents a set of interaction structures, which have identical interaction base pairs. Figure 1 illustrates an example state.

Steps, paths, and direct paths. Moreover, the abstraction from intra-molecular structure allows focusing on elementary changes of the interaction site to study interaction dynamics. As elementary steps, we consider the addition or removal of an interaction base pair in I_C. Since states correspond to contiguous subsets of I_C, these changes happen only at the left or right end of the interaction.

Thus, on the path from the seed to the full interaction, one extends from $S\langle k, l\rangle$ to either $S\langle k-1, l\rangle$ or $S\langle k, l+1\rangle$. A *path* from x_1 to x_m is a sequence of states $x_1, \ldots, x_m$, where all successive states x_i, x_{i+1} differ by a single base pair. It is called *direct path*, if additionally all steps from x_i to x_{i+1} enlarge the interaction site (see Fig. 2).

2.2 Energy Model

Based on the free energies of interaction structures R in the Turner nearest neighbor energy model (Matthews et al., 2004), we assign free energies $E(k, l)$ to the individual states $S\langle k, l\rangle$. We consider two different variants of our model that both define the energy $E(k, l)$ of a state $S\langle k, l\rangle$ as the sum of a hybridization energy and a "unpairing" energy that is required to make the interaction site accessible:

$$E(k, l) = E^{\mathrm{hyb}}(k, l) + E^{\mathrm{unp}}(k, l). \tag{1}$$

$E^{\mathrm{hyb}}(k, l)$ is common to both variants and defined as the sum of the free energy contributions of the individual loops defined by $I_C^{k..l}$ according to the nearest neighbor model.

By defining E^{unp} in specific ways (see details below), the two variants cover extreme limit cases in the relation of inter- and intra-molecular folding speed. In the first variant, the *equilibrated model*, we assume that intra-molecular structure forms infinitely faster than inter-molecular structure, whereas we consider frozen intra-molecular structure in the second variant (*frozen model*). The true behavior is expected to between these extremes with similarly or equally fast formation of intra- and inter-molecular base pairs. Therefore, one can gain interesting insights by comparing these limit cases that can be effectively studied based on our models.

The equilibrated model. By our first model, which we also call *adiabatic model*, we describe the limit case, where the intramolecular structure of each state $S\langle k, l\rangle$ equilibrates

instantaneously after every change of inter-molecular structure. This is expressed by defining $E^{\text{unp}}(k,l)$ as the difference of two terms: 1) the sum of constrained ensemble energies of both RNAs, where the k,l-specific interaction regions are constrained to be free from intra-molecular structure and 2) the sum of the RNA's unconstrained ensemble energies. The "unpairing" energy $E^{\text{unp}}(k,l)$ is required to make the interaction site accessible for the formation of interaction.

Note that this model is closely related to the idea of RNAup by Mückstein et al. (2006), who compute the free energy of an RNA–RNA complex as the contribution of the interaction structure plus the cost of making the interaction site accessible in both intramolecular structures. The unpairing energies $E^{\text{unp}}(k,l)$ can be obtained efficiently by RNAup or RNAplfold Bernhart et al. (2011). The latter requires only cubic time in total with possible improvements to handle long sequences; both tools are implemented in the Vienna RNA package (Lorenz et al., 2011).

The frozen model. As contrasting limit case, the *frozen model* assumes that during the formation of the candidate interaction, the intra-molecular structure changes as little as possible with respect to given initial intra-molecular start structure of both RNAs, e.g. their minimum free energy (MFE) structures.

To this end, E^{unp} is defined as the energy difference between the initial intra-molecular structures of C and their parsimonious modifications, where the interaction site is cleared from all intra-molecular base pairs.

2.3 Targeted interaction energy landscapes

Summarizing, we describe RNA interaction landscapes as energy landscapes (Flamm and Hofacker, 2008), which consists of a state space, a neighborhood and an energy function. The *targeted energy landscape* for a pair of RNAs, given the targeted candidate interaction C, is defined by

- the state space $\mathcal{X}_C$,
- a neighborhood $\mathcal{N}$, characterized by changes of single interaction base pair as defined below, and
- the energy function $\mathcal{E} : S\langle k,l\rangle \mapsto E(k,l)$.

The energy function $\mathcal{E}$ is determined according to the concrete model (equilibrated or frozen, as described previously). The neighborhood $\mathcal{N}$specifies the elementary steps, also called moves, that change a state by adding or removing a single base pair of I_C. Formally, $\mathcal{N}$ is a function from states $S\langle k,l\rangle$ to the set of its *neighbor states*, defined by $\mathcal{N}(S\langle k,l\rangle) = \{S\langle k-1,l\rangle, S\langle k,l+1\rangle, S\langle k+1,l\rangle, S\langle k,l-1\rangle\} \cap \mathcal{X}_C$.

Along with the interaction energy landscapes, we introduce informative visualizations of RNA–RNA interaction formation. Since any interaction structure along direct paths can be uniquely described by two indices, the entire space $\mathcal{X}_C$ of states $S\langle k,l\rangle$ can be arranged in two dimensions, with k on the y-axis and l on the x-axis. This enables a heat-map like representation, where the free energy of an interaction structure is color-encoded. Fig. 3 provides an example, which is discussed in Sec. 3.

2.4 Efficiently computing energy barriers

According to the Arrhenius equation, the folding rate along a particular path p is dominated by the highest energy difference between its start and any other state on the path; let's call the latter *path saddle energy*. Consequently, we study "energy barriers" as quality scores for evaluating transitions.

In particular, we are interested in seed-specific energy barriers on transitions from seeds $s = (s_{\text{l}}, s_{\text{r}})$ to the full interaction. Here, we have to consider all possible paths between the two states. Since we study interaction growth between two states $x = S\langle s_{\text{l}}, s_{\text{r}}\rangle$ and $y = S\langle 1,n\rangle$, we define the *saddle energy* between x and y as the lowest path saddle energy among all *direct* paths from x to y. Then, the *energy barrier* from x to y is the difference between their saddle energy and the energy of the starting state x.

We denote the *saddle energy from seed s to $S\langle k,l\rangle$* as $B_s(k,l)$ and show how it can be computed efficiently by a dynamic programming algorithm: We initialize by $B_s(s_l,s_r) = E(s_l,s_r)$; then $B_s(k,l)$ can be calculated recursively for all other $k \leq s_l$ and $l \geq s_r$ by

$$B_s(k,l) = \max \begin{cases} \min \begin{cases} B_s(k+1,l) & \text{if } k < s_{\mathrm{l}} \\ B_s(k,l-1) & \text{if } l > s_{\mathrm{r}} \end{cases} \\ E(k,l) \end{cases} \tag{2}$$

This recursion tailors an efficient dynamic programming algorithm that computes saddle energies, and therefore the desired energy barriers.

Algorithmic details and correctness. The saddle energy from the seed s to $S\langle s_l,s_r\rangle$ is the seed energy $E(s_l,s_r)$. Recall that successive states in direct paths are connected by moves that grow the interaction site. Therefore, any path from s to a non-seed $S\langle k,l\rangle$, $k \leq s_l$ and $l \geq s_r$, extends a path from s to $S\langle k+1,l\rangle$ if $k < s_l$ or $S\langle k,l-1\rangle$ if $l > s_r$ (see Fig. 2). $B_s(k,l)$ is thus at least the minimum of $B_s(k+1,l)$ if $k < s_l$ and $B_s(k,l-1)$ if $l > s_r$. The maximization of Eq. 2 is finally justified by $B_s(k,l) \geq E(k,l)$.

The correctness of the barrier energy algorithm follows by induction from the correctness and the completeness of the case distinction in Eq. 2.

Algorithmic complexity. All possible values of $E(k,l)$ can be precomputed and stored in a matrix of $O(n^2)$, where n is the number of interaction base pairs in the candidate structure C. This optimization is also implemented in our software. Following precomputation, each seed-specific energy barrier is computed in $O(n^2)$ time and space, since the DP algorithm evaluates the recurrence Eq. 2 for all $1 \leq k \leq s_{\mathrm{l}}$, $s_{\mathrm{r}} \leq l \leq n$, where every single evaluation takes constant time.

Finding best seeeds. While finding the seed with optimal barrier energy would naively require cubic time by running the algorithm for each of the seeds, we accomplish the same in $O(n^2)$ time. For this purpose we evaluate a modification of Eq. 2 by DP and find the best seed by trace-back. In this modification, we drop the dependency on a specific s and instead initialize with 0 if $k = l$ (see Supplement).

Precomputation of state energies. In both model variants, i.e. the equilibrated and the frozen model, we perform the energy precomputation in $O(N^3)$ time depending on the lengths N and M of the interacting RNAs, assuming w.l.o.g. $N \geq M$. In more detail, all hybridization energies $E^{\mathrm{hyb}}(k,l)$ are calculated by a dynamic programming algorithm in time $O(N^2)$, since $E^{\mathrm{hyb}}(k,l)$ can be recursively obtained from $E^{\mathrm{hyb}}(k,l-1)$ with termination at $k = l$. Recall that, moreover, all required unpairing energies $E^{\mathrm{unp}}(k,l)$ can be calculated efficiently.

2.5 The stochastic targeted interaction process

To delve deeper into the folding dynamics, we model the interaction formation as a stochastic, continuous-time Markov process based on the targeted kinetics model. Note that proper abstractions in our model keep the state space small such that we can solve the master equation of the transition system exactly.

A Markov process is characterized by the Master equation based on a specific rate matrix and an initial probability distribution of states (e.g. starting from a specific seed). Solving this equation provides us with detailed information of the system's behavior over time. Concretely, we obtain the probabilities $P_x(t)$ of the system being in state x at time t, i.e. the probability of the two RNAs forming an interaction represented by state x at time t.

The Master equation has the form

$$\frac{P_x(t)}{dt} = \sum_{x \neq y} [P_y(t)k_{yx} - P_x(t)k_{xy}], \tag{3}$$

where the transition rate from state x to state y is k_{xy}, such that $P_x(t)k_{xy}$ expresses the flux from x to y. In these terms, the Master equation states that the probability change is equal to the difference of influx and outflux.

Recall that the energy landscape of our targeted model already defines the state space $\mathcal{X}_C$, the neighborhood $\mathcal{N}$, and the energy $\mathcal{E}$. To specify the Markov process, it remains to define the transition rates k_{xy}.

Since we consider only elementary steps, we define all transition rates between neighboring states $y \in \mathcal{N}(x)$ as Metropolis rates with prefactor 1:

$$k_{xy} = \exp(-\frac{\mathcal{E}^t(x,y) - \mathcal{E}(x)}{RT}) \tag{4}$$

where $\mathcal{E}^t(x,y) := \max(\mathcal{E}(x), \mathcal{E}(y))$ and all other rates are set to 0.

We consider two extensions of this fundamental process to provide insights into specific aspects of the interaction.

Absorbing full interaction state. Treating $S\langle 1,n\rangle$ as absorbing state allows us to model systems where the full interaction is removed from the process. This addresses biological systems, where the full interaction complex is sequestered due to performing its biological function and possibly further processed. We approximate this by adding another state S_f connected to the full interaction with significantly lower free energy, ensuring an almost 100% chance for any interaction reaching the full interaction to get trapped in this state. The energy difference between the full interaction state $S\langle 1,n\rangle$ and the trapped state S_f should be high enough to make the reverse rate from S_f to $S\langle 1,n\rangle$ negligibly small while maintaining numerical stability and detailed balance.

Modeling dissociation. Similarly, it is interesting to study dissociation of the interaction complex. At low concentrations of the RNAs, it is reasonable to assume that any complex that dissociates does effectively not contribute to the dynamics anymore, since association takes much longer than the conformational changes of the interaction complex. Analogously to the trapping state for the full interaction, we add an absorbing state to the dissociated state $S_\varnothing$.

Computing rates, solving of the Master Equation and visualization. To solve the Master equation, we employ Treekin by Wolfinger et al. (2004), which implements an efficient numerical solving algorithm based on matrix diagonalization. Concretely, we automate the calculation of the rate matrix for the targeted process, solution of the master equation by running Treekin, and the annotation and visualization of the results (Fig. 3). Additionally, we support model variations and modifications by absorbing states as previously described. Recall that the calculation of all the energies and therefore the rates in our models is efficient, as we elaborated in Section 2.4.

2.6 Features characterizing RNA–RNA interactions

Our models of interaction kinetics, along with the developed methods, allow us to characterize specific RNA–RNA interactions by a series of thermodynamic and kinetic features. As *thermodynamic features*, we study the energy of the full interaction E_{full}. This energy is the sum of further features: the hybridization energy $E^{\text{hyb}}_{\text{full}}$, and the unpairing energies E^{unp} of the full interaction site in the equilibrated model; further broken down for respective sequences a and b as $E^{\text{unp},\,a}$ and $E^{\text{unp},\,b}$.

Seed and barrier features. An important set of kinetic *barrier features* is based on the efficient computation of energy barriers by dynamic programming. We consider the minimum barrier energies B as well as corresponding maximum rates r_{barrier} from all possible seeds to the full interaction. As further class of kinetic features, *seed features* describe the seed interactions. Here we consider the energy, hybridization energy, and unpairing energy. In this way, we define features such as seed energy E_{seed}, seed hybridization energy $E^{\text{hyb}}_{\text{seed}}$, and seed unpairing energies $E^{\text{unp}}_{\text{seed}}$, $E^{\text{unp},\,a}_{\text{seed}}$, $E^{\text{unp},\,b}_{\text{seed}}$; each time we consider the best value across all seeds. Further corresponding features $p^{\text{unp}}_{\text{seed}}, \ldots$ can be derived by transforming unpairing energies E^{unp} into unpairing probabilities $p^{\text{unp}} = -RT\ln(E^{\text{unp}})$.

Features from the kinetic process. A variety of promising kinetic *process features* can be calculated after solving the stochastic interaction process. Most directly, we compute

probabilities of the full interaction and the dissociation state at time t. Moreover, we determine the mean interaction energies $\hat{E}(t)$ at time t. More precisely, we define the minimum over the processes starting at all possible seeds as the feature $\hat{E}(t)$. The mean interaction energies are defined as average of the state energies weighted by their probabilities at time t, formally $\hat{E}(t) = \sum_{x \in \mathcal{X}_C} P_x(t)\mathcal{E}(x)$.

The evaluation of the Markov process yields the probability of all states over time. Of special interest are the probability of the full interaction $P_{\text{full}}(t)$ and the dissociated state $P_{\text{diss}}(t)$ at time t. By setting both states as absorbing, we obtain the probability $P_{\text{diss}}(t = \infty)$ of dissociating before reaching full interactions (again, minimizing over all seeds). In practice, we approximate $P_{\text{diss}}(t = \infty)$ by the dissociation probability after a long simulation time, e.g. $t = 10^8$.

Feature variants and more fine-grained analysis.

While we defined all seed-dependent features as always representing the *optimal* value over all possible seed, the presented ideas and definitions can be extended in at least two ways. First, in addition to optimal values, we also considered *average* values as features. Second, the underlying values for the individual seed interactions serve fine-grained analysis of interactions in case studies.

3 Case study: the interaction of RyhB and sodB

Applying our model and developed toolset, we study the formation of the interaction of the small RNA RyhB with the 5' untranslated region of the mRNA of sodB in *E. coli*. The iron-regulated RhyB is known to regulate the expression of iron-storage and iron-using proteins, such as sodB, under iron-limiting conditions. In this way, iron is preserved for essential processes (Massé and Gottesman, 2002; Massé et al., 2005).

Our analysis starts with an IntaRNA-predicted candidate interaction structure consisting of 21 base pairs, which is in agreement with previous experimental studies (Massé and Gottesman, 2002). Fig. 3A shows the resulting energy landscapes of this candidate interaction, based on which we can identify and assess potential pathways from seeds to the full interaction and suggest mechanistic hypotheses of the formation of the interaction.

By solving the kinetic process, we obtain further direct insights, e.g., into the remarkable effect of specific interaction start sites (seeds). Fig. 3B shows the behavior of the process starting from the seed of base pairs 2–4 versus the one starting with base pairs 19–21. In the first case, part of the population extends to base pairs 1–5, where it apparently runs into a kinetic block and ultimately dissociates without ever reaching an interaction stability below –1kcal/mol. In contrast, the seed of base pairs 19–21 is less accessible but more stable. It quickly extends to stable partial interactions 14–21 (–7.35kcal/mol) and 13–21 (–8.05kcal/mol) and is predicted to reach the full interaction rather than to dissociate ($P_{\text{diss}}(10^8) < 0.01$). (More detailed results in Supplement)

4 Evaluation of the model and feature importance

The model quality cannot be directly evaluated by comparing predictions with experimentally measured (time-resolved) kinetic data because there is not sufficient reliable unbiased data, due to the challenges of informative experiments. The few available data sets are usually artificially designed and/or feature limited complexity in their folding paths. Therefore, they likely cover only a small fraction of interaction mechanisms and are prone to miss important and so far unknown features in interaction formation.

Lacking suitable experimental data, we propose to compare confirmed biologically evolved RNA interactions to a set of randomized, decoy interactions. After predicting interactions for both, biologically relevant pairs and randomized decoy pairs, we use our model to analyze their thermodynamic and kinetic features. First of all, this lets us assess the quality of the kinetic model. Second, differences in feature distributions between confirmed and decoy RNA–RNA interactions can indicate informative features. Third, we were mainly interested

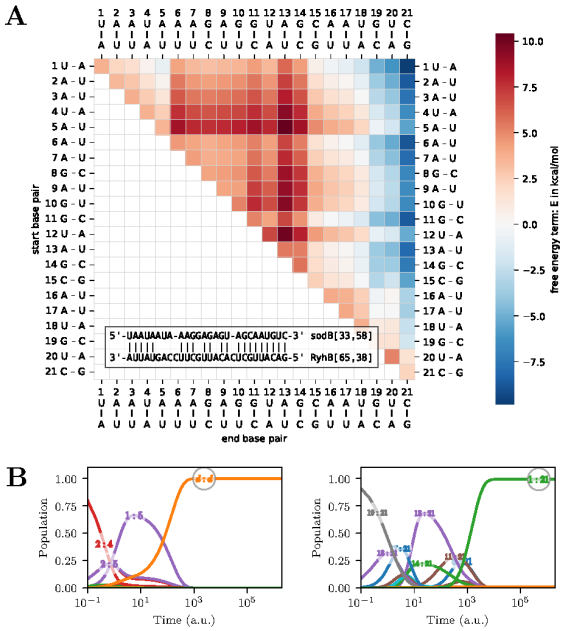


Figure 3: The interaction of RyhB and the 5' UTR of sodB in *E. coli* targeted at their MFE interaction. **A**) Energy landscape of the sub-interactions (states) in our targeted model of the RyhB–sodB interaction. States $S\langle i, j\rangle$ are defined by their first base pair i (row) and their last base pair j (column), with the their free energy ΔG in kcal/mol indicated by the red to blue color scale. Cells on the diagonal correspond to single base pair interactions. Moving from one cell to its right (upper) cell this matrix corresponds to extending the interaction by one base pair on the right (left). **B)** Population of sub-interactions $P_x(t)$ over time t, when starting from seed interaction of base pairs 2–4 (left) or from 19–21 (right). The former process dissociates ("d:d"), while the latter is finally absorbed in the full interaction ("1:21").

in identifying kinetic features that provide additional information beyond thermodynamic stability. To this end, we trained classifiers on both thermodynamic and kinetic features, evaluating whether the addition of kinetic features provided new, predictive information beyond thermodynamic stability alone (Sec. 4.2).

4.1 A dataset of confirmed and decoy interactions

A reliable evaluation requires high quality datasets with experimentally confirmed RNA–RNA interactions. Consequently, we avoided highly specialized interactions involving ribosomal RNA or tRNA, as they may not represent general RNA–RNA interactions. Moreover, we could not use high-throughput cross-linking datasets due to their inherent noise. Instead, we manually curated a dataset of sRNA-mRNA interactions in *E. coli*. We started with data from (Richter and Backofen, 2012) and extended it with information from RNAinter (Kang et al., 2022), as RegulonDB (Salgado et al., 2023), as well as single publications on specific sRNAs. The resulting data set includes the names and sequences of sRNAs, their known mRNA regulatory targets represented by their 5' untranslated regions (UTRs) (from the transcription start site [TSS] to TLS+100 or TLS-200 to TLS+100 if TSS is unknown), experimentally confirmed base-pair interactions where available, and predicted minimum free energy (MFE) interactions obtained via IntaRNA as lists of base pairs.

We generated sets of decoy interactions by dinucleotide shuffling on the 5' UTRs of *E. coli* mRNA, the sRNAs, or both. Additionally, interactions were predicted between sRNAs and all extracted 5' UTRs of *E. coli* mRNAs to offer a baseline for off-target interactions

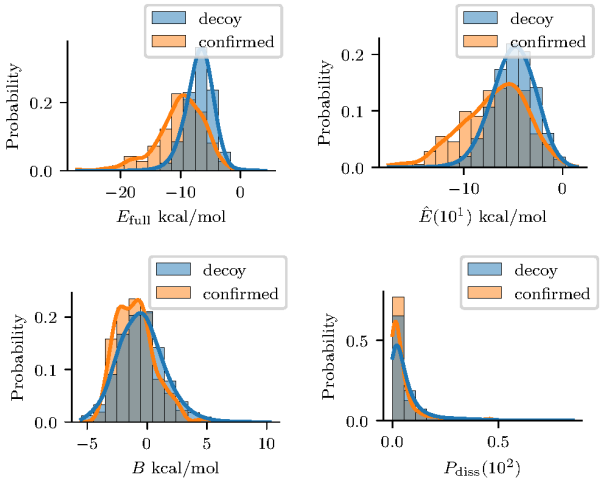


Figure 4: Distributions of selected features on confirmed (orange) and decoy interactions (blue). We show the features full energy E_{full}, minimum mean energy $\hat{E}(10^1)$ after time 10^1, minimum barrier energy B, and minimum dissociation probability $P_{\text{diss}}(10^2)$ across all possible seeds.

and highlighting potential so far unknown interactions.

The data sets along with complete references and curation details are provided in supplemental materials. Moreover, we provide supporting scripts on GitHub.

4.2 ML classifier

Thermodynamic and kinetic features were computed for each known interaction and decoy pair to evaluate the classifier's ability to distinguish true interactions from decoy interactions.

Because any given sRNA typically has far fewer target mRNAs than non-target mRNAs we tested various unbalanced ratios of true-to-decoy interactions. In our 10-fold cross-validation we take care to prevent data leakage, by keeping all interactions of the same sRNA either exclusively in the test or training set.

We implemented three classifiers—Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), and Random Forests (RF)—to classify interactions as true or decoy based on various feature subsets. Classifiers were first trained based on individual features, providing baselines. Subsequently, we trained on combinations of thermodynamic stability E_{full} with one or two other features in order to identify sets of features that are more informative than thermodynamics alone. The hyper-parameters were optimized to small data sets to prevent overfitting (Supplement). Scripts, trained models, and documentation are provided on GitHub.

4.3 Evaluation results

To assess the reliability and robustness of our results, we evaluated the features from our model in different settings. As such, we performed experiments for different data sets e.g. generating decoys in different ways (Sec. 4.1), compared the both intramolecular refolding variants of the model; and a range of parameter settings. Full details are provided in the supplement.

In the main text, we specifically report results based on exact interaction prediction by IntaRNA (without seeds heuristic) up to maximum interaction length of 32; moreover, we use seeds of length 3 in the RRIKinDP analysis, and define dissociation and the full interaction as absorbing states of the process. Concretely, we studied a set of 222 confirmed interactions in

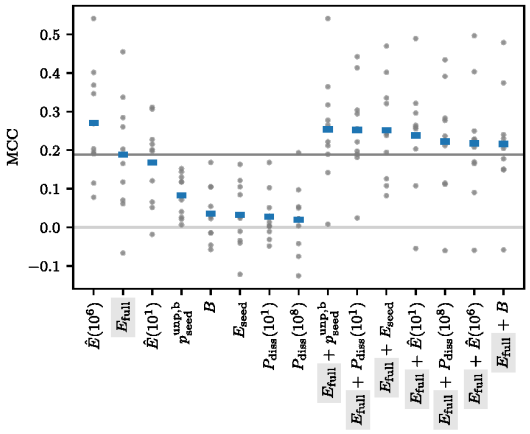


Figure 5: Random Forest classification performance of selected features and feature combinations; mean and single MCCs from 10-fold cross validation marked by a blue and gray markers respectively. The darker horizontal line represents the mean performance of the classifier using only the free energy of the full interaction (E_{full}), serving as base line for comparison.

E. coli, comprising 29 native sRNAs and 5'UTRs of the corresponding mRNAs. To generate decoys, we shuffled the UTRs of 4261 mRNAs of *E. coli*.

Recall that, for the purpose of classification, all seed-dependent features were determined as value at the best possible seed (Sec. 2.6). For example, for a given interaction, the feature E_{seed} is the minimal energy of all possible seeds and B is the minimum over the barriers starting from all possible seeds.

Figure 4, compares the distribution of selected features on the 222 confirmed interactions and 500 decoy interactions per sRNA, generated by di-nucleotide shuffling of the mRNAs. We observe that the thermodynamic feature of full energy, as computed by IntaRNA, cannot clearly separate the confirmed from the decoy interactions. At the same time, several kinetic features carry potentially additional information. Good examples of such features are interaction energies after limited time and barrier energies.

Figure 5 reports the observed classification performance (MCCs) of selected feature combinations. To assess the performance, we average the results from RF classification with 10-fold cross-validation on a dataset of confirmed interactions and 100 randomly shuffled mRNAs per sRNA. Note that while we studied three different classifiers, RF most clearly distinguished the performance of kinetic features and feature combinations while SVM sowed no significant differences, possibly due to limitations of the available data (see Discussion and Supplement). Of particular interest is the improvement over the thermodynamic base line E_{full}—which can be directly interpreted as classification performance of RNAup, which is equivalent to IntaRNA in its non-heuristic, exact mode. Remarkably, even single kinetic features derived from the kinetic process, notably mean energy $\hat{E}(t)$ after limited time t, let us improve over previous thermodynamics-based approaches. Namely, we observe a meaningful increase of the MCC by 42% based on $\hat{E}(10^6)$. Other kinetic features such as the barrier energy, the minimum seed energy, or the dissociation probability, while performing insufficiently as single features, strongly boost the classification in combination with E_{full}. Our systematic study of all feature pair and triplet combinations with full energy (Supplement) identifies several sets of kinetic features that inform the RF classifier beyond pure thermodynamics.

Features of the 'frozen' model seem to be generally less informative than features of the adiabatic model (results reported in the Supplement). This observation suggests that interaction formation requires at least small local rearrangements of intramolecular structure.

Notably, analogous experiments based on an alternative decoy set that randomly pairs

sRNAs with native mRNAs, yield highly similar results. Tables with kinetic and thermodynamic features for all sRNA–mRNA pairs are provided on GitHub for inspection of high scoring and potentially so far unknown regulatory pairs.

5 Discussion

Biomolecular structures have to be not only thermodynamically stable, but also kinetically accessible. The role of kinetics in RNA–RNA interaction formation has, however, so far received little attention. In this study, we introduce a first model of RNA–RNA interaction kinetics that supports complex RNA–RNA interactions, such as kissing hairpins. Given a target interaction, e.g. predicted one the basis of thermodynamic stability alone, we introduce a reduced conformation space, containing all direct paths to the target interaction, enabling the efficient calculation of various kinetic features on a large scale.

The model provides mechanistic insights into the formation of concrete RNA–RNA interactions by visualizing interaction energy landscapes. By integrating the dynamics on this energy landscape we can predict the probability of various states, e.g. corresponding to full and partial interactions or even dissociation, as a function of time. The case study of RyhB–sodB interaction in *E. coli* exemplifies this approach, showing how specific start sites (seeds) influence the kinetic behavior. The visualization technique developed here has already proven effective in previous work (Waldl et al., 2025; Beckmann et al., 2024; Mrozowich et al., 2023).

To evaluate our model we tested whether kinetic features derived from the model can help identify true RNA–RNA interactions. To this end we used a comprehensive set of confirmed sRNA–mRNA interactions from literature and contrasted them with a background of decoy interactions, obtained by prediction of interactions with shuffled RNA sequences. By comparing the interaction kinetics of both data sets, we identified specific kinetic features that are characteristic of true interactions and cannot be explained by thermodynamic stability alone.

To assess the relevance of kinetic features, we trained simple classifiers to distinguish true interactions from decoys, using either thermodynamic stability of the interaction alone or in combination with kinetic features. Several features, such as accessibility of the seed interaction, the dissociation probability, or energy barriers, can indeed improve the classification. Since all features can be efficiently calculated, they can readily be used to improve the accuracy of large-scale RNA targeting screens.

While our model provides concrete testable predictions, validation on available experimental data remains challenging. Even for well-established RNA–RNA interactions the exact boundaries (corresponding to our target interaction) are usually not known and direct time-resolved measurements are almost non-existent. Moreover, experiments usually only provide positive interactions, but no confirmed negatives, such as a set of interactions that are predicted thermodynamically, but that do not form in the experiment. The set of sRNA–mRNA interactions used in this work is not helpful in this respect, since it is likely to miss many biologically functional interactions.

In conclusion, our RNA–RNA interaction kinetics model offers a significant advancement in the computational study of RNA interactions. By incorporating kinetic factors, we provide a more detailed view RNA interactions, paving the way for improved target prediction tools and a better understanding of RNA function.

Acknowledgements

This work was supported by the Austrian Science Foundation (FWF): I-2874-N28 “Prediction of RNA-RNAinteractions”, W-1207 “DK RNA Biology“, I-45 20 “Deciphering Complex RNA structure by probing and predictions“ and F-80 “RNAdeco”; moreover, by the French ANR: ANR-21-CE45-0034 “INSSANE”; Additional funding was provided by the Novo Nordisk Foundation, grant NNF21OC0066551 “MATOMIC”.

References

Alkan, C., Karakoç, E., Nadeau, J. H. et al. RNA-RNA interaction prediction and antisense RNA target search. *J. Comput. Biol.*, 13(2):267–282, Mar. 2006. ISSN 1066-5277. doi: 10.1089/cmb.2006.13.267.

Alkan, F., Wenzel, A., Palasca, O. et al. RIsearch2: suffix array-based large-scale prediction of RNA–RNA interactions and siRNA off-targets. *Nucleic Acids Research*, 45(8):e60, May 2017. doi: 10.1093/nar/gkw1325.

Aw, J. G. A., Shen, Y., Wilm, A. et al. In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Mol. Cell*, 62(4):603–617, May 2016. doi: 10.1016/j.molcel.2016.04.028.

Beckmann, I. K., Waldl, M., Will, S., Hofacker, I. L. 3D feasibility of 2D RNA–RNA interaction paths by stepwise folding simulations. *RNA*, 30(2):113, Feb. 2024. doi: 10.1261/rna.079756.123.

Bernhart, S. H., Tafer, H., Mückstein, U. et al. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.*, 1(1):1–10, Jan. 2006. ISSN 1748-7188. doi: 10.1186/1748-7188-1-3.

Bernhart, S. H., Mückstein, U., Hofacker, I. L. RNA Accessibility in cubic time. *Algorithms Mol. Biol.*, 6(1):1–7, Dec. 2011. doi: 10.1186/1748-7188-6-3.

Busch, A., Richter, A. S., Backofen, R. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–2856, 10 2008. doi: 10.1093/bioinformatics/btn544.

Cai, Z., Cao, C., Ji, L. et al. RIC-seq for global in situ profiling of RNA–RNA spatial interactions. *Nature*, 582(7812):432–437, June 2020. doi: 10.1038/s41586-020-2249-1.

Flamm, C. Hofacker, I. L. Beyond energy minimization: approaches to the kinetic folding of RNA. *Monatsh. Chem.*, 139(4):447–457, Apr. 2008. doi: 10.1007/s00706-008-0895-3.

Flamm, C., Fontana, W., Hofacker, I. L., Schuster, P. RNA folding at elementary step resolution. *RNA*, 6(3):325–338, Mar. 2000. doi: 10.1017/S1355838200992161.

Kang, J., Tang, Q., He, J. et al. RNAInter v4.0: RNA interactome repository with redefined confidence scoring system and improved accessibility. *Nucleic Acids Res.*, 50(D1):D326–D332, Jan. 2022. doi: 10.1093/nar/gkab997.

Lorenz, R., Bernhart, S. H., Siederdissen, C. et al. ViennaRNA Package 2.0. *Algorithms Mol. Biol.*, 6(1):1–14, Dec. 2011. doi: 10.1186/1748-7188-6-26.

Mann, M., Wright, P. R., Backofen, R. IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions. *Nucleic Acids Research*, 45(W1):W435–W439, 05 2017. doi: 10.1093/nar/gkx279.

Massé, E. Gottesman, S. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. *Proc. Natl. Acad. Sci. U.S.A.*, 99(7):4620–4625, Apr. 2002. doi: 10.1073/pnas.032066599.

Massé, E., Vanderpool, C. K., Gottesman, S. Effect of RyhB small RNA on global iron use in Escherichia coli. *J. Bacteriol.*, 187(20):6962–6971, Oct. 2005. doi: 10.1128/JB.187.20.6962-6971.2005.

Matthews, D. H., Disney, M. D., Childs, J. L. et al. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceeding of the National Academy of Science of the USA*, 101:7287–7292, May 2004. doi: 10.1073/pnas.0401799101.

Mrozowich, T., Park, S. M., Waldl, M. et al. Investigating RNA–RNA interactions through computational and biophysical analysis. *Nucleic Acids Res.*, 51(9):4588–4601, May 2023. doi: 10.1093/nar/gkad223.

Mückstein, U., Tafer, H., Hackermüller, J. et al. Thermodynamics of RNA–RNA binding. *Bioinformatics*, 22(10):1177–1182, May 2006. doi: 10.1093/bioinformatics/btl024.

Rehmsmeier, M., Steffen, P., Höchsmann, M., Giegerich, R. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, Oct. 2004. ISSN 1355-8382. doi: 10.1261/rna.5248604.

Richter, A. S. Backofen, R. Accessibility and conservation: general features of bacterial small RNA-mRNA interactions? *RNA Biol.*, 9(7):954–965, July 2012. doi: 10.4161/rna.20294.

Salari, R., Möhl, M., Will, S. et al. Time and Space Efficient RNA-RNA Interaction Prediction via Sparse Folding. In *Research in Computational Molecular Biology*, pages 473–490. Springer, Berlin, Germany, 2010. ISBN 978-3-642-12683-3. doi: 10.1007/978-3-642-12683-3_31.

Salgado, H., Gama-Castro, S., Lara, P. et al. RegulonDB v12.0: a comprehensive resource of transcriptional regulation in E. coli K-12. *Nucleic Acids Res.*, 52(D1):D255, Nov. 2023. doi: 10.1093/nar/gkad1072.

Waldl, M., Yao, H.-T., Hofacker, I. L. Sequence Design for RNA-RNA Interactions. In *RNA Design: Methods and Protocols*, pages 1–16. 2025. doi: 10.1007/978-1-0716-4079-1_1.

Wolfinger, M. T., Svrcek-Seiler, W. A., Flamm, C. et al. Efficient computation of RNA folding dynamics. *J. Phys. A: Math. Gen.*, 37(17):4731, Apr. 2004. doi: 10.1088/0305-4470/37/17/005.

Ziv, O., Gabryelska, M. M., Lun, A. T. L. et al. COMRADES determines in vivo RNA structures and interactions. *Nat. Methods*, 15(10):785–788, Oct. 2018. doi: 10.1038/s41592-018-0121-0.

5.2 Local RNA folding revisited

Maria Waldl, Thomas Spicher, Ronny Lorenz, Irene K. Beckmann, Ivo L. Hofacker, Sarah Von Löhneysen and Peter F. Stadler (2023) 'Local RNA folding revisited',
Journal of Bioinformatics and Computational Biology,
21(04), p. 2350016. doi: 10.1142/S0219720023500166.

Summary: This paper provides a formalized framework for describing local folding approaches as for example implemented in RNAplfold. In such approaches, the RNA is folded within overlapping sequence windows. An overall base pair expectation value, approximating base pair probabilities, is computed by averaging over the probability of the base pair within each window that spans it. Such approaches provide a good approximation because most functional RNA elements within large transcripts are determined by local secondary structures rather than global folding patterns. Therefore local folding is a practical approximation of the more general global folding approach: Especially for long RNAs it often provides more accurate results while also reducing the computational complexity from $\mathcal{O}(n^3)$ to $\mathcal{O}(nL^2)$, where n is the sequence length and L the length of the local folding window. To enable more detailed work on localized folding, the ViennaRNA package was extended to enable stochastic backtracing in DP matrices of the McCaskill algorithm on sequence windows rather than full sequences. Stochastic sampling can for example be applied to compute dependencies between structural elements that are not accessible from the loop based combinatorial structure decomposition, e.g. the mutual information between sequence positions or conditional unpaired probabilities. In addition, the influence of parameters like window length, the size of stochastic samples and the maximum span of considered base pairs were (re)evaluated.

This work contributes to improving RNA-RNA interaction predictions and the kinetic model: Reducing the computational complexity of RNA folding enables fast accessibility prediction needed in large scale interaction screens. The evaluation of local folding parameters provides a guide on how to set parameters for accessibility computations in the interaction model. In addition, in large scale screens the boundaries of RNA molecules are not always known. By averaging over windows one can dampen the influence of artificial sequence ends that cut through structure elements.

Author's contributions: MW contributed to the framing of the research question and to large parts of the method design; organized the project; contributed significant parts of the data collection, software implementation and analysis as well as the interpretation of the results; wrote significant parts of the manuscript and contributed critical review to the other parts.

OPEN ACCESS
Journal of Bioinformatics and Computational Biology
Vol. 21, No. 4 (2023) 2350016 (19 pages)

DOI: 10.1142/S0219720023500166

World Scientific
www.worldscientific.com

Local RNA folding revisited

Maria Waldl[*,††,‡‡], Thomas Spicher[†,††,§§], Ronny Lorenz[‡],
Irene K. Beckmann[§,¶¶] and Ivo L. Hofacker[¶,||||]

Institute for Theoretical Chemistry
University of Vienna
Währingerstraße 17, 1090 Wien, Austria
[*]*maria@tbi.univie.ac.at*
[†]*tspicher@tbi.univie.ac.at*
[‡]*ronny@tbi.univie.ac.at*
[§]*irene@tbi.univie.ac.at*
[¶]*ivo@tbi.univie.ac.at*

Sarah Von Löhneysen[‖] and Peter F. Stadler[**,***]

Institute of Computer Science and Interdisciplinary Center
for Bioinformatics, Leipzig University
Härtelstraße 16-18, D-04107 Leipzig, Germany
[‖]*lsarah@bioinf.uni-leipzig.de*
[**]*studla@bioinf.uni-leipzig.de*

Received 30 March 2023
Accepted 5 June 2023
Published 28 July 2023

[††] These authors contributed equally to this work.

[‡‡] MW is also affiliated with Vienna Doctoral School in Chemistry (DoSChem), University of Vienna, Währinger Str. 42, 1090 Vienna, Austria; Center for Anatomy and Cell Biology, Medical University of Vienna, 1090 Vienna, Austria.

[§§] TS is also affiliated with UniVie Doctoral School Computer Science (DoCS), University of Vienna, Währinger Str. 29, 1090 Vienna, Austria.

[¶¶] IKB is also affiliated with Vienna BioCenter PhD Program, Doctoral School of the University of Vienna and Medical University of Vienna, A-1030, Vienna, Austria.

[||||] ILH is also affiliated with Faculty of Computer Science, Research Group Bioinformatics and Computational Biology, University of Vienna, Währingerstraße 29, 1090 Vienna, Austria.

[***] PFS is also affiliated with Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany; Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria; Facultad de Ciencias, Universidad National de Colombia, Bogotá, Colombia; Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA.

[**] Corresponding author.

Most of the functional RNA elements located within large transcripts are local. Local folding therefore serves a practically useful approximation to global structure prediction. Due to the sensitivity of RNA secondary structure prediction to the exact definition of sequence ends, accuracy can be increased by averaging local structure predictions over multiple, overlapping sequence windows. These averages can be computed efficiently by dynamic programming. Here we revisit the local folding problem, present a concise mathematical formalization that generalizes previous approaches and show that correct Boltzmann samples can be obtained by local stochastic backtracing in McCaskill's algorithms but not from local folding recursions. Corresponding new features are implemented in the `ViennaRNA` package to improve the support of local folding. Applications include the computation of maximum expected accuracy structures from `RNAplfold` data and a mutual information measure to quantify the sensitivity of individual sequence positions.

Keywords: RNA folding; partition function; local structure.

1. Introduction

Long-range base pairs are a notorious difficulty for RNA secondary structure prediction algorithms[10,40] and may be subject to kinetic folding effects.[29] Nevertheless, there is ample evidence that long-range base pairs do exist in nature, exemplified e.g. by the panhandle structure of RNA viruses[1] and the long-range base pairings associated with pre-mRNA processing.[16] Statistical arguments, furthermore, show that such configurations are in fact to be expected in random RNA sequences,[5,12,37] a fact that was also confirmed experimentally.[21]

Despite these specific exceptions, local RNA secondary structures are of particular importance in practice. The Rfam database,[17] for instance, lists hundreds of evolutionarily conserved RNA elements, including riboswitches, thermoregulators, leader sequences, frameshift elements, and internal ribosomal entry sites. Even in the absence of sequence conservation, local secondary structures also serve as a key predictor for protein binding affinity.[31] Local structures thus are of key importance in particular in long RNAs such as complete mRNAs. Moreover, local folding seems to be a good way to handle structure prediction for RNA sequences whose ends are not known with precision.

Several tools[2,29] therefore modify the standard RNA energy model[34] and treat base pairs (k, l) in dependence of their *span* $l - k + 1$. Since long range base pairs are so poorly predicted by global energy minimization,[10] this can actually improve overall predicition accuracy. In the simplest case one may simply omit pairs exceeding some cut-off value L. Several implementations for this strategy are available, including `RNALfold`[13] and `Rfold`.[19] A positive side effect is that computation complexity of the folding algorithms drop from $O(n^3)$ to $O(nL^2)$, where n denotes the length of the RNA sequence. A down-side of these approaches, however, is that they are still not truely local since partition function terms of the form $Z_{1,k-1}$ and $Z_{l+1,n}$ still appear in the pertinent recursions, in other words, the probability of the base pair (k, l) also depends on the sequence outside the intervall $[k..l]$.

`RNAplfold`[3] and its variants[4,20] approximate these long-range contributions by truncated terms that cover only a fixed-length window instead of the entire length of

the input sequence. To compensate at least in part for the approximation error, base pairing probabilities are then averaged over all overlapping windows that contain the base pair (k, l). The averages π_{kl} computed by `RNAplfold`, however, have some unexpected properties rooted in their definition as an average over sets of windows. For example, the sum $\sum_{l>k}\pi_{kl}$ is not bounded by 1. In this contribution we provide a more formal account of `RNAplfold` in terms of a statistical mechanics model that generalizes the secondary structure model. The main virtue of this approach is that it clarifies the meaning of `RNAplfold`'s π_{kl} and makes it transparent how other observables are to be computed in this setting. The probability q_k that a base k is unpaired is of particular practical utility in this context. Revisting the theory behind `RNAplfold` also shows how hard and soft constraints[24] can be handled consistently in the setting of `RNAplfold`. This is relevant for instance to devise assisted folding algorithms.[25,35]

Complementary to the efficient recursive computation of observables such as π_{kl} and q_k we consider the problem of producing Boltzmann-weighted samples of localized structures. In conventional RNA folding algorithms, such samples are obtained with the help of stochastic backtracing[8,32] in partition function calculations. As we shall see, local structures cannot be correctly sampled from the recursion used in `RNAplfold`. Instead, the forward recursions of McCaskill's partition function algorithm[27] possibly truncated to a limited base pair span, can be used directly.

The purpose of this contribution is two-fold. First, we revisit the formal foundations of `RNAplfold` and clarify what exactly is computed when local structures in overlapping windows are averaged. Second, we introduce extensions to the algorithms implemented in the `ViennaRNA` package to support local structure predictions systematically. For instance, access to stochastic sampling of local structures has been added. In Sec. 3 we then provide some illustrative showcase applications.

2. Theory

2.1. *The PL model*

Consider a fixed RNA sequence of length n. Furthermore, we define a set $\mathfrak{I}$ of sub-intervals $\langle i\rangle \subseteq [1..n]$. For each interval $\langle i\rangle \in \mathfrak{I}$ we consider the ensemble of possible RNA secondary structures $\Omega(\langle i\rangle)$ on the sequence fragment $\langle i\rangle$, see Fig. 1. With each secondary structure $x_{\langle i\rangle} \subseteq \Omega(\langle i\rangle)$ we associate an energy $E(x_{\langle i\rangle})$.

Definition 1. A *configuration* x is a set $\{x_{\langle i\rangle} | \langle i\rangle \in \mathfrak{I}\}$ with associated energy $E(x) := \sum_{\langle i\rangle\in\mathfrak{I}} E(x_{\langle i\rangle})$.

A configuration as defined above is a collection of local structures, one for each sub-sequence $\langle i\rangle \in \mathfrak{I}$. Note that at this point we make no further assumption on $\mathfrak{I}$. In particular, the sub-intervals may overlap, and may fail to cover $[1, n]$. In the special case $\mathfrak{I} = \{[1..n]\}$, a configuration is simply a global RNA structure. The *PL model* therefore contains the standard model of RNA folding as a special case. The global

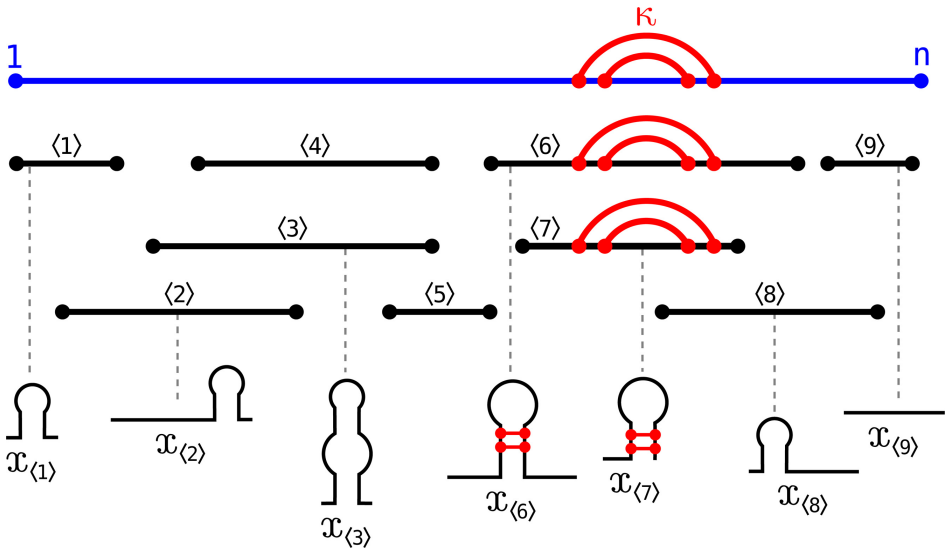


Fig. 1. The PL model prescribes a fixed sets of intervals or subsequenes $\mathfrak{I}$, here labeled $\langle 1\rangle$ through $\langle 9\rangle$, for the input sequence shown on top. In this example, in addition two base pairs (red arcs) are specified as a hard constraint κ. A configuration x of the PL model comprises a secondary structure $x_{\langle i\rangle}$ for each subsequence $\langle i\rangle \in \mathfrak{I}$, shown at the bottom. The constraint κ is inherited by all intervals that completely contain the constraint. Since the energy $E(x)$ of the configuration x is the sum of the folding energies of the individual intervals, x is a mininum energy configuration if and only if each local fold $x_{\langle i\rangle}$ is a mininum energy fold on the corresponding subsequence $\langle i\rangle$.

folding model with restricted base pair span is also covered by the PL model. In this case $E(x) = \infty$ whenever $x_{[1,n]}$ contains as base pair with a span exceeding the locality constraint. Importantly, the local components $x_{\langle i\rangle}$ of x are considered to be independent. The PL model is especially suitable for situations where the sequence is a genomic region and $\mathfrak{I}$ represents the set of RNA transcripts that could be generated from that region. Alternatively, we may think of $\mathfrak{I}$ as a collection of fragments of an RNA with structures determined independently for each fragment. In either setting, the PL model considers RNA structures as "composites" of such a collection of local structures.

In the following we provide a formal analysis of the PL model and show how constraints such as enforced or forbidden basepairs are accommodated. As an immediate consequence of Def. 1, the partition function of the PL model can be expressed in the following form:

$$Z = \sum_x \exp(-E(x)/RT) = \prod_{\langle i\rangle\in\mathfrak{I}} \sum_{x_{\langle i\rangle}} \exp(-E(x_{\langle i\rangle})/RT) = \prod_{\langle i\rangle\in\mathfrak{I}} Z_{\langle i\rangle} \qquad (1)$$

Additivity of the energies of PL configurations yields $\exp(-E(x)/RT) = \prod_{\langle i\rangle\in\mathfrak{I}} \exp(-E(x)/RT)$. Since the sum over all PL configurations ($\sum_x$) is the sum over all combinations of secondary structures on the individual intervals and their contributions to Z are multiplicative, we obtain the second equality. The exponential functions in the third term are simply the Boltzmann factor for local structures $x_{\langle i\rangle}$.

Their sum is the partition function of the local secondary structures on interval $\langle i\rangle$. Equation (1) thus shows that the partition function of the PL model is the product of the partition functions of the contributing local structure ensembles. Correspondingly, the probability of a given PL model configuration x is

$$\mathbb{P}[x] := \frac{\exp(-E(x)/RT)}{Z} = \prod_{\langle i\rangle\in\mathfrak{I}} \frac{\exp(-E(x_{\langle i\rangle}))}{Z_{\langle i\rangle}} = \prod_{\langle i\rangle\in\mathfrak{I}} \mathbb{P}[x_{\langle i\rangle}] \tag{2}$$

The probabilities $\mathbb{P}[x_{\langle i\rangle}]$ of the local folds $x_{\langle i\rangle}$ on the individual intervals $\langle i\rangle \in \mathfrak{I}$ contribute independently to the probability of PL configuration.

In its most general form, a *constraint* κ is a combination of enforced and forbidden base pairs.[24] Denote by $\mathrm{span}(\kappa)$ the "footprint" of κ, i.e., the minimal interval that contains κ. Naturally, then consider the partition function $Z^{[\kappa]}$ of the "restricted ensemble" comprising all structures that satisfy the constraint κ in every interval $\langle i\rangle$ such that $\mathrm{span}(\kappa) \subseteq \langle i\rangle$, see Fig. 1. The constrained partition function therefore is

$$Z^{[\kappa]} = \prod_{\langle i\rangle\in\mathfrak{I}} Z^{[\kappa]}_{\langle i\rangle} \tag{3}$$

where $Z^{[\kappa]}_{\langle i\rangle}$ is the restricted partition function on interval $\langle i\rangle$. Here we will only consider the simple case of constraints that only affect intervals $\langle i\rangle$ with $\mathrm{span}(\kappa) \subseteq \langle i\rangle$. In this case we have $Z^{[\kappa]}_{\langle i\rangle} = Z_{\langle i\rangle}$ whenever $\mathrm{span}(\kappa) \not\subseteq \langle i\rangle$. We note in passing that one might also consider more elaborate models of constraints where κ also affects intervals with $\mathrm{span}(\kappa) \cap \langle i\rangle \neq \emptyset$. In our setting, the probability of picking a configuration that satisfies the constraint κ is given by

$$\begin{aligned}\mathbb{P}[\kappa] :=\ & \frac{Z^{[\kappa]}}{Z} = \frac{\prod_{\langle i\rangle\in\mathfrak{I}} Z^{[\kappa]}_{\langle i\rangle}}{\prod_{\langle i\rangle\in\mathfrak{I}} Z_{\langle i\rangle}} = \prod_{\langle i\rangle\in\mathfrak{I}} \frac{Z^{[\kappa]}_{\langle i\rangle}}{Z_{\langle i\rangle}} = \prod_{\substack{\langle i\rangle\in\mathfrak{I}\\ \mathrm{span}(\kappa)\subseteq\langle i\rangle}} \frac{Z^{[\kappa]}_{\langle i\rangle}}{Z_{\langle i\rangle}} \\ =\ & \prod_{\substack{\langle i\rangle\in\mathfrak{I}\\ \mathrm{span}(\kappa)\subseteq\langle i\rangle}} \mathbb{P}_{\langle i\rangle}[\kappa]\end{aligned} \tag{4}$$

where $\mathbb{P}_{\langle i\rangle}[\kappa]$ is the probability of the occurrence of κ in the "usual" RNAfold partition function computed from interval $\langle i\rangle$. Considering a base pair (k,l) as constraint, we see that Eq. (4) does **not** yield the "probabilities" computed by `RNAplfold`.

2.2. `RNAplfold` *revisited*

In order to connect the quantities computated by `RNAplfold` to the configurations of the PL model introduced above, we start from the probabilities $p^{\langle i\rangle}_{kl} := \mathbb{P}_{\langle i\rangle}[(k,l)]$ of finding the base pair (k,l) in the interval $\langle i\rangle$ under the usual RNA folding model.

Moreover, we write $Z^{\langle i\rangle}_{\min\langle i\rangle,k-1}$ for the partition function of the ensemble of RNA secondary structure on the sub-sequences from the 5'end $\min\langle i\rangle$ of the interval $\langle i\rangle$ to $k-1$. Correspondingly, $Z^{\langle i\rangle}_{l+1,\max\langle i\rangle}$ denotes the partition function from $l+1$ to the upper end $\max\langle i\rangle$ of the interval $\langle i\rangle$. These terms are well-defined whenever

their bounds are contained in $\langle i\rangle$. Otherwise, no such (local) structure may exist. We therefore *define* $Z^{\langle i\rangle}_{\min\langle i\rangle,k-1} = 0$ for $k-1 < \min\langle i\rangle$, and $Z^{\langle i\rangle}_{l+1,\max\langle i\rangle} = 0$ for $l+1 > \max\langle i\rangle$, respectively. The partition function Z^B_{kl} for the structures enclosed by the base pair (k,l) is independent of the choice of $\langle i\rangle$ as long as $[k,l] \subseteq \langle i\rangle$. Similarly, the conditional partition function to find the base pair (k,l) immediate interior of the pair (r,s) is also independent of $\langle i\rangle$ as long as $[r,s] \in \langle i\rangle$ and thus in particular whenever the probability $\mathbb{P}_{\langle i\rangle}[(r,s)] > 0$ of the enclosing pair (r,s) does not vanish. Note that in practice the evaluation of the conditional probability $\Xi_{\langle i\rangle}(k,l|r,s)$ of observing the base pair (k,l) provided the base pair (r,s) is present involves the multi-loop decompostion in exactly the same way as in global RNA folding. Then we have

$$\mathbb{P}_{\langle i\rangle}[(k,l)] = \frac{Z^{\langle i\rangle}_{\min\langle i\rangle,k-1} Z^B_{kl} Z^{\langle i\rangle}_{\max\langle i\rangle,*}}{Z_{\langle i\rangle}} + \sum_{rs} \Xi(k,l|r,s)\mathbb{P}_{\langle i\rangle}[(r,s)] \tag{5}$$

Let us now define the observable

$$\mathbb{E}[k,l] := \sum_{\langle i\rangle\in\mathfrak{I}} \mathbb{P}_{\langle i\rangle}[(k,l)] \tag{6}$$

as the expected number of intervals $\langle i\rangle \in \mathfrak{I}$ in which the base pair (k,l) is observed in configuration sampled from the PL model. Summing Eq. (5) over the intervals yield

$$\mathbb{E}[k,l] = \sum_{\langle i\rangle\in\mathfrak{I}} \frac{Z^{\langle i\rangle}_{\min\langle i\rangle,k-1} Z^B_{kl} Z^{\langle i\rangle}_{l+1,\max\langle i\rangle}}{Z_{\langle i\rangle}} + \sum_{rs} \Xi(k,l|r,s)\mathbb{E}[r,s] \tag{7}$$

By definition $\mathbb{E}[k,l]$ is *not* a probability but an expected value.

In `RNAplfold`, the expected values $\mathbb{E}[k,l]$ are normalized with respect to the number of intervals in which a given base pair *can* appear. Thus, we introduce

$$h_{kl} := |\{\langle i\rangle \in \mathfrak{I} | (k,l) \in \langle i\rangle\}| \tag{8}$$

and define the normalized observables $\pi_{kl} := \mathbb{E}[k,l]/h_{kl}$ if $h_{kl} > 0$. Then we can re-write Eq. (7) as

$$\pi_{kl} = \frac{1}{h_{kl}}\mathbb{E}[k,l] = \frac{1}{h_{kl}} \sum_{\langle i\rangle\in\mathfrak{I}} \frac{Z^{\langle i\rangle}_{\min\langle i\rangle,k-1} Z^B_{kl} Z^{\langle i\rangle}_{l+1,\max\langle i\rangle}}{Z_{\langle i\rangle}} + \sum_{rs} \Xi(k,l|r,s)\,\frac{h_{rs}}{h_{kl}}\,\pi_{rs} \tag{9}$$

Equation (9) recovers exactly the recursions[3] of `RNAplfold` provided $\mathfrak{I}$ is defined as the set of intervals with fixed length W and the base pair span is limited to $L \le W$. In this case, we can easily compute the number h_{rs} of intervals $\langle i\rangle \in \mathfrak{I}$ that contain the base pair (r,s) in explicit form:

$$h_{rs} = \begin{cases} W-(s-r) & \text{if } s \ge W \text{ and } r \le n-W+1 \\ r & \text{if } s \le W \\ n-s+1 & \text{if } r > n-W+1 \\ n-W+1 & \text{if } r > n-W+1 \text{ and } s \le W \end{cases} \tag{10}$$

The quantity π_{kl} can therefore be interpreted as the probability of observing the basepair (k, l) in an interval of length W that is picked at random from the set of all intervals of length W that contains both k and l. Note that the window length W corresponds to $L + 1$ in the publication[3] describing `RNAplfold` since there intervals $\langle i\rangle = [i \,..\, i + L]$ are considered. These have length $W = L + 1$.

2.3. *The general case and the probability of being unpaired*

In general we may consider arbitrary secondary structure features μ and *define* under which conditions μ *properly occurs* within the interval $\langle i\rangle$. In the case of base pairs $\mu = (k, l)$ in `RNAplfold` we require the spanning interval $\text{span}(\mu) := [k, l] \subseteq \langle i\rangle$. In an attempt to remove border effects Lange *et al.*[20] required that the base pair (k, l) is a minimum distance away from $\min\langle i\rangle$ and $\max\langle i\rangle$. The definition of $\mathbb{E}[\mu]$, the expected number of occurrences, is naturally modified as

$$\mathbb{E}[\mu] = \sum_{\substack{\langle i\rangle\in\mathfrak{I}\\ \text{span}(\mu)\subseteq\langle i\rangle}} \mathbb{P}_{\langle i\rangle}[\mu] \tag{11}$$

where $\mathbb{P}_{\langle i\rangle}[\mu] = Z_{\langle i\rangle}[\mu]/Z_{\langle i\rangle}$ is the probability of observing feature μ in the ensemble of local structure on interval $\langle i\rangle$. Note that $Z_{\langle i\rangle}[\mu]$ is the partition function of $\langle i\rangle$ constrained to secondary structures containing the feature μ. Correspondingly, $h_\mu := |\{\langle i\rangle \in \mathfrak{I} | \text{span}(\mu) \subseteq \langle i\rangle\}|$ counts the intervals in which the feature μ can properly occur. Thus we obtain $\pi_\mu := \mathbb{E}[\mu]/h_\mu$ as generalization of the `RNAplfold` "probabilities" for arbitrary secondary structure features μ that can be expressed as hard constraints. Again, π_μ is not a probability but a scaled expected value.

Based on this general discussion it is now clear how the "probability of being unpaired" can be evaluated. Using the simplified notation $p_{jk}^{\langle i\rangle} := \mathbb{P}_{\langle i\rangle}[(j, k)]$, the probability that position k is unpaired in interval $\langle i\rangle$ is given by

$$q_k^{\langle i\rangle} := 1 - \sum_{j<k} p_{jk}^{\langle i\rangle} - \sum_{j>k} p_{kj}^{\langle i\rangle}. \tag{12}$$

The expected number of occurrences of unpaired k is

$$\mathbb{E}[k] = \sum_{\langle i\rangle\in\mathfrak{I}} q_k^{\langle i\rangle} = h_k - \sum_{j<k} \mathbb{E}[j, k] - \sum_{j>k} \mathbb{E}[k, j], \tag{13}$$

where h_k is the number of intervals in $\mathfrak{I}$ that contain the base k. The key observation is that h_k is **not** the same as the number of intervals that feature the contributing base pairs. Using the `RNAplfold` "probabilities", i.e., the normalized expectations, therefore requires an additional combinatorial correction:

$$q_k := \frac{\mathbb{E}[k]}{h_k} = 1 - \sum_{j<k} \frac{h_{jk}}{h_k}\pi_{jk} - \sum_{j>k} \frac{h_{kj}}{h_k}\pi_{kj} \tag{14}$$

The pre-factors h_{jk}/h_k in this expression correct the "strange" behavior that arises from using the `RNAplfold` "probabilities" naïvely. The numbers of intervals that

contain position k are easily obtained as the special case $k = r = s$ from Eq. (10):

$$h_k = \begin{cases} W & \text{if } k \geq W \text{ and } k \leq n - W + 1 \\ r & \text{if } k < W \\ n - s + 1 & \text{if } k > n - W + 1 \end{cases} \tag{15}$$

Disregarding the special cases at either end of the input sequence, the generic correction factor is $h_{kj}/h_k = (W - |k - j|)/W = 1 - |k - j|/W$, reflecting the fact that base pairs with longer ranges need to be down-weighted because they appear in fewer intervals.

Within the `ViennaRNA` package, the values of q_k can also be computed as *accessibilities* for intervals of length 1 using `RNAplfold -u 1`. The algorithm[4] for computing accessiblities, i.e., the probabilities that intervals of given lengths are unpaired, requires moderate overhead compared to the computation of the partition function and scaled expected values π_{kl}. Thus it may be preferrable to use Eq. (14) to compute the q_k. We verified empirically that the two methods yield identical results up to round-off errors expected for floating point calculations.

The computation of q_k, Eq. (14), is also useful to obtain better Maximum Expected Accuracy (MEA) structures from `RNAplfold` data. As noted in earlier work,[9,26] one naturally uses a cost function of the form $2\pi_{kl}$ for each base pair (k, l) and q_k for an unpaired base k.

2.4. *Sampling PL configurations*

PL configurations can be sampled with the correct distribution by independently generating a random structure $x_{\langle i \rangle}$ for each interval $\langle i \rangle \in \mathfrak{I}$. This is rather inefficient, however, since the independent computation of the partition functions requires $O(|\langle i \rangle|^3) = O(L^3)$ operations for each interval.

The key observation for our purpose is that the partition functions on the interval $\langle i \rangle$ **and** all its subintervals only depend on the interval $[i, k]$ under consideration but are independent of the choice of the interval system $\mathfrak{I}$. To see this, consider the forward recursion of McCaskill's algorithm[27] for (global) folding of a given RNA sequence:

$$Z_{i,k} = Z_{i+1,k} + \sum_{j=i+4}^{k} Z^B_{i,j} Z_{j+1,k} \tag{16}$$

where $Z_{i,k}$ is the partition function over all substructures on the interval $[i, k]$, $Z^B_{i,j}$ is the partition function over all substructures enclosed by the base pair (i, j) on the interval $[i, k]$; For simplicity, the convention $Z_{i+1,i} = 1$ is used. The recursion is initialized by $Z_{i,i} = 1$ for all i, representing a single unpaired base. By construction, $Z_{i,k}$ and Z^B_{ij} do not depend on the details of an enclosing interval $[r, s]$ as long as $[i, j] \subseteq [i, k] \subseteq [r, s]$. In particular, therefore backtracing from $Z_{r,s}$ covers exactly all local structures on the interval $[r, s]$.

Stochastic backtracing in RNA folding[8,32] operates on these dynamic programming matrices. Operating on the interval $[i, k]$, one determines the probability that i is unpaired and that (i, j) is a base pairs as

$$p_i^\circ = Z_{i+1,k}/Z_{i,k} \quad \text{and} \quad p_{(i,j)} = Z_{i,j}^B Z_{j+1,k}/Z_{i,k}. \tag{17}$$

One of the alternatives is selected with the corresponding probability and backtracing then continues, in the first case, on $[i + 1, k]$ and in the second case both with a structure on $[i, j]$ enclosed by the pair (i, j) and the "tail" on $[j + 1, k]$. Since p_i° and $p_{(i,j)}$ are independent of the enclosing interval $[r, s]$, the same is true for the backtracing results. It therefore suffices to compute McCaskill's forward recursion for $s - r + 1 \leq L$, i.e., up to the maximum base pair span of the intervals under consideration. The part $x_{\langle i \rangle}$ of a configuration x is correctly sampled by backtracing on the interval $\langle i \rangle$.

Local stochastic backtracing is available in the `ViennaRNA` package from version 2.5.0 on, which provides the function `vrna_pbacktrack_sub()` returning a sample of Boltzmann-weighted structures for a user-defined inverval. Starting from version 2.5.1, the probabilities of unpaired positions can also be accessed directly for use in MEA computations. The current implementation is not optimized for memory consumption and uses the complete, triangular matrix to store the required partition functions. A memory efficient version for limited basepair span is planned for a future release of the `ViennaRNA` package. The accuracy of the sampling depends only on the number of intervals that contribute to the estimator for the quantity of interest. For the probability q_k that position k is unpaired, for example, this is given by Ws, the product of the window size W and the number s of structures drawn from each window, see Fig. 2.

It is important to note that as a consequence of the considerations above, PL configurations are **not** correctly sampled by stochastic backtracing on the arrays

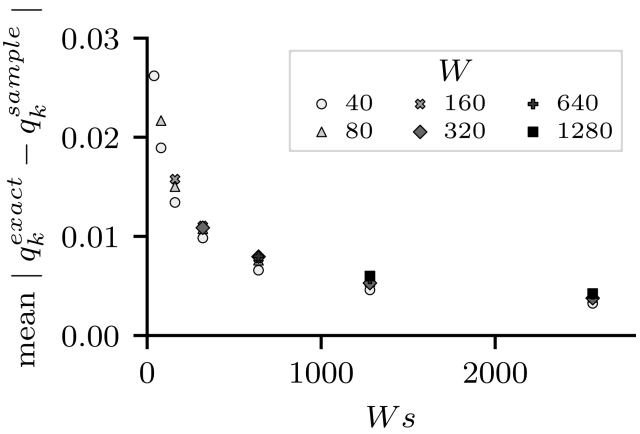


Fig. 2. Accuracy of q_k values obtained from sampling as a function of Ws, the number of intervals contribute to the estimate of q_k for window size W and s local structures per window. Different symbols refer to different values of W.

stored in `RNAplfold` because the π_{kl} are neither probabilities nor scaled partition functions of configurations.

It is a tempting idea, furthermore, to utilize the recursions of `RNAplfold` to construct a span-restricted variant of the *assisted folding* algorithms.[25,35] Here, an optimization procedure is used to minimize the discrepancy between predicted and experimentally derived probabilities of individual positions being unpaired. This is conveniently quantified in terms of "excess energies" ε_k that favor position k to be unpaired. Gradient optimization in this context requires terms of the form

$$\left.\frac{\partial q_k}{\partial \epsilon_j}\right|_{\epsilon=0} = \frac{1}{h_k}\sum_{\langle i\rangle\in\mathfrak{I}} \left.\frac{\partial q_k^{\langle i\rangle}}{\partial \epsilon_j}\right|_{\epsilon=0} = \frac{1}{h_k}\sum_{\langle i\rangle\in\mathfrak{I}} \frac{1}{RT} q_k^{\langle i\rangle}(0)[q_j^{\langle i\rangle}(0) - q_j[k]^{\langle i\rangle}(0)] \tag{18}$$

Here, (0) refer to the current vector of excess energies and $q_j[k]^{\langle i\rangle}$ refer to the probability that j is unpaired given that k is unpaired in interval $\langle i\rangle$. We note that $q_j[k]^{\langle i\rangle} = q_j$ if $j \in \langle i\rangle$ and $k \notin \langle i\rangle$ since the constraint on k outside of the interval $\langle i\rangle$ has no effect on structures within $\langle i\rangle$. Furthermore, by definition, we have $q_j[k]^{\langle i\rangle} = 0$ whenever $j \notin \langle i\rangle$. The summations over the quadratic terms, i.e., $\sum_{\langle i\rangle\in\mathfrak{I}} q_k^{\langle i\rangle}(0) q_j^{\langle i\rangle}(0)$, etc., unfortunately cannot be expressed in terms of q_k and q_j. The expected values computed by `RNAplfold` therefore cannot be used to obtain a corresponding span-restricted version of assisted folding. In constrast, both $q_k^{\langle i\rangle}$ and the constrained quantities $q_j[k]^{\langle i\rangle}$ are readily available from samples of local structures obtained by stochastic backtracing on the partition functions of the global folding model initialized at the ends of the interval $\langle i\rangle$.

3. Computational Applications

The differences in the definition of the PL model and global RNA folding have a noticable effect on the distribution of quantities such as q_k. The average local single-position accessibility q_k converges to the probability p_k° that base k is unpaired as the window size W approaches the sequence length, i.e., in the limit that only a single (global) window is used. The PLfold accessibilities q_k are a good approximation of p_k° only for window sizes W close to n as shown by the example in Fig. 3. We observe that q_k and p_k° values are well correlated but appreciable scatter remains. This behaviour is independent of the particular input sequence.

Comparing the distribution of q_k and p_k° we find that a systematic discrepancy persists to very large window sizes. While the distribution of the p_k° values for global folding are extremely concentrated at 0 and 1, i.e., most nucleotides are predicted to be either nearly always paired or always unpaired, we obtain a much more even distribution for the q_k values. The latter distributions become more focussed on the extremes as the window size increased. Qualitatively this behavior is independent of the base pair span. On average, the PL model yields a moderate overestimate of accessibility. Depending on the intended application, it may be useful, therefore, to

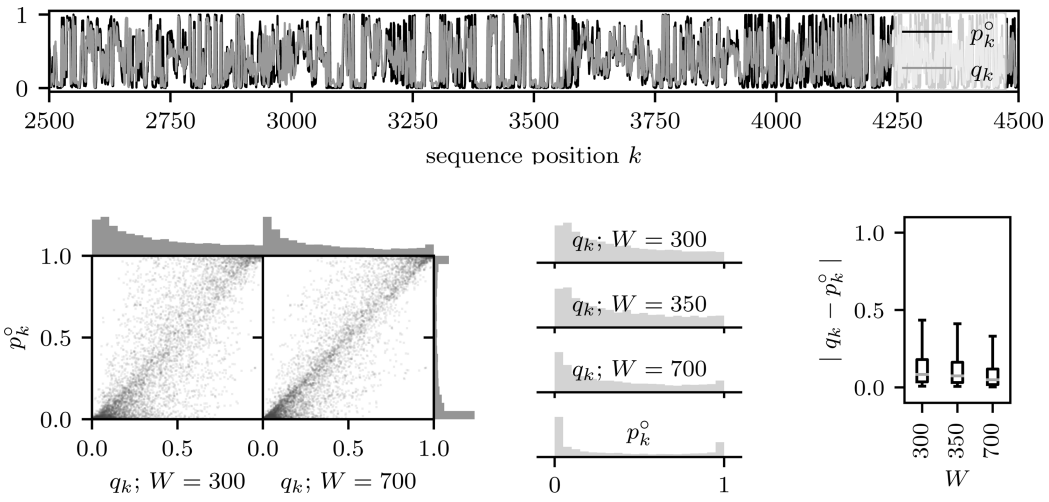


Fig. 3. Comparison of local and global accessibilities. *Top*: Comparison of q_k (estimated with $s = 100$, $W = 350$, $L = 300$, black line), and p_k° (global folding using `RNAfold`, gray line). The panel shows a region of the mir17HG (NG_032702.1 [5000-11756]). The same restriction on the base pair span ($L = 300$) was used for both `RNAplfold` and `RNAfold`. The curves appear to be well correlated. *Bottom left*: Scatter plot of q_k versus p_k° for two different window sizes. Distributions of q_k values are indicated above, the distribution of p_k° values (which is independent of W) is shown on the right. *Bottom center*: Comparison of the distribution of q_k values for three different window size and p_k° (last panel). For larger W, the distribution of q_k becomes more similar to p_k°. *Bottom right*: Distribution of the absolute differences $|q_k - p_k^\circ|$ for different window sizes.

compensate for this effect. The broader distribution of the accessibilities in the PL model is caused by the different window positions.

A shortcoming of `RNAplfold` is that it computes the expected fraction π_{kl} of structure that harbour a given base pair only, but does not predict a concrete secondary structure. This weakness can be overcome by using the values of π_{kl} and the corresponding expectations q_k for unpaired position to compute a Maximum Expected Accuracy structure.[26] We therefore implemented a MEA computation that accepts lists of both pair probabilities π_{kl} and q_k as input. The feature is available in `ViennaRNA` (version 2.5.1 and higher). As an illustrative example, Fig. 4 shows the MEA structure of a large portion of the human miRNA host gene MIR17HG, which harbours the miR-17-92 cluster. This gene has received ample attention in the literature.[28,33,39] The six miRNA precursor hairpins (miR-17, miR-18a, miR-19a, miR-20a, miR-19b-1, and mir-92a-1) are clearly visible in the predicted MEA structure. The computation of MEA structures enhances the practical usefulness of local folding approaches.

The MEA approach can easily be modified to compensate for the bias of PL model. The simplest approach is the use $(1 + \alpha/2)\pi_{kl}$ and $(1 - \alpha/2)q_k$ for the paired and unpaired contributions, respectively. A natural choice for the bias α is the median of the discrepancy shown in the lower-right panel of Fig. 3. Similar weighting factors have been proposed already in earlier work on MEA secondary structures.[26]

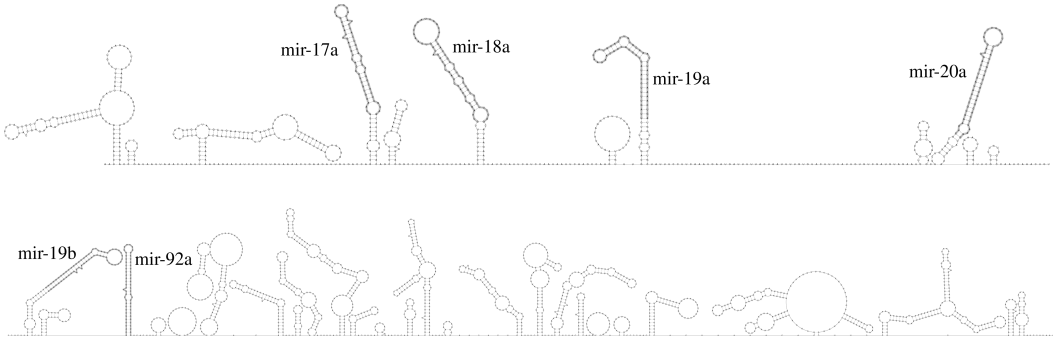


Fig. 4. Maximum expected accuracy (MEA) structure of the mir-17-92 cluster computed from local structure predictions for a region in the miRNA host gene MIR17HG. The six miRNAs are highlighted in bold and annotated. The drawing was generated using the `RNApuzzler` algorithm[36] as implemented in `ViennaRNA` and covers the interval [7500-9500] in the NCBI reference sequence NG_032702.1.

A third practical use-case for local sampling is the computation of the effect of constraining the structure at position j on the q_k. Beyond the potential applications to assisted folding for long sequences, it is also of interest to consider such intramolecular dependencies. As a simple measure of the influence of a position k on its surrounding we compute the cummulative mutual information $M_k := \sum_j I(k;j)$, where $I(k;j)$ is the mutual information of the pairing states of positions k and j as computed from samples of local structures obtained by stochastic backtracing. In principle $I(k;j)$ can be computed from partition functions constrained to the four combinations of pairing states at k and j. However, this requires quadratically many separate computations, hence sampling is typically much more efficient.

We note, furthermore, that $I(k;j)$ can be estimated in different, inequivalent ways. In particular, it is possible to obtain estimates for individual windows $\langle i \rangle \in \mathfrak{I}$, which may then be averaged over different windows. Alternatively, it is possible to estimate the joint probabilities $p(k;j)$ that two positions are simultaneously paired (not necessarily with each other) directly from samples generated from the PLfold model. These can then be converted into corresponding information measures. For details we refer to the Tutorial provided as Supplemental Material to this contribution.

Inspection of the mutual information can help to understand details of local structural stability and structural alternatives. Possible applications parallel the use of information theoretic measures to characterize the Boltzmann ensemble of global secondary structures.[22,30] Figure 5 again uses the miRNA host gene MIR17HG as an example. The microRNA precursor hairpins appear as regions with significantly depleted values of M_k. More precisely, the precursors show very low values of $I(k,j)$ for paired positions, with some increase in the hairpin loop. MicroRNA precursors thus appear as closely spaced double-minima, each extending about 20nt, i.e., length of the helix in precursor. The r.h.s. panel in Fig. 5 shows the full mutual information for a single precursor hairpin, highlighting the two regions with extremely low values

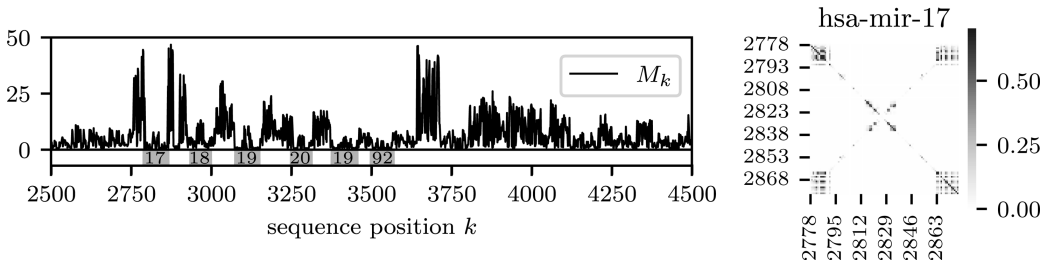


Fig. 5. Mutual information of the miRNA host gene MIR17HG. L.h.s.: cummulative mutual information M_k in the region close to the six microRNA precursor hairpins indicated by gray bars. The values of M_k computed with a basepair span of $L = 150$ show noticeable minima consistent with the unambiguous structures of precursor hairpins. R.h.s.: Mutual information $I(k; j)$ at the positions of the pre-mir-17. While the flanking sequence shows a diffuse pattern, $I(k; j)$ is strongly reduced except for a moderate signal along the backbone (NW-SE) and along the basepairs of the hairpin (SW-NE).

corresponding to basepaired positions, and the slight increase in the hairpin, which appears as a small $\times$ in the center of the plot. In contrast, the flanking regions (corners) show a diffuse pattern of mutual information.

Local folding computations can be combined freely with hard constraints in the `ViennaRNA` package. A detailed example has been added to the Tutorial.

4. Discussion

We introduced the PL model as rigorous formalization of local folding in overlapping windows. The model provides a mathematical grounding for `RNAplfold`[3] and `LocalFold`[20] and makes it easy to incorporate hard constraints.[24] We show that proper Boltzmann sampling of local structures cannot be formulated as a stochastic backtracing from the averages computed in `RNAplfold`. Instead, such samples are obtained by backtracing on the matrices holding partial partition function for global RNA folding, i.e., McCaskill's algorithm[27] starting from the boundaries of intervals $\langle i \rangle \in \mathfrak{I}$. In order to make it easier for users to work with this type of local structural ensembles, `ViennaRNA` (versions 2.5.1 an higher) now provides access to this type of local stochastic backtracing. The sampling approach makes it possible to estimate quantities that cannot be readily computed from restricted partition functions, and for which the computation of restricted partition functions would require complex, computationally expensive algorithms. Mutual information measures, Fig. 5 serve as just one example. The most recent versions of `ViennaRNA` furthermore provide easier access to MEA computations using `RNAplfold` π_{kl} and q_k values.

A general advantage of local folding, i.e., limited base pair spans, is that the running time is asympotically linear in sequence length. This is also true for the subsequent MEA computation. Catalogs of local structure, Fig. 4, thus can now be obtained efficiently. In contrast to the strict locality of `RNAplfold` and other local folding algorithms, `LinearFold`[15,23] achieves linear running time by pruning the search space using beam search. This results in non-local structures with unlimited

base pair span that approximate optimal global folding of given transcript. Local folding algorithms, in contrast, exclude long-range base pairs and thus can provide useful secondary structure information also in cases where transcript boundaries are unknown. In particular, it is meaningfully possible in the local setting to scan only parts of transcript to operate on genomic regions directly.

Local folding approaches create intrinsic biases, however, that — depending on the application — may require special consideration. Most importantly, RNA secondary have a strong tendency towards long-range base pairs. In some cases, these are a functional requirement, as in case of the panhandle structures common in RNA virus genomes.[14,21] This is, however, also a combinatorial phenomenon affecting also random sequences.[5,12,37] This leads to a bias for base pairs with span close to the window size. In practical computations, therefore, it is strongly advisable to use the constraint framework to limit the base pair span L to a value smaller than the window size W. On the other hand, for $L \ll W$, there is a bias towards very large accessibilities close to the borders of window.[20] This effect is explained by structures in the center of the window leaving no reachable pairing partners for the "left-overs" at the borders. Upon average formation, this leads to a systematic overestimation of accessiblities. For such applications, `LocalFold`[20] suggests to omit positions with b nucleotide from either boundary from the computation of averages. In general, this approach has to deal with base pairs (k, l) where one nucleotide is inside and one nucleotide is outside the "relevant interval" $[p + b, q - b]$. Using samples from stochastic backtracing, the effect of alternative choices is easily evaluated. If the local folding results are used to computate a MEA structure, it is also possible to compensate for the effect by slightly downweighting unpaired bases in comparison to base pairs.[26] In practise, the biases are most pronounced if L is close to W.

The most prominent application of `plfold` is the prediction of local RNA structures in large sequences. To this end, we make it easy to process π_{kl} and q_k predictions into a single MEA structure. This approach may be combined with incorporating experimental probing data. The current implementations already support the "bonus energy" approach[6,7,18,38] via the soft-constraint framework provided by `ViennaRNA`. In a different type of applications, stochastic backtracing followed by the computation of mutual information measures for structure predictions with different hard constraints can be used to determining "keystone" nucleotides whose pairing status has a large influence on the overall structure.

A number of questions arose from the present work and remain open for future research and software development. For instance, differences between "hard" limits on base pair spans and a "soft" decrease of enery contributions for long-range base pairs deserves more detailed attention. The current implementation of local stochastic backtracing is not yet optimized for memory usage. For fixed window size, the quadratic matrices in the forward recursion can be limited to $|j - i| + 1 \leq W$. If, furthermore, the sample size for each interval interval is pre-determined, stochastic backtracing can be interleaved with the forward recursions in analogy to implementation of `RNAplfold`: After computing for the current j, all entries Z_{ij}, all

samples for intervals $[i', j]$ with endpoint j are backtraced. Once this step is completed, all rows $i'' < j - W$, where W is the maximum interval size, will never be used again, and thus can be removed from memory. This reduced memory requirements to $O(W^2)$. Since samples from non-overlapping intervals are independent by construction, it suffices to keep sampled sequences within a 5' end $i \geq j - 2W$ for evaluation. This will limit the memory requirement for sampling to $O(W(W + ys))$, where s is the sample size per interval, and y denotes the number of intervals per endpoint, e.g. $y = 1$ for `RNAplfold`.

Stochastic backtracing could also be used to extend the idea of assisted folding, i.e., the reconciliation of inaccuracies in both the energy model and the experimental data by minimizing a total error functional,[11,25,35] to local instead of global structure prediction.

Availability

`ViennaRNA` versions 2.5.0 and 2.5.1 are availble at https://github.com/ViennaRNA/ and the official `ViennaRNA` website https://www.tbi.univie.ac.at/RNA/.

A collection of python scripts detailing the applications described in this contribution serves as tutorial describing e.g. `RNAplfold`-based MEA secondary structure predictions, sampling with local stochastic backtracing, and the extraction of feature frequencies from such samples. The Tutorial in particular contains the code necessary to reproduce (up to formatting details) the figures shown in this contribution. It is available at https://github.com/ViennaRNA/PLfoldSampling.

Acknowledgements

This work was supported in part by the German Federal Ministery of Education in extension of the German Network for Bioinformatics Infrastructure de.NBI/RBC with transition funding in 2022 provided through the Forschungs Zentrum Jülich, and by the Austrian FWF (I-2874-N28, W 1207 DK RNA Biology, I-4520 and F 80 RNAdeco).

References

1. Alvarez DE, Lodeiro MF, Ludueña SJ, Pietrasanta LI, Gamarnik AV, Long-range RNA-RNA interactions circularize the Dengue virus genome, *J Virol* **79**:6631–6643, 2005. doi: 10.1128/JVI.79.11.6631-6643.2005.
2. Amman F, Bernhart SH, Doose G, Hofacker IL, Qin J, Stadler PF, The Students of the Bioinformatics II Lab Class 2013, Will S, The trouble with long-range base pairs in RNA folding, Setubal JC, Almeida NF (eds.), *Advances in Bioinformatics and Computational Biology, 8th BSB*, Lect. Notes Comp. Sci., Vol. 8213, pp. 1–11, 2013. doi:10.1007/978-3-319-02624-4_1.
3. Bernhart S, Hofacker IL, Stadler PF, Local RNA base pairing probabilities in large sequences, *Bioinformatics* **22**:614–615, 2006. doi: 10.1093/bioinformatics/btk014.

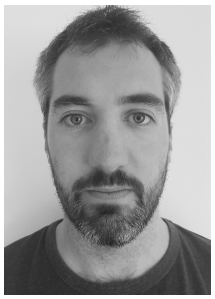
4. Bernhart SH, Mückstein U, Hofacker IL, RNA accessibility in cubic time, *Alg Mol Biol* **6**:3, 2011. doi: 10.1186/1748-7188-6-3.
5. Clote P, Ponty Y, Steyaert JM, Expected distance between terminal nucleotides of RNA secondary structures, *J Math Biol* **65**:581–599, 2012. doi: 10.1007/s00285-011-0467-8.
6. Cordero P, Kladwang W, VanLang CC, Das R, Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference, *Biochemistry* **51**:7037–7039, 2012. doi: 10.1021/bi3008802.
7. Deigan KE, Li TW, Mathews DH, Weeks KM, Accurate SHAPE-directed RNA structure determination, *Proc Natl Acad Sci USA* **106**:97–102, 2009. doi: 10.1073/pnas.08069291.
8. Ding Y, Lawrence CE, Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond, *Nucleic Acids Res* **29**:1034–1046, 2001. doi: 10.1093/nar/29.5.1034.
9. Do C, Woods D, Batzoglou S, CONTRAfold: RNA secondary structure prediction without physics-based models, *Bioinformatics* **22**:e90–98, 2006. doi: 10.1093/bioinformatics/btl246.
10. Doshi K, Cannone J, Cobaugh C, Gutell R, Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction, *BMC Bioinformatics* **5**:105, 2004. doi: 10.1186/1471-2105-5-105.
11. Eddy SR, Computational analysis of conserved RNA secondary structure in transcriptomes and genomes, *Annu Rev Biophys* **43**:433–456, 2014. doi: 10.1146/annurev-biophys-051013-022950.
12. Fang LT, The end-to-end distance of RNA as a randomly self-paired polymer, *J Theor Biol* **280**:101–107, 2011. doi: 10.1016/j.jtbi.2011.04.010.
13. Hofacker IL, Priwitzer B, Stadler PF, Prediction of locally stable RNA secondary structures for genome-wide surveys, *Bioinformatics* **20**:191–198, 2004. doi: 10.1093/bioinformatics/btg388.
14. Hsu MT, Parvin JD, Gupta S, Krystal M, Palese P, Genomic RNAs of influenza viruses are held in a circular conformation in virions and in infected cells by a terminal panhandle, *Proc Natl Acad Sci USA* **84**:8140–8144, 1987. doi: 10.1073/pnas.84.22.8140.
15. Huang L, Zhang H, Deng D, Zhao K, Liu K, Hendrix DA, Mathews DH, LinearFold: Linear-time approximate RNA folding by 5'-to-3' dynamic programming and beam search, *Bioinformatics* **35**(14):i295–i304, 2019. doi: 10.1093/bioinformatics/btz375.
16. Kalmykova S, Kalinina M, Denisov S, Mironov A, Skvortsov D, Guigó R, Pervouchine D, Conserved long-range base pairings are associated with pre-mRNA processing of human genes, *Nature Comm* **12**:2300, 2021. doi: 10.1038/s41467-021-22549-7.
17. Kalvari I, Nawrocki EP, Ontiveros-Palacios N, Argasinska J, Lamkiewicz K, Marz M, Griffiths-Jones S, Toffano-Nioche C, Gautheret D, Weinberg Z, Rivas E, Eddy SR, Finn RD, Bateman A, Petrov AI, Rfam 14: expanded coverage of metagenomic, viral and microRNA families, *Nucleic Acids Res* **49**:D192–D200, 2020. doi: 10.1093/nar/gkaa1047.
18. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, Segal E, Genome-wide measurement of RNA secondary structure in yeast, *Nature* **467**:103–107, 2010. doi: 10.1038/nature09322.
19. Kiryu H, Kin T, Asai K, Rfold: An exact algorithm for computing local base pairing probabilities, *Bioinformatics* **24**:367–373, 2008. doi: 10.1093/bioinformatics/btm591.
20. Lange SJ, Daniel M, Möhl M, Joshua JN, Brown CM, Backofen R, Global or local? Predicting secondary structure and accessibility in mRNAs, *Nucleic Acids Res* **40**:5215–5226, 2012. doi: 10.1093/nar/gks181.
21. Leija-Martínez N, Casas-Flores S, Cadena-Nava RD, Roca JA, Mendez-Cabañas JA, Gomez E, Ruiz-Garcia J, The separation between the 5'-3' ends in long RNA molecules is

short and nearly constant, *Nucleic Acids Res* **42**:13963–13968, 2014. doi: 10.1093/nar/gku1249.

22. Lin L, McKerrow WH, Richards B, Phonsom C, Lawrence CE, Characterization and visualization of RNA secondary structure Boltzmann ensemble via information theory, *BMC Bioinformatics* **19**:82, 2018. doi: 10.1186/s12859-018-2078-5.
23. Zhang, He, Zhang, Liang, MathewsMathews, David H and Huang, Liang, LinearPartition: linear-time approximation of RNA folding partition function L, base-pairing probabilities, Zhang, he and zhang, liang and mathewsmathews, david h and huang, liang, *Bioinformatics* **36**(S1):i258–i267, 2020. doi:10.1093/bioinformatics/btaa460.
24. Lorenz R, Hofacker IL, Stadler PF, RNA folding with hard and soft constraints, *Alg Mol Biol* **11**:8, 2016. doi: 10.1186/s13015-016-0070-z.
25. Lorenz R, Luntzer D, Hofacker IL, Stadler PF, Wolfinger MT, SHAPE directed RNA folding, *Bioinformatics* **32**:145–147, 2016. doi: 10.1093/bioinformatics/btv523.
26. Lu Y, Sze SH, Improving accuracy of multiple sequence alignment algorithms based on alignment of neighboring residues, *Nucl Acids Res* **37**:463–472, 2009. doi: 10.1093/nar/gkn945.
27. McCaskill JS, The equilibrium partition function and base pair binding probabilities for RNA secondary structure, *Biopolymers* **29**:1105–1119, 1990. doi: 10.1002/bip.360290621.
28. Mogilyansky E, Rigoutsos I, The miR-17/92 cluster: A comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease, *Cell Death Differentiation* **20**:1603–1614, 2013. doi: 10.1038/cdd.2013.125.
29. Proctor JRP, Meyer IM, CoFold: An RNA secondary structure prediction method that takes co-transcriptional folding into account, *Nucleic Acids Res* **41**:e102, 2013. doi: 10.1093/nar/gkt174.
30. Schroeder SJ, Challenges and approaches to predicting RNA with multiple functional structures, *RNA* **24**:1615–1624, 2018. doi: 10.1261/rna.067827.118.
31. Sutandy FXR, Ebersberger S, Huang L, Busch A, Bach M, Kang HS, Fallmann J, Maticzka D, Backofen R, Stadler PF, Zarnack K, Sattler M, Legewie S, König J, In vitro iCLIP-based modeling uncovers how splicing factor U2AF2 relies on regulation by cofactors, *Genome Res* **28**:699–713, 2018. doi: 10.1101/gr.229757.117.
32. Tacker M, Stadler PF, Bornberg-Bauer EG, Hofacker IL, Schuster P, Algorithm independent properties of RNA structure prediction, *Eur Biophy J* **25**:115–130, 1996. doi: 10.1007/s002490050023.
33. Tanzer A, Stadler PF, Molecular evolution of a microRNA cluster, *J Mol Biol* **339**:327–335, 2004. doi: 10.1016/j.jmb.2004.03.065.
34. Turner DH, Mathews DH, NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure, *Nucl Acids Res* **38**:D280–D282, 2010. doi: 10.1093/nar/gkp892.
35. Washietl S, Hofacker IL, Stadler PF, Kellis M, RNA folding with soft constraints: Reconciliation of probing data and thermodynamic secondary structure prediction, *Nucleic Acids Res* **40**:4261–4272, 2012. doi: 10.1093/nar/gks009.
36. Wiegreffe D, Alexander D, Stadler PF, Zeckzer D, RNApuzzler: Efficient outerplanar drawing of RNA-secondary structures, *Bioinformatics* **35**:1342–1349, 2019. doi: 10.1093/bioinformatics/bty817.
37. Yoffe AM, Prinsen P, Gelbart WM, Ben-Shaul A, The ends of a large RNA molecule are necessarily close, *Nucl Acids Res* **39**:292–299, 2011. doi: 10.1093/nar/gkq642.
38. Zarringhalam K, Meyer MM, Dotu I, Chuang JH, Clote P, Integrating chemical footprinting data into RNA secondary structure prediction, *PLOS ONE* **7**(10):e45160, 2012. doi: 10.1371/journal.pone.0045160.

39. Zhao W, Gupta A, Krawczyk J, Gupta S, The miR-17-92 cluster: Yin and Yang in human cancers, *Cancer Treat Res Commun* **33**:100647, 2022. doi: 10.1016/j.ctarc.2022.100647.
40. Zhao Y, Wang J, Zeng C, Xiao Y, Evaluation of RNA secondary structure prediction for both base-pairing and topology, *Biophysics Reports* **4**:123–132, 2018. doi: 10.1007/s41048-018-0058-y.

Maria Waldl holds a M.Sc. degree in Chemistry and is currently finishing her Ph.D. thesis in RNA Bioinformatics at the University of Vienna in the group of Ivo L. Hofacker. Her research interest is centered around RNA structure prediction with a focus on RNA folding kinetics, RNA-RNA interactions, evolution of non-coding RNAs and integration with experimental data.



Thomas Spicher holds a Master degree in Bioinformatics. He currently is a PhD student in Ivo L. Hofacker's Theoretical Biochemistry Group at the University of Vienna. His main research activities focus on the analysis and prediction of RNA structures, concentrating on the contribution of modified nucleotides to those structures.

Ronny Lorenz is a bioinformatician who received his PhD in Molecular Biology at the University of Vienna in 2014. He is a senior scientist at the Department of Theoretical Chemistry of the University of Vienna, Austria. His main research interests are various algorithmic aspects in the field of RNA secondary structure prediction and the development of novel methods to simulate RNA folding kinetics. Since 2010, he is the leading developer of the ViennaRNA Package.

Irene K. Beckmann holds a M.Sc. degree in Molecular Biology (2018). She currently is a PhD candidate at the Institute for Theoretical Chemistry at the University of Vienna, with specialization on Bioinformatics. Her primary area of research focus on developing computational tools to predict RNA interactions and their dynamics in 3D.

Ivo L. Hofacker did his Ph.D. in Physics and is a Full Professor at the University of Vienna in the Faculty of Chemistry, Department for Theoretical Chemistry, as well as Faculty of Computer Science, where he heads the Research Group Bioinformatics and Computational Biology. Since 2022 he is a corresponding member of the Austrian Academy of Sciences. His research interests focus on the prediction of RNA structures, noncoding RNA annotation, and the analysis of folding landscapes.

Sarah von Löhneysen holds M.Sc. degrees in Biochemistry and Bioinformatics. She currently is a Ph.D. student in Peter Stadler's group at the University of Leipzig. The central focus of her research lies on the integration of experimental data into RNA structure prediction.

Peter F. Stadler received a PhD in Chemistry from U. Vienna in 1990 and then worked as Assistant and Associate Professor for Theoretical Chemistry at the same School. In 2002 he moved to Leipzig as Full Professor for Bioinformatics. He is External Professor at the Santa Fe Institute since 2004, Corresponding Member Abroad of Austrian Academy of Sciences, External Scientific Member of the Max Planck Society, and a Honorary Professor of the Universidad Nacional de Colombia.

5.3 Investigating RNA–RNA interactions through computational and biophysical analysis

Tyler Mrozowich, Sean M. Park, Maria Waldl, Amy Henrickson, Scott Tersteeg, Corey R. Nelson, Anneke De Klerk, Borries Demeler, Ivo L. Hofacker, Michael T. Wolfinger and Trushar R. Patel (2023) 'Investigating RNA–RNA interactions through computational and biophysical analysis', Nucleic Acids Research, 51(9), pp. 4588–4601. doi: 10.1093/nar/gkad223.

Summary: This study presents an novel approach that integrates biophysical techniques and computational analyses to identify and characterize RNA–RNA interactions.

As a model system, the study focuses on a critical long-range RNA-RNA interaction between the 3' and 5' terminal regions (TRs) of flavivirus genomes. This interaction is essential for recruiting the viral RNA polymerase, yet it had not been previously identified in the Japanese encephalitis virus (JEV), a medically significant member of the Flaviviridae family.

To identify putative interactions between the 3' and 5' TRs, the study applied thermodynamic, kinetic, and covariance-based computational methods. These predictions guided the selection of representative sequences and the design of mutated constructs for experimental validation. Biophysical experiments, including mass and binding affinity measurements, aligned with computational expectations, strengthening also the mechanistic hypotheses suggested by the computational model.

This work demonstrates how kinetic predictions can successfully guide the discovery of functionally relevant RNA-RNA interactions. By integrating computational and biophysical approaches, it provides empirical validation for kinetic modeling and establishes a framework for future investigations of RNA interactions.

Author's contributions: MW extended the research question to include kinetic aspects, planned the evaluation of interaction kinetics in related viruses via bioinformatics methods; contributed to the collection of virus sequence data and analysed them for kinetic and covariance features; presented bioinformatics results and provided interpretations; wrote the corresponding sections in the paper; provided relevant literature; contributed to revisions and provided critical review for the whole manuscript.

4588–4601 *Nucleic Acids Research, 2023, Vol. 51, No. 9* *Published online 31 March 2023*
https://doi.org/10.1093/nar/gkad223

Investigating RNA–RNA interactions through computational and biophysical analysis

Tyler Mrozowich [1,†], Sean M. Park [1,†], Maria Waldl [2,3,4], Amy Henrickson [1], Scott Tersteeg [1], Corey R. Nelson [1], Anneke De Klerk [1], Borries Demeler [1,5], Ivo L. Hofacker [2,6], Michael T. Wolfinger [2,6,7] and Trushar R. Patel [1,8,9,*]

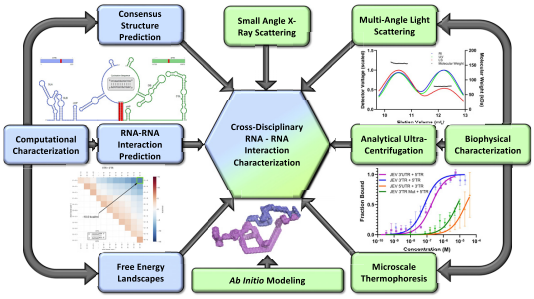
[1]Department of Chemistry and Biochemistry, Alberta RNA Research and Training Institute, University of Lethbridge, 4401 University Drive, Lethbridge, ABT1K 3M4, Canada, [2]Department of Theoretical Chemistry, University of Vienna, Währinger Strasse 17, 1090, Vienna, Austria, [3]Center of Anatomy & Cell Biology, Division of Cell & Developmental Biology, Medical, University of Vienna, Schwarzspanierstrasse 17, 1090, Vienna, Austria, [4]Vienna Doctoral School in Chemistry (DoSChem), University of Vienna, Währinger Strasse 42, 1090 Vienna, Austria, [5]Department of Chemistry and Biochemistry, University of Montana, Missoula, MT59812, USA, [6]Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Währinger Strasse 29, 1090, Vienna Austria, [7]RNA Forecast e.U., 1100 Vienna, Austria, [8]Li Ka Shing Institute of Virology, University of Alberta, Edmonton T6G 2E1, Alberta, Canada and [9]Department of Microbiology, Immunology & Infectious Diseases, Cumming School of Medicine, University of Calgary, Calgary T2N 4N1, Canada

Received December 31, 2022; Revised February 24, 2023; Editorial Decision March 13, 2023; Accepted March 15, 2023

ABSTRACT

Numerous viruses utilize essential long-range RNA–RNA genome interactions, specifically flaviviruses. Using Japanese encephalitis virus (JEV) as a model system, we computationally predicted and then biophysically validated and characterized its long-range RNA–RNA genomic interaction. Using multiple RNA computation assessment programs, we determine the primary RNA–RNA interacting site among JEV isolates and numerous related viruses. Following *in vitro* transcription of RNA, we provide, for the first time, characterization of an RNA–RNA interaction using size-exclusion chromatography coupled with multi-angle light scattering and analytical ultracentrifugation. Next, we demonstrate that the 5′ and 3′ terminal regions of JEV interact with nM affinity using microscale thermophoresis, and this affinity is significantly reduced when the conserved cyclization sequence is not present. Furthermore, we perform computational kinetic analyses validating the cyclization sequence as the primary driver of this RNA–RNA interaction. Finally, we examined the 3D structure of the interaction using small-angle X-ray scattering, revealing a flexible yet stable interaction. This pathway can be adapted and utilized to study various viral and human long-non-coding RNA–RNA interactions and determine their binding affinities, a critical pharmacological property of designing potential therapeutics.

GRAPHICAL ABSTRACT



INTRODUCTION

Computational and biophysical methods to obtain structure, properties and dynamics of protein–ligand interactions are well developed as a foundational component of many pharmaceutical discovery pipelines. As a result, nearly all currently approved drugs target one of ∼700 disease-related proteins, despite an increasing number of diseases attributed to the >98% of the human genome,

*To whom correspondence should be addressed. Tel: +1 403 332 4443; Fax: +1 403 329 2057; Email trushar.patel@uleth.ca
†The authors wish it to be known that, in their opinion, these authors should be regarded as Joint First Authors.

which is non-coding (1). Among the limitations in targeting these elements is a lack of foundational techniques for characterizing RNA–RNA interactions in the same depth as protein–ligand interactions. The importance of RNA characterization has only gained traction as of late, highlighting an ever-increasing need for techniques capable of doing so (2–5). In this study, we describe the novel application of well-established biophysical modalities with computational studies for extensive characterization of RNA–RNA interactions using a flaviviral system theorized to contain crucial intragenomic RNA–RNA interactions, Japanese encephalitis virus (JEV).

JEV is a mosquito-borne flavivirus in the genus Flavivirus (family *Flaviviridae*), which contains several pathogenic viruses such as Dengue virus (DENV), West Nile virus (WNV), Zika virus (ZIKV) and yellow fever virus (YFV). JEV is the leading cause of viral encephalitis in Southeast Asia and the Western Pacific, with approximately 68 000 cases globally each year (6). Currently, no approved treatments are available following infection by JEV or other flaviviral infections (7), which lends to the importance of further investigation. JEV is transmitted through competent mosquito vectors of the genera Aedes and Culex (8), suggesting that flaviviral infections will likely become more prevalent as global temperatures rise and vector populations expand (9). Like other flaviviruses, JEV is an enveloped virus with a single-stranded (+)-sense RNA genome of approximately 11 000 nt in length (10). A single open reading frame encodes for a polyprotein and is flanked by highly structured 5′ and 3′ untranslated regions (UTRs) (11). The genome has a type I cap at the 5′ end (m7GpppAMP) and lacks polyadenylation at the 3′-terminus. The single open reading frame is cleaved post-translationally into three structural and seven non-structural proteins (12,13). During replication, the 5′ and 3′ terminal regions (TRs) in flaviviruses undergo long-range intragenomic RNA–RNA interactions, thereby forming a so-called panhandle structure that mediates recruitment of the viral RNA-dependent RNA polymerase (NS5) (14). Removal of the TRs has shown inhibition of viral replication (14–17). A cyclization sequence of 11 nt is complementary in the 5′ and 3′ TRs, which facilitates this interaction (18). Furthermore, these terminal regions show binding with a variety of human host proteins (19), including but not limited to numerous DEAD-box helicases (20). In WNV and DENV, the 5′–3′ long-range interaction has been previously demonstrated (16–18,21–22). Genome cyclization in JEV has been computationally predicted previously (23); however, detailed experimental verification is still missing.

In this study, we explore the novel application of complementary computational and biophysical techniques that have not been used to characterize RNA–RNA interactions. Through a computational approach, we first identified an isolate of JEV that we hypothesized to interact with high affinity, then evaluated a consensus duplex structure of the 5′ and 3′ TRs of 20 different flaviviruses and measured their conservation. With this knowledge and utilizing various biophysical characterization techniques, we directly demonstrate for the first time that JEV 5′–3′ TRs interact *in vitro* with nanomolar affinity and with 1:1 stoichiometry. Furthermore, we isolated and identified the RNA–RNA complex using size-exclusion chromatography coupled with multi-angle light scattering (SEC-MALS). We additionally use analytical ultracentrifugation (AUC) as an orthogonal biophysical validation as evidence of JEV 5′–3′ TR interaction. We provide computational evidence which complements our experimental data showing that the cyclization sequence interaction is kinetically favorable and will outcompete potential homo-dimer RNA interactions. Finally, we present a low-resolution *ab initio* model showing the potential architectural arrangement of this RNA–RNA interaction in solution. This characterization can be used as a foundation in potential pharmaceutical therapies to inhibit viral replication through cyclization interruption and to help understand the replication pathway/mechanism by which NS5 replicates the viral genome.

MATERIALS AND METHODS

Computational assessment of RNA–RNA interaction genome cyclization

Putative interaction sites between 3′ and 5′ TRs of 109 JEV isolates were predicted with IntaRNA v3.2.0, RNAup v2.4.18 and RNAcofold (24–28). A consensus secondary structure of the 5′/ 3′ TRs in 20 phylogenetically related flaviviruses was computed with RNAalifold v2.4.18 (29) from the ViennaRNA Package v2.4.18 (30), based on a structural nucleotide multiple sequence alignment computed with LocARNA v2.0.0RC8 (31).

We evaluated whether the predicted TR interactions are kinetically feasible by applying a novel direct path model. In this model, the full target interaction is formed starting from an initial seed interaction, which folds into the target interaction by adding or removing target base pairs. All substructures on a folding path consist of consecutive base pairs of the target interaction. Consequently, paths are 'direct' since they cannot contain detours via non-target interaction base pairs. We compute the free energy of each substructure analogous to the RNAup energy model. The two main contributions are the cost of making the interaction site unpaired in the two intramolecular structures and the stabilizing contribution of the interaction base pairs. Both contributions were obtained from the ViennaRNA library. We can model interaction formation as a Markov process based on this direct paths model and the energy model/function. The overall rate at which an interaction is formed is determined by the energy barrier along the direct folding path (activation energy). Since the model's possible substructures (aka states) can be completely described by their first base pair i and last base pair j, they form a 2D energy landscape, and any direct path from a first base pair to the full interaction can be drawn, as shown in Supplementary Figure S2. Plots of these energy landscapes allow for the visual identification of barriers along the folding path.

Preparation and purification of non-coding RNA

JEV 3′ TR, 3′ TR Mut and 5′ TR constructs of JEV were designed based on the GenBank sequence of KR265316.1, while JEV 5′ and 3′ UTR were based on KT957419.1. cDNA sequences were prepared in pUC57 plasmids under

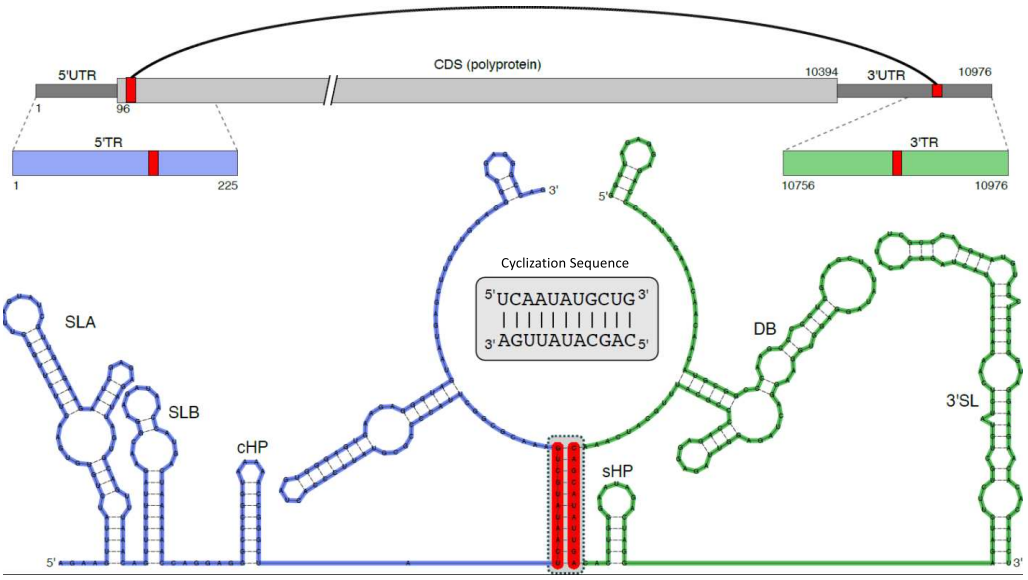


Figure 1. JEV structural organization and theoretical secondary structure prediction. Visual representation of JEV genome organization and different construct regions which were designed and studied. Predicted secondary structure of the cyclization between JEV 5′ TR and JEV 3′ TR from constraining co-folding, highlighted red portions represents the 11 conserved nucleotides which form a stable base pairing interaction to facilitate the long-range genome interaction.

the control of T7 RNA polymerase. The cDNA was flanked by two additional G nucleotides on the 5′ end and an XbaI restriction enzyme cut site (T^CTAGA) on the 3′ end. The 3′ TR and 5′ TR constructs are graphically represented in Figure 1A to enhance clarity on which portion of the terminal region is involved in each experiment. All RNA construct sequences are listed on the next page with an underlined region representing the theoretical cyclization sequence (CS) which forms the 11-nucleotide base pairing complement between the 5′ and 3′ TR. Briefly, JEV 5′ UTR serves as a cyclization sequence truncation mutant, having nucleotides 97–224 removed compared to JEV 5′ TR. JEV 3′ TR Mut is an exact copy of JEV 3′ TR, with only the 11-nucleotide cyclization sequence mutated to no longer base pair to JEV 5′ TR.

RNA was prepared via *in vitro* transcription reaction using T7 RNA polymerase (purified in-house) followed by size-exclusion chromatography (SEC) purification using a Superdex 200 Increase GL 10/300 (Global Life Science Solutions USA LLC, Marlborough, MA, USA) in JEV RNA Buffer (10 mM Bis-tris pH 5.0, 100 mM NaCl, 15 mM KCl 15 mM $MgCl_2$, 10% glycerol) via an ÄKTA pure FPLC (Global Life Science Solutions USA LLC, Marlborough, MA, USA) with a flow rate of 0.5 mL/min. Urea-polyacrylamide gel electrophoresis (urea-PAGE) was utilized to analyze SEC peak fractions. We mixed 10 μL of each fraction with 2 μL of denaturing RNA loading dye and loaded it into a 1.0 cm well PAGE casting plate (Bio-Rad Laboratories, Mississauga, ON, Canada). Urea-PAGE (7.5%) was run at room temperature, 300 V, for 25 min (20 min for JEV 5′ TR) in 0.5× TBE (Tris-Borate-EDTA) buffer (heated), followed by staining with Sybr safe (Thermofisher Scientific, Saint-Laurant, QC, Canada) and visualization. Fractions containing a single band were deemed acceptable and used in subsequent experiments.

Light scattering

Multi-angle light scattering (MALS) experiments were performed on a Dawn® (Wyatt Technology Corporation, Santa Barbara, CA, USA) multi-angle light scattering instrument with 18 detector angles utilizing a 658 nm laser. Furthermore, an Optilab® (Wyatt Technology Corporation, Santa Barbara, CA, USA) refractometer was also positioned downstream to measure the solvent refractive index and absolute concentration of solutes. These instruments were positioned in line with an SEC column (Superdex 200 increase 10/300 GL, Global Life Science Solutions, USA LLC, Marlborough, MA, USA) attached to an ÄKTA pure FPLC (SEC-MALS). All experiments were performed at ambient room temperature (20 °C) with the same flow rate and buffer as previous SEC experiments described above. The refractive index of the solvent was defined as 1.3308 (measured by in-line Optilab® refractometer), while the dn/dc (refractive index increment) value of 0.1720 mL/g was used for all RNAs (32). The final concentration of 3′ TR and 5′ TR RNA used was 1.3 and 3.5 μM, respectively, in a combined volume of 500 μL. Samples were incubated for 3 h prior to loading. Data were analyzed using Astra v8.0.0.25, and absolute molecular weight (M_w) was

Name	Genomic Location	Length (NTs)	Sequence (5′–3′)	M_w (kDa)
JEV 5′ TR	1–224	228	GGAGAAGUUUAUCUGUGUGAACUUCUUGGCUU AGUAUCGUUGAGAAGAAUCGAGAGAUUAGUGCAG UUUAAACAGUUUUUUAGAACGGAAGAUAACCAUG ACUAAAAAACCAGGAGGGCCCGGUAAAAACCGGG CUAUCAAUAUGCUGAAACGCGGCUUACCCCGCGU AUUCCCACUAGUGGGAGUGAAGAGGGUAGUAAUG AGCUUGUUGGACGGCAGAGGGCCAGU	73.9
JEV 5′ UTR	1–96	100	GGAGAAGUUUUAUCGUGUGAACUUCUUGGCUU AGUAUCGUUGAGAAGAAUCGAGAGAUUAGUGCAG UUUAAACAGUUUUUUAGAACGGAAGAACAACCU	32.2
JEV 3′ TR	10758–10976	221	GGUUAGAGGAGACCCCGUGGAAACAACAACAUGC GGCCCAAGCCCCCUCGAAGCUGUAUAGGAGGUGG AAGGACUAGAGGUUAGAGGAGACCCCGCAUUUGC AUCAAACAGCAUAUUGACACCUGGGAAUAGACUAG GAGAUCUUCUGCUCUAUCUCAACAUCACUUACUA GGCACAUAUCGCCGAAGUAUGUAGCUGGUGGUAG GAAGAACACACGAUCU	71.4
JEV 3′ TR Mut	10758–10976	221	GGUUAGAGGAGACCCCGUGGAAACAACAACAUGC GGCCCAAGCCCCCUCGAAGCUGUAUAGGAGGUGG AAGGACUAGAGGUUAGAGGAGACCCCGCAUUUGC AUCAAAAUCUUCGAACUCACCUGGGAAUAGACU AGGAGAUCUUCUGCUCUAUCUCAACAUCACUUAC UAGGCACAUAUCGCCGAAGUAUGUAGCUGGUGGU AGGAAGAACACACGAUCU	71.4
JEV 3′ UTR	10395–10965	573	GGACAGGAUAAAGUCAUAUGUGUAAUGUGAGAUA AGAAAAUGUGCAUGUGGAGUCAGGCCAGCAAAAG CUGCCACCGGAUACUGAGUAGACGGUGCUGCCUG CGUCUCAGUCCCAGGAGGACUGGGUUAACAAAUC UGACAACGGAAGGUGGGAAAGCCCUCAGAACCGU CUCGGAAGCAGGUCCCUGCUCACCGGAAGUUGAA AGACCAACGUCAGGCCACAAUUCUGUGCCACUCC GCUGGGGAGUGCGGCCUGCGCAGCCCCAGGAGGA CUGGGUUAACAAAGCCGUUGAGGCCCCCACGGCC CAAGCCUCGUCUAAGAUGCAAUAGACUAGGUGUA AGGACUAGAGGUUAGAGGAGACCCCGUGGAAACA ACAUUGUGCGGCCCAAACCCCCUCGAAGCUGUAG AGGAGGUGGAAGGACUAGAGGUUAGAGGAGACCC CGCAUUUGCAUCAAAACAGCAUAUUGACACCUG GGAAUAGACUGGGAGAUCUUCUGCUCUAUCUCAA CAUCAGCUACUAGGCACAGAGCGCCGAAGUAUGU AGCUGGUGGUGAGGAAGAAACCAGGAUCUU	185.6

calculated using Equation 1 for each elution point, where: $R(\theta)$ is Rayleigh's ratio, K is the polymer constant, and c is the concentration of the solution.

$$M_w = \frac{R(\theta)}{K^* c} \quad (1)$$

Analytical ultracentrifugation

AUC data for 5′ TR and 3′ TR were collected using a Beckman Optima AUC centrifuge and an AN50-Ti rotor at 20 °C at the Canadian Center for Hydrodynamics at the University of Lethbridge. We measured the 5′ TR (214 nM, 0.5 OD at 260 nm), 3′ TR (224 nM, 0.5 OD at 260 nm), and a 1:1 mixture of 5′ TR and 3′ TR samples in standard Beckman Coulter cell housings equipped with Epon-2 channel centerpieces, and fitted with sapphire windows, in JEV RNA Buffer. We centrifuged samples at 25 000 rpm and collected scans at 20-s intervals. We used the UltraScan-III package (33) to analyze all data via supercomputer calculations in-house. We analyzed the SV-AUC data using 2D spectrum analysis (2DSA) (34,35) with simultaneous removal of time-invariant noise, meniscus and bottom positions fitted, followed by enhanced van Holde–Weischet analysis (36). We estimated the buffer density and viscosity corrections with UltraScan (1.0269 g/cm^3 and 1.293 cP, respectively). All hydrodynamic parameters were corrected to standard conditions at 20 °C and water.

Fluorescent labeling of RNA

Purified RNA was subjected to labeling at the 5′ end by the fluorophore Alexa 488 (Thermofisher Scientific, Saint-Laurant, QC, Canada). One milligram of A488 was resuspended in 175 μL of 0.2 M KCl. About 7.5 μL concentrated RNA (>100 μM) was added to 1.25 mg 1-ethyl-3-(3-dimethylamino) propyl carbodiimide hydrochloride (EDC) prior to the addition of 10 μL resuspended A488. Samples were vortexed until contents were dissolved entirely before adding 20 μL of 0.1 M imidazole, pH 6. Reactions were incubated in a 37°C water bath overnight in the absence of light, followed by the removal of free dye through 10 kDa Vivaspin® 500 centrifugal concentrators (Sartorius Stedim Biotech, Göttingen, Lower Saxony, Germany). After removing all free dye, labeled RNA was diluted in JEV RNA Buffer, and fluorescence checks were conducted using microscale thermophoresis (MST).

Microscale thermophoresis

A 2-fold serial dilution was performed on the RNA ligand, either 3′ TR, 3′ TR Mut or 3′ UTR, where the highest concentration in the assay was 21, 9.8 and 5.5 μM, respectively. A constant amount of fluorescently labeled RNA Target, 5′ TR or 5′ UTR was added to each serial dilution of RNA ligand resulting in a final concentration of 25 and 52 nM, respectively. Mixtures were incubated at room temperature for 3 h and then loaded into a Nanotemper Technologies Monolith® NT.115 instrument (Munich, Germany) using standard capillaries. Thermophoresis was measured at room temperature (22°C) and performed using 100% excitation power along with medium IR-Laser power. Initial fluorescence migration was measured from −1.0 to 0 s and used to normalize the measured fluorescent migration time (9.0 to 10.0 s). Three independent replicates were analyzed using MO.Affinity Analysis software v2.1.3 and fit to the standard K_D fit model, which describes a 1:1 stoichiometric molecular interaction according to the law of mass action. The dissociation constant (K_D) is estimated by fitting equation 1., where $F(c)$ is the fraction bound at a given ligand concentration c. Unbound is the F_{norm} signal of the isolated target; Bound is the F_{norm} signal of the complex, while c_{target} is the final concentration of the target in the specific assay.

$$F\ (c) = \mathbf{Unbound} + (\mathbf{Bound} - \mathbf{Unbound}) \times \frac{c + c_{\mathbf{target}} + K_d - \sqrt{\left(c + c_{\mathbf{target}} + K_d\right)^2 - 4\ c\ c_{\mathbf{target}}}}{2\ c_{\mathbf{target}}} \quad (2)$$

Small-angle X-ray scattering

RNA sample data collection was performed on the B21 HPLC-SAXS beamline at Diamond Light Source (Didcot, Oxfordshire, UK), as reported elsewhere (37). An Agilent 1200 (Agilent Technologies, Stockport, UK) HPLC was utilized through connection to a specialized flow-cell, whereas 50 μL of each purified RNA (5′ TR, 3′ UTR and 5′ TR + 3′ UTR, respectively) was injected into a JEV RNA buffer equilibrated Shodex 403KW-4F HPLC column (Showa Denko America Inc., New York, NY, USA) with a flow rate of 0.160 mL/min. Concentrations were 1.2 and 1.1 mg/mL for 5′ TR and 3′ UTR, respectively, and the complex was a 1:1 volume mixture of both. Frames were exposed to synchrotron radiation (X-rays) for 3 s for a total of ∼600 frames. The resulting data were buffer subtracted using Chromixs (38) for each sample peak, and then data analysis was performed using the ATSAS suite of programs (39). The radius of gyration (R_g) was evaluated through Guinier analysis, while additionally determining sample quality (40) and relative foldedness of RNA molecules were determined via dimensionless Kratky analysis (41). GNOM was used to perform paired distance distribution $P(r)$ analysis to obtain real space R_g and maximum particle dimension (D_{max}) measurements (42,43). Using $P(r)$ derived information, 20 models were generated for 5′ TR and 100 models for 3′ UTR using DAMMIN (44). Following simulated annealing via DAMMIN, 5′ TR models were averaged and filtered to produce a single representative model using DAMAVER and DAMFILT (44,45). Given the size of 3′ UTR, we decided to cluster representative models instead of generating a singular averaged model via DAMCLUST (45). Model reconstruction for the 5′ TR and 3′ UTR complex was generated through MONSA (44) using data input from 5′ TR, 3′ UTR and data from 5′ TR + 3′ UTR. Hundred models were generated and then partitioned into representative clusters via DAMCLUST.

RESULTS AND DISCUSSION

Computational analysis of the cyclization RNA–RNA interacting element in JEV and related flaviviruses

We performed computational analysis of the potential long-range interaction between the 5' and 3′ TRs in >100 JEV isolates. This work revealed that all isolates interact via the canonical 11nt cyclization sequences in their 5′ and 3′ TRs. Based on this work, we selected JEV isolate KR265316.1 as a model to assess long-range RNA–RNA interactions *in silico* and *in vitro*. The Flaviviral TRs are known to harbor functional RNA elements such as stem-loops A and B in the 5′ UTR, a short conserved hairpin (cHP) at the beginning of the coding region, and multiple *cis*-regulatory elements in the 3′ UTR, such as dumbbell and terminal 3′ stem-loop (3′ SL) structures (46,47). These conserved elements exert crucial roles in the flaviviral life cycle; therefore, we required them to be formed by constraint co-folding of the 5′ and 3′ TRs. As presented in Figure 1, the resulting duplex structure suggests that the canonical 11 nt cyclization sequence in the 5′ and 3′ TRs is responsible for mediating the interaction. Furthermore, the other canonical secondary structures, such as SLA, SLB and cHP in the 5′ TR with sHP, the 3′ SL and the 3′ DB in the 3′ TR, are also present (Figure 1).

We performed a comparative genomics assay in phylogenetically related viruses to further assess flaviviral long-range RNA–RNA interactions. To this end, we analyzed the propensity of duplex formation between 5′ and 3′ TRs in 20 mosquito-borne flaviviruses (Supplementary Figure S2) utilizing consensus structure evaluation of the terminal genomic regions (Figure 2). Interestingly, the tendency to form a long-range interaction via the canonical cyclization motif is more pronounced in the consensus duplex than in the single sequence (JEV only) analysis. Specifically, the 11 bp duplex is formed in the consensus structure without any constraints on the known 5′ and 3′ UTR elements (Figure 2). All 20 sampled mosquito-borne flaviviruses have the capacity to form the exact same long-range interaction. The almost perfect conservation of the cyclization sequences suggests that this region is not only functionally essential but also suggests that there may be evolutionary importance to the sequence conservation.

Additionally, a kinetic analysis of the canonical cyclization structure suggests that it is also kinetically favored in all investigated JEV isolates. Exemplarily, the energy landscape of the cyclization structure from isolate KR265316.1 is shown in Figure 6 (discussed later). Independent of the selected start base pair, every interaction extension step leads to a more stable structure. Thus, there are no barriers along the folding paths, and the interaction can form fast. We then

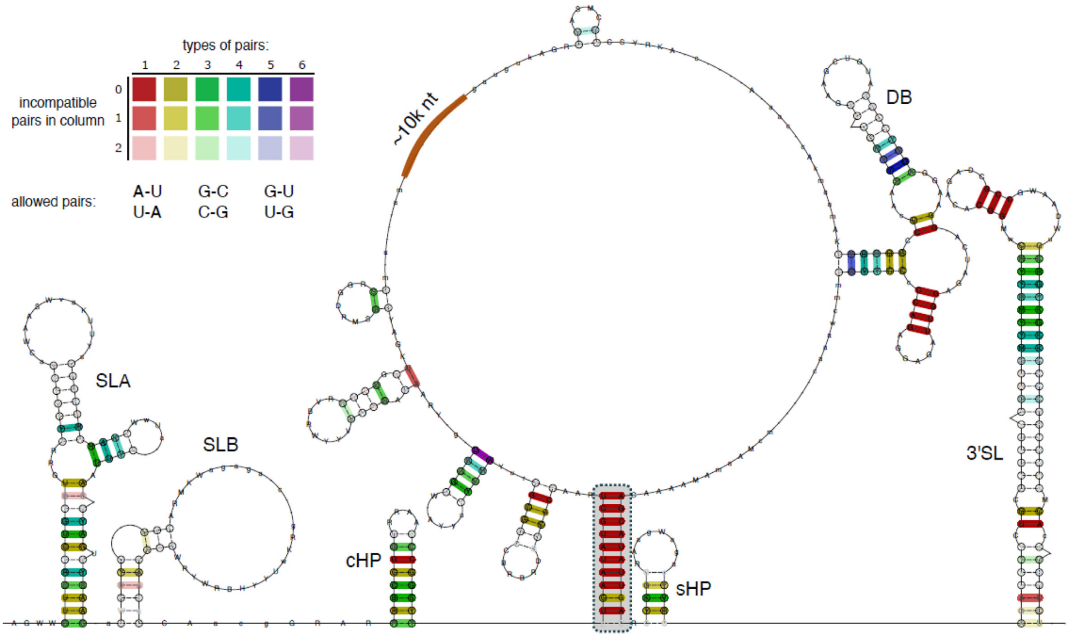


Figure 2. Consensus secondary structure of the 5′ TR/3′ TR long-range interaction, computed from a structural multiple sequence alignment of 20 mosquito-borne flaviviruses (Supplementary Figure S1). Coloring of base pairs follows the RNAalifold schema and indicates different covariation levels, ranging from red (no covariation, full primary sequence conservation) to violet (full covariation, all six possible combinations of base pairs) at corresponding columns of the underlying alignment. The duplex formed by the almost fully sequence-conserved 11 nt cyclization sequences (highlighted in gray) represents the only long-range interaction in the consensus structure. Canonical 5'TR elements (SLA, SLB, and cHP), as well as 3'TR elements (DB, sHP and 3'SL) are predicted to fold in the consensus structure, indicating that they are energetically more favorable than an extended long-range interaction duplex structure.

sought to validate the *in silico*-derived cyclization interaction in JEV through extensive biophysical characterization *in vitro*.

In vitro transcription and purification of JEV non-coding RNA for interaction studies

JEV viral non-coding RNAs were purified immediately after *in vitro* transcription using SEC, like in previous works (20,48,49). The elution profile for 5′ TR (Figure 3A) indicates that the RNA elutes at approximately 11.5 mL as a single monodispersed species, evident by the typical Gaussian distribution. 3′ TR elutes as a multimodal distribution with an observable peak at ∼12 mL, consistent with its size compared to 5′ TR (Figure 1A). The 5′ TR and 3′ TR transcripts have very similar molecular weights, 73.5 and 71.3 kDa, respectively, which is reflected in the chromatogram as both monomeric peaks appear to have a similar elution volume ∼11.5–12.0 mL (Figure 3A). Higher-order oligomeric species for 3′ TR can be observed at ∼8–11 mL, and these elution fractions we avoided for downstream experiments. To determine if the SEC elution fractions contained the correct-sized RNA species, we utilized urea-PAGE. Figure 3B (left side) shows 3′ TR fractions from the right side of the ∼12 mL peak containing the appropriate length RNA species (227 nt). Only selected fractions (12.5–13.5 mL) were pooled to avoid contamination from any oligomeric species (Figure 3A, red). 5′ TR elution fractions (Figure 3B, right side) demonstrate, as expected, a single species of the correct RNA size (221 nt) across the peak. These fractions were pooled similarly to 3′ TR and highlighted for clarification (Figure 3A, green). 3′ TR Mut, 3′ UTR and 5′ UTR (additional RNA used later) were transcribed and purified identically to the above RNA, and only fractions containing a single band were pooled and used in downstream experiments.

Biophysical analysis of the RNA–RNA interacting complex

Due to the highly conserved nature of Flavivirus TRs, it is theorized that JEV will utilize similar 3′–5′ long-range TR interactions as a necessary step in viral RNA amplification, like other members (50). This interaction has been shown in various Flaviviridae family members (16–18,21,22). SEC can separate biomolecules based on their sizes, and MALS allows for the absolute molecular weight determination of biomacromolecules in solution (51); we utilized SEC-MALS to investigate if the 5' and 3′ TRs form a complex. Figure 4A represents the elution profiles of JEV 3′ TR, which indicates that it is almost entirely monodispersed eluting at ∼12 mL with a minor peak of potential oligomeric assembly at ∼11.5 mL. The elution profile of 3′ TR is consistent with our initial purification using SEC, where JEV 3′ TR oligomerizes into higher-order species.

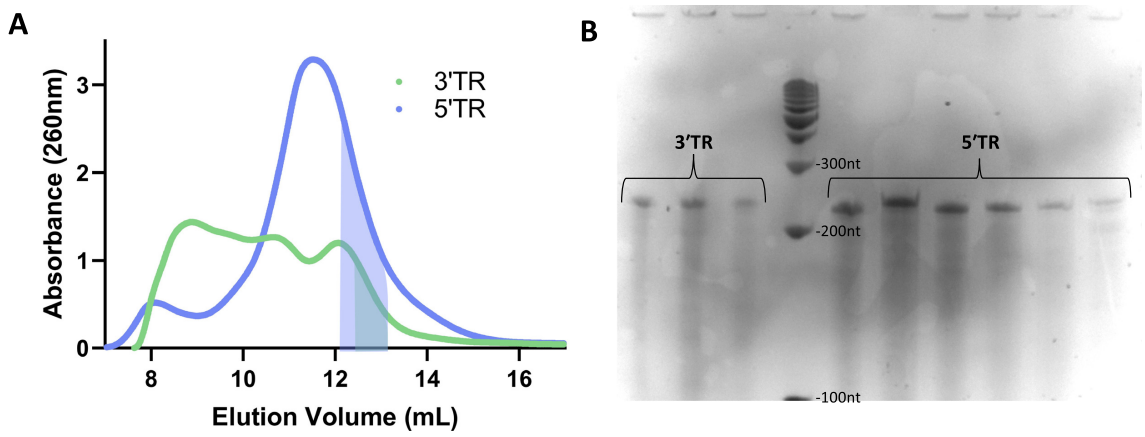


Figure 3. Purification of JEV TR RNA. (**A**) Size-exclusion chromatogram representing the purification of both RNA. Shaded boxes represent the region which was collected for downstream experiments to avoid any potential oligomeric species. (**B**) Urea PAGE of associated size exclusion chromatography fractions showing a single size of RNA (∼220 nt) which is the correct size of the expected RNA.

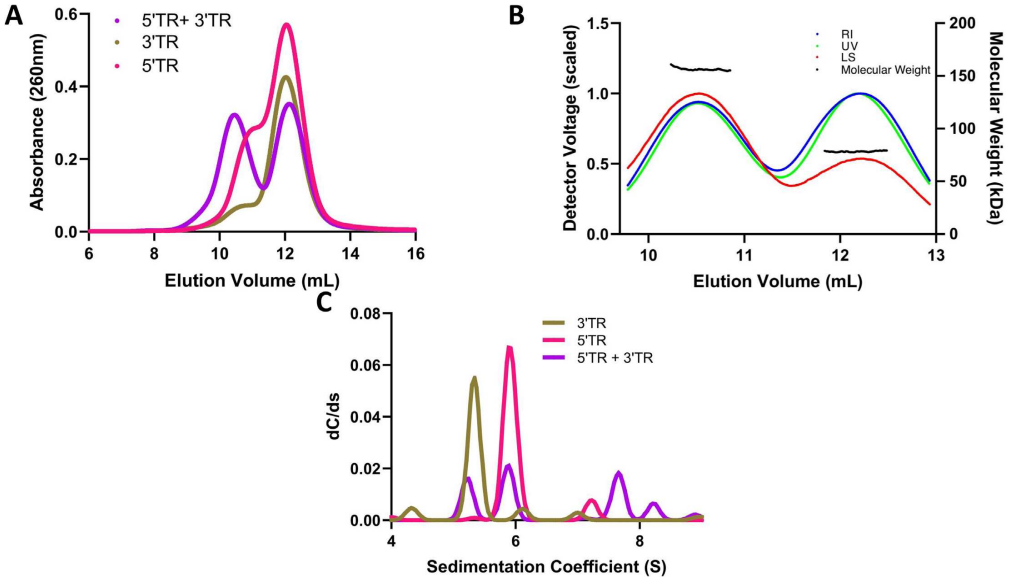


Figure 4. Light scattering analysis of JEV TR RNA cyclization. (**A**) Multiple size exclusion chromatography runs associated with SEC-MALS. (**B**) MALS traces of each peak from the 5′ TR + 3′ TR run, and the absolute molecular weight across them. (**C**) Sedimentation distribution profiles of JEV 5′ TR and 3′ TR obtained from sedimentation velocity-analytical ultracentrifugation. Sedimentation coefficient values are corrected to standard solvent conditions (20 °C, water)

JEV 5′ TR also elutes at ∼12 mL, consistent with previous purifications; however, it also contains a shoulder at ∼11.5 mL (Figure 4A).

We then mixed both 3′ TR and 5′ TR and performed SEC-MALS experiments. The RNA–RNA mix resulted in a bimodal SEC chromatogram distribution, with an additional peak eluting at ∼10.5 mL (Figure 4A, purple). Furthermore, we see a peak at ∼12 mL where excess 5′ TR RNA elutes as a monomer, consistent with the concentrations we mixed (3.5 μM of JEV 5′ TR versus 1.3 μM of JEV 3′ TR). These results indicate that the TR complex has a different elution profile upon binding than any oligomeric assembly or monomeric species. Successful separation of the TR RNA complex allowed us to calculate the absolute molecular weight of each species. As presented in Figure 4B, the SEC-MALS-derived molecular weights are consistent across both peaks, indicating that both are homogeneous species. Excess unbound monomer(s) elute as a single

peak and are characterized at ∼75 kDa, which is expected considering the similarity in their theoretical weights (73.9 and 71.4 kDa). The molecular weight of the complex (∼150 kDa) is double that of each TR in isolation (∼75 kDa) and very similar to the predicted molecular weight of the complex (145.3 kDa). This change in molecular weight is indicative of a 1:1 stoichiometric interaction. This stoichiometric determination is vital for biological relevance, considering these terminal regions exist on the same genome and come together to cyclize.

AUC is a powerful biophysical technique often used to study the purity of biomolecules in solution (50). AUC subjects biomolecules to extremely high centrifugal force (up to 250 000 × *g*), separating them based on size, anisotropy and density while monitoring sedimentation via an optical system. While SEC-MALS can provide us with absolute molecular weight determination, since both 5′ TR and 3′ TR are of similar molecular weight, we needed further validation to rule out potential RNA self-oligomerization as an explanation of the 5′ TR-3′ TR peak in figure 3A. AUC has been proven to be a useful technique to characterize RNA (48,49,52), but never an RNA–RNA interaction. Therefore, we utilized the specialized capabilities of AUC to provide us with an orthogonal technique that could validate our SEC-MALS results. While both 5′ TR and 3′ TR are similar in length, the sequences differ significantly, which will cause a change in secondary structure and, ultimately, tertiary structure. If these potential tertiary structural differences are significant enough, there should be a difference between the 5′ TR and 3′ TR sedimentation. Both 5′ TR and 3′ TR show highly monodispersed sedimentation profiles, evident by a single Gaussian distribution. 5′ TR is represented by a primary sedimentation profile at ∼5.9 S, while 3′ TR is represented at ∼5.3 S (Figure 4C). These distributions confirm that not only are the TRs homogeneous, but they also sediment differently based on tertiary structure differences. This difference in sedimentation profiles allowed us to validate our previous SEC-MALS results by mixing 5′ TR and 3′ TR in a 1:1 ratio and performing another sedimentation velocity experiment. As with previous SEC-MALS experiments, the AUC results show an additional sedimentation peak at ∼7.6 S, which is not present in the individual TR experiments allowing us to confirm JEV 5′–3′ TR interaction (Figure 4C). Additionally, since the concentrations of both RNA were ∼200 nM, we can also conclude that the binding affinity (K_D) is likely in the low nanomolar range. We believe this experiment provides the first evidence of an RNA–RNA interaction via AUC, showing that AUC can be an essential tool for showing RNA–RNA interaction while simultaneously determining sample quality and oligomeric state. Taking the AUC and SEC-MALS results together, we can confidently conclude that the additional peak(s) formed in both experiments is the complex of 5′ TR and 3′ TR, which provides experimental validation of computational predictions.

Determining the RNA–RNA interaction affinity

With the above-mentioned relative binding affinity and direct evidence of interaction, we sought to quantify the affinity of the 5′ and 3′ TRs using MST. MST allows for studying the interaction between a serially diluted ligand and a fluorescently labeled target by measuring the change in fluorescent migration following excitation by an infrared laser (53). Using the difference between the 'cold' and 'hot' areas of the MST traces (Figure 5A), MST can determine the K_D of the biomolecular interaction. Additionally, Figure 5A demonstrates that aggregation of the target was not observed (54) and that monomeric fractions were utilized in the assay. Our analysis indicates that the TRs interact with a K_D of 60 ± 9 nM (Figure 5B), which agrees with the previously reported K_D of 32 ± 1 nM found in WNV using isothermal titration calorimetry (22), which is reasonably similar, validating our findings.

Consequently, the binding affinity further validates our AUC experiment, which was performed with a ∼200 nM concentration of JEV TR RNA. To our knowledge, few studies (55,56) have shown evidence of RNA–RNA interactions characterized using MST, highlighting the growing applicability of the technique. Next, we needed to validate whether the CS is the primary driver of this RNA–RNA interaction. We computationally predicted the RNA duplex secondary structure (Supplementary Figure S3), showing numerous potential base-pairing sites across the interaction, but the longest uninterrupted stretch of nucleotides was the CS. We investigated the degree to which the inclusion of the CS impacts binding affinity by exploring various fragments of the JEV TRs. JEV 3'UTR, the full-length 3′ non-coding RNA at 573 nt, and JEV 5′ TR, which both still include the CS, were assayed and found to interact with a K_D of 169 ± 18 nM (Figure 5B). As expected, the theoretically determined highest binding isolates of JEV (5′ TR and 3′ TR) bound with a slightly higher affinity than the 3′ UTR and 5′ TR but are within the same magnitude. When excluding the CS through a truncation (JEV 5′ UTR-97 nt), the construct lacking 97–224 compared to JEV 5′ TR showed that the interaction was almost non-existent with a K_D of 23 ± 6 μM. Finally, we performed the same assay with a mutant version of 3′ TR, 3′ TR Mut. This construct is identical to JEV 3′ TR but has the 11 nucleotides of the CS mutated to no longer base pair to the 5′ CS. Therefore, we believed that including the additional long-range binding element would result in a binding tighter than the 5′ UTR–3′ TR interaction but still considerably weaker than the canonical CS included. The interaction was as expected and was significantly lower than the 5′ TR–3′ TR interaction, with a K_D of 7 ± 1.4 μM (Figure 5B). This change in affinity demonstrates that while the theoretical duplex interaction may have considerable base pairing, the primary driver of the interaction is the cyclization sequence.

RNA–RNA kinetic and thermodynamic studies

As observed in Figure 3A and Figure 4A, both TRs are likely to self-associate even though careful consideration was taken to avoid this. Therefore, we utilized computational analysis to explain the self-association of TRs potentially. This kinetic model builds upon the Turner nearest neighbor energy model for RNA secondary structures (57). While our predictions find potential homodimer interactions with similar thermodynamic stability as the hetero duplex (canonical CS), these homodimer interactions

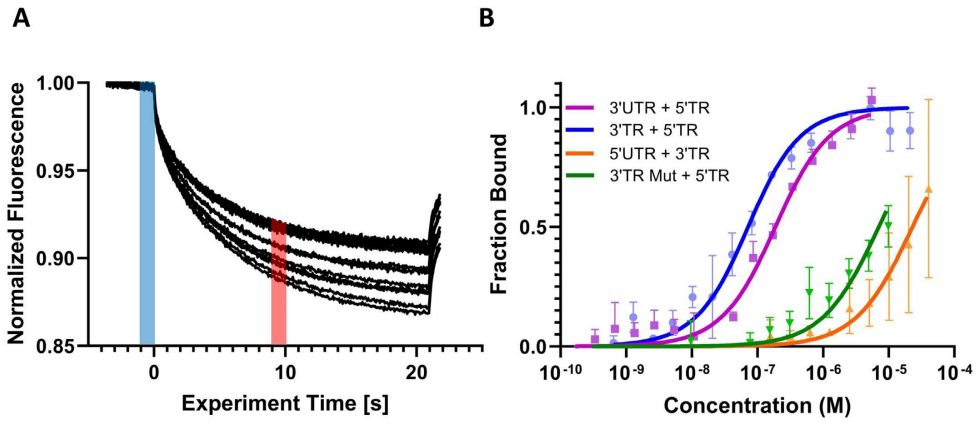


Figure 5. Affinity analysis of JEV TR RNA cyclization. (**A**) MST raw data traces for JEV 3′ TR + 5′ TR. Blue line represents the 'cold' time and the red line represent the 'hot' region, and the difference between the two is used to calculate the ΔF_{norm}. (**B**) Microscale thermophoresis measurements of different combinations of JEV TRs representing concentration versus fraction bound. Measurement ran on 'high' MST power.

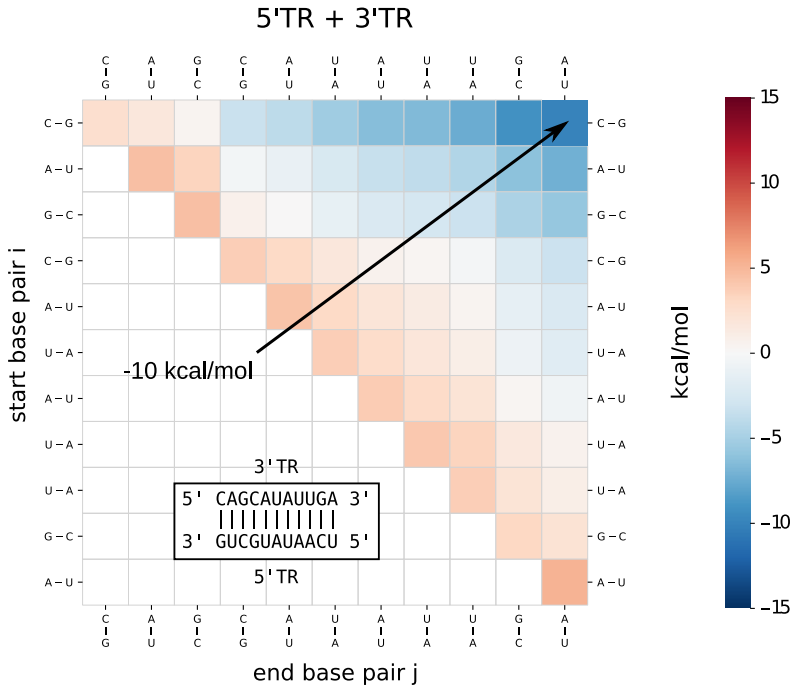


Figure 6. Energy landscape of the predicted 3′ TR and 5′ TR CS interaction. The energy landscape represents all possible substructures that can occur along a direct path from a single base pair interaction to the full CS interaction.

are quite extended and would require extensive refolding. Based on a kinetic analysis (Supplementary Figure S2), we assume that homodimers would form only short, less stable interactions, which would readily dissociate in favor of the more stable heterodimer (3′-5′ complex) when mixed.

To help explain our MST observations, we compared the measured dissociation constants to the predicted interactions for the different TR and UTR constructs. We calculated binding free energies at 37°C either from the measured dissociation constants or from predictions using RNAup (predictions are in parentheses). For the three constructs,

the 5′ TR and 3′ TR, 5′ TR and 3′ UTR, and 5′ UTR and 3′ TR complex, we obtained ΔGs of −10.2 (−10.1) kcal/mol, −9.6 (−7.6) kcal/mol and −6.6 (−7.2) kcal/mol, respectively. The more considerable discrepancy for the second construct (5′ TR and 3′ UTR) can be explained because the thermodynamic model predicts a refolding (compared to the known consensus structure) when the longer 3′ UTR is used, occluding the CS. We also investigated potential additional interaction sites to the known CS. The most promising is a kinetically favorable interaction corresponding to the known upstream AUG Region (UAR) (58) with a predicted stability of −3.6 kcal/mol. This interaction could stabilize the CS interaction but is not strong enough to explain the duplex formation alone. Moreover, in a cellular environment, other factors could likely influence binding kinetics and thermodynamics; however, this evidence shows that this interaction is spontaneous and favorable even in the absence of other cellular factors. One would assume that the contact between the 5′ and 3′ TRs would be rare in solution due to the distance between them, but that once a single contact has formed, the rest would spontaneously 'zip up.' This phenomenon is likely consistent within the cell regarding the amount of (−) sense RNA versus (+) sense RNA (59). Most (+) sense viral RNA is translated into proteins and not used to replicate the virus.

3D RNA–RNA interaction analysis through small-angle X-ray scattering

Having characterized the cyclization RNA–RNA binding affinity, we turned to small-angle X-ray scattering to understand the 3D interaction in solution. Using SEC-SAXS, it is possible to differentiate species based on size and shape prior to actual SAXS measurements, providing increased confidence in monodispersity and allowing for the characterization of a potential complex from two mixed samples (60–63). MST experiments revealed that the binding affinity of the 5′ TR and 3′ UTR were comparable to the 5′ TR–3′ TR interaction, so we decided to perform structural experiments on the larger RNA construct (3′ UTR) to gain a better characterization of how this RNA interaction is arranged in solution and its relative flexibility.

Three SAXS data sets were collected: 5′ TR, 3′ UTR and 5′ TR + 3′ UTR. Each was merged and represented in Supplementary Figure S4a as relative intensity versu scattering angle. Guinier analysis was performed on each data set to determine the radius of gyration (R_g) for each RNA resulting in 71.61 ± 0.106 Å and 111.5 ± 0.33 Å for 5′ TR and 3′ UTR, respectively (Supplementary Figure S4b). Complex (5′ TR + 3′ UTR) Guinier analysis reveals an R_g of 127.6 ± 0.34 Å, an increase compared to each RNA. The linear regression of each sample demonstrates that each sample is monodispersed and free of any electrostatic interactions between similar molecules (64–66). Intensity data were transformed into dimensionless Kratky analysis data (Supplementary Figure S4c) to evaluate approximate foldedness and conformation (67). Kratky analysis (Supplementary Figure S4c) for all three RNA suggests extended conformations of folded RNA based on the relative plateauing of each data set (64). Finally, each data set was evaluated using paired-distance distribution ($p(r)$) analysis in which reciprocal space data is converted into real-space electron density data via indirect Fourier transformations (68). $P(r)$ analysis presents real space R_g values of 71.8 ± 0.104 Å, 112.0 ± 0.329 Å and 128.1 ± 0.355 Å for the 5′ TR, 3′ UTR, and the complex, respectively (Supplementary Figure S4d). Real-space R_g values have a very high level of agreement with the reciprocal-space values, indicating the validity of the entire data set. Furthermore, the shape of each $P(r)$ plot indicates an elongated conformation, whereas a Gaussian distribution with a peak at $D_{max}/2$ would be indicative of a spherical, globular molecule. The maximum distance for each RNA data set was 224, 345 and 400 Å for the 5′ TR, 3′ UTR and complex, respectively. Notably, the D_{max} values illustrate that while the complex is larger than either individual RNA, likely that the interaction is not end-to-end. The maximal distance for an end-to-end interaction would be considerably >400 Å, suggesting there may be considerable overlap between the 5′ TR and 3′ UTR.

Ab inito modeling through DAMMIN was then performed on 5′ TR and 3′ UTR data sets to generate low-resolution 3D structures. Hundred models were generated for each RNA showing favorable agreement via χ^2 values of 1.15 and 1.12 for the 5′ TR and 3′ UTR, respectively. The 5′ TR was filtered and merged into a singular representative structure with a normalized spatial discrepancy (NSD) of 0.937 ± 0.020, indicating a good fit of each structure to the representative filtered model (Figure 7A) (69). Represented in blue, the 5′ TR is an elongated RNA structure like other RNA of similar size (49,52,70). The 3′ UTR, however, could not be filtered and averaged together to get a singular representative model, even though the average χ^2 value was 1.12, suggesting that each model strongly agreed with the original scattering data. We, therefore, chose to cluster the 100 respective models from the 3′ UTR, resulting in 22 distinctly different clusters represented by the pie chart in Figure 7B. This suggests that the 3′ UTR has flexible regions, similar to other long non-coding RNA (71,72). The three largest clusters of models are color coded to the pie chart, and while each is classified as a distinct structure, they contain a similar overall architecture: each having a large center pocket and large portions of electron density toward each terminus.

Importantly, since the affinity of the 5′ TR–3′ UTR interaction was sufficiently low, we believed that the complex would remain intact throughout SEC-SAXS, allowing for peak separation and, ultimately, data collection. Creation of 3D models required multiple distinct inputs, differing from the terminal regions individually. Using MONSA, multiphase bead modeling can simultaneously fit multiple scattering curves to a single data set (44,73). Therefore, discerning which portion of the total scattering data was contributed by each RNA element required inputting data collected from the 5′ TR, the 3′ TR and the complex. Importantly, when evaluating χ^2 values for the complex, each input is scored based on its agreement with the raw scattering data with approximate values of 1.20, 1.40 and 1.32 for the 5′ TR, 3′ UTR and complex, respectively. These values show that each distinct scattering curve fits into the data set and has a good agreement. Like the 3′ UTR on its own, a singular averaged model cannot accurately represent the RNA, which led us to cluster the model representations (Figure 7C) similarly. This clustering resulted in 11 distinct repre-

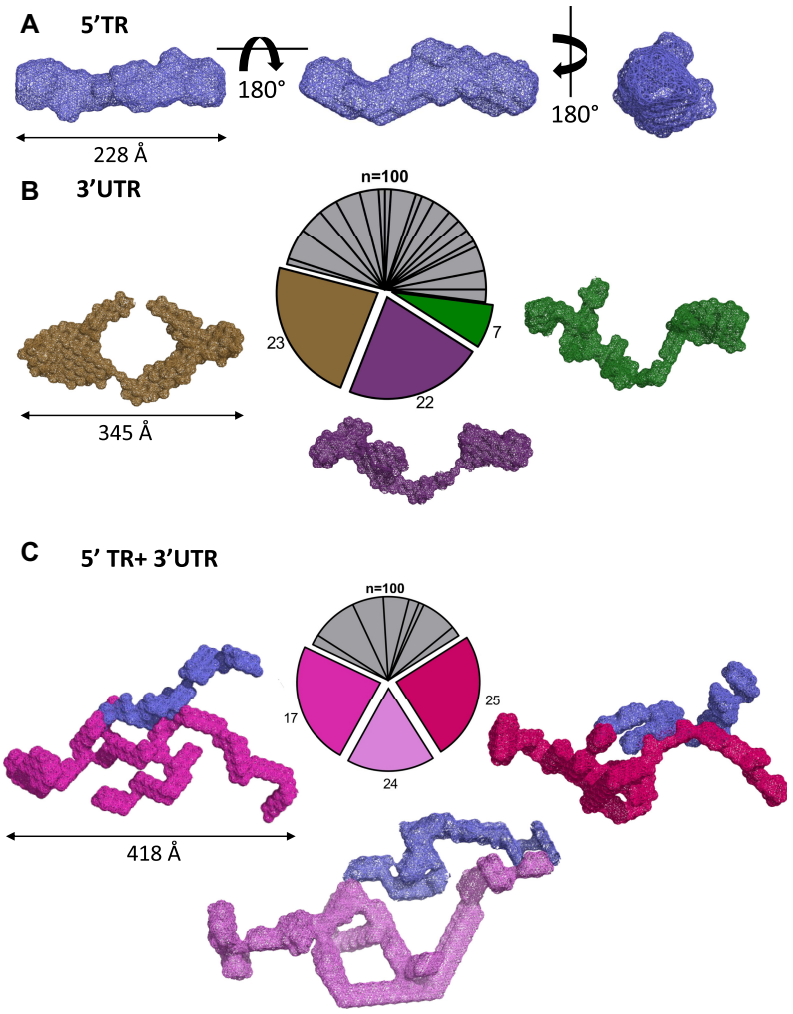


Figure 7. Small-angle X-ray scattering *ab inito* model reconstructions of JEV RNA. (**A**) JEV 5' TR RNA, rotations are 180° about both the *X*- and *Y*-axis and showing D_{max}. (**B**) JEV 3' UTR RNA illustrating relative conformational fluidity through clustering analysis of 100 *ab inito* models. Shown models represent three of the largest clusters of models. (**C**) JEV 5′ TR + 3′ UTR RNA–RNA complex. JEV 5′ TR is represented in blue, while 3′ UTR is represented in shades of red/purple.

sentations of the solution structure, half as many clusters as the 3′ UTR alone, with 66% of the total structures falling into three clusters. Clustering provides evidence that the interaction causes a change in the conformational space and a subsequent reduction in the flexibility of the 3′ UTR. Figure 7C represents the top 3 conformational clusters, with the 5′ TR in a consistent blue color and the 3′ UTR in pink/red. The 5′ TR remains relatively unchanged upon interaction with the 3′ UTR, still adopting an elongated shape similar to its previous conformation (Figure 7A). Most of the conformational change to the complex is changes to the 3′ UTR. Overall, the central pocket, which was visible in the top 3′ UTR conformations, is no longer visible, and the electron density has been re-distributed from being concentrated around the ends of the molecule (see Figure 7B). Furthermore, the 3D complex shows that there is likely not considerable 5′–3′ duplex formation, with generally 1–2 contact points, one of which is likely the CS. Structural information about this complex can also inform on the potential mechanism by which the flaviviral RNA-dependent RNA polymerase NS5 interacts with the 5′ TR SLA yet transcribes the negative sense genome 3′ to 5′ (14,74–76). However, the nature of SAXS being low-resolution means we cannot accurately pinpoint specific RNA motifs such as the CS or the 3′ or 5′ ends. Future directions will need to be focused on gaining a higher-resolution model of the interaction through

either computational modeling with SAXS envelopes or cryo-EM imaging.

CONCLUSION

Our study demonstrates that combining computational and biophysical approaches can provide detailed insights into RNA–RNA interactions that govern many fundamental biological processes. Our SEC-MALS and AUC data also suggest, for the first time, that these techniques can be used to study RNA–RNA interactions in solution while confirming that MST is a powerful technique for affinity determination of RNA–RNA interactions. We demonstrate that computational analysis, such as kinetic landscapes, can provide vital evidence for predicting RNA–RNA interactions and even sites of interactions. Additionally, our binding affinity data presents that the CS is the primary driver of terminal region interaction with a nanomolar affinity. This binding affinity is critical to developing potential inhibitory therapeutics targeting the cyclization sequence. Additionally, we present the 3D low-resolution structure complex structure in solution, providing evidence that it is likely flexible and dynamic, with a small amount of duplex formation between the 5′ TR and 3′ UTR. Our work directly contributes to understanding the Flaviviral cyclization conservation that could help develop therapeutics with a potential for multi-virus inhibition. Moreover, we believe our comprehensive computational and biophysical pipeline can be applied to virtually any RNA–RNA interacting system in solution so long as proper sample purification and quality control are forefronts. We hope this study may influence researchers to further investigate the importance of the cyclization sequence of JEV and other flaviviruses such as ZIKV and DENV *in vivo*. Additionally, *in vivo* experiments investigating sequence conservation of the CS could be instrumental to understanding the evolutionary relationship(s) of this family of viruses.

DATA AVAILABILITY

We selected complete JEV isolates (including 3′ and 5′ UTR) from NCBI GenBank and used the first 225 nts (5′ TR) and the last 221 nts (3′ TR) to perform interaction predictions. The corresponding accession numbers are included in the supplementary information. SAXS data has been deposited to SASDB under accession numbers SASDQL9, SASDQM9 and SASDQN9 for JEV 5′ TR, JEV 3′ UTR and JEV 5′ TR + 3′ UTR complex, respectively.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

T.R.P. is a Canada Research Chair in RNA and Protein Biophysics and B.D. is the Chair 150 Research Chair in Biophysics. The SAXS data collection was supported by DIAMOND Light Source (SM26855), UK. We would additionally like to thank all of the B21 beamline scientists at Diamond Light Source for their continual support.

FUNDING

Natural Sciences and Engineering Research Council (NSERC) PGS-D award (to T.M.); NSERC Undergraduate Student Research Award (to S.M.P.); NSERC CGS-D award (to A.H.); Alberta Innovates Graduate student award (to C.R.N.); Canada 150 Research Chairs program C150-2017-00015 (to B.D.); Canada Foundation for Innovation CFI-37589 (to B.D.); National Institutes of Health 1R01GM120600 (to B.D.); NSERC Discovery RGPIN-2019-05637 (to B.D.); NSF/XSEDE TG-MCB070039N (to B.D.); Austrian Science Fund (FWF) grant numbers I-2874-N28 (to I.L.H), W-1207 RNA Biology (to I.L.H), F 80 RNAdeco (to I.L.H.), I 6440-N (to M.T.W., in part); NSERC Discovery RGPIN-2017-04003 (to T.R.P.); NSERC RTI 2020–00090 (to B.D. and T.R.P.); Canada Foundation for Innovation CFI-37155 (to T.R.P.); Canada Foundation for Innovation CFI-41008 (to B.D. and T.R.P.). Funding for open access charge: NSERC/Univ. of Lethbridge Startup Grants.
Conflict of interest statement. None declared.

REFERENCES

1. Zhang,F. and Lupski,J.R. (2015) Non-coding genetic variants in human disease. *Hum. Mol. Genet.*, **24**, R102–R110.
2. Balas,M.M., Hartwick,E.W., Barrington,C., Roberts,J.T., Wu,S.K., Bettcher,R., Griffin,A.M., Kieft,J.S. and Johnson,A.M. (2021) Establishing RNA-RNA interactions remodels lncRNA structure and promotes PRC2 activity. *Sci. Adv.*, **7**, eabc9191.
3. Aguilar,R., Spencer,K.B., Kesner,B., Rizvi,N.F., Badmalia,M.D., Mrozowich,T., Mortison,J.D., Rivera,C., Smith,G.F., Burchard,J. *et al.* (2022) Targeting Xist with compounds that disrupt RNA structure and X inactivation. *Nature*, **604**, 160–166.
4. Jain,A. and Vale,R.D. (2017) RNA phase transitions in repeat expansion disorders. *Nature*, **546**, 243–247.
5. Van Treeck,B. and Parker,R. (2018) Emerging roles for intermolecular RNA-RNA interactions in RNP assemblies. *Cell*, **174**, 791–802.
6. Heffelfinger,J.D., Li,X., Batmunkh,N., Grabovac,V., Diorditsa,S., Liyanage,J.B., Pattamadilok,S., Bahl,S., Vannice,K.S., Hyde,T.B. *et al.* (2017) Japanese Encephalitis Surveillance and Immunization - Asia and Western Pacific Regions, 2016. *MMWR Morb. Mortal. Wkly. Rep.*, **66**, 579–583.
7. Wang,Q.-Y. and Shi,P.-Y. (2015) Flavivirus entry inhibitors. *ACS Infect. Dis.*, **1**, 428–434.
8. Huang,Y.J., Higgs,S., Horne,K.M. and Vanlandingham,D.L. (2014) Flavivirus-mosquito interactions. *Viruses*, **6**, 4703–4730.
9. Tjaden,N.B., Caminade,C., Beierkuhnlein,C. and Thomas,S.M. (2018) Mosquito-borne diseases: advances in modelling climate-change impacts. *Trends Parasitol.*, **34**, 227–245.
10. Lindenbach,B., Thiel,H.J. and Rice,C.M. (2007) Flaviviridae: the viruses and their replication. *Fields Virology*, **1**, 1101–1151.
11. Ng,W.C., Soto-Acosta,R., Bradrick,S.S., Garcia-Blanco,M.A. and Ooi,E.E. (2017) The 5′ and 3′ untranslated regions of the flaviviral genome. *Viruses*, **9**, 137.
12. Bollati,M., Alvarez,K., Assenberg,R., Baronti,C., Canard,B., Cook,S., Coutard,B., Decroly,E., de Lamballerie,X., Gould,E.A. *et al.* (2010) Structure and functionality in flavivirus NS-proteins: perspectives for drug design. *Antiviral Res.*, **87**, 125–148.
13. Mandl,C.W., Holzmann,H., Kunz,C. and Heinz,F.X. (1993) Complete genomic sequence of Powassan Virus: evaluation of genetic elements in tick-borne versus Mosquito-borne flaviviruses. *Virology*, **194**, 173–184.
14. Filomatori,C.V., Lodeiro,M.F., Alvarez,D.E., Samsa,M.M., Pietrasanta,L. and Gamarnik,A.V. (2006) A 5' RNA element promotes dengue virus RNA synthesis on a circular genome. *Genes Dev.*, **20**, 2238–2249.
15. You,S. and Padmanabhan,R. (1999) A novel in vitro replication system for Dengue Virus: initiation of RNA synthesis at the 3′-end of

exogenous viraL RNA templates requires 5′- and 3′-terminal complementary sequence motifs of the viral RNA*. *J. Biol. Chem.*, **274**, 33714–33722.
16. Zhang,B., Dong,H., Stein,D.A., Iversen,P.L. and Shi,P.-Y. (2008) West Nile virus genome cyclization and RNA replication require two pairs of long-distance RNA interactions. *Virology*, **373**, 1–13.
17. Alvarez,D.E., Lodeiro,M.F., Ludueña,S.J., Pietrasanta,L.I. and Gamarnik,A.V. (2005) Long-range RNA-RNA interactions circularize the dengue virus genome. *J. Virol.*, **79**, 6631–6643.
18. Hahn,C.S., Hahn,Y.S., Rice,C.M., Lee,E., Dalgarno,L., Strauss,E.G. and Strauss,J.H. (1987) Conserved elements in the 3' untranslated region of flavivirus rnas and potential cyclization sequences. *J. Mol. Biol.*, **198**, 33–41.
19. Meier-Stephenson,V., Mrozowich,T., Pham,M. and Patel,T.R. (2018) DEAD-box helicases: the yin and yang roles in viral infections. *Biotechnol. Genet. Eng. Rev.*, **34**, 3–32.
20. Nelson,C., Mrozowich,T., Gemmill,D.L., Park,S.M. and Patel,T.R. (2021) Human DDX3X unwinds Japanese encephalitis and Zika viral 5' Terminal regions. *Int. J. Mol. Sci.*, **22**, 413.
21. Men,R., Bray,M., Clark,D., Chanock,R.M. and Lai,C.J. (1996) Dengue type 4 virus mutants containing deletions in the 3' non-coding region of the RNA genome: analysis of growth restriction in cell culture and altered viremia pattern and immunogenicity in rhesus monkeys. *J. Virol.*, **70**, 3930–3937.
22. Deo,S., Patel,T.R., Chojnowski,G., Koul,A., Dzananovic,E., McEleney,K., Bujnicki,J.M. and McKenna,S.A. (2015) Characterization of the termini of the West Nile virus genome and their interactions with the small isoform of the 2' 5'-oligoadenylate synthetase family. *J. Struct. Biol.*, **190**, 236–249.
23. Thurner,C., Witwer,C., Hofacker,I.L. and Stadler,P.F. (2004) Conserved RNA secondary structures in Flaviviridae genomes. *J. Gen. Virol.*, **85**, 1113–1124.
24. Mann,M., Wright,P.R. and Backofen,R. (2017) IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions. *Nucleic Acids Res.*, **45**, W435–W439.
25. Busch,A., Richter,A.S. and Backofen,R. (2008) IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, **24**, 2849–2856.
26. Mückstein,U., Tafer,H., Hackermüller,J., Bernhart,S.H., Stadler,P.F. and Hofacker,I.L. (2006) Thermodynamics of RNA–RNA binding. *Bioinformatics*, **22**, 1177–1182.
27. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte Für Chemie / Chemical Monthly*, **125**, 167–188.
28. Bernhart,S.H., Tafer,H., Mückstein,U., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorith. Mol. Biol.*, **1**, 3.
29. Bernhart,S.H., Hofacker,I.L., Will,S., Gruber,A.R. and Stadler,P.F. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinf.*, **9**, 474.
30. Lorenz,R., Bernhart,S.H., Höner zu Siederdissen,C., Tafer,H., Flamm,C., Stadler,P.F. and Hofacker,I.L. (2011) ViennaRNA package 2.0. *Algorithm. Mol. Biol.*, **6**, 26.
31. Will,S., Reiche,K., Hofacker,I.L., Stadler,P.F. and Backofen,R. (2007) Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, **3**, e65.
32. Pam Wang,R.A., Chen,M. and Legaspi,K. (2020) *Wyatt Technologies*. pp.1–4.
33. Demeler,B. and Gorbet,G.E. (2016) In: Uchiyama,S., Arisaka,F., Stafford,W.F. and Laue,T. (eds). *Analytical Ultracentrifugation: Instrumentation, Software, and Applications*. Springer Japan, Tokyo, pp. 119–143.
34. Brookes,E., Cao,W. and Demeler,B. (2010) A two-dimensional spectrum analysis for sedimentation velocity experiments of mixtures with heterogeneity in molecular weight and shape. *Eur. Biophys. J.*, **39**, 405–414.
35. Brookes,E., Boppana,R.V. and Demeler,B. (2006) Computing Large Sparse Multivariate Optimization Problems with an Application in Biophysics. *Supercomputing '06 ACM 0-7695-2700-0/06*. IEEE, Tampa, FL.
36. Demeler,B. and van Holde,K.E. (2004) Sedimentation velocity analysis of highly heterogeneous systems. *Anal. Biochem.*, **335**, 279–288.
37. Meier,M., Moya-Torres,A., Krahn,N.J., McDougall,M.D., Orriss,G.L., McRae,E.K.S., Booy,E.P., McEleney,K., Patel,T.R., McKenna,S.A. *et al.* (2018) Structure and hydrodynamics of a DNA G-quadruplex with a cytosine bulge. *Nucleic Acids Res.*, **46**, 5319–5331.
38. Panjkovich,A. and Svergun,D.I. (2017) CHROMIXS: automatic and interactive analysis of chromatography-coupled small-angle X-ray scattering data. *Bioinformatics*, **34**, 1944–1946.
39. Manalastas-Cantos,K., Konarev,P., Hajizadeh,N., Kikhney,A., Petoukhov,M., Molodenskiy,D., Panjkovich,A., Mertens,H., Gruzinov,A., Borges,C. *et al.* (2021) ATSAS 3.0: expanded functionality and new tools for small-angle scattering data analysis. *J. Appl. Crystallogr.*, **54**, 343–355.
40. Putnam,C.D. (2016) Guinier peak analysis for visual and automated inspection of small-angle X-ray scattering data. *J. Appl. Crystallogr.*, **49**, 1412–1419.
41. Burke,J.E. and Butcher,S.E. (2012) Nucleic acid structure characterization by small angle X-ray scattering (SAXS). *Curr. Protoc. Nucleic Acid Chem.*, **Chapter 7**, Unit7.18.
42. Semenyuk,A.V. and Svergun,D.I. (1991) GNOM– a program package for small-angle scattering data processing. *J. Appl. Crystallogr.*, **24**, 537–540.
43. Svergun,D.I. (1992) Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Crystallogr.*, **25**, 495–503.
44. Svergun,D.I. (1999) Restoring low resolution structure of biological macromolecules from solution scattering using simulated annealing. *Biophys. J.*, **76**, 2879–2886.
45. Volkov,V.V. and Svergun,D.I. (2003) Uniqueness of ab initio shape determination in small-angle scattering. *J. Appl. Crystallogr.*, **36**, 860–864.
46. Ochsenreiter,R., Hofacker,I.L. and Wolfinger,M.T. (2019) Functional RNA structures in the 3′UTR of tick-borne, insect-specific and No-known-vector flaviviruses. *Viruses*, **11**, 298.
47. Michael T. Wolfinger,R.O. and Hofacker,I.L. (2021) In: Marz,M. and Frishman,D. (ed). *Virus Bioinformatics*. 1st edn., Chapman and Hall/CRC, pp. 36.
48. Nelson,C.R., Mrozowich,T., Park,S.M., D'souza,S., Henrickson,A., Vigar,J.R.J., Wieden,H.-J., Owens,R.J., Demeler,B. and Patel,T.R. (2021) Human DDX17 unwinds Rift Valley fever virus non-coding rnas. *Int. J. Mol. Sci.*, **22**, 54.
49. Mrozowich,T., Henrickson,A., Demeler,B. and Patel,T.R. (2020) Nanoscale structure determination of Murray Valley encephalitis and Powassan Virus non-coding rnas. *Viruses*, **12**, 190.
50. Patel,T.R., Winzor,D.J. and Scott,D.J. (2016) Analytical ultracentrifugation: a versatile tool for the characterisation of macromolecular complexes in solution. *Methods*, **95**, 55–61.
51. Wyatt,P.J. (1993) Light scattering and the absolute characterization of macromolecules. *Anal. Chim. Acta*, **272**, 1–40.
52. D'Souza,M.H., Mrozowich,T., Badmalia,M.D., Geeraert,M., Frederickson,A., Henrickson,A., Demeler,B., Wolfinger,M.T. and Patel,T.R. (2022) Biophysical characterisation of human LincRNA-p21 sense and antisense alu inverted repeats. *Nucleic Acids Res.*, **50**, 5881–5898.
53. Wienken,C.J., Baaske,P., Rothbauer,U., Braun,D. and Duhr,S. (2010) Protein-binding assays in biological liquids using microscale thermophoresis. *Nat. Commun.*, **1**, 100.
54. Mrozowich,T., MeierStephenson,V. and Patel,T.R. (2019) Microscale thermophoresis: warming up to a new biomolecular interaction technique. *The Biochemist*, **41**, 8–12.
55. Swain,M., Ageeli,A.A., Kasprzak,W.K., Li,M., Miller,J.T., Sztuba-Solinska,J., Schneekloth,J.S., Koirala,D., Piccirili,J., Fraboni,A.J. *et al.* (2021) Dynamic bulge nucleotides in the KSHV PAN ENE triple helix provide a unique binding platform for small molecule ligands. *Nucleic Acids Res.*, **49**, 13179–13193.
56. Jordan,B., Nickel,L. and Schmitz,R.A. (2022) In: *Prokaryotic Gene Regulation: Methods and Protocols*. Springer, pp. 291–303.
57. Turner,D.H. and Mathews,D.H. (2010) NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res.*, **38**, D280–D282.
58. Friebe,P., Shi,P.-Y. and Harris,E. (2011) The 5' and 3' downstream AUG region elements are required for mosquito-borne flavivirus RNA replication. *J. Virol.*, **85**, 1900–1905.

59. Chu,P.W.G. and Westaway,E.G. (1985) Replication strategy of Kunjin Virus: evidence for recycling role of replicative form RNA as template in semiconservative and asymmetric replication. *Virology*, **140**, 68–79.
60. Brosey,C.A. and Tainer,J.A. (2019) Evolving SAXS versatility: solution X-ray scattering for macromolecular architecture, functional landscapes, and integrative structural biology. *Curr. Opin. Struct. Biol.*, **58**, 197–213.
61. Pérez,J. and Vachette,P. (2017) In: Chaudhuri,B., Muñoz,I.G., Qian,S. and Urban,V.S. (eds). *Biological Small Angle Scattering: Techniques, Strategies and Tips*. Springer Singapore, Singapore, pp. 183–199.
62. O'Brien,D.P., Brier,S., Ladant,D., Durand,D., Chenal,A. and Vachette,P. (2018) SEC-SAXS and HDX-MS: a powerful combination. The case of the calcium-binding domain of a bacterial toxin. *Biotechnol. Appl. Biochem.*, **65**, 62–68.
63. Graewert,M.A., Da Vela,S., Gräwert,T.W., Molodenskiy,D.S., Blanchet,C.E., Svergun,D.I. and Jeffries,C.M. (2020) Adding size exclusion chromatography (SEC) and light scattering (LS) devices to obtain high-quality small angle X-ray scattering (SAXS) data. *Cryst.*, **10**, 975.
64. Putnam,C.D., Hammel,M., Hura,G.L. and Tainer,J.A. (2007) X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. *Q. Rev. Biophys.*, **40**, 191–285.
65. Grant,T.D., Luft,J.R., Carter,L.G., Matsui,T., Weiss,T.M., Martel,A. and Snell,E.H. (2015) The accurate assessment of small-angle X-ray scattering data. *Acta. Crystallogr. D Biol. Crystallogr.*, **71**, 45–56.
66. Cantara,W.A., Olson,E.D. and Musier-Forsyth,K. (2017) Analysis of RNA structure using small-angle X-ray scattering. *Methods*, **113**, 46–55.
67. Rambo,R.P. and Tainer,J.A. (2011) Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. *Biopolymers*, **95**, 559–571.
68. Kikhney,A.G. and Svergun,D.I. (2015) A practical guide to small angle X-ray scattering (SAXS) of flexible and intrinsically disordered proteins. *FEBS Lett.*, **589**, 2570–2577.
69. Kozin,M.B., Svergun,D.I. and Embl,B. (2001) Automated matching of high- and low-resolution structural models. *J. Appl. Cryst.*, **34**, 33–41.
70. Jones,C.P., Cantara,W.A., Olson,E.D. and Musier-Forsyth,K. (2014) Small-angle X-ray scattering-derived structure of the HIV-1 5' UTR reveals 3D tRNA mimicry. *Proc. Nat. Acad. Sci. USA*, **111**, 3395–3400.
71. Spokoini-Stern,R., Stamov,D., Jessel,H., Aharoni,L., Haschke,H., Giron,J., Unger,R., Segal,E., Abu-Horowitz,A. and Bachelet,I. (2020) Visualizing the structure and motion of the long non-coding RNA HOTAIR. *RNA*, **26**, 629–636.
72. Chillón,I. and Marcia,M. (2020) The molecular structure of long non-coding rnas: emerging patterns and functional implications. *Crit. Rev. Biochem. Mol. Biol.*, **55**, 662–690.
73. Svergun,D.I. and Nierhaus,K.H. (2000) A map of protein-rRNA distribution in the 70 s Escherichia coli ribosome*. *J. Biol. Chem.*, **275**, 14432–14439.
74. Wang,S., Chan,K.W.K., Tan,M.J.A., Flory,C., Luo,D., Lescar,J., Forwood,J.K. and Vasudevan,S.G. (2022) A conserved arginine in NS5 binds genomic 3' stem-loop RNA for primer-independent initiation of flavivirus RNA replication. *RNA*, **28**, 177–193.
75. Lee,E., Bujalowski,P.J., Teramoto,T., Gottipati,K., Scott,S.D., Padmanabhan,R. and Choi,K.H. (2021) Structures of flavivirus RNA promoters suggest two binding modes with NS5 polymerase. *Nat. Commun.*, **12**, 2530.
76. Fajardo,T. Jr, Sanford,T.J., Mears,H.V., Jasper,A., Storrie,S., Mansur,D.S. and Sweeney,T.R. (2020) The flavivirus polymerase NS5 regulates translation of viral genomic RNA. *Nucleic Acids Res.*, **48**, 5081–5093.

5.4 TERribly Difficult: Searching for Telomerase RNAs in Saccharomycetes

Maria Waldl, Bernhard C. Thiel, Roman Ochsenreiter, Alexander Holzenleiter, João Victor de Araujo Oliveira, Maria Emília M. T. Walter, Michael T. Wolfinger and Peter F. Stadler (2018)
'TERribly Difficult: Searching for Telomerase RNAs in Saccharomycetes',
Genes, 9(8), p. 372. doi: 10.3390/genes9080372.

Summary: Unlike e.g. their mammalian counterparts, yeast telomerase RNAs (TERs) exhibit highly diverse sequences, making them difficult to identify purely based on sequence information.

In this paper, we combined a wide range of bioinformatics methods, including structure-based and synteny-based approaches, to annotate 27 previously unidentified TERs in *Saccharomycetacea*.

An interesting aspect was the identification of a biologically functional pseudoknot and two mid-to-long range interaction within TERs using RNA-RNA interaction prediction tools. During this process, thermodynamic predictions suggested a wide range of possible interaction lengths, raising questions about the kinetic and 3D feasibility of these interactions. While structure based findings were not included in the homology search at the time of publication, this challenge provided motivation for subsequent studies presented in this thesis. Using the newly developed RRIkinDP model, the TER annotation approach presented in this paper, could now be extended to include the pseudoknot, the template boundary element and the core-enclosing helix.

Author's contributions: MW contributed to the design of the methods; coordinated large parts of the project; contributed to significant parts of the literature search, data collection, software implementation and analysis as well as the interpretation of the results; wrote relevant parts of the manuscript and contributed critical review.

genes

MDPI

Article

TERribly Difficult: Searching for Telomerase RNAs in Saccharomycetes

Maria Waldl [1,†], Bernhard C. Thiel [1,†], Roman Ochsenreiter [1], Alexander Holzenleiter [2,3], João Victor de Araujo Oliveira [4], Maria Emília M. T. Walter [4], Michael T. Wolfinger [1,5,*] and Peter F. Stadler [1,6,7,8,*]

1 Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria; maria@tbi.univie.ac.at (M.W.); thiel@tbi.univie.ac.at (B.C.T.); romanoch@tbi.univie.ac.at (R.O.)
2 BioInformatics Group, Fakultät CB Hochschule Mittweida, Technikumplatz 17, D-09648 Mittweida, Germany; alexander.holzenleiter@web.de
3 Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
4 Departamento de Ciência da Computação, Instituto de Ciências Exatas, Universidade de Brasília, Campus Universitário–Asa Norte, Brasília, DF CEP: 70910-900, Brazil; joaovicers@gmail.com (J.V.d.A.O.); mariaemilia@unb.br (M.E.M.T.W.)
5 Center for Anatomy and Cell Biology, Medical University of Vienna, Währingerstraße 13, 1090 Vienna, Austria
6 German Centre for Integrative Biodiversity Research (iDiv) Halle-Jena-Leipzig, Competence Center for Scalable Data Services and Solutions, and Leipzig Research Center for Civilization Diseases, Universität Leipzig, D-04107 Leipzig, Germany
7 Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany
8 Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA
* Correspondence: michael.wolfinger@univie.ac.at (M.T.W.); studla@bioinf.uni-leipzig.de (P.F.S.)
† These authors contributed equally to this work.

Received: 15 May 2018; Accepted: 18 July 2018; Published: 26 July 2018

Abstract: The telomerase RNA in yeasts is large, usually >1000 nt, and contains functional elements that have been extensively studied experimentally in several disparate species. Nevertheless, they are very difficult to detect by homology-based methods and so far have escaped annotation in the majority of the genomes of Saccharomycotina. This is a consequence of sequences that evolve rapidly at nucleotide level, are subject to large variations in size, and are highly plastic with respect to their secondary structures. Here, we report on a survey that was aimed at closing this gap in RNA annotation. Despite considerable efforts and the combination of a variety of different methods, it was only partially successful. While 27 new telomerase RNAs were identified, we had to restrict our efforts to the subgroup *Saccharomycetacea* because even this narrow subgroup was diverse enough to require different search models for different phylogenetic subgroups. More distant branches of the Saccharomycotina remain without annotated telomerase RNA.

Keywords: non-coding RNA; telomerase RNA; secondary structure; synteny; homology search; yeast

1. Introduction

The linear chromosomes of eukaryotes require a specialized mechanism for completing duplication. Most commonly this is achieved by a special reverse transcriptase, telomerase, that carries a specific RNA that includes the template with telomeric sequence [1]. Most likely, this constitutes the ancestral state in eukaryotes. Despite its crucial function, telomerase has been lost several times in both animals (in particular insects) and possibly also in some plants [2]. In some cases, the ancestral telomere

structure has been replaced by tandem arrays of DNA sequences that look much like heterochromatin and can be elongated by gene conversion. Specialized telomere-specific retrotransposons are at work in *Drosophila* [3].

The telomerase (holo)enzyme consists of two main components, a specialized reverse transcriptase (TERT) and a RNA component (TER) that provides the template sequence. In addition, there are usually multiple clade-specific accessory protein components [4,5]. Four conserved regions in TER (Figure 1) are essential for telomerase activity: a core-enclosing helixig (CEH), template boundary element (TBE), the template sequence and a pseudoknot, are, in this order along the RNA, part of the the catalytic core [6]. The trans activating domain is involved in binding of TERT [7]. An Area of Required Connectivity (ARC) has been identified as the fifth essential element in *Saccharomyces cerevisiae* [6]. It is to some extent conserved in human telomerase RNA [8]. The ARC is not a localized sequence-structure feature but rather constitutes a combination of CEH, TBE, and some interactions between them. Hence, it is at least difficult to utilize the ARC for homology search. The three-way junction (TWJ) structure of this region region is widely conserved at least between animal and fungal telomerase RNAs (TER) [9]. It has been shown that, despite its conservation, the TWJ is not critical for TER function in *S. cerevisiae* [6,10,11]. The precisely defined template within *TER* is processively copied by TERT and regenerated, releasing a single-stranded DNA product [12].

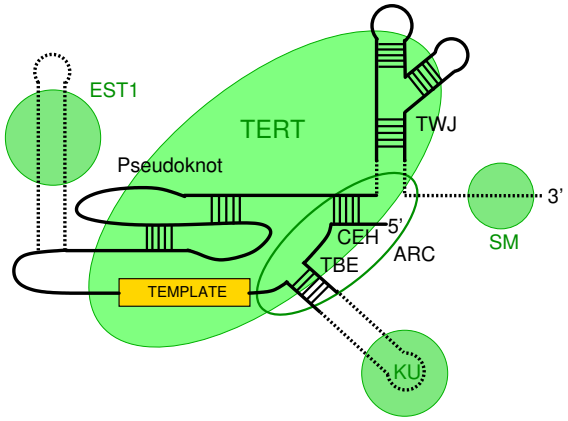


Figure 1. Schematic organization of telomerase RNA (TER). Contact regions for important binding sites are indicated by green circles (EST1, SM, and KU). The green ellipse denotes the contact region with the reverse transcriptase (TERT). Other major features are the template, the pseudoknot region, the template boundary element (TBE), the three-way junction (TWJ), and the Area of Required Connectivity (ARC). Adapted from [13].

Telomerase RNA is highly divergent. The TER in ciliates [14], human [15], and budding yeast [16–18] have a length of about 150 nt, 438 nt, and 1160 nt, respectively. A *TER* more than 2 kb in length has been reported for *Candida glabrata* [19], which, interestingly, seems to lack a TWJ. TERs in other kingdoms of eukaryotes have been discovered only quite recently in plants [20,21], excavates [22,23] and alveolates [24,25].

Despite their deeply conserved primary function and architectural similarities that seem to extend across eukaryotic kingdoms, TERs have turned out be very difficult to find by homology searches even within phylogenetically relatively narrow groups. Within the animal kingdom, even surveys of vertebrates turned out to be non-trivial [26]. Echinoderm TERs were found by deep sequencing of *Strongylocentrotus purpuratus* RNA pulled down with the TERT protein [27] after homology based searches remained unsuccessful. This opened the door to identifying TERs from other sea urchins, brittle stars, and a crinoid [28]. However, no TER from a protostome is known.

Within Fungi, the situation is similar: Thus far, TERs have been reported only for Ascomycota, while no candidates are known in Basidiomycota and any of the basal divisions. The TERs of

Pezizomycotina and Taphrinomycotina share core features of vertebrate TERs. In particular, they have a fairly well-conserved secondary structure of the pseudoknot and the TWJ, and at least in these regions the sequence is sufficiently conserved for successful homology-based identification of TERs within these clades [29–31]. The TERs known for Saccharomycetes, the relatives of budding yeast, on the other hand, are sometimes remarkably large and present little similarity in sequence and secondary structure to vertebrate or ciliate TERs.

To-date, yeast TERs have been reported for three phylogenetically narrow subgroups (*Saccharomyces* spp. [16,17], *Kluyveromyces* spp. [9,32,33], and *Candida* spp. [34,35]), as well as some individual species such as *Candida glabrata* [19] and *Hansenula polymorpha* [36]. These sequences are already too diverse for reliable sequence alignments. It is not surprising, therefore, that simple sequence-based homology searches have not been successful in identifying TER in the majority of the saccharomycete genome sequences to-date. Even protein binding sites that are functionally important in budding yeast [37,38] are not widely conserved. For instance, Ku or Sm binding sites seem to be absent in the TERs of filamentous fungi [4,29].

The obvious alternative is to increase the set of known TERs by finding homologs that are sufficiently similar to one of known yeast TERs, to allow the construction of multiple alignments of phylogenetically narrow subgroup. From these alignments, conserved elements can be extracted, which in turn form the basis for searches with tools such as `fragrep` [39] or `infernal` [40]. This strategy has been successful in previous searches for *TER* genes in both animals [26] and fungi [29], but thus far has not been successfully applied to Saccharomycetes. Here, sequences are highly divergent, so that plausible multiple sequence alignments and secondary structure models derived from them can be obtained only for quite narrow phylogenetic groups. The best studied case is *Saccharomyces* sensu stricto [17,41].

Until very recently, a phylogenetically local approach to homology search was also hampered by the lack of a trustworthy phylogeny of the Saccharomycotina. Recent updates in the International Code of Nomenclature for algae, fungi and plants [42,43] have substantially restructured the classification of fungi in general and of Saccharomycotina in particular. With large-scale efforts to sequence fungal genomes underway, first phylogenomic studies provide a trustworthy backbone of Saccharomycotina phylogeny [44], which we largely confirmed with an independent analysis.

2. Materials and Methods

2.1. Phylogenomics of Ascomycotes

Annotated protein sequences for 72 yeast species were downloaded from the National Center for Biotechnology Information `(NCBI) refseq` database (https://www.ncbi.nlm.nih.gov/refseq/). Initially, `ProteinOrtho` [45,46] was used to identify an initial set of 21,289 ortholog groups. Only 193 of these contained representatives of all 72 species. We therefore included all 1666 ortholog groups that covered at least 67 species. We used `OMA` (2.2.1, with default settings) [47,48] to decompose the `ProteinOrtho` groups further into clusters of 1-1 orthologs. This resulted in 6295 groups, of which 841 contained at least 67 species. This conservatively filtered dataset was then processed with `Gblocks` [49] (version 0.91b, with default parameters) to remove uninformative and potentially error-prone parts of the alignment, resulting in a dataset comprising 72 species and 248,581 characters. Phylogenetic trees were estimated with `RAxML` [50].

2.2. Ascomycote Telomerase RNAs

Telomerase RNA regions have been published for several *Saccharomyces* [16,17], *Kluyveromyces* [9,32,33], and *Candida* [19,34,35] species. Most of these published TER regions are collected in the telomerase database [51], which therefore provided a good starting point for our research. These sequences, however, are too diverse to construct multiple sequence alignments beyond the three genera individually. This effectively prohibits the automated discovery of novel TERs beyond close relatives

with the help of either `BLAST` [52] (using sequence information alone) or `infernal` (relying on a combination of sequence and secondary structure information).

Therefore, we explored different strategies to overcome the limitations imposed by the extremely poor sequence conservation of saccaromycete telomerase RNAs. The basic idea is to use common features of the TERs to extract candidates from the genomes that can be analyzed and then inspected further using different techniques.

First, we attempted to learn TER-specific sequence patterns using `MEME/GLAM2` [53], and also several machine learning techniques using *k*-mer distributions within sequence windows of the size of the known TERs. All attempts to learn from a training set covering the Saccharomycetaceae or all Saccharomycotina species failed.

There are several possible reasons. Machine learning methods crucially depend on training and test sets, both positive and negative. In our case, we have few positive samples, these have poorly defined features, and are very diverse as far as their sequences are concerned. It is unclear in this setting how a negative training set should be properly designed. The obvious choice of picking genomic sequence at random may be confounded by unintended strong signals, such as coding potential or repetitive sequence elements. It would appear that at the very least a more a careful construction of the positive and negative sets, and an appropriate normalization or scaling of the feature data will be required to make progress in this direction. Restricting the training phase to a more narrow phylogenetic range to reduce the inherent diversity of the training data, on the other hand, is infeasible due to the small number of known TER sequences.

The `EDeN` motif finder [54] was applied to 24 known TERs as positive set and 48 shuffled sequences as negative data. Only trivial sequence motifs such a poly-U stretch, presumably corresponding to part of the U-rich pseudoknot region, were found. Unsupervised clustering also remained unsuccessful.

2.3. Synteny-Based Homology Search

As an alternative strategy, we established a semi-automated workflow that aims at first extracting partially conserved RNA sequence-structure elements, which are then used to identify candidate loci. In response to the negative results of a direct pattern-based approach, we systematically used synteny to narrow down the search space in the initial phase. Starting from a whole genome alignment of phylogenetically related species, we used the positions of protein coding genes whose homologs are known to be adjacent in a closely related species to delimit the syntenic regions that are likely to contain a *TER* gene. These candidate regions were then analyzed in detail by means of pairwise or multiple sequence alignments. Whenever a global alignment of the entire candidate syntenic region did not yield a plausible alignment, we attempted to identify conserved motifs inside the syntenic region (usually the SM binding site and/or the template region, which is sometimes conserved between close relatives). Typically, these motifs were also sufficient to determine the correct reading direction of the TER candidate.

To identify known features in the candidate TER regions, we first constructed `infernal` [40] covariance models restricted to subgroups of Saccharomycetaceae covering only substructures, such as the Ku hairpin, Est1 binding site, and TWJ in the *Saccharomyces* and *Kluyveromyces* species. The alignments underlying the `infernal` models were constructed with the help of many software tools, including `locARNA` [55], `MAFFT` [56], `mauve` [57], `MEME` [53] and `fragrep` [39], as well as manual curation. Default settings were used for all tools. The consensus structure for the Ku binding motive corresponds to a recent crystal structure [58]. These models were then analyzed with `CMCws` [59] and used for precise localization of conserved TER elements in species that were: (a) taxonomically closely related, but not/only partially annotated in the literature (*Saccharomyces uvarum*, *Saccharomyces* sp. "*boulardii*", *Saccharomyces* sp. *M14*, and *Saccharomyces eubayanus*); or (b) phylogenetically located in the subtree spanned by the *Saccharomyces* and *Kluyveromyces* species (Figure 2). Both the `ViennaNGS` [60] suite and custom Perl/Python scripts were used for handling and conversion of genomic annotation data.

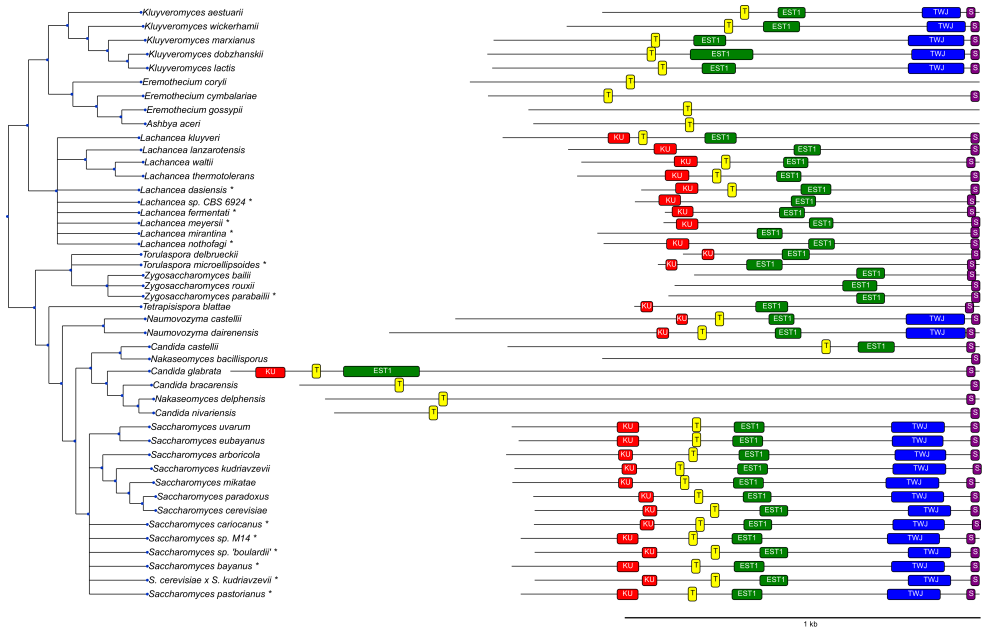


Figure 2. Features identified in TER sequences. (**KU**) ku binding hairpin, (**T**) template region, (**EST1**) Est1 binding site, (**TWJ**) three-way junction, and (**SM1**) SM1 binding site. Elements not shown are either not present in the corresponding species (e.g., the TWJ in *Candida glabrata*) or could not be located with reasonable certainty. Species marked by * are not part of the phylogenetic tree and were placed next to their closest related neighbor based on the similarity of their TER sequences.

We then extracted a sequence corresponding to the most closely related TER sequence as initial estimate of the full-length TER gene. We used `MAFFT` [56] to produce initial sequence-based alignments of candidate regions, which were then realigned with `locARNA` [55] to obtain RNA structural alignments. The latter was used with its free-end-gaps option, in particular in those cases where `MAFFT` was not sensitive enough to reliably estimate the TER boundaries. Conversely, `MAFFT` was able to identify and correctly align highly conserved subsequences, providing reliable anchors for the more divergent sequence regions. While `locARNA` is good at finding locally conserved structures in the whole alignment, we expected only parts of the TER sequences to be structurally conserved. Typically, multiple iterations of refinement of the TER boundaries were required to obtain the final TER candidate sequence.

Following this approach, we could localize TER regions for several members of the *Saccharomycetacea* clade. Subsequent alignment of candidate regions with known TERs allowed for exact localization of TERs.

2.4. Search for Candidates Using Telomere Template Sequences

The scope of the synteny-based approach is limited because fungal genomes are subject to frequent genome rearrangements at the time-scales of interest. We therefore attempted to identify candidate regions containing the template sequence for the telomere repeats (see [61] for a comprehensive review of the characteristics of different telomeric repeats). In genomes for which these sequences have not been reported, we searched chromosome ends for telomeric repeats. Unfortunately, most genome assemblies are not on chromosome level or do not include the telomere regions, hence we only succeeded to newly identify the template region of *Ashbya aceri* and *Eremothecium cymbalariae* this way. For the latter species, the pertinent information is available in [62], although the telomeric repeat is not explicitly reported. In addition, we used the published telomere sequences from the telomerase database [51].

We used the concatenation of two copies of telomeric repeat sequence as query for a `BLAST` [52] search against the whole genome (in case of longer, complex repeats) or against the syntenic region for

shorter repeats. Other template regions were identified with by aligning them to known sequences and/or BLAST searches of known template regions in closely related species. A typical feature of the template region, which helped us to verify our hits, is the fact that it usually contains a few nucleotides repeated at both the beginning and the end of the template region [19].

2.5. BLAST Pipeline

BLAST [52] is by far the most commonly used tool for homology search. While it has been reported to have limited sensitivity for telomerase RNAs in previous studies [26,27,39], it has contributed significantly to the identification of the TER sequences in other ascomycete clades [29,31]. Here, we used a set of known TER regions as BLAST queries that comprises all Saccharomycetales TER regions that we found in literature, as well as all TERs newly identified in the contribution. As targets for BLASTn (with default parameters) we used the full genomes of species that are featured at the NCBI refseq database within the Saccharomycetales group (Taxonomy ID: 4892). The resulting BLAST hits were then filtered for E-values ($E < 0.1$), a minimum alignment length of 25 nt and a minimum identity of 60%. In addition, all hits on known telomeric regions were excluded. From the hits in genomes with known TERs, we computed the empirical false positive rate and found that the alignment length proved to be the most informative parameter. It has therefore been used to evaluate the reliability of hits, given their score.

The BLAST pipeline also contributed to the identification of the TER boundaries in some of the unannotated genomes. In cases where we initially chose the boundaries of our queries too generously and included neighboring coding regions or regulatory elements, the BLAST pipeline returned "false positive" hits. Thus, whenever multiple false positive hits in the beginning or the end of the query sequence occurred, we rechecked and, if necessary, improved the boundaries of the TER region.

3. Results

3.1. Phylogenomics of Saccharomycotina

The phylogenetic trees obtained of our phylogenomic analysis of the Saccharomycetales is essentially congruent with the one reported by Shen et al. [44] (see Figure A1 for more details). For consistency, we adopted the phylogenetic tree from the same publication as the basis for presenting our results.

3.2. Survey of Telomerase RNA Genes in Saccharomycotina

We initially screened 52 ascomycote genomes. Predominantly sequence-based methods (BLAST, but also MEME, GLAM2, and infernal) only contributed TERs from close relatives of baker's yeast. The BLAST pipeline was applied to all 185 NCBI genomes Saccharomycetales, the subclade containing all known Saccharomycotina genomes. With the exception of the *TER* in *Ogataea parapolymorpha*, a very close relative of the known *Ogataea polymorpha TER* [36] all new sequences we found within the Saccharomycetaceae. We therefore restricted a more detailed analysis to this clade.

We found credible TER sequences in 46 of the 53 Saccharomycetaceae. Most of these TER sequences could be detected only after a short candidate region had been identified based on synteny. To our knowledge, at least 27 of these have not been reported previously.

3.3. Features of Telomerase RNA in Saccharomycetacea

To better understand the TER and its evolutionary constraints, at least within the Saccharomycetacea, we performed a detailed analysis of their structural features. Table 1 summarizes the results of the homology search and the functional features of the candidate *TER* genes. A graphical overview is given in Figure 2.

Table 1. Overview of conserved TER substructures in Saccharomycetacea, as identified by the combined synteny/covariance model pipeline. The 3′ end is defined as 10 nt downstream of the SM binding site. The 5′ end is approximate. Citations refer to publication in which the sequence and/or the coordinates of respective features are reported explicitly. These annotations form the basis of Figure 2.

Species	Accession	Strand	TER Coordinates	Ku Binding Site	Template Region	Est1 Binding Site	TWJ	SM1
K. aestuarii	AEAS01000245.1	neg	16,338–17,322 [32]		16,940–16,966	16,794–16,862 [32]	16,378–16,485 [9]	16,350–16,359
K. wickerhamii	AEAV01000432.1	pos	250–1327 [32]		662–693	765–858 [32]	1183–1290 [9]	1307–1316
K. marxianus	NC_036029.1	pos	506,443–507,711 [32]		506,855–506,888	506,967–507,049 [32]	507,518–50,7671 [9]	507,691–507,700
K. dobzhanskii	CCBQ010000012.1	pos	461,805–463,090 [32]		462,224–462,257	462,337–462,499 [32]	462,905–463,051 [9]	463,070–463,079
K. lactis	NC_006038.1	pos	611,456–612,727 [33]	absent [63]	611,890–611,919 [64]	612,006–612,090 [32]	612,532–612,687 [9]	612,708–612,716 [64]
E. coryli	AZAH01000001.1	neg	269,038–270,368		269,938–269,968			
E. cymbalariae	NC_016454.1	pos	54,147–54,960		54,451–54,480			
E. gossypii	NC_005782.2	neg	677,871–679,048 [65]		678,276–678,305 [35]			
Ashbya aceri	CP006020.1	neg	693,543–694,708		693,942–693,973			
L. kluyveri	CM000690.1	pos	348,600–349,844	348,876–348,930	348,957–348,982 [66]	349,129–349,208		349,825–349,833
L. lanzarotensis	NW_019212880.1	pos	854,162–855,236	854,389–854,444		854,754–854,820		855,217–855,225
L. waltii	AADM01000270.1	neg	134,961–136,000	135,698–135,756 [19]	135,613–135,636	135,409–135,470		134,973–134,981
L. thermotolerans	NC_013079.1	pos	702,500–703,549	702,730–702,791 [19]	702,853–702,876	703,022–703,083		703,530–703,538
L. dasiensis	LT598456.1	pos	682,034–682,916	682,124–682,181	682,261–682,283			682,900–682,905
L. sp. *CBS 6924*	LT598470.1	neg	441,802–442,700	442,582–442,638		442,229–442,292		441,811–441,820
L. fermentati	LT598488.1	neg	306,329–307,150	307,076–307,129		306,786–306,850		306,339–306,348
L. meyersii	LT598477.1	pos	575,851–576,676	575,886–575,941		576,233–576,294		576,657–576,666
L. mirantina	LT598468.1	pos	690,800–691,797			691,218–691,282		691,777–691,786
L. nothofagi	LT598449.1	pos	388,401–389,382	388,567–388,624		388,937–389,004		389,362–389,371
T. delbrueckii	NC_016504.1	pos	709,007–709,780	709,057–709,086		709,267–709,336		709,761–709,770
T. microellipsoides	FYBL01000005.1	neg	426,211–427,050	427,000–427,028		426,726–426,817		426,221–426,229
Z. bailii	HG316456.1	neg	712,655–713,400			712,902–712,974		712,665–712,673
Z. rouxii	NC_012990.1	pos	297,087–297,883			297,527–297,616		297,865–297,873
Z. parabailii	CP019499.1	pos	455,564–455,975 [67]			455,656–455,728		455,957–455,965
T. blattae	NC_020193.1	neg	404,150–405,050	405,003–405,033		404,650–404,733		404,165–404,173
N. castellii	NC_016499.1	pos	381,827–383,194 [64]	382,404–382,432 [19]	382,506–382,519 [64]	382,647–382,710 [64]	382,994–383,155 [64]	383,176–383,184 [64]
N. dairenensis	NC_016479.1	neg	1,519,837–1,521,377	1,520,648–1,520,678	1,520,550–1,520,562	1,520,303–1,520,369	1,519,864–1,520,027	1,519,849–1,519,857
C. castellii	CAPW01000002.1	neg	272,769–274,000		273,158–273,179	272,992–273,085		272,781–272,789
N. bacillisporus	CAPX01000073.1	pos	1230–2215					2197–2204
C. glabrata	NC_006032.2	neg	419,194–421,150 [19]	421,007–421,081 [19]	420,914–420,932 [19]	420,657–420,852 [19]		419,206–419,214 [19]
C. bracarensis	CAPU01000044.1	pos	2586–4361		2836–2854			4342–4350
N. delphensis	CAPT01000167.1	neg	254,761–256,469		256,151–256,169			254,773–254,781
C. nivariensis	CAPV01000033.1	pos	87,530–89,215		87,780–87,798			89,196–89,204
S. uvarum	NOWY01000011.1	pos	45,720–46,940	45,996–46,050	46,193–46,203	46,301–46,377	46,703–46,848	46,921–46,929
S. eubayanus	NC_030979.1	pos	476,134–47,7336	476,392–476,446	476,588–476,598	476,694–476,770	477,100–477,240	477,317–477,325
S. arboricola	NC_026172.1	pos	287,410–288,645	287,705–287,739	287,888–287,898	288,019–288,096	288,417–288,558	288,626–288,634
S. kudriavzevii	AY639012.1	pos	1–1215 [17]	284–320 [17]	424–434	585–662	981–1128 [9]	1201–1209

Table 1. *Cont.*

Species	Accession	Strand	TER Coordinates	Ku Binding Site	Template Region	Est1 Binding Site	TWJ	SM1
S. mikatae	AABZ01000048.1	neg	18,591–19,809 [17]	19,497–19,532 [17]	19,349–19,356	19,156–19,232	18,687–18,833 [9]	18,603–18,611
S. paradoxus	CP020294.1	pos	307,733–308,897 [17]	308,010–308,045 [17]	308,154–308,161	308,281–308,353	308,660–308,803 [9]	308,878–308,886
S. cerevisiae	NC_001134.8	pos	307,597–308,757 [17,18]	307,880–307,914 [68]	308,057–308,064 [64]	308,185–308,256 [32]	308,563–308,682 [9]	308,737–308,746 [64]
S. pastorianus	AZCJ01000004.1	neg	478,773–479,970 [17]	479,664–479,718 [17]	479,512–479,520	479,340–479,417	478,866–479,012 [9]	478,785–478,793
S. cer. x S. kud.	AGVY01000004.1	pos	284,183–285,344	284,465–284,501	284,645–284,655	284,772–284,843	285,150–285,269	285,325–285,333
S. bayanus	AACG02000058.1	pos	58,142–59,362 [17]	58,418–58,472 [17]	58,613–58,620	58,723–58,799	59,125–59,270 [9]	59,343–59,351
S. sp. *'boulardii'*	CM003558.1	pos	287,536–288,696	287,818–287,854	287,998–288,008	288,124–288,195	288,502–288,621	288,677–288,685
S. sp. *M14*	MVPU01000005.1	neg	473,800–474,997	474,691–474,745	474,537–474,547	474,368–474,444	473,894–474,038	473,812–473,820
S. cariocanus	AY639010.1	pos	1–1163 [17]	278–313 [17]	424–434	549–621	928–1072 [9]	1147–1155

The exact genomic positions marking the 3′ and 5′ ends of the *TER* RNA are difficult to determine without additional experimental evidence. The 5′ ends are therefore approximate. The 3′ end of the mature TER is produced by splicing in most Ascomycota [31,69,70]. This mechanism, however, was lost at some point during the evolution of the Saccharomycotina. It has been reported in the *Candida* group and for *Ogataea angusta* (previously *H. polymorpha*), but it is missing in *Saccharomyces* and *Kluyveromyces* [31]; hence, we expect that the splicing-based 3′-end processing was lost prior to the divergence of Saccharomycetacea. Indeed, no indication of a splice site was found for any of the TER sequences included in Table 1. We therefore used a position 10 nt downstream of the SM binding motif as approximation of the 3′ end in Table 1.

Several of the features listed in Table 1 have been discussed in some detail in the literature. Not all of them were found in all the candidates reported here. This may, in some cases, be explained by sequences that are too divergent to be detected. In other cases, most likely the function is not preserved. Unfortunately, many studies report neither complete sequences nor coordinates, making it effectively impossible to accurately compare their results with the annotation reported here. References are included in Table 1 if sufficient information was included to locate the features unambiguously.

No Ku binding hairpin was recovered in *Kluyveromyces* or the *Eremothecium* species. This is not unexpected since there is experimental evidence that neither the Ku binding hairpin nor its function is present in *Kluyveromyces lactis* [63]. The putative Ku binding hairpin reported for *C. glabrata* in [19] lacks experimental support and contains long insertions that made it impossible to include it in our covariance model. Furthermore, this region of the TER sequence is very poorly conserved in the closest relatives of *C. glabrata*. While the *TER* of *C. glabrata* is among the longest known members of this gene family [19], its close relative *Candida castellii* features a *TER* that has been shortened drastically in its 3′ half, with only ∼200 nt separating the EST1 and SM1 binding sites. Furthermore, the sequence `GCUA`, which is conserved in most known Ku binding sites, is not present within 600 nt upstream of the template region. The most likely explanation is that the TER of *C. castellii* (which, similar to *C. glabrata*, does not belong to the monophylogenetic genus *Candida*) (Appendix A) does not bind Ku. Of course, we cannot rule out without further experimental data that the motif has diverged beyond our ability to recognize it.

In a few species, we failed to identify the template region. In these cases (*Lachancea*, *Zygosaccharomyces* and *Torulaspora* species and *Nakaseomyces bacillisporus*), the telomeric repeat sequence is not known and seems to be very different from both the fungal consensus sequence `TTAGGG` [29] and the telomeric sequences found in closely related species.

The EST1 binding site forms a hairpin that is similar to the P3 domain of the RNase P and RNAse MRP RNAs [71]. We were not able to find an EST1 binding site in *Eremothecium* species, *Lachancea dasiensis* and in the *C. glabrata* group, even though it has been published for *C. glabrata*. While an EST1 binding site is present even in the more distantly related genus *Candida* [35], this motif is intrinsically too variable to be unambiguously recognizable in distant relatives. This pertains to both its sequence and the its base-pairing patterns.

Consistent with [19], we found no plausible secondary structure for the TWJ in *C. glabrata*, although the respective region of the sequence contains the highly conserved sequence `AATA`. It is worth noting in this context that the telomerase of the ciliate *Tetrahymena* has a stem-loop structure in place of the threeway junction [72]. TERs of the *C. glabrata* group thus may also have a functional trans-activation domain, albeit with an aberrant structure. Our TWJ covariance model, which was constructed from *Kluyveromyces* and *Saccharomyces* sequences only, also failed to detect a TWJ in *Eremothecium* and *Lachancea*. It remains an open question whether TERs of these species have a TWJ with a diverged structure that is just beyond our ability to detect, or whether trans-activation is achieved by different means. This is not implausible, given that the TWJ has been shown to be dispensable for TER function [6,10,11].

The sequence of the SM binding motif `AATTTTTGG` is perfectly conserved throughout much of the Saccharomycetaceae, with the notable exception of *K. lactis* [64] and additional small variations in other

Kluyoveromyces species (Figure 3). We could not find this motif in species of the genus *Eremothecium* and the highly related species *Ashba acerii*.

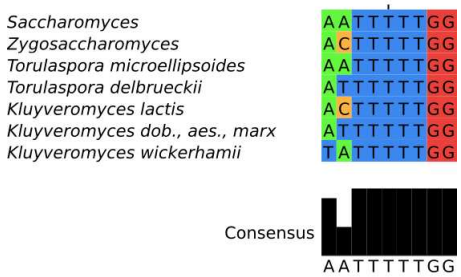


Figure 3. Alignment of the core SM-binding site motif: The common pattern of most Saccharomycetaceae (**top**); and species-specific variants (**bottom**).

4. Discussion

Although we succeeded in detecting 27 previously unknown TER sequences in Saccharomycetaceae, the main take-home message of this contribution is that homology search can be a terribly difficult problem. Although yeast TERs are quite long and fulfill a well-conserved function, their sequences are very poorly conserved. In this respect, yeast TER behaves similar to the the majority of long non-coding RNAs, which are also poorly conserved in sequence but often are evolutionary quite well conserved as functional entitied (see [73] for a recent review).

The "`BLAST` graph" in Figure 4 highlights the practical problem. Sequence comparison methods identify homology only in closely related species. A comparison of Figure 4 and a corresponding graph based on the previously published TER sequences only (see Supplementary Materials) shows that the larger set of queries identifies many additional connections and thus improves the situation at least within the Saccharomycetacea. Even within the clade, however, we have been unable to confirm the candidate hits in *Kasachstania*. The tree in Figure A1 indicates longer branch lengths leading to *Kasachstania*; it appears that the accelerated evolution of these genomes is already sufficient to hide the *TER* genes from our homology search methods.

While the direct sequence-based search against complete genomes was not very successful, we observed that the synteny-based approach worked remarkably well. This is not entirely unexpected, since the restriction to the interval between a pair of coding genes effectively reduces the size of the target from several million nucleotides to a few thousand. Unfortunately, the applicability of synteny-based methods is limited to relatively narrow phylogenetic scales. On longer time-scales, genome rearrangements are likely to disrupt syntenic conservation. A systematic exploitation of synteny similar to the work described here for Saccharomycetacea would most likely be successful in a survey for *TER* in the *Candida* group. In fact, synteny has been employed to find some of the known TERs in this clade [17,35].

The study presented here was largely conducted using publicly available tools complemented by some custom scripting. It also highlights the need for customized tools to conduct difficult homology searches. In particular, specific alignment tools and viewers to efficiently evaluate the synteny-based candidates relative to known template sequences and alignments of the better conserved regions would facilitate the manual curation efforts, which we found to be indispensable.

Finally, it remains on open question whether direct machine learning methods can be adapted as homology search tools, and, if so, whether such a strategy can be more effective than sequence comparison methods. It is likely that such efforts failed so far because of the difficulties inherent in the construction of a suitable negative training set that is not confounded by frequent genomic features such as coding sequence. Furthermore, the small number of positive samples was presumably insufficient to capture the full variability of TER sequences.

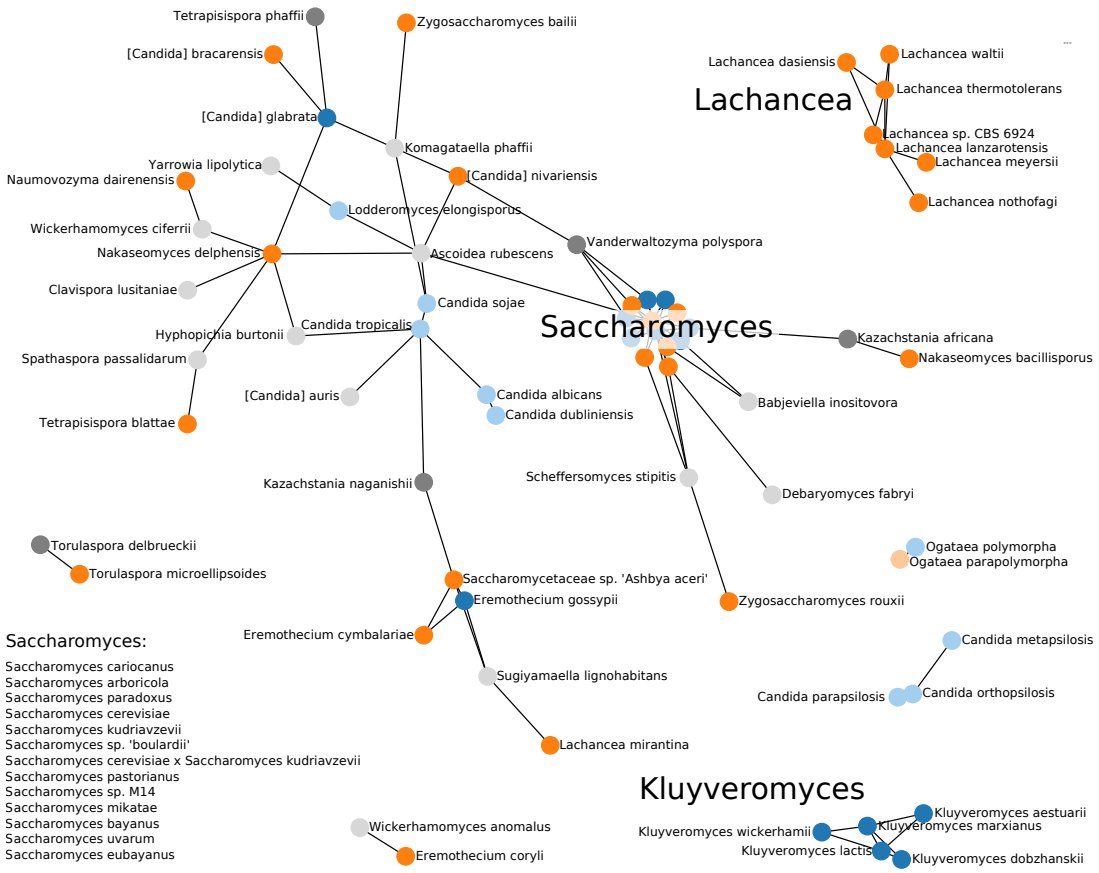


Figure 4. Summary of the BLAST-based survey of *TER* genes. Blue nodes show TERs described in the literature, orange nodes represent TERs that we identified, and grey nodes are additional candidates for which we could not validate characteristic features. TERs outside the Saccharomycetaceae group are presented in light colors. The length of the edges are weighted by the inverse of the length of the BLAST hit. Note that distances in drawing between nodes not connected by an edge are not indicative of their evolutionary distance.

Complementarily, a phylogenetically dense sample of TERs that are sufficiently similar to support global sequence alignments might help to better understand the rapid divergence of TER sequences. This might be helpful not only to identify informative features for machine learning applications, but also to design modified sequence comparison algorithms that better reflect the peculiarities of rapidly evolving long non-coding RNAs. In this contribution, we have provided such a set of TERs for the Saccharomycetaceae.

Supplementary Materials: Machine readable Supplemental Information, in particular accession numbers, TER sequences, alignments of conserved features, and covariance models are available at http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/18-048/.

Author Contributions: P.F.S. and M.T.W. conceived the study. M.W., B.C.T., R.O., and M.T.W. analyzed the TER sequences and structures. A.H. conducted the phylogenomic analysis. J.V.d.A.O. and M.E.M.T.W. contributed machine learning approaches. All authors contributed to writing the paper and approved of the submitted version.

Funding: This research was funded in part by the Doktoratskolleg RNA Biology at Univ. Vienna, the Austrian Science Fund (grant numbers SFB 4305-B09, I 2874-N28, I-1303 B21), Sinergia (grant number CRSII3_154471/1), and the German Federal Ministery for Education and Research (grant number BMBF 031A538A, de.NBI/RBC).

Acknowledgments: The research reported here is the outcome of a PhD level seminar taught by P.F.S. at Univ. Vienna in the fall term 2017/2018.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Phylogenomics of the Saccharomycetales

The maximum likelihood tree obtained from 841 orthologous groups of proteins present in at least 67 of the 72 species is shown in Figure A1. The phylogeny is nearly identical to the tree reported in [44]. In particular, it provides strong support for monophyletic Saccharomycetacea (comprising in particular the genera *Saccharomyces* and *Kluyveromyces*), and the *Candida* group. Noteworthy, "*Candida glabrata*" is nested within the Saccharomycetacea as a close relative of *Saccharomyces* rather than appearing as member of the *Candida* clade.

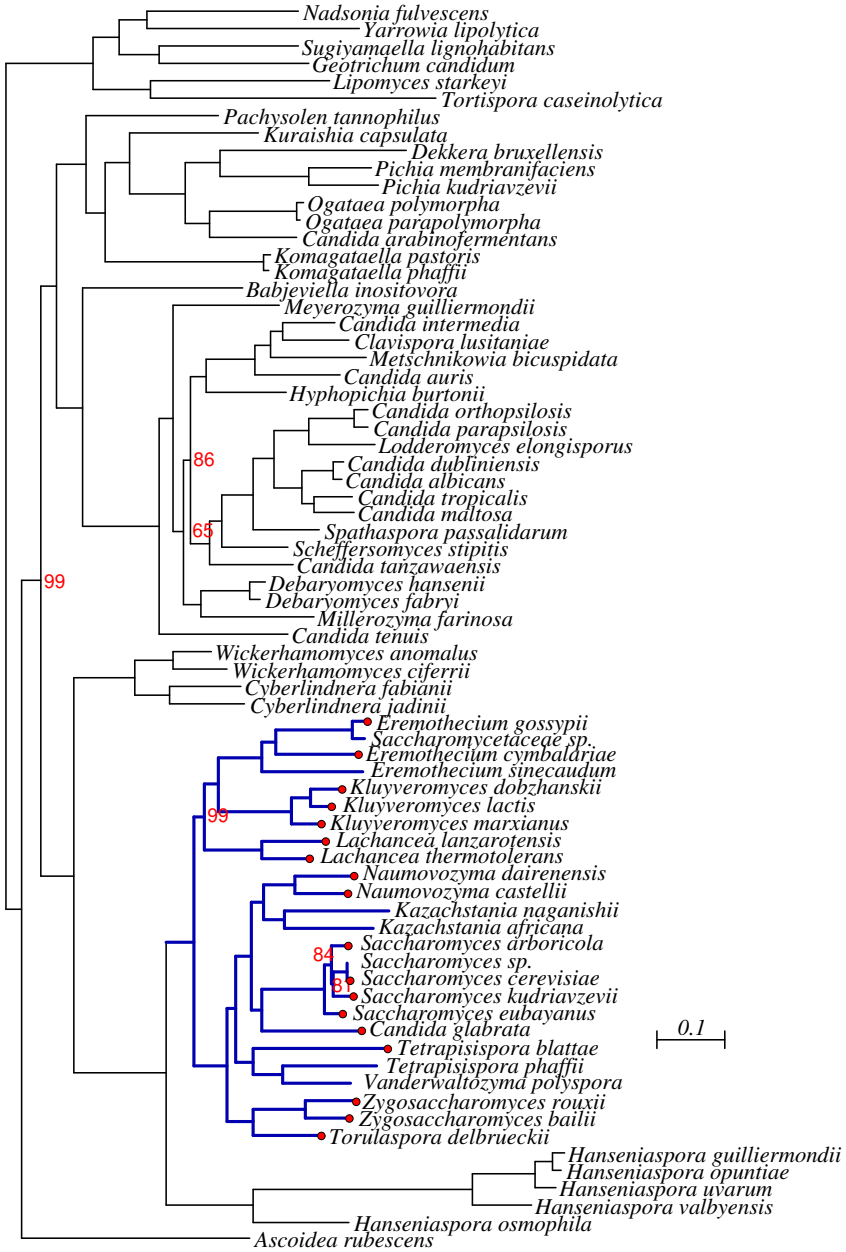


Figure A1. Phylogeny of the Saccharomycetales. Bootstrap support is 100% unless otherwise indicated. The Saccharomycetacea are indicted in dark blue. A red dot at tip of the tree indicates a TER sequences listed in Table 1.

References

1. Greider, C.W.; Blackburn, E.H. Identification of a specific telomere terminal transferase activity in *Tetrahymena* extracts. *Cell* **1985**, *43*, 405–413. [CrossRef]
2. Mason, J.M.; Reddy, H.M.; Frydrychova, R.C. Telomere maintenance in organisms without telomerase. In *DNA Replication—Current Advances*; Seligmann, H., Ed.; InTech: Rijeka, Croatia, 2011; Chapter 15.
3. Pardue, M.; Rashkova, S.; Casacuberta, E.; DeBaryshe, P.G.; George, J.A.; Traverse, K. Two retrotransposons maintain telomeres in Drosophila. *Chromosome Res.* **2005**, *13*, 443–453. [CrossRef] [PubMed]
4. Podlevsky, J.D.; Chen, J. Evolutionary perspectives of telomerase RNA structure and function. *RNA Biol.* **2016**, *13*, 720–732. [CrossRef] [PubMed]
5. Musgrove, C.; Jansson, L.I.; Stone, M.D. New perspectives on telomerase RNA structure and function. *Wiley Interdiscip. Rev. RNA* **2018**, *9*, e1456. [CrossRef] [PubMed]
6. Mefford, M.A.; Rafiq, Q.; Zappulla, D.C. RNA connectivity requirements between conserved elements in the core of the yeast telomerase RNP. *EMBO J.* **2013**, *32*, 2980–2993. [CrossRef] [PubMed]
7. Chen, J.; Greider, C.W. An emerging consensus for telomerase RNA structure. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 14683–14684. [CrossRef] [PubMed]
8. Mefford, M.A.; Zappulla, D.C. Physical connectivity mapping by circular permutation of human telomerase RNA reveals new regions critical for activity and processivity. *Mol. Cell. Biol.* **2015**. [CrossRef] [PubMed]
9. Brown, Y.; Abraham, M.; Pearl, S.; Kabaha, M.M.; Elboher, E.; Tsfati, Y. A critical three-way junction is conserved in budding yeast and vertebrate telomerase RNAs. *Nucleic Acids Res.* **2007**, *35*, 6280–6289. [CrossRef] [PubMed]
10. Livengood, A.J.; Zaug, A.J.; Cech, T.R. Essential regions of *Saccharomyces cerevisiae* telomerase RNA: Separate elements for Est1p and Est2p interaction. *Mol. Cell. Biol.* **2002**, *22*, 2366–2374. [CrossRef] [PubMed]
11. Zappulla, D.C.; Goodrich, K.; Cech, T.R. A miniature yeast telomerase RNA functions in vivo and reconstitutes activity in vitro. *Nat. Struct. Mol. Biol.* **2005**, *12*, 1072. [CrossRef] [PubMed]
12. Wu, R.A.; Upton, H.E.; Vogan, J.M.; Collins, K. Telomerase mechanism of telomere synthesis. *Annu. Rev. Biochem.* **2017**, *86*, 439–460. [CrossRef] [PubMed]
13. Webb, C.J.; Zakian, V.A. Telomerase RNA is more than a DNA template. *RNA Biol.* **2016**, *13*, 683–689. [CrossRef] [PubMed]
14. Greider, C.W.; Blackburn, E.H. A telomeric sequence in the RNA of *Tetrahymena* telomerase required for telomere repeat synthesis. *Nature* **1989**, *337*, 331–337. [CrossRef] [PubMed]
15. Feng, J.; Funk, W.D.; Wang, S.S.; Weinrich, S.L.; Avilion, A.A.; Chiu, C.P.; Adams, R.R.; Chang, E.; Allsopp, R.C.; Yu, J. The RNA component of human telomerase. *Science* **1995**, *269*, 1236–1241. [CrossRef] [PubMed]
16. Singer, M.S.; Gottschling, D.E. TLC1: Template RNA component of *Saccharomyces cerevisiae* telomerase. *Science* **1994**, *266*, 404–409. [CrossRef] [PubMed]
17. Dandjinou, A.T.; Lévesque, N.; Larose, S.; Lucier, J.F.; Abou Elela, S.; Wellinger, R.J. A phylogenetically based secondary structure for the yeast telomerase RNA. *Curr. Biol.* **2004**, *14*, 1148–1158. [CrossRef] [PubMed]
18. Bosoy, D.; Peng, Y.P.; Mian, I.S.; Lue, N.F. Conserved N-terminal motifs of telomerase reverse transcriptase required for ribonucleoprotein assembly in vivo. *J. Biol. Chem.* **2003**, *278*, 3882–3890. [CrossRef] [PubMed]
19. Kachouri-Lafond, R.; Dujon, B.; Gilson, E.; Westhof, E.; Fairhead, C.; Teixeira, M.T. Large telomerase RNA, telomere length heterogeneity and escape from senescence in *Candida glabrata*. *FEBS Lett.* **2009**, *583*, 3605–3610. [CrossRef] [PubMed]
20. Cifuentes-Rojas, C.; Kannan, K.; Tseng, L.; Shippen, D.E. Two RNA subunits and POT1a are components of *Arabidopsis telomerase*. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 73–78. [CrossRef] [PubMed]
21. Beilstein, M.A.; Brinegar, A.E.; Shippen, D.E. Evolution of the *Arabidopsis* telomerase RNA. *Front. Genet.* **2012**, *3*, 188. [CrossRef] [PubMed]
22. Gupta, S.K.; Kolet, L.; Doniger, T.; Biswas, V.K.; Unger, R.; Tzfati, Y.; Michaeli, S. The *Trypanosoma brucei* telomerase RNA (TER) homologue binds core proteins of the C/D snoRNA family. *FEBS Lett.* **2013**, *587*, 1399–1404. [CrossRef] [PubMed]
23. Sandhu, R.; Sanford, S.; Basu, S.; Park, M.; Pandya, U.M.; Li, B.; Chakrabarti, K. A trans-spliced telomerase RNA dictates telomere synthesis in *Trypanosoma brucei*. *Cell Res.* **2013**, *23*, 537–551. [CrossRef] [PubMed]

24. Chakrabarti, K.; Pearson, M.; Grate, L.; Sterne-Weiler, T.; Deans, J.; Donohue, J.P.; Ares, M., Jr. Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis. *RNA* **2007**, *13*, 1923–1939. [CrossRef] [PubMed]
25. Religa, A.A.; Ramesar, J.; Janse, C.J.; Scherf, A.; Waters, A.P. *P. berghei* telomerase subunit TERT is essential for parasite survival. *PLoS ONE* **2014**, *9*, e108930. [CrossRef] [PubMed]
26. Xie, M.; Mosig, A.; Qi, X.; Li, Y.; Stadler, P.F.; Chen, J.J.L. Size variation and structural conservation of vertebrate telomerase RNA. *J. Biol. Chem.* **2008**, *283*, 2049–2059. [CrossRef] [PubMed]
27. Li, Y.; Marz, M.; Qi, X.; Podlevsky, J.D.; Hoffmann, S.; Stadler, P.F.; Chen, J.J.L. Identification of Purple Sea urchin telomerase RNA using a next-generation sequencing based approach. *RNA* **2013**, *19*, 852–860. [CrossRef] [PubMed]
28. Podlevsky, J.D.; Li, Y.; Chen, J.J. Structure and function of echinoderm telomerase RNA. *RNA* **2016**, *22*, 204–215. [CrossRef] [PubMed]
29. Qi, X.; Li, Y.; Honda, S.; Hoffmann, S.; Marz, M.; Mosig, A.; Podlevskya, J.D.; Stadler, P.F.; Selker, E.U.; Chen, J.J.L. The common ancestral core of vertebrate and fungal telomerase RNAs. *Nucleic Acids Res.* **2013**, *41*, 450–462. [CrossRef] [PubMed]
30. Kuprys, P.V.; Davis, S.M.; Hauer, T.M.; Meltser, M.; Tzfati, Y.; Kirk, K.E. Identification of telomerase RNAs from filamentous fungi reveals conservation with vertebrates and yeasts. *PLoS ONE* **2013**, *8*, e58661. [CrossRef] [PubMed]
31. Qi, X.; Rand, D.P.; Podlevsky, J.D.; Li, Y.; Mosig, A.; Stadler, P.F.; Chen, J.J. Prevalent and distinct spliceosomal 3′-end processing mechanisms for fungal telomerase RNA. *Nat. Commun.* **2015**, *6*, 6105. [CrossRef] [PubMed]
32. Seto, A.G.; Livengood, A.J.; Tzfati, Y.; Blackburn, E.H.; Cech, T.R. A bulged stem tethers Est1p to telomerase RNA in budding yeast. *Genes Dev.* **2002**, *16*, 2800–2812. [CrossRef] [PubMed]
33. McEachern, M.J.; Blackburn, E.H. Runaway telomere elongation caused by telomerase RNA gene mutations. *Nature* **1995**, *376*, 403–409. [CrossRef] [PubMed]
34. Hsu, M.; McEachern, M.J.; Dandjinou, A.T.; Tzfati, Y.; Orr, E.; Blackburn, E.H.; Lue, N.F. Telomerase core components protect *Candida* telomeres from aberrant overhang accumulation. *Proc. Natl. Acad. Sci. USA* **2007**, *104*, 11682–11687. [CrossRef] [PubMed]
35. Gunisova, S.; Elboher, E.; Nosek, J.; Gorkovoy, V.; Brown, Y.; Jean-François, L.; Laterreur, N.; Wellinger, R.J.; Tzfati, Y.; Tomaska, L. Identification and comparative analysis of telomerase RNAs from *Candida* species reveal conservation of functional elements. *RNA* **2009**, *15*, 546–559. [CrossRef] [PubMed]
36. Smekalova, E.M.; Malyavko, A.N.; Zvereva, M.I.; Mardanov, A.V.; Ravin, N.V.; Skryabin, K.G.; Westhof, E.; Dontsova, O.A. Specific features of telomerase RNA from *Hansenula polymorpha*. *RNA* **2013**, *19*, 1563–1574. [CrossRef] [PubMed]
37. Stellwagen, A.E. Ku interacts with telomerase RNA to promote telomere addition at native and broken chromosome ends. *Genes Dev.* **2003**, *17*, 2384–2395. [CrossRef] [PubMed]
38. Zappulla, D.C.; Goodrich, K.J.; Arthur, J.R.; Gurski, L.A.; Denham, E.M.; Stellwagen, A.E.; Cech, T.R. Ku can contribute to telomere lengthening in yeast at multiple positions in the telomerase RNP. *RNA* **2011**, *17*, 298–311. [CrossRef] [PubMed]
39. Mosig, A.; Chen, J.L.; Stadler, P.F. Homology Search with Fragmented Nucleic Acid Sequence Patterns. In *Lecture Notes in Computer Science, Proceedings of the International Workshop on Algorithms in Bioinformatics (WABI 2007), Philadelphia, PA, USA, 8–9 September 2007*; Giancarlo, R., Hannenhalli, S., Eds.; Springer: Berlin/Heidelberg, Germany, 2007; Volume 4645, pp. 335–345.
40. Nawrocki, E.P.; Eddy, S.R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **2013**, *29*, 2933–2935. [CrossRef] [PubMed]
41. Zappulla, D.C.; Cech, T.R. Yeast telomerase RNA: A flexible scaffold for protein subunits. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 10024–10029. [CrossRef] [PubMed]
42. Hawksworth, D.L. A new dawn for the naming of fungi: Impacts of decisions made in Melbourne in July 2011 on the future publication and regulation of fungal names. *IMA Fungus* **2011**, *2*, 155–162. [CrossRef] [PubMed]
43. McNeill, J.; Barrie, F.R.; Buck, W.R.; Demoulin, V.; Greuter, W.; Hawksworth, D.L.; Herendeen, P.S.; Knapp, S.; Marhold, K.; Prado, J.; et al. (Eds.) International Code of Nomenclature for algae, fungi and plants. In *Regnum Vegetabile*; Koeltz Scientific Books: Oberreifenberg, Germany, 2012; Volume 154.

44. Shen, X.X.; Zhou, X.; Kominek, J.; Kurtzman, C.P.; Hittinger, C.T.; Rokas, A. Reconstructing the backbone of the Saccharomycotina yeast phylogeny using genome-scale data. *G3 (Bethesda)* **2016**, *6*, 3927–3939. [CrossRef] [PubMed]
45. Lechner, M.; Findeiß, S.; Steiner, L.; Marz, M.; Stadler, P.F.; Prohaska, S.J. Proteinortho: Detection of (co-)orthologs in large-scale analysis. *BMC Bioinform.* **2011**, *12*, 124. [CrossRef] [PubMed]
46. Lechner, M.; Hernandez-Rosales, M.; Doerr, D.; Wieseke, N.; Thévenin, A.; Stoye, J.; Hartmann, R.K.; Prohaska, S.J.; Stadler, P.F. Orthology detection combining clustering and synteny for very large datasets. *PLoS ONE* **2014**, *9*, e105015. [CrossRef] [PubMed]
47. Roth, A.C.J.; Gonnet, G.H.; Dessimoz, C. Algorithm of OMA for large-scale orthology inference. *BMC Bioinform.* **2008**, *9*, 518. [CrossRef] [PubMed]
48. Altenhoff, A.M.; Gil, M.; Gonnet, G.H.; Dessimoz, C. Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS ONE* **2013**, *8*, e53786. [CrossRef] [PubMed]
49. Talavera, G.; Castresana, J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **2007**, *56*, 564–577. [CrossRef] [PubMed]
50. Stamatakis, A. RAxML Version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **2014**, *30*, 1312–1313. [CrossRef] [PubMed]
51. Podlevsky, J.D.; Bley, C.J.; Omana, R.V.; Qi, X.; Chen, J.J.L. The telomerase database. *Nucleic Acids Res.* **2007**, *36*, D339–D343. [CrossRef] [PubMed]
52. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410. [CrossRef]
53. Bailey, T.L.; Boden, M.; Buske, F.A.; Frith, M.; Grant, C.E.; Clementi, L.; Ren, J.; Li, W.W.; Noble, W.S. MEME SUITE: Tools for motif discovery and searching. *Nucleic Acids Res.* **2009**, *37*, W202–W208. [CrossRef] [PubMed]
54. Costa, F.C.; De Grave, K. Fast neighborhood subgraph pairwise distance Kernel. In Proceedings of the 27th International Conference on International Conference on Machine Learning (ICML'10), Haifa, Israel, 21–24 June 2010; Fürnkranz, J., Joachims, T., Eds.; Omnipress: Madison, WI, USA, 2010; pp. 255–262.
55. Will, S.; Reiche, K.; Hofacker, I.L.; Stadler, P.F.; Backofen, R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comp. Biol.* **2007**, *3*, e65. [CrossRef] [PubMed]
56. Katoh, K.; Standley, D.M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **2013**, *30*, 772–780. [CrossRef] [PubMed]
57. Darling, A.E.; Mau, B.; Perna, N.T. progressiveMauve: Multiple genome alignment with gene gain, loss and rearrangement. *PLoS ONE* **2010**, *5*, e11147. [CrossRef] [PubMed]
58. Chen, H.; Xue, J.; Churikov, D.; Hass, E.P.; Shi, S.; Lemon, L.D.; Luciano, P.; Bertuch, A.A.; Zappulla, D.C.; Géli, V.; et al. Structural insights into yeast telomerase recruitment to telomeres. *Cell* **2018**, *172*, 331–343. [CrossRef] [PubMed]
59. Eggenhofer, F.; Hofacker, I.L.; Höner zu Siederdissen, C. CMCompare webserver: Comparing RNA families via covariance models. *Nucleic Acids Res.* **2012**, *41*, W499–W503. [CrossRef] [PubMed]
60. Wolfinger, M.T.; Fallmann, J.; Eggenhofer, F.; Amman, F. ViennaNGS: A toolbox for building efficient next-generation sequencing analysis pipelines. *F1000Research* **2015**, *4*, 50. [PubMed]
61. Teixeira, M.T.; Gilson, E. Telomere maintenance, function and evolution: The yeast paradigm. *Chromosome Res.* **2005**, *13*, 535–548. [CrossRef] [PubMed]
62. Wendland, J.; Walther, A. Genome evolution in the *Eremothecium* clade of the *Saccharomyces* complex revealed by comparative genomics. *Genes Genomes Genet.* **2011**, *1*, 539–548. [CrossRef] [PubMed]
63. Kabaha, M.M.; Zhitomirsky, B.; Schwartz, I.; Tzfati, Y. The 5′ arm of *Kluyveromyces lactis* telomerase RNA is critical for telomerase function. *Mol. Cell. Biol.* **2008**, *28*, 1875–1882. [CrossRef] [PubMed]
64. Telomerase Database-Secondary Structures. Available online: http://telomerase.asu.edu/structures.html#secondary (accessed on 30 April 2018).
65. Dietrich, F.S. The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science* **2004**, *304*, 304–307. [CrossRef] [PubMed]
66. Genome Resources for Yeast Chromosomes Database-TLC1 (Telomerase RNA Template). Available online: http://gryc.inra.fr/index.php?page=locus&seqid=SAKL0D04356r (accessed on 30 April 2018).

67. Ortiz-Merino, R.A.; Kuanyshev, N.; Braun-Galleani, S.; Byrne, K.P.; Porro, D.; Branduardi, P.; Wolfe, K.H. Evolutionary restoration of fertility in an interspecies hybrid yeast, by whole-genome duplication after a failed mating-type switch. *PLoS Biol.* **2017**, *15*, e2002128. [CrossRef] [PubMed]
68. Peterson, S.E.; Stellwagen, A.E.; Diede, S.J.; Singer, M.S.; Haimberger, Z.W.; Johnson, C.O.; Tzoneva, M.; Gottschling, D.E. The function of a stem-loop in telomerase RNA is linked to the DNA repair protein Ku. *Nat. Genet.* **2001**, *27*, 64. [CrossRef] [PubMed]
69. Box, J.A.; Bunch, J.T.; Tang, W.; Baumann, P. Spliceosomal cleavage generates the 3′ end of telomerase RNA. *Nature* **2008**, *456*, 910–914. [CrossRef] [PubMed]
70. Kannan, R.; Helston, R.M.; Dannebaum, R.O.; Baumann, P. Diverse mechanisms for spliceosome-mediated 3′ end processing of telomerase RNA. *Nat. Commun.* **2015**, *6*, 6104. [CrossRef] [PubMed]
71. Lemieux, B.; Laterreur, N.; Perederina, A.; Noël, J.F.; Dubois, M.L.; Krasilnikov, A.S.; Wellinger, R.J. Active yeast telomerase shares subunits with ribonucleoproteins RNase P and RNase MRP. *Cell* **2016**, *165*, 1171–1181. [CrossRef] [PubMed]
72. Singh, M.; Wang, Z.; Koo, B.K.; Patel, A.; Cascio, D.; Collins, K.; Feigon, J. Structural basis for telomerase RNA recognition and RNP assembly by the holoenzyme La family protein p65. *Mol. Cell* **2012**, *47*, 16–26. [CrossRef] [PubMed]
73. Nitsche, A.; Stadler, P.F. Evolutionary clues in lncRNAs. *Wiley Interdiscip. Rev. RNA* **2017**, *8*, e1376. [CrossRef] [PubMed]

5.5 3D feasibility of 2D RNA–RNA interaction paths by stepwise folding simulations

Irene K. Beckmann, Maria Waldl, Sebastian Will and Ivo L. Hofacker (2024)
'3D feasibility of 2D RNA–RNA interaction paths by stepwise folding simulations'
RNA, 30(2), pp. 113–123. doi:10.1261/rna.079756.123.

Summary: A functional RNA-RNA interaction must not only be thermodynamically and kinetically favorable in two dimensions, but must also be feasible in the three-dimensional space that the molecules inhabit. These spatial constraints apply not only to the final structure, but to every step of the folding process. Current approaches to 3D structure prediction are limited in the size of RNAs that can be feasibly analyzed. Even for single RNAs, exhaustive exploration of the folding process in the 3D structure space is computationally infeasible.

This study presents a computational pipeline designed to assess whether each step in a proposed 2D RNA interaction formation trajectory is viable in a coarse-grained 3D model. Simulating the folding process by extending interactions incrementally by one or two base pairs at each step, this approach allows the identification of sterically unfavorable intermediates and hindrances that may impede the formation of longer interactions.

For example, the analysis shows that interactions appear to reach a length of approximately 6 base pairs quickly. Further extension appear to require longer timeframes and extensive refolding,suggesting inherent steric limitations. In addition, the context of an interaction may limit its length, for example if it is situated in a kissing hairpin.

By integrating steric feasibility assessments into 2D RNA–RNA interaction predictions, this approach enables efficient filtering of interaction candidates, reducing false positives and improving the reliability of computational models.

The computational pipeline is available as open-source implementation at `www.github.com/irenekb/RRI-3D`, allowing researchers to apply and adapt the tool for various RNA interaction analyses.

Author's contributions: MW contributed to the ideas and research questions and developed the methodology together with the co-authors; provided supervision of the projects; contributed data and literature on RNA-RNA interactions that were analysed; analysed 2D interaction energy landscapes to provide plausible 2D trajectories as input for 3D simulations; contributed to the interpretation of the results and provided feedback on the software implementation; contributed substantive revisions and critical review on the manuscript.

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.

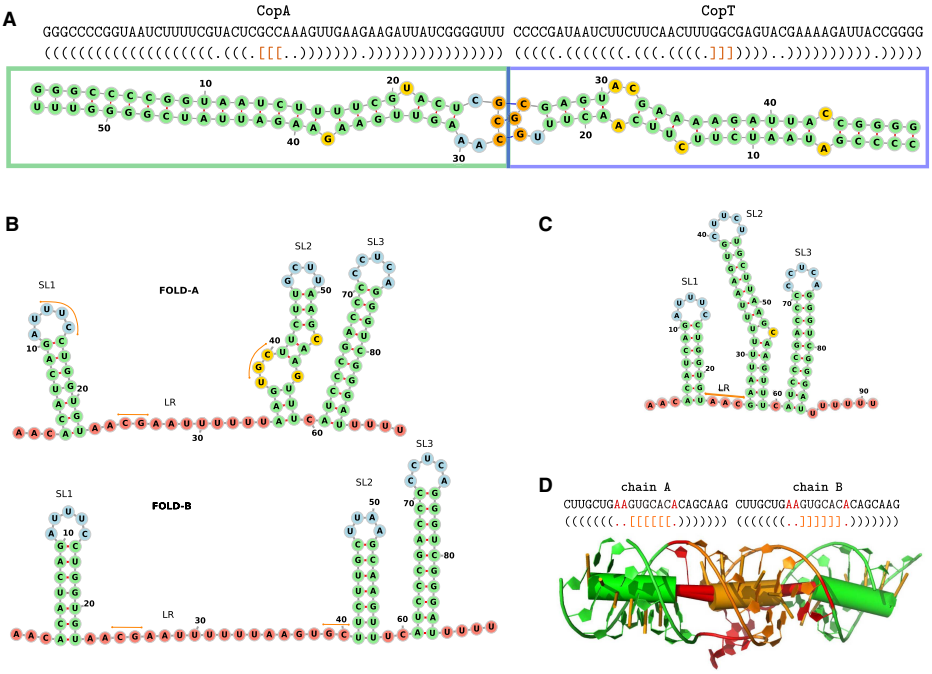


FIGURE 1. (*A*) Sequence and secondary structure representation of the 54 nt part of CopA (green box) and the corresponding 47 nt CopT (blue box) counterpart used in our simulation study. The three central residues of the hairpins (CCG/GGC) are the initial interaction site (marked in orange). (*B*) Fold-A 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

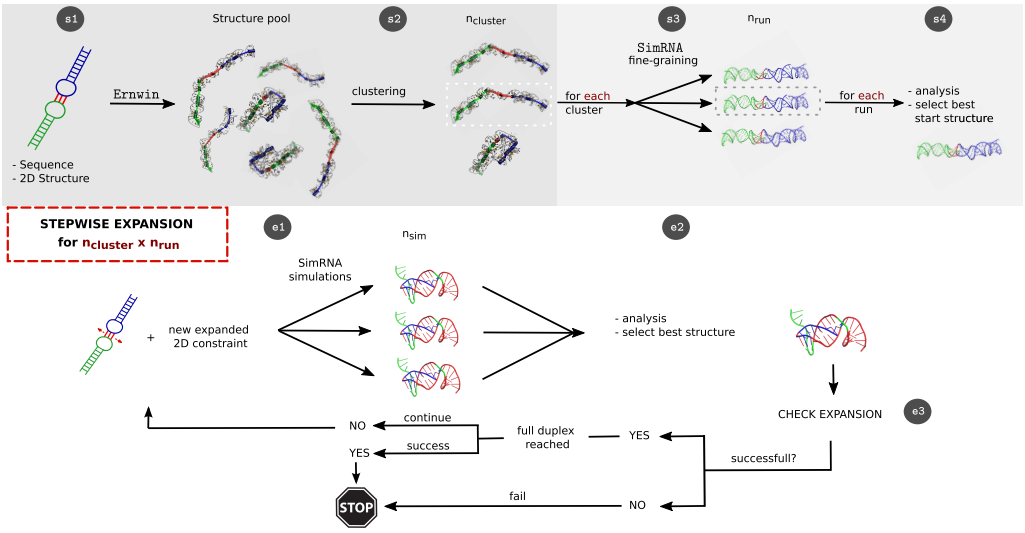


FIGURE 2. Graphical representation of the developed pipeline: An initial interaction is extended until the full duplex, a predefined target interaction, is reached or an extension of the interaction side is no longer possible due to steric and kinetic effects. (*s1*) Starting from a sequence and the 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.

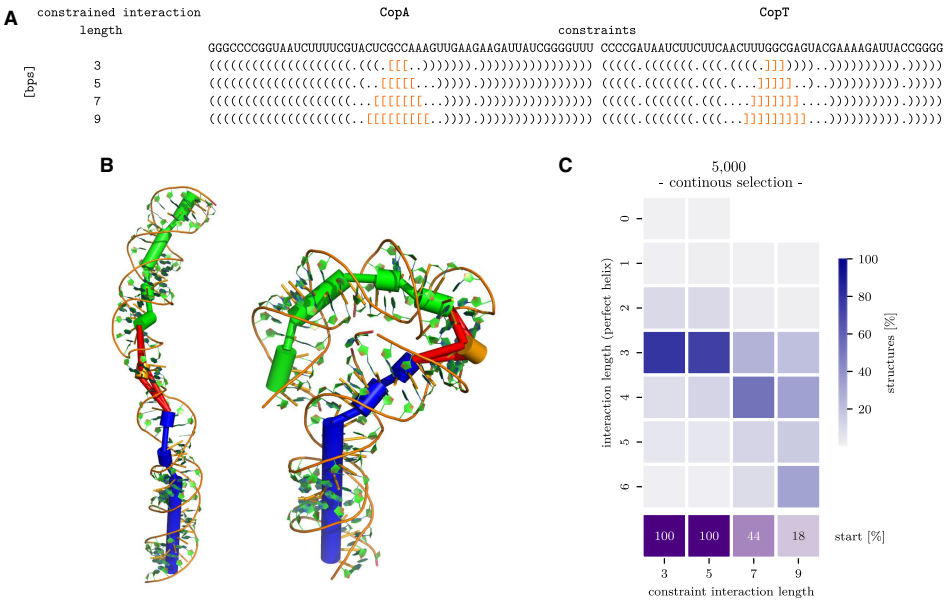


FIGURE 3. (*A*) CopA–CopT sequence and the dot-bracket representation (interaction in orange) of the constrained 2D path (full 2D path in Supplemental Fig. S1). (*B*) Two representative starting clusters with very different 3D conformations. CopA is shown in green, CopT in blue, and the interaction region 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.

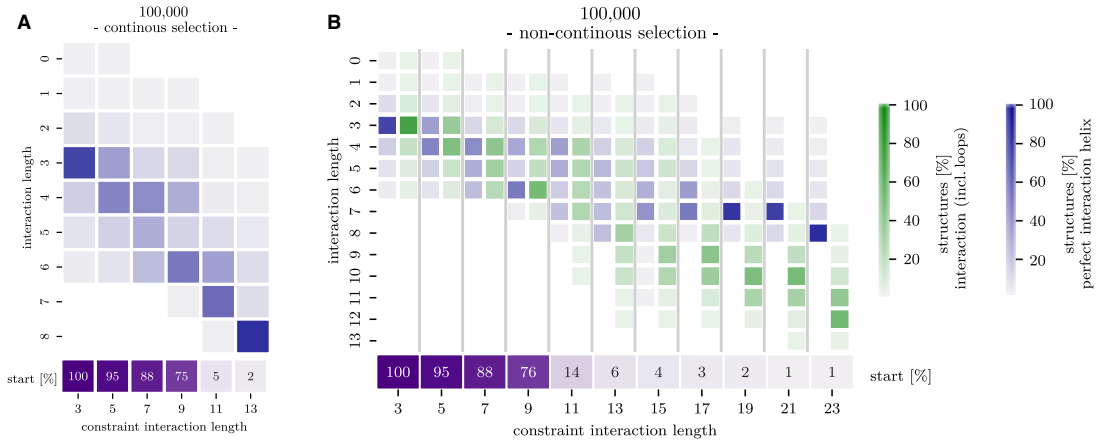


FIGURE 4. Distribution of CopA–CopT interaction lengths for two different pipeline settings and long SimRNA runs (n_{step} = 100,000); see Figure 3 for description. (*A*) Setting I, extensions form perfect helices. (*B*) Setting II, with loops in the interaction region allowed. Histograms are shown for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

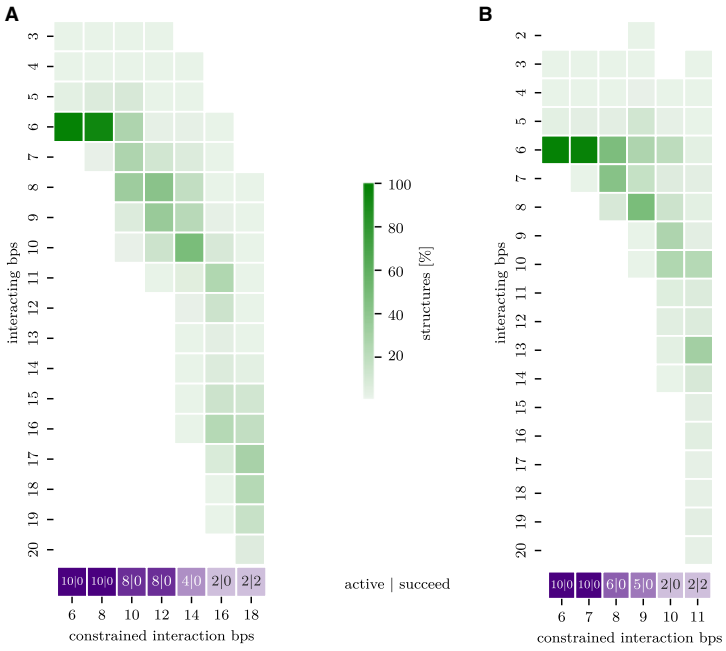


FIGURE 5. (*A*) Setting I. Symmetric interaction extension of the HIV-1 DIS interaction site with $n_{step} = 1$ million. Each column shows the histogram (green) of interaction lengths including loops (*y*-axis) at the corresponding constrained interaction length in bp (*x*-axis). For every elongation 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

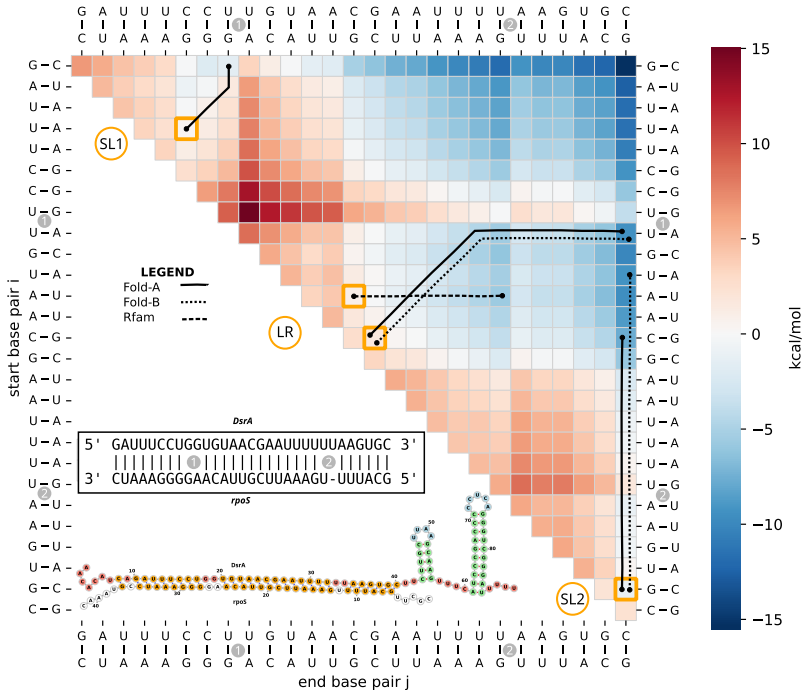


FIGURE 6. Interaction of DsrA–*rpoS* in *E. coli* as a direct path energy landscape following RRIkinDP (Waldl et al. 2023) and comparison of different formation scenarios. RRIkinDP, a novel 2D structure-based tool for studying RNA–RNA interaction formation on direct paths, was used to generate the energy 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 ($n_{\mathrm{cluster}} \times n_{\mathrm{run}} \times n_{\mathrm{sim}}$) of independent simulations in order to sample diverse starting conformations and alternative pathways. As an example, a typical CopA–CopT simulation with $n_{\mathrm{cluster}} = 10$, $n_{\mathrm{run}} = 5$, $n_{\mathrm{sim}} = 5$, $n_{\mathrm{step}} = 10{,}000$ 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***

5.6 Sequence design for RNA-RNA interactions

Maria Waldl, Hua-Ting Yao and Ivo Hofacker (2025)
'Sequence Design for RNA-RNA Interactions',
in A. Churkin and D. Barash (eds) RNA Design: Methods and Protocols.
New York, NY: Springer US, pp. 1–16. doi: 10.1007/978-1-0716-4079-1_1.

Summary: Designing RNA sequences that can specifically and efficiently interact with target RNAs is an essential aspect in synthetic biology and RNA-based therapeutics. This book chapter presents an approach for designing RNA sequences that form desired RNA-RNA interactions with favorable kinetic properties in a tutorial format.

A key contribution of this work is demonstrating how the RRIkinDP model can be incorporated into the existing Infrared RNA design framework. In this approach, RRIkinDP is used as part of a scoring function that is evaluated at each iteration of the optimization process. The computational efficiency of RRIkinDP enables the practical use of kinetic features in RNA interaction design.

Beyond sequence design, this approach enables experimental validation of the RRIkinDP kinetic model. By generating RNA sequences with precisely defined thermodynamic and kinetic properties, it allows direct testing of whether designed interactions occur as predicted. Comparing these sequences to those designed purely by thermodynamic criteria helps assess the model's predictive power. Additionally, using designed rather than natural sequences minimizes unknown biases, providing a clearer evaluation of kinetic effects on RNA-RNA interactions.

Author's contributions: MW conceptualized the idea of showcasing how RRIkinDP enabels the design of RNA-RNA interactions with sepcific kinetic properties by integrating it with the Infrared framework. MW led the design of the approach and tutorial. All three authors contributed to the development and review of the accompanying code, as well as the drafting, revision, and finalization of the manuscript.

Chapter 1

Sequence Design for RNA-RNA Interactions

Maria Waldl, Hua-Ting Yao, and Ivo L. Hofacker

Abstract

The design of RNA sequences with desired structural properties presents a challenging computational problem with promising applications in biotechnology and biomedicine. Most regulatory RNAs function by forming RNA-RNA interactions, e.g., in order to regulate mRNA expression. It is therefore natural to consider problems where a sequence is designed to form a desired RNA-RNA interaction and switch between structures upon binding. This contribution demonstrates the use of the `Infrared` framework to design interacting sequences. Specifically, we consider the regulation of the rpoS mRNA by the sRNA DsrA and design artificial 5′ UTRs that place a downstream protein coding gene under control of DsrA. The design process is explained step by step in a Jupyter notebook, accompanied by Python code. The text discusses setting up design constraints for sampling sequences in `Infrared`, computing quality measures, constructing a suitable cost function, as well as the optimization procedure. We show that not only thermodynamic but also kinetic folding features can be relevant. Kinetics of interaction formation can be estimated efficiently using the `RRIkinDP` tool, and the chapter explains how to include kinetic folding features from `RRIkinDP` directly in the cost function. The protocol implemented in our Jupyter notebook can easily be extended to consider additional requirements or adapted to novel design scenarios.

Key words RNA sequence design, RNA-RNA interactions, RNA structure, RNA folding kinetics

1 Introduction

In recent years, there has been notable advancement in RNA design tools, addressing various design challenges such as negative design problems, multi-structure design, and complex constraints. Tools like `RNAblueprint` [4], `Redprint` [5], and `Infrared` [13] exemplify this progress, offering sophisticated algorithms and user-friendly interfaces for designing RNA molecules with precise control over their structural and functional properties.

RNA-RNA interactions play critical roles in regulating gene expression, RNA processing, and other fundamental biological processes. These interactions involve base pairing between

Maria Waldl and Hua-Ting Yao shared first authors.

Alexander Churkin and Danny Barash (eds.), *RNA Design: Methods and Protocols*, Methods in Molecular Biology, vol. 2847,
https://doi.org/10.1007/978-1-0716-4079-1_1,

complementary regions of two RNA molecules, resulting in the formation of complex secondary and tertiary structures. Understanding the thermodynamic and kinetic features governing RNA-RNA interactions is crucial for unraveling their functional implications and engineering RNA molecules with tailored properties. Since RNA-RNA interactions can be treated as an extension of RNA secondary structures, they are arguably more amenable to rational design than interactions with proteins or small molecules and may thus provide an ideal avenue for engineering artificial regulators of gene expression.

In this chapter, we also present the integration of `RRIkinDP` [11], a computational tool that allows to evaluate not only thermodynamic but also kinetic features of RNA-RNA interactions, into RNA design schemes such as the `Infrared` framework. `RRIkinDP` offers efficient evaluation of interaction features, making it fast enough to directly integrate into the design process. This allows researchers to engineer RNA sequences with functional characteristics beyond equilibrium properties.

`Infrared` provides a versatile framework for the rapid development of task-specific RNA design tools. Leveraging the capabilities of `Infrared`, our approach builds upon existing design methodologies to optimize various aspects of RNA sequences, including accessibility, seed stability, barrier optimization, alternative structures, and dynamic switching through interaction formation. Integration into the constraint framework within `Infrared` further enhances computational efficiency, enabling the design of complex RNA molecules with reduced computational costs.

In contrast to earlier tools for multi-structure design [3, 9] that allowed only to specify a list of target structures, but provided little choice of cost function, optimization strategy and additional design constraints, `Infrared` provides a framework in which arbitrary design constraints can be formulated, making it equally suitable for interaction design and classical multi-structure design.

The design approach outlined in this chapter is presented in the form of a Jupyter notebook, available at https://github.com/ViennaRNA/RRIdesign. This notebook provides a comprehensive guide to implementing our design methodology and offers insights into the rational engineering of RNA sequences for specific applications.

By elucidating the integration of `RRIkinDP` into RNA design frameworks and showcasing the capabilities of `Infrared`, this chapter aims to provide researchers with valuable insights and methodologies for engineering RNA molecules with enhanced functionality and versatility.

1.1 Theory

There are several common approaches to predicting RNA-RNA interactions as extensions of RNA secondary structure prediction. The `ViennaRNA` package, which we're using in this contribution, provides the programs `RNAcofold` and `RNAup`. `RNAcofold` performs the prediction by concatenating the two RNA strands and then runs a standard folding algorithm except that loops containing the break point between the two strands are treated specially. This results in a pseudo-knot free joint structure and therefore excludes structures containing, e.g., kissing hairpins. `RNAup` [7] views interaction formation as a two-step process consisting of freeing up the interaction sites on both molecules and then forming the interaction between single-stranded regions. The binding free energy can then be expressed as

$$\Delta G_{bind} = \Delta G_{open} + \Delta G_{interact}, \tag{1}$$

where $\Delta G_{open} \geq 0$ is the free energy required for making the binding site accessible and $\Delta G_{interact} \leq 0$ the free energy gained from forming interaction helices. The algorithm proceeds by precomputing ΔG_{open} for every possible region $[i \ldots j]$ before then finding the interaction minimizing ΔG_{bind}. `IntaRNA` in addition requires a perfectly matching seed interaction, which can speed up the interaction search. The main restriction of `RNAup` and `IntaRNA` is that there can only be one interaction region. Note that the unrestricted RNA-RNA interaction problem is NP-hard similarly to the problem of predicting secondary structures with pseudo-knots.

A putative interaction may be thermodynamically stable but hard to reach by the dynamic folding process due to high energy barriers. `RRIkinDP` is a tool to analyze the energy landscape for structure formation and can extract useful features, such as the highest energy barrier. By restricting itself to direct trajectories, i.e., shortest paths from initial contact to full interaction, the problem can be solved efficiently using dynamic programming. In addition to the evaluation of folding paths, the energy landscape provides information on possible short initial seed interactions that can stabilize the first contact between the two RNAs and thereby initiate the interaction formation. Further information on the feasibility of an interaction is provided by the accessibility of the interaction site, or the seed interaction site, where a high accessibility indicates that it is easy (not much energy is needed) to make the interaction site unpaired in the intramolecular structure and therefor available to form interaction base pairs. Furthermore, the structure of the landscape can reveal shorter suboptimal interactions that are kinetically more favorable than the full interaction predicted by thermodynamic criteria.

The perhaps most important ingredient for any successful design is the cost function that will be optimized. Thermodynamic features in the cost function should typically be based on partition function calculations. A simple, commonly used, cost function for designing bistable sequences folding into two target structures S_1, S_2 might read

$$C(x) = E(x, S_1) - G(x) + E(x, S_2) - G(x), \tag{2}$$

where $G(x) = -RT \ln Z$ is the ensemble free energy derived from the partition function Z over all possible structures of sequence x and $E(x, S)$ is the free energy of structure S on sequence x. Minimizing the cost function above is equivalent to maximizing the product of the Boltzmann probabilities $p(S_1) \cdot p(S_2)$ of the two structures.

The above example assumes that we want to precisely specify both target structures down to each individual base pair. However, in a typical application, we may want to only specify structures in part, thus giving the design process more freedom. Partial specification can easily be incorporated by using constrained folding. Given a partial structure P that specifies some base pairs as well as some unpaired positions (but leaves other positions free), we can easily compute a constrained partition function Z^C, which sums only over those structures that fulfill the constraints, and the corresponding free energy $E^C(x, P) = -RT \ln Z^C$. Simply substituting $E(x, S)$ with the constrained version yields a suitable cost function for partially specified structures. The opening energies used by `RNAup` for interaction prediction (see above) are another example of features that can be easily computed using a constrained partition function.

In prokaryotes, translation initiation starts with binding of the ribosome via an RNA-RNA interaction between the Shine-Delgarno sequence of the mRNA and a CCUCC element near the end of the 16S rRNA. The opening energy of the ribosome binding site (RBS) is therefore an important determinant of translation efficiency and can be used to predict the effect of small RNA (sRNA) regulation [1]. In our application example, we will start from the well-known interaction between the bacterial sRNA DsrA and its target mRNA rpoS. The rpoS mRNA is normally translated poorly, since the RBS, which we assume as a 30nt region starting with the Shine-Dalgarno motif, is sequestered in a stable secondary structure. Binding of DsrA further upstream leads to a refolding of the untranslated region (UTR) making the RBS accessible, thus upregulating translation. In this chapter, we will as a showcase design the $5'$ UTR of a messenger RNA coding sequence, e.g., GFP to put it under the control of the DsrA sRNA. We note that one could analogously design an artificial sRNA to target a given (fixed) mRNA.

Our design criteria will therefore be: (i) sRNA and mRNA should bind strongly at reasonable concentrations; (ii) the mRNA should exhibit poor RBS accessibility without the sRNA; (iii) in the bound state, the RBS is highly accessible. Since the RBS reaches into the coding sequence, we will include the first codons into the designed sequence while requiring a fixed encoded amino acid sequence. In a second step, as a showcase for the design of more complex interaction features, we will also add (iv) a stable interaction seed as well as (v) a low folding barrier along the interaction formation path as design criteria.

In practice, a design problem will often have numerous additional constraints, e.g., to allow for handling of the sequences (such as primer binding sites, restriction sites, etc.) or to ensure that the designed sequence is compatible with the expression system. The `Infrared` framework allows to easily incorporate such constraints into the design process. Many sRNA-mRNA interactions, including the DsrA-rpoS, are facilitated by Hfq [10]. Hfq has distinct binding sites for sRNAs and mRNAs, and while we keep the DsrA sequence (and thus its Hfq binding site) fixed, a corresponding binding motif might be designed into the mRNA. Finally, since not all requirements of sRNA regulation are fully understood, designing a fully functional sequence will likely require multiple rounds of computational design and experimental testing.

2 Materials

Our RNA-RNA interaction design approach relies on several key components: the `ViennaRNA` library for free energy evaluations, `RNAup` and `IntaRNA` for interaction prediction, `RRIkinDP` for computing kinetic and thermodynamic RNA-RNA interaction features, and the design framework `Infrared`. The design approach and examples presented in this book chapter are provided as Jupyter notebooks. This section outlines the steps to install these tools within a Conda environment.

2.1 Installing Conda Package Manager

We recommend installing the necessary dependencies using the Conda package manager. Instructions for installing Conda, in the form of Miniconda, on MacOS and Linux can be found here: https://docs.anaconda.com/free/miniconda/miniconda-install.

2.2 Installation of RRIkinDP, ViennaRNA, and Infrared

After installing and activating Conda, the required dependencies can be easily installed using the `conda install` command. We recommend setting up a new environment and installing the dependencies at the same time to avoid compatibility issues between different package versions, e.g., with the following commands:

```
$ conda create -name rridesign \
     conda-forge::'infrared>=1.2' \
     bioconda::'viennarna>=2.6.2' \
     bioconda::intarna \
     bioconda::'rrikindp>=0.0.2'
$ conda activate rridesign
```

For additional installation instructions, see the Notes section at the end of the chapter.

2.3 Design Approach and Example Scripts

The Jupyter notebook with the design approach and example files can be downloaded from https://github.com/ViennaRNA/RRIdesign, e.g., with `git clone`:

```
$ git clone git@github.com:viennarna/RRIdesign.git
```

To run the Jupyter notebook, ensure a notebook interface is installed, e.g., JupyterLab via Conda:

```
$ conda install conda-forge::jupyterlab
```

Navigate to the downloaded git repository and start JupyterLab:

```
$ cd RRIdesign
$ jupyter lab
```

This will open JupyterLab in your web browser, where you can access the notebook `RRIdesign.ipynb` containing all the scripts described in the Methods section.

2.4 RNA Sequence Data

For illustration, we use the DsrA-rpoS interaction from Escherichia coli. The corresponding sequences can be obtained from the RefSeq Genomes data base entry of the Escherichia coli strain K-12 substr. MG1655 at NCBI [8]:

- DsrA: NC_000913.3(2025313-2025225)
- rpoS 5′ UTR fragment (TSS -150nts +27nts): NC_000913.3 (2867701-2867525)

As an example of a coding sequence that can be controlled by DsrA via a designed mRNA 5′ UTR, we use the GFP coding sequence from Aequorea victoria, GenBank: M62654.1.

3 Methods

In this section, we will briefly describe functions included in the Jupyter notebook and accompanying python module.

3.1 The DsrA-rpoS Example

The first section of the notebook examines the interaction between the DsrA and rpoS sequences as a natural example of translational control via sRNA–mRNA interaction. We use a piece of the rpoS mRNA comprising 150nt upstream of the start codon as well as the first 9 codons (27nt) of the coding sequence after the start codon AUG.

Note that results will vary depending on the extent of the mRNA region. In the first step, we predict the minimum free energy (MFE) secondary structures of dsrA and rpoS. For convenience, the `rri` module contains the `prediction_unbound` function which returns the MFE structures of both sequences. A second helper function `draw_RNAplot` calls the `RNAplot` program of the `ViennaRNA` package, highlighting the Shine-Delgarno motif, the AUG, as well as the 30nt RBS. The resulting plot should look as shown in Fig. 1, where the region downstream of the RBS is sequestered by a stem structure. It has been shown that this stem structure inhibits the translation of rpoS [2].

The `rri.prediction_bound` helper function computes the rpoS structure in the bound state, by first predicting the interaction using `RNAup`, then performing a structure prediction for rpoS under the constraint, that the region of interaction with DsrA cannot form intramolecular structure. Therefore, the intramolecular stem that blocks the RBS has to open upon DsrA binding [6]. The resulting interaction structure of both molecules is shown in Fig. 2. Note that for simplicity, we do not include intramolecular DsrA structure. The interaction between DsrA and the rpoS 5′ UTR is predicted to range from position 10 to position 40 in the DsrA and from position 25 to position 54 in the rpoS fragment with a free energy of about -12 kcal/mol.

Binding of DsrA has not resulted in a completely single-stranded RBS. This is expected, since 30nt long regions are almost never completely free of base pairs. However, we can quantify the

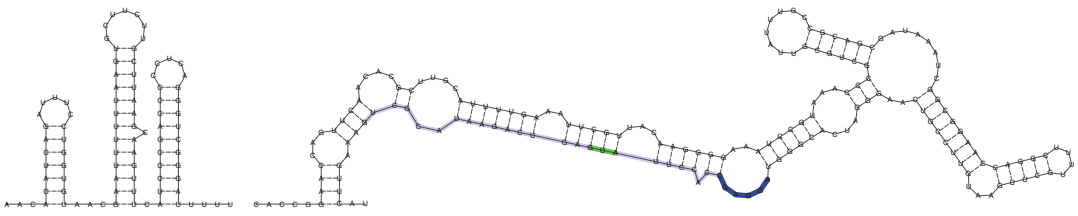


Fig. 1 Predicted MFE structure for DsrA (left) and rpoS (right).The Shine-Delgarno sequence is highlighted in blue, the AUG in green, and the RBS in gray

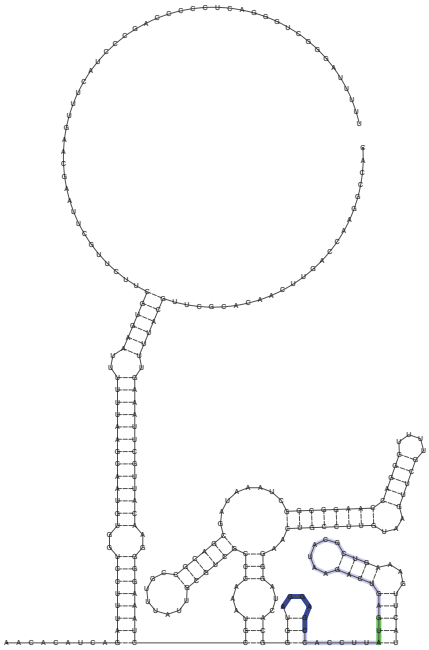


Fig. 2 Predicted MFE structure for rpoS after binding to DsrA. The Shine-Delgarno sequence is highlighted in blue, the AUG in green, and the RBS in gray

RBS accessibility by computing the opening energies ΔG_{open} before and after binding. ΔG_{open} is the difference between two ensemble free energies

$$\Delta G_{open} = \Delta G^{RBS} - \Delta G \geq 0 \tag{3}$$

where ΔG denotes the ensemble free energy of the mRNA over all possible structures, while ΔG^{RBS} is the corresponding free energy restricted to structures with unpaired RBS. To compute the binding energy after binding, we use the additional constraint that positions in the binding region cannot form (intramolecular) base pairs. Using the helper function `rri.dGopen`, we obtain $\Delta G^{free}_{open} = 11.23$ kcal/mol before and $\Delta G^{bound}_{open} = 6.34$ kcal/mol after binding.

In addition to the binding energy and the RBS accessibility, we computed kinetic features of the native DsrA-rpoS interaction. Figure 3 represents the energy landscape of all intermediate interaction. From this we can, e.g., search for the optimal starting seed interaction as well as a direct folding path that involves the lowest possible energy barrier. Assuming a seed length of 5 nucleotides for the first stable interaction, the landscape suggests that the optimal start point for the DsrA-rpoS interaction is region from position 12 to 16 with a free energy of 0.68 kcal/mol. From this seed, the full interaction can be reached without crossing any barrier states with a less favorable energy than the seed interaction.

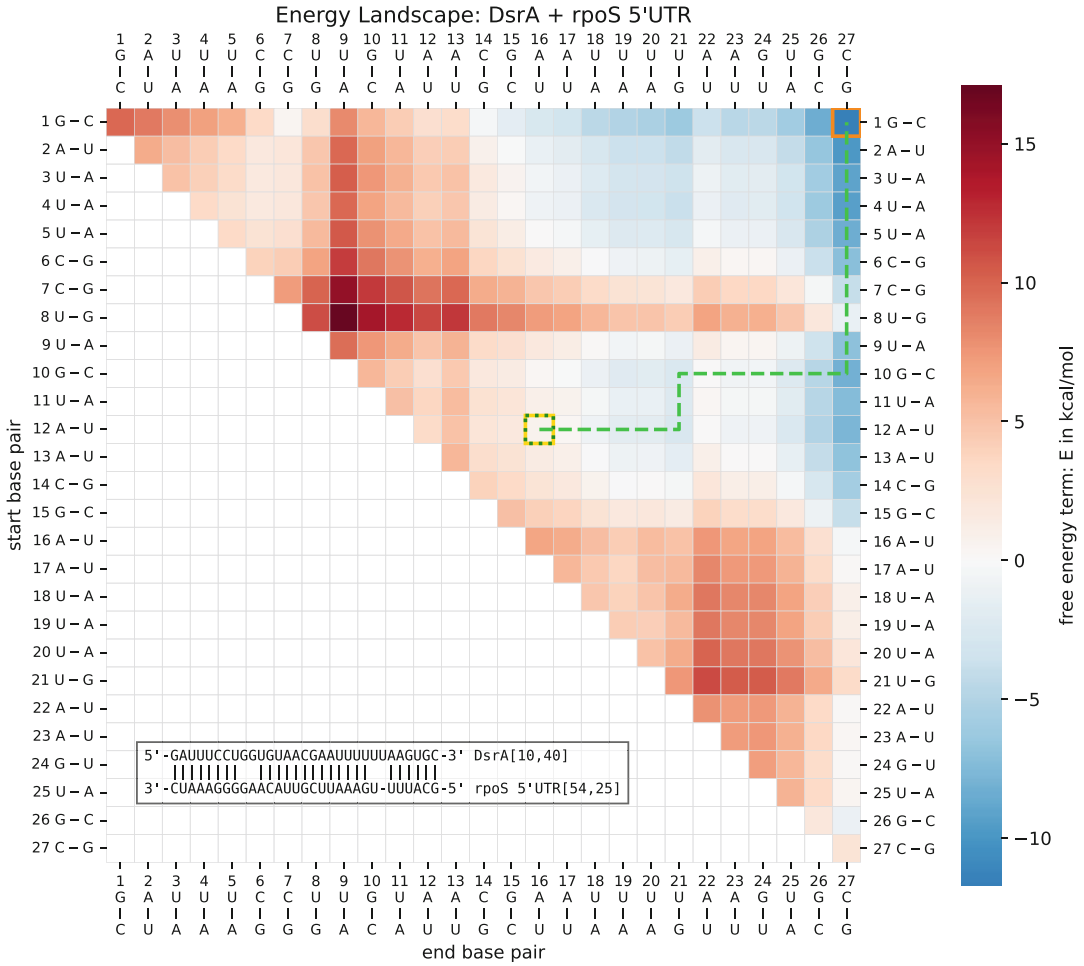


Fig. 3 Energy landscape of the predicted minimum free energy interaction between DsrA and the 5′ UTR of rpoS in E. coli The energy landscape represents all possible sub-interactions $_{k,l}$ from the *k*th to the *l*th base pair with in the full DsrA-rpoS interaction, which is shown in the box at the bottom left corner. The cell in the *k*th row and *l*th column represents the binding energy of sub-interaction $_{k,l}$ in kcal/mol indicated by the red to blue color scale, where red (blue) signifies a high (low) energy. Assuming a seed interaction of length *m*, the folding path starts with a cell located on the *m*th diagonal, the diagonal starting with $(1, m)$. The possible moves are either to the right or to the top. Moving one cell to the right (top) in this matrix corresponds to extending the interaction by one base pair on the right (left) side of the sub-interaction. An ideal path should encode a sequence of decreasing free energy from the start to the end. The final interaction is the sub-interaction corresponding to the end of the path. A minimum barrier folding path with seed length $m = 5$ is shown as a green dashed line, starting with $_{12,16}$ (yellow) and ending at the full interaction (orange). The barrier state (dark green) is the same as the starting state

3.2 Design Model

`Infrared` provides different solvers for user-defined problems described as the object `Model`. In the context of RNA design, a model typically contains a set of variables, each with 4 possible values $\{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}\}$ and representing one position in the RNA; a set of constraints, each restricts nucleotide composition for a subset of variables; and a set of functions defining different (pseudo-)

energies of sequences. From the model, valid RNA sequences satisfying the constraints are sampled with a Boltzmann weight defined by the energies. The constraints can overlap, allowing us, e.g., to specify multiple target structures. Without an energy function, the sampling will be uniform over all valid sequences. `Infrared` also allows to target a specific value for each energy in the output samples.

In this section, we will explain step by step the `Model` object construction to design a similar sRNA-mRNA system. First, we need to import the framework and its RNA extension.

```
import infrared as ir
import infrared.rna as irrna
```

To simplify the task, we focus only on designing 150nt upstream of the start codon. The sRNA and coding sequences are fixed to the DsrA sequence and the first 9 codons of GFP. We also insert a Shine-Dalgarno motif at the same location as in the rpoS example. Let `seq` be the concatenation of the described sRNA and mRNA sequence constraints expressed as a IUPAC sequence, we create an `Infrared` design model as

```
model = ir.Model(len(seq), 4)
for i, x in enumerate(seq):
    model.add_constraints(
     ir.ValueIn(i, irrna.iupacvalues(x))
    )
```

The first line initiates a model of n variables with 4 values where n is the total length of sRNA and mRNA. The last two lines restrict the possible values for each variable given by the IUPAC sequence.

Next, we extract two structure constraints, `mRNA_str` and `inter_str`, from the DsrA-rpoS system to guide the conformation before and after binding. The first constraint `mRNA_str` imposes the structure of the RBS region as shown in Fig. 1. The second one `inter_str` describes the binding region and overlaps with `mRNA_str` to destabilize RBS conformation after binding (Fig. 2). The constraints are expressed in well-balanced dot-bracket notation and added to the model using `add_constraints`.

```
for name, dbn in [
    ('M', mRNA_str),
    ('I', inter_str),
]:
    bps = irrna.parse(dbn)
    model.add_constraints(
     irrna.BPComp(i, j) for (i, j) in bps
   )
```

We further define the energy of folding to each structure with a simple base pair energy model using `add_functions`

```
model.add_functions(
  [
    irrna.BPEnergy(
        i, j, (i-1, j+1) not in bps
    )
    for (i,j) in bps
  ],
  f"energy{name}",
)
```

and define a more complex Turner energy with `add_feature`.

```
model.add_feature(
  f'Energy{name}',
  f'energy{name}',
  lambda sample, ss=ss:  RNA.energy_of_struct(
    irrna.ass_to_seq(sample), ss
  ),
)
```

Here the base pair energy is used as weight to sample an RNA sequence from the Boltzmann distribution.

```
sampler = ir.Sampler(model)
sampler.sample()
```

The additional features allow us to efficiently control the final structure energy of the sampled sequences. For instance, the code below returns a sequence forming the interaction `inter_str` with an energy of approximately -15 kcal/mol and the `mRNA_str` structure with -7 kcal/mol with a tolerance of 1.5 kcal/mol under Turner energy model. We invite readers to read `Infrared` RNA design tutorial for more details [12].

```
sampler = ir.Sampler(model)
sampler.set_target(-15, 1.5, 'EnergyI')
sampler.set_target(-7, 1.5, 'EnergyM')
sampler.targeted_sample()
```

From the sampling procedure, we obtain RNA sequences that satisfy the imposed sequence and structure constraints, thus already conforming to the criteria of *positive* design. However, this does not guarantee that candidate sequences actually form the desired structures and fulfill other design objectives (binding conformation, accessibility) while avoiding alternative structures. Avoiding

unwanted features is known as *negative* design. To address it, we deploy local optimization to search the neighborhood of sampled sequences. This requires a move set that defines the local neighborhood of sequences that still fulfill the constraints, as well as a cost function to minimize, which will be discussed in the next section.

At each step, a connected component (a set of related positions) defined by structure constraints is randomly and uniformly selected, and the corresponding subsequence is resampled. This could result in a point mutation or larger change, depending on the size of the connected component. In addition, we ensure the new sequence is distinct to the previous one. The new sequence is accepted if its cost is lower or, otherwise, with an acceptance probability controlled by temperature. We provide two optimization routines in the notebook: `mc_optimize` performs a Markov Chain Monte Carlo (MCMC) simulation at constant temperature, while `sa_optimize` performs a simulated annealing, i.e., an MCMC with a cooling schedule.

3.3 Cost Function

The last piece to complete the sRNA-mRNA system design is the cost function to minimize during optimization. We consider two cost functions to address 5 design criteria listed in Subheading 1:

(i) sRNA and mRNA should bind strongly;

(ii) the mRNA should exhibit poor RBS accessibility without the sRNA;

(iii) in the bound state the RBS is highly accessible;

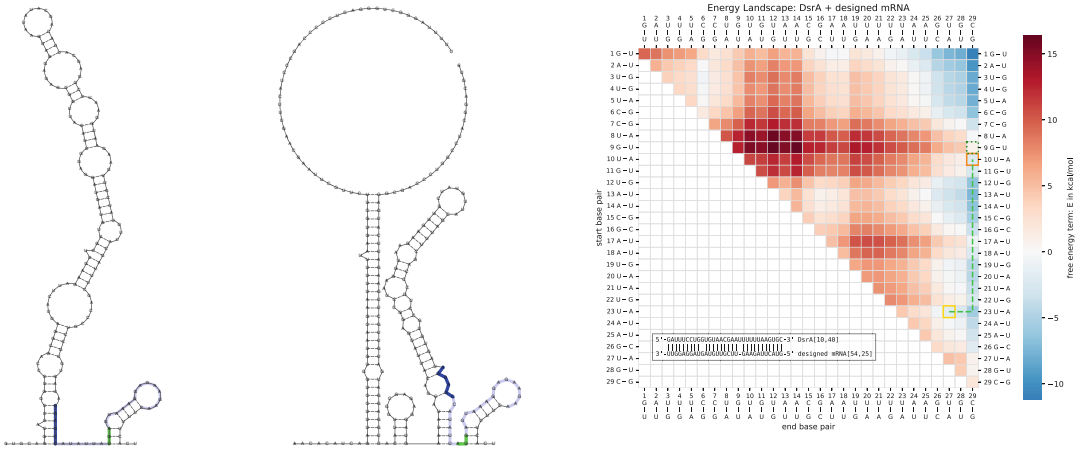
(iv) there is a suitable, accessible, seed interaction;

(v) the energy barrier for interaction formation is low.

An equilibrium thermodynamics cost function will employ only the first three design criteria, while a cost function with kinetics will additionally cover the last two. In this section, we present in detail these two cost functions and show the designs obtained at the end of optimization. Note that the designed sequences will vary depending on the sampled starting sequences and the partial resampling while running the Jupyter notebook.

3.3.1 Thermodynamics Cost Function with RBS Accessibility Optimization

Our first cost function will be based on three free energy values: (i) the binding energy as computed by `RNAup/IntaRNA` ΔG_{bind}, (ii) the RBS opening energy for the unbound mRNA ΔG_{open}^{free}, as well as (iii) for the sRNA bound mRNA ΔG_{open}^{bound}. We can find suitable target values θ for these free energies by comparing to known sRNA – mRNA interactions. Combining several optimization criteria can be tricky, since the optimization procedure may end up over-optimizing some (easy) criteria while ignoring hard to optimize ones. To avoid this, we make use of a `LogSumExp` function, which computes a smooth maximum over a list of values,



(a) (b) (c)

Fig. 4 Predicted MFE structure for designed mRNA with thermodynamics cost function before (**a**) and after (**b**) binding to DsrA and its energy landscape (**c**) The predicted RBS opening free energy is 12.29 kcal/mol before and 5.01 kcal/mol after binding

$\mathrm{LogSumExp}(x_1, \ldots, x_n) = \ln \sum \exp x_i$. Note that the feature values x_i may need to be scaled and shifted first. As cost function, we can use, e.g.,

$$C(x) = \mathrm{LogSumExp}\Big((\Delta G_{bind} - \theta_{bind}), (-\Delta G^{free}_{open} + \theta^{free}_{open}), (\Delta G^{bound}_{open} - \theta^{bound}_{open}) \Big) \tag{4}$$

The second term has opposite sign, since we want to minimize ΔG_{bind} and ΔG^{bound}_{open} while maximizing ΔG^{free}_{open}. See the notebook for chosen target values.

Figure 4 visualizes one mRNA design obtained after 1 000 steps using the thermodynamic cost function. The predicted values are -12.97 kcal/mol for binding energy, 12.29 kcal/mol for opening energy before binding, and 5.01 kcal/mol for opening energy after binding. The difference of opening energy can also be observed on MFE predictions. The Shine-Dalgarno is sequestered in an almost perfect helix before binding, after binding it is part of three consecutive loops separated by an isolated base pair.

3.3.2 *Adding Kinetic Features to the Cost Function*

The mRNA UTRs designed with the thermodynamic cost function feature the targeted binding and opening energies. However, when evaluating kinetic features, we find that the designed sequences might often not fold into the full interaction within reasonable time, e.g., due to high barriers along the folding path. For instance, the minimum barrier energy of the design shown in Fig. 4c is about 2.7 kcal/mol higher than the binding energy of seed interaction.

To include kinetic features into the design optimization, we convert the design criteria (iv), a stable seed interaction and (v) a low barrier state energy, into two energies.

Given a fixed seed interaction length m, we define the first energy as the minimum barrier energy $\Delta G_{barrier}$ among all barriers that folding paths reach starting with a seed interaction of length m (design criteria (v)). The second energy is the binding energy ΔG_{seed} of seed interaction from which the path has a barrier energy at $\Delta G_{barrier}$ to measure the stability of seed interaction (design criteria (iv)). Both these energies are available after calling `rri.EnergyLandscape` from the `RRIkinDP` module to analyze the interaction formation landscape. Combining with Eq. 4, we obtain the kinetic cost function as

$$\begin{aligned} C(x) \quad = \text{LogSumExp}\Big(& (\Delta G_{bind} - \theta_{bind}), (-\Delta G_{open}^{free} + \theta G_{open}^{free}), \\ & (\Delta G_{open}^{bound} - \theta_{open}^{bound}), \\ & (\Delta G_{seed} - \theta_{seed}), (\Delta G_{barrier} - \theta_{barrier}) \Big) \end{aligned} \tag{5}$$

An example sequence, obtained after 1 000 optimization steps, is shown in Fig. 5. As with the thermodynamic cost function, we obtain a sequence where the RBS region is sequestered in the

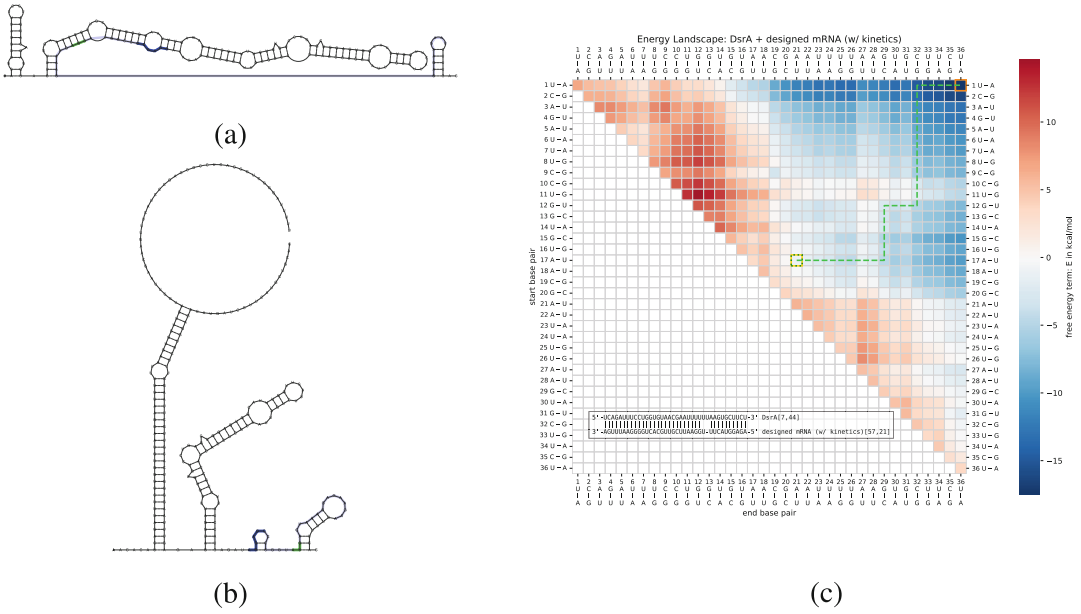


Fig. 5 Predicted MFE structure for designed mRNA with kinetic cost function before (**a**) and after (**b**) binding to DsrA and its energy landscape (**c**) The predicted RBS opening free energy is 12.58 kcal/mol before and 4.88 kcal/mol after binding. `RRIkinDP` predicts the interaction starts with the seed region from 17th to 21th base pairs

unbound state and weakly structured after binding with DsrA. The RBS opening energies are 12.58/4.88 kcal/mol before/after binding. The landscape, however, has notably lower energy barriers than the example from Fig. 4 as well as the native DsrA-ropS interaction (Fig. 3). The interaction is predicted to have a free energy around − 17 kcal/mol, the optimal 5nt seed region consists of the 17th to 21th base pairs (out of 36 base pairs in the interaction).

4 Notes

By default, Conda is set up with the newest Python version, which may lead to compatibility issues with certain packages. To ensure compatibility, you can create a new Conda environment with a specific Python version and install the required packages:

```
$ conda create -name rridesign python=3.10
$ conda activate rridesign
$ conda install conda-forge::'infrared>=1.2' \
    bioconda::'viennarna>=2.6.2' \
    bioconda::intarna \
    bioconda::'rrikindp>=0.0.2'
```

Installing packages using Conda may sometimes be slow or fail, because Conda is unable to resolve dependencies. Mamba, a drop in replacement for Conda, is in general faster and better at resolving dependencies. Installation instructions can be found here: https://mamba.readthedocs.io/en/latest/installation/mamba-installation.html. We highly recommend using Mamba, especially if you anticipate combining this tutorial with other tools.

References

1. Amman F, Flamm C, Hofacker IL (2012) Modelling translation initiation under the influence of sRNA. Int J Mol Sci 13:16223–16240. https://doi.org/10.3390/ijms131216223
2. Brown L, Elliott T (1997) Mutations that increase expression of the rpoS gene and decrease its dependence on hfq function in Salmonella typhimurium. J Bacteriol 179(3):656–662. https://doi.org/10.1128/jb.179.3.656-662.1997
3. Flamm C, Hofacker IL, Maurer-Stroh S, Stadler PF, Zehl M (2001) Design of multistable RNA molecules. RNA 7(2):254–65. https://doi.org/10.1017/s1355838201000863
4. Hammer S, Tschiatschek B, Flamm C, Hofacker IL, Findeiß S (2017) RNAblueprint: flexible multiple target nucleic acid sequence design. Bioinformatics 33(18):2850–2858. https://doi.org/10.1093/bioinformatics/btx263
5. Hammer S, Wang W, Will S, Ponty Y (2019) Fixed-parameter tractable sampling for RNA design with multiple target structures. BMC Bioinform 20(1):209. https://doi.org/10.1186/s12859-019-2784-7
6. Lease RA, Cusick ME, Belfort M (1998) Riboregulation in Escherichia coli: DsrA RNA acts by RNA:RNA interactions at multiple loci. Proc Natl Acad Sci 95(21):12456–12461. https://doi.org/10.1073/pnas.95.21.12456
7. Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF, Hofacker IL (2006) Thermodynamics of RNA-RNA binding.

Bioinformatics 22(10):1177–82. https://doi.org/10.1093/bioinformatics/btl024

8. O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD (2015) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44(D1):D733–D745. https://doi.org/10.1093/nar/gkv1189
9. Höner zu Siederdissen C, Hammer S, Abfalter I, Hofacker IL, Flamm C, Stadler PF (2013) Computational design of RNAs with complex energy landscapes. Biopolymers 99(12):1124–36. https://doi.org/10.1002/bip.22337
10. Soper T, Woodson SA (2008) The rpoS mRNA leader recruits Hfq to facilitate annealing with DsrA sRNA. RNA 15:1907–1917. https://doi.org/10.1261/rna.1110608
11. Waldl M, Beckmann IK, Will S, Hofacker IL (2023) Modeling kinetics of RNA-RNA interactions on direct paths. bioRxiv. https://doi.org/10.1101/2023.07.28.548983. https://github.com/mwaldl/RRIkinDP
12. Yao HT, Ponty Y, Will S (2022) Developing complex RNA design applications in the Infrared framework. RNA Folding-Methods and Protocols
13. Yao HT, Marchand B, Berkemer SJ, Ponty Y, Will S (2024) Infrared: a declarative tree decomposition-powered framework for bioinformatics. Allgorithms Mol Biol. https://doi.org/10.21203/rs.3.rs-3366298/v1

5.7 Additional publications

In addition to the 6 paper presented in this thesis on RNA-RNA interactions, I worked on three papers related to RNA structure in general:

- Maria Waldl, Sebastian Will, Michael T. Wolfinger, Ivo L. Hofacker and Peter F. Stadler (2020) 'Bi-alignments as Models of Incongruent Evolution of RNA Sequence and Secondary Structure', in P. Cazzaniga et al. (eds) Computational Intelligence Methods for Bioinformatics and Biostatistics. Cham: Springer International Publishing, pp. 159–170. doi:10.1007/978-3-030-63061-4_15.
- Tom Rappol, Maria Waldl, Anastasia Chugunova, Ivo L. Hofacker, Andrea Pauli, and Elisa Vilardo (2024) 'tRNA expression and modification landscapes, and their dynamics during zebrafish embryo development', Nucleic Acids Research, 52(17), pp. 10575–10594. doi: 10.1093/nar/gkae595.
- Leonhard P. Sidl, Maximilian Faissner, Manuel Uhlir, Cristian A. Velandia Huerto, Maria Waldl, Hua-Ting Yao, Ivo L. Hofacker and Peter F. Stadler (2025) 'Computational Complexity, Algorithmic Scope, and Evolution', Journal of Physics: Complexity, 6(1), p. 015013. doi: 10.1088/2632-072X/adb928.

6 Discussion and Future Directions

The research presented in this thesis builds upon six papers, each contributing to the understanding of the RNA-RNA interaction formation process. Together, these works address the two main research questions: (RQ1) How do folding kinetics and steric constraints influence RNA-RNA interaction formation? and (RQ2) Can a reduced structure space be defined to accurately capture these dynamics while simplifying the interaction formation problem? This section will first discuss the contributions of each paper to the research questions, then summarize the contributions of the full thesis and in the end an outlook is presented.

RRIkinDP: Direct Paths RNA-RNA Interaction Kinetics Model (Waldl et al., 2025)

The RRIkinDP paper forms the conceptual and methodological core of this thesis. It directly addresses RQ1, as it introduces a kinetic folding model and defines kinetic interaction features on top of it. The comparison of kinetic features of experimentally confirmed RNA interactions and a randomized control suggests that the model captures features of RNA-RNA interactions beyond thermodynamic stability. It also contributes to RQ2: The structure space was restricted to direct paths from a first contact interaction to a candidate interaction. The expressivity of this model suggests that this restricted state space is sufficient to describe some of the properties of biologically functional interactions.

A case study on the interaction between the small RNA RyhB and its mRNA target sodB in E. coli illustrates the model's expressiveness. While standard 2D structure prediction suggests an extended interaction, the kinetic model favors a shorter interaction covering the start codon. This aligns with experimental findings. IntaRNA and RNAup predict an extension of the interaction in 3' direction with respect to rhyB. This region is predicted to be highly accessible and is known to bind the Hfq chaperone, while an direct hybridization via base pairs so far has not been shown [GT04]. Although RRIkinDP can not model steric constraints from the kissing hairpin configuration at the start of the interaction formation, it reveals kinetic barriers for any extensions of the interaction that would require either a full unfolding of the the hairpins or the threading of one RNA through the hairpinloop of the other RNA. Further sRNA-mRNA interactions examined (though not included in the main paper text) showed similar qualitative agreement with experimental data, motivating the model's use for investigating unknown RNA-RNA interactions, as e.g. explored in [MPW$^+$22].

In addition, the paper compares two kinetic boundary cases: a frozen model, where only base pairs in direct conflict with the interaction are broken, and an instantaneous refolding model, which assumes immediate equilibration of intramolecular structure given an accessible interaction site. The latter performs better in identifying known interactions, suggesting that some intramolecular refolding must occur on the timescale of interaction formation and can not be fully ignored. The reality might be between the two extreme representations of the behaviour of intramolecular structure and needs further investigation. A detailed evaluation of intramolecular structure changes typically induced by interaction formation on biologically functional interactions was hampered by the limited availablity of experimantal interaction structure data at a base pair resolution.

A preliminary application to antisense RNAs further indicates an important role of kinetics. Here, perfect long duplexes, though thermodynamically optimal, were found to be kinetically inaccessible due to stable intramolecular folds that are assumed to form before the two RNAs come in contact. Instead, multiple shorter, kinetically favorable sub-interactions were identified, often differing from thermodynamic most favourable sub-interactions.

The model allows detailed evaluation of these interactions e.g. by seed accessibility and folding barrier analyses, computed by the dynamic programming approach. However, often kinetic features obtained from the Markov process often provide more information on functional interactions than the evaluation of the folding barrier alone. Despite its efficiency (most simulations complete in seconds to minutes), computational cost becomes a limiting factor when solving the underlying master equation for very long interactions due to increasing numerical precision requirements. Future work could improve scalability by rescaling energy landscapes or by coarse-graining the folding space into local minima connected by barriers—potentially enabling even faster computation of the dynamics.

In summary, RRIkinDP provides a kinetic framework that bridges the gap between static thermodynamic models and the dynamic nature of RNA-RNA interaction formation. It strongly supports RQ1 and demonstrates that a reduced yet informative structure space (RQ2) can capture biologically meaningful dynamics.

Local RNA Folding Revisited (Waldl & Spicher et al., 2023)

This paper contributes to RQ2 (reducing the structure space) by exploring how local RNA folding constraints can effectively approximate RNA secondary structure while limiting computational cost. This aligns with the goal of identifying functional interactions by computing the accessibility of interaction sites and seed regions without the need to explore the entire intramolecular folding landscape of long RNAs. Therefore, local folding approaches—including RNAplfold—make predictions more scalable, opening the path toward genome-wide interaction screens [BMH11]. This paper revisits and extends the theoretical and practical underpinnings of such approaches.

Given that 2D structure prediction based on the McCaskill algorithm relies on context-

free grammars, the probability of certain conditional structural features are not directly accessible [McC90a]. In particular conditional accessibility (the probability that one region is unpaired given that another is paired/unpaired) is essential for modeling kinetics of multi-site RNA–RNA interactions, where multiple seeds may need to simultaneously extend and the interaction sites need to become accessible.

To address this, the paper combines sliding-window local folding with Boltzmann-weighted structure sampling, enabling efficient estimation of conditional accessibility while preserving linear runtime scaling. The impact of different parameter choices (e.g., window size, maximum base pair span, and sample size) is evaluated and recommendations for parameter sets to support reliable predictions are provided.

This new sampling method provide a foundation for extending the RRIkinDP model to more complex, multi-site interaction scenarios. Overall, the work offers both theoretical and practical contributions toward simplifying the intramolecular RNA structure landscape without sacrificing functional insight—directly addressing RQ2.

Investigating RNA-RNA Interactions Through Computational and Biophysical Analysis (Mrozowich et al., 2022)

This study addresses RQ1 by integrating computational modeling with biophysical experiments to investigate the stability and formation dynamics of RNA–RNA interactions in flaviviruses, particularly the Japanese encephalitis virus (JEV). It provides empirical support for kinetic modeling by comparing predicted interaction properties with experimental binding data.

In the JEV example, thermodynamic predictions from RNAup aligned well with measured binding affinities – for functionally validated interactions at the cyclisation site as well as for off target interactions in mutant constructs. The kinetic model refined these predictions by identifying likely interaction starts and barriers to full duplex formation.

Additionally, thermodynamic models suggested that homo-dimers of terminal sequences could form extensive duplexes, potentially hindering hetero-dimer formation. However, the kinetic model predicted—and experiments confirmed—that only shorter regions formed within the experimental timeframe. This finding supports the idea that kinetic accessibility, not just thermodynamic stability, governs interaction formation.

Following an established approach, evolutionary conservation and covariation of interaction sites across JEV isolates were used to identify the long-range interaction between the 5' and 3' terminal regions of the genome. Notably, the kinetic model consistently identified the functional interaction as kinetically accessible across these homologs, suggesting that predictions capture conserved dynamic properties rather than sequence-specific artifacts. Such findings suggest that integrating comparative approaches directly with the RRIkinDP model could to enhance the reliability of predictions.

Beyond qualitatively validating the kinetic model, the study illustrates how computa-

tional predictions can inform experimental design. For instance, structure predictions helped identify RNA subsequences that retain native folding when isolated, minimizing experimental artifacts caused by refolding upon truncation. While this was done to some degree in this study, especially the 3'UTR construct could be further optimized. This supports a broader methodological point: incorporating computational modeling early in the experimental workflow can improve construct design and interpretation.

TERribly Difficult: Searching for Telomerase RNAs in Saccharomycetes (Waldl et al., 2018)

The telomerase RNA paper served as an initial motivation for investigating the dynamics of RNA-RNA interactions in the context of structural complexity and formation constraints. The pseudoknot and the middle-to-long-range interaction that forms the core-enclosing helix (CEH) in yeast telomerase RNA (TER) are representative examples of complex intramolecular RNA-RNA interactions. These elements can be explored with interaction prediction methods and raise important questions about the physical feasibility of long-range helix formation.

From the CEH, a long helical structure extends toward the terminal arm, with considerable variation in length across related yeast species. Despite this variability, these helices consistently contain bulges and interior loops. The structural elements connected to the two ends of the helix span several hundred nucleotides, raising mechanistic questions about how such long-range helices can form. A perfect RNA helix undergoes a 360° twist every 10-11 base pairs; thus, formation of an extended CEH without structural interruptions would imply that the long flanking regions would need to rotate in tandem with the twist. This prompted the hypothesis that bulges and interior loops may act as flexible regions that compensate for helical torsion and enable interaction formation without large-scale rotation of the attached domains.

While this paper did not directly address these questions, it provided key motivation for subsequent studies in this thesis. For instance, the long interaction predicted in the CopA–CopT system was selected as a model case for 3D simulation, as it was supported by available experimental structural data. There, we observed that the predicted 2D interaction was interrupted by a long bulge in 3D space, suggesting that extended helices predicted by thermodynamic models may not fully form due to steric constraints—supporting the need to examine folding kinetics and 3D feasibility (RQ1).

Additionally, using tools such as IntaRNA and RNAup, we were able to reliably predict the TER pseudoknot. Follow-up 3D simulations and comparisons with the human TER crystal structure revealed further stabilization of this motif via non-canonical base pairs—highlighting the role of structural motifs beyond canonical helices in interaction stability.

At the time of publication, a kinetic model of RNA-RNA interaction formation had not yet been established, and comparative methods for analyzing homologous interactions

with high sequence and structural divergence did not provide conclusive insights. As a result, these predictions were not included in the homology search approach of the original paper, but they directly motivated the development of models and methodologies presented in this thesis. In this way, the study contributed observations that shaped both research questions.

Feasibility of 2D RNA-RNA Interaction Paths by Stepwise Folding Simulations (Beckmann et al., 2024)

This study contributes to both RQ1 (steric and kinetic influences on RNA interactions) and RQ2 (reducing the structure space) by embedding 2D folding pathways into 3D simulations. This approach tests the structural feasibility of RNA–RNA interaction formation paths by simulating their stepwise formation in three dimensions, enabling the identification steric constraints and energetic barriers that are not apparent in 2D models.

Key findings include that short interactions of 5–6 base pairs can typically form quickly, whereas extending these duplexes is often hindered by the structural context, for instance, when interaction sites are embedded within hairpin loops. Even when long, perfectly complementary regions are available, the 3D simulations frequently stabilize intermediate structures that contain interior loops. This underscores how steric effects can alter the course of interaction formation and lead to divergent predictions compared to 2D simulations.

Crucially, using 2D folding paths as guides allows the 3D simulation to focus on a relevant subspace of conformations. This guided strategy improves sampling efficiency and helps identify sterically infeasible intermediates. These findings can inform future improvements to 2D kinetic models by translating them to constraints that exclude implausible states, thereby enhancing prediction reliability.

To build on the presented results and enable broader conclusions, a larger and more diverse set of RNA–RNA interactions needs to be analyzed. However, this is currently limited by the scarcity of high-resolution 3D structure data, especially for interactions outside of RNA-proteins complexes.

Future work could also apply the guided folding strategy to alternative 3D simulation frameworks. For example, comparing results to more fine-grained, physics-based models may reveal how well the knowledge-based potentials used in this study (and data driven models in general) generalize to novel RNAs. Additionally, while the SimRNA move set supports rapid exploration of the structure space, it may not always reflect physically realistic transitions—for instance, it can introduce knots into the backbone. More detailed models may thus offer more interpretable folding trajectories and more immediate insights into RNA interaction kinetics.

Sequence Design for RNA-RNA Interactions (Waldl, Yao & Hofacker, 2025)

This contribution primarily presents an application of the kinetic model in the context of RNA sequence design, rather than directly contributing to either research question. However, it lays important groundwork for future investigations into and evaluation of the RRIkinDP model and its features.

By enabling the de novo design of RNA sequences with predefined thermodynamic and kinetic interaction features with respect to the model, the approach makes it possible to construct synthetic interaction pairs that are free from biological and experimental (selection) biases. These designed sequences can be tailored to systematically probe specific kinetic hypotheses. Testing such synthetic interactions in in vitro experiments could allow for detailed validation and refinement of the kinetic model. This, in turn, could improve our understanding of the kinetic mechanisms governing RNA–RNA interactions (RQ1), and inform which aspects of the structure space are essential to model (RQ2).

The presented design approach optimizes sequences for both thermodynamically and kinetically favorable interactions, requiring the evaluation of the kinetic model at each step. This showcases that the current implementation is computationally efficient enough to design sequences with interaction features similar to those of known functional pairs within minutes. This efficiency opens the door for integrating kinetic modeling into broader synthetic biology design workflows. For instance, the study demonstrates a use case where interaction formation triggers a structural rearrangement in the 5' UTR of an mRNA, regulating access to the ribosome binding site. Such designs could eventually support regulatory systems in which the expression of a target protein is controlled by the presence and concentration of a small RNA.

6.1 Discussion summary

This thesis was motivated by the fundamental observation that many RNA functions depend on interactions with other RNAs. Existing computational models of RNA–RNA interactions often focus on thermodynamic stability, overlooking the kinetic and steric constraints that shape the interaction formation process. The goal of this work was to explore these dynamic aspects and to develop computational approaches that more accurately capture the formation of RNA–RNA interactions.

Contributions to RQ1 — kinetic and steric influences on RNA–RNA interactions: The presented papers suggest that kinetics and steric constraints do indeed play a critical role in RNA–RNA interaction formation. The RRIkinDP model introduces kinetic features that provide new information that cannot be explained by thermodynamic stability. Notably, successful interactions appear to require the rapid formation of a stable minimal base-pairing core (typically 5–6 nts), and the accessibility of initial contact side (seed region)

emerges as a key predictive feature. These computational predictions are supported by experimental findings, such as the cyclization site in Japanese Encephalitis Virus (JEV), where the predicted kinetically accessible interaction, but not the thermodynamically more stable and extended one, was observed in biophysical experiments.

The 3D simulation work further validates the relevance of steric and kinetic constraints. It shows that many seemingly favorable 2D predictions are not feasible in 3D, and that extensions to long (perfect duplex) interactions are often hindered by local structural context. Together, these results support the conclusion that steric and kinetic effects significantly shape the interaction landscape, making their consideration essential for accurate prediction.

Contributions to RQ2 — restrictions on the state space: The RRIkinDP model shows that a restricted state space focused on direct folding paths can successfully capture features of RNA–RNA interaction formation. This suggests that RNAs in nature may also by-and-large follow direct interaction formation paths. Furthermore, the 3D simulations suggest additional constraints that could refine the 2D state space, such as the frequent inability to extend interactions beyond 6 consecutive stacking pairs. The local folding paper contributes a detailed evaluation of parameter sets to enable efficient accessibility computation under the assumption that longe range base pairs only provide a small contribution to the feasibility of interactions. Collectively, these results provide first insights into the essential state space for interaction formation and allow the implementation of efficient interaction kinetics prediction tools, enabling in turn the presented RNA-RNA interaction design approach and future large-scale interaction screens.

Overall, this thesis provides insights into kinetic and steric aspects of RNA-RNA interaction formation, as well as conceptual and technical contributions to computational models that capture this process.

The combination of efficient algorithms based on 2D representations, kinetic models and 3D embedding of folding paths, with validation on experimental data and selected 3D simulations, as well as the interaction design approach outlines a modular framework for studying interaction formation, that can be integrated and extended in the future.

The methods, findings, and limitations (highlighted in the discussions of the individual papers) form a basis for the future exploration, refinement, and integration of new experimental data.

6.2 Outlook

Building on the discussion of individual papers and overall results, several concrete next steps emerge that can refine our models and deepen insights into the thesis research questions:

Next steps

Expanding the RRIkinDP state space Extend the RRIkinDP direct-path model to include local detours and multi-site interactions. Specif ideas include allowing parallel formation of multiple seed regions (e.g., the OxyS–fhlA sRNA–mRNA pair), extending the state space to target not just the most stable interaction structure but multiple sub-optimal interaction structures and permit the local relaxation of unfavorable base pairs in the "frozen" RRIkinDP variant. This will test to what extent natural RNAs exploit interaction formation pathways beyond simple direct routes.

Further Validation of the Kinetic Model and Features: Investigate how well RRIkinDP features generalize by experimentally testing interactions in diverse RNA systems (e.g., riboswitches, lncRNAs, snRNAs, long-range interactions, pseudoknots). Use the sequence-design approach developed here to craft constructs with targeted kinetic signatures, for in vitro verification. This could show how broadly kinetic predictors apply beyond sRNA–mRNA pairs.

Leveraging Cross-linking and 3D Datasets: While cross-linking methods (e.g. PARIS, SPLASH) provide transcriptome-wide data on RNA–RNA contacts, they lack the base-pair resolution needed to validate kinetic models directly. Therefore, methods that can predict interaction candidate structures based on cross-linking reads should be benchmarked and extended. When applying this predicted interactions in the evaluation of kinetic features, it is crucial to quantify how noise and false positives in these datasets affect the reliability of the identification of kinetic features. Finally, agreement between 2D/3D computational predictions and experimental SAXS-derived structures, such as those generated collaboratively in Mrozowich et al. should be assessed [MPW$^+$22].

Formalizing homology-based approaches: Building on observations in JEV, incorporate evolutionary conservation into kinetic modeling by investigating kinetic features across homologous RNAs. Develop algorithms to compare not only thermodynamically predicted structures but also kinetic features and folding paths, e.g. building on existing approaches for covariance analysis, tools like CopraRNA, or using the trajectory-alignment method introduced in Waldl (2017) [WRP$^+$13, Wal17]. This will leverage functional conservation to boost prediction robustness.

Detailed Evaluation of Simulated 3D Folding Paths: Systematically analyze intrinsic 3D characteristics (e.g., backbone angles, steric clashes) from simulated folding trajectories, not just 2D features of the 3D simulations. This could yield additional insights into steric constraints and might result in additional "3D filters" that can be applied to 2D kinetic models to exclude sterically infeasible states.

While these next steps build directly on the presented methods, they also point toward

broader open questions and long-term goals that require conceptual, technical, and experimental advances. These are discussed in the following section.

Open Questions

This thesis advances computational methods and insights into the dynamics of RNA–RNA interactions. It also exposes several long-term challenges in modeling, data integration, and biological interpretation. Some of these more long-term questions are discussed here:

This work began with a central question: *Why do we need more accurate RNA–RNA interaction models?* Revisiting this question helps highlighting which additional factors may be relevant for further research:

At the most immediate level, more accurate predictions of RNA-RNA interactions are crucial for understanding RNA function, particularly for novel transcripts. Ideally they are not just limited to predicting interaction targets but provide insights into conditions and mechanisms of interaction formation. Thereby predictions contribute to our understanding of living systems and evolution in general, possibly also enabling the design of interventions and synthetic application.

An open question is under what conditions an interaction is not possible and how *off-target interactions* are avoided in nature. The risks of unintended off-target effects is obvious when considering therapeutic use cases. This could motivate research into *negative design*, that not only generates a favourable target interactions but ensures that there are no alternative favourable interactions.

Despite the complexity of RNA folding, it remains unclear whether a relatively small number of *physical or evolutionary principles* govern most biologically relevant interactions. Do generalizable rules govern RNA-RNA interactions? To what extent do RNAs follow direct folding pathways do detours through metastable states sample the structure space? How do co-transcriptional folding and existing intramolecular structures influence interaction formation? To what level can interaction formation be described through a *generalized model* and at what point is a *differentiation by interaction class* necessary. Analysing individual interaction mechanisms and their evolutionary conservation in even more detail could provide new hypothesises for further testing.

Environmental factors likely play a significant role in shaping interaction dynamics. Most current prediction models—like those used in this thesis—rely primarily on intrinsic properties of RNAs and corresponding thermodynamic parameters, neglecting external influences such as RNA-binding proteins, metabolites, competition effects, concentration, or subcellular localization. A key challenge is to assess the extent to which these factors impact folding and interaction formation, and where models, in vitro and in vivo behaviors diverge.

Another question is whether *different levels of structural representation* can be systematically *integrated* to improve interaction models. This thesis explored 3D simulations as a source for constraints on more efficient 2D models. A direct extension could use

local 3D motives, including non-canonical base pairs and pseudoknots, as constraints or guiding potential in 2D. Further directions could integrate more fine grained models from molecular dynamics and quantum chemistry as well as course-grained representations (eg helix level) with existing 2D and 3D model. Selecting important features for integrating fine-grained accuracy with efficient state space exploration will be essential especially for kinetic modeling.

Finally, a limiting factor is *data availability*: Datasets for RNA structure and interactions are scarce compared to those for proteins. Moreover, current data may be biased, for example by research focus on certain RNA classes, experimental accessibility, or difficulties in observing transient and dynamic conformations. Time-resolved data on RNA folding remains limited to relatively simple structures. Progress in this area will require not only further development of experimental techniques towards higher throughput and better structure resolution but also clarity on *how much* and *what type* of data are needed to support robust model development, especially for machine learning-based approaches. Designing RNAs to test mechanistic hypotheses and generating curated datasets that include structural, kinetic, and functional annotations across diverse interaction types will be essential to advance this field.

6.3 Conclusions

This thesis set out to deepen our understanding of RNA–RNA interaction formation by explicitly incorporating kinetic and steric constraints into computational models. The key contributions include:

- **Novel mechanistic insights:** A range of kinetic features of functional interactions are defined, including a rapid extension to an energetically favourable core interaction, good seed accessibility and low dissociation probability.
- **Reduced yet expressive state spaces:** The RRIkinDP model demonstrated that focusing on direct folding paths captures essential kinetic features, while local-folding approximations further prune long-range intramolecular structures without major changes in predictive power.
- **3D simulations:** By embedding 2D folding trajectories into coarse-grained 3D simulations, we identified sterically infeasible intermediates and additional constraints, such as the difficulty of extending duplexes beyond six stacked pairs, that refine 2D kinetic predictions.
- **Experimental corroboration:** Biophysical measurements on the Japanese encephalitis virus cyclization site and TER constructs confirmed that interactions predicted to be kinetically accessible indeed form, whereas thermodynamically favorable but kinetically blocked pathways do not.
- **Practical software tools:** The open-source RRIkinDP implementation in C++ with Python bindings enables other researchers to take advantage of the new model.

The local-folding screens, the 3D simulation pipeline and the implementation of the interaction design approach are also available on github for public use.

Combined, these contributions provide a modular framework for modeling RNA–RNA interactions in a manner that is both more biophysically realistic and computationally tractable. They lay the basis for improved computational screening for functional interactions, synthetic RNA design, and further mechanistic studies. This thesis also lays the foundation and provides motivation for the open questions outlined in the Outlook section. Addressing these questions can further refine our understanding of RNA kinetics and enable the development of more accurate models that reflect the dynamic nature of RNA interactions in living systems.

Bibliography

[AAD+24] Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016):493–500, June 2024. Publisher: Nature Publishing Group.

[ASW+16] Jong Ghut Ashley Aw, Yang Shen, Andreas Wilm, Miao Sun, Xin Ni Lim, Kum-Loong Boon, Sidika Tapsin, Yun-Shen Chan, Cheng-Peow Tan, Adelene Y. L. Sim, Tong Zhang, Teodorus Theo Susanto, Zhiyan Fu, Niranjan Nagarajan, and Yue Wan. In Vivo Mapping of Eukaryotic RNA Interactomes Reveals Principles of Higher-Order Organization and Regulation. *Molecular Cell*, 62(4):603–617, May 2016.

[AWP+17] Ferhat Alkan, Anne Wenzel, Oana Palasca, Peter Kerpedjiev, Anders Frost Rudebeck, Peter F. Stadler, Ivo L. Hofacker, and Jan Gorodkin. RIsearch2: suffix array-based large-scale prediction of RNA–RNA interactions and siRNA off-targets. *Nucleic Acids Research*, 45(8):e60, May 2017.

[Bar04] David P. Bartel. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell*, 116(2):281–297, January 2004. Publisher: Elsevier.

[BCLW21] Carlos J. Bustamante, Yann R. Chemla, Shixin Liu, and Michelle D. Wang. Optical tweezers in single-molecule biophysics. *Nature Reviews Methods Primers*, 1(1):1–29, March 2021. Publisher: Nature Publishing Group.

[BHW+08] Stephan H. Bernhart, Ivo L. Hofacker, Sebastian Will, Andreas R. Gruber, and Peter F. Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9(1):474, November 2008.

[BK22] Steve L. Bonilla and Jeffrey S. Kieft. The promise of cryo-EM to explore

RNA structural dynamics. *Journal of Molecular Biology*, 434(18):167802, September 2022.

[BLD+16] Michal J. Boniecki, Grzegorz Lach, Wayne K. Dawson, Konrad Tomala, Pawel Lukasz, Tomasz Soltysinski, Kristian M. Rother, and Janusz M. Bujnicki. SimRNA: a coarse-grained method for RNA folding simulations and 3D structure prediction. *Nucleic Acids Research*, 44(7):e63, April 2016.

[BLH23] Stefan Badelt, Ronny Lorenz, and Ivo L Hofacker. DrTransformer: heuristic cotranscriptional RNA folding using the nearest neighbor energy model. *Bioinformatics*, 39(1):btad034, January 2023.

[BMH11] Stephan H Bernhart, Ullrike Mückstein, and Ivo L Hofacker. RNA Accessibility in cubic time. *Algorithms for Molecular Biology : AMB*, 6:3, March 2011.

[BPFC+23] Danny Bergeron, Hermes Paraqindes, Étienne Fafard-Couture, Gabrielle Deschamps-Francoeur, Laurence Faucher-Giguère, Philia Bouchard-Bourelle, Sherif Abou Elela, Frédéric Catez, Virginie Marcel, and Michelle S Scott. snoDB 2.0: an enhanced interactive database, specializing in human snoRNAs. *Nucleic Acids Research*, 51(D1):D291–D296, January 2023.

[BRB08a] Anke Busch, Andreas S. Richter, and Rolf Backofen. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–2856, December 2008.

[BRB08b] Anke Busch, Andreas S. Richter, and Rolf Backofen. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–2856, 10 2008.

[BTM+06] Stephan H. Bernhart, Hakim Tafer, Ulrike Mückstein, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms for Molecular Biology*, 1(1):3, March 2006.

[CCC12] Jarred M. Callura, Charles R. Cantor, and James J. Collins. Genetic switchboard for synthetic biology applications. *Proceedings of the National Academy of Sciences*, 109(15):5850–5855, April 2012. Publisher: Proceedings of the National Academy of Sciences.

[CCJ+20] Zhaokui Cai, Changchang Cao, Lei Ji, Rong Ye, Di Wang, Cong Xia, Sui Wang, Zongchang Du, Naijing Hu, Xiaohua Yu, Juan Chen, Lei Wang, Xianguang Yang, Shunmin He, and Yuanchao Xue. RIC-seq for global in situ profiling of RNA–RNA spatial interactions. *Nature*, 582(7812):432–437, June 2020. Publisher: Nature Publishing Group.

[CLW06] Jasmine HP Chan, Shuhui Lim, and WS Fred Wong. Antisense Oligonucleotides: From Design to Therapeutic Application. *Clinical and Experimental Pharmacology and Physiology*, 33(5-6):533–540,

2006. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1440-1681.2006.04403.x.

[CMR+23] Shani Cohen, Eden Maximof, Shay Rokach, Mor Tadeski, and Isana Veksler-Lublinsky. sInterBase: a comprehensive database of Escherichia coli sRNA–mRNA interactions. *Bioinformatics*, 39(4):btad172, April 2023.

[CRI70] CRICK FRANCIS. Central Dogma of Molecular Biology. *Nature*, 227(5258):561–563, aug 1970. 10.1038/227561a0.

[CT97] K. Y. Chang and I. Tinoco. The structure of an RNA "kissing" hairpin complex of the HIV TAR hairpin loop and its complement. *Journal of Molecular Biology*, 269(1):52–66, May 1997.

[DBS+07] Robert M. Dirks, Justin S. Bois, Joseph M. Schaeffer, Erik Winfree, and Niles A. Pierce. Thermodynamic Analysis of Interacting Nucleic Acid Strands. *SIAM Review*, 49(1):65–88, January 2007. Publisher: Society for Industrial and Applied Mathematics.

[DC02] Jennifer A. Doudna and Thomas R. Cech. The chemical repertoire of natural ribozymes. *Nature*, 418(6894):222–228, July 2002. Publisher: Nature Publishing Group.

[DC20] Philipp J. Dexheimer and Luisa Cochella. MicroRNAs: From Mechanism to Organism. *Frontiers in Cell and Developmental Biology*, 8, June 2020. Publisher: Frontiers.

[DCR+22] Fangyuan Ding, Simona Cocco, Saurabh Raj, Maria Manosas, Thao Thi Thu Nguyen, Michelle M Spiering, David Bensimon, Jean-Francois Allemand, and Vincent Croquette. Displacement and dissociation of oligonucleotides during DNA hairpin closure under strain. *Nucleic Acids Research*, 50(21):12082–12093, December 2022.

[DFH+23] Jie Deng, Xianyang Fang, Lin Huang, Shanshan Li, Lilei Xu, Keqiong Ye, Jinsong Zhang, Kaiming Zhang, and Qiangfeng Cliff Zhang. RNA structure determination: From 2D to 3D. *Fundamental Research*, 3(5):727–737, September 2023.

[ED94] Sean R. Eddy and Richard Durbin. RNA sequence analysis using covariance models. *Nucleic Acids Research*, 22(11):2079–2088, June 1994.

[Edd01] Sean R. Eddy. Non–coding RNA genes and the modern RNA world. *Nature Reviews Genetics*, 2(12):919–929, December 2001. Publisher: Nature Publishing Group.

[FFHS00] Christoph Flamm, Walter Fontana, Ivo L. Hofacker, and Peter Schuster. RNA folding at elementary step resolution. *RNA*, 6(3):325–338, March 2000.

[FRWS03] Boris Fürtig, Christian Richter, Jens Wöhnert, and Harald Schwalbe. NMR Spectroscopy of RNA. *ChemBioChem*, 4(10):936–962, 2003. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cbic.200300700.

[FW09] Mohammad Ali Faghihi and Claes Wahlestedt. Regulatory roles of natural antisense transcripts. *Nature Reviews Molecular Cell Biology*, 10(9):637–643, September 2009. Publisher: Nature Publishing Group.

[FWW+22] Christoph Flamm, Julia Wielach, Michael T. Wolfinger, Stefan Badelt, Ronny Lorenz, and Ivo L. Hofacker. Caveats to Deep Learning Approaches to RNA Secondary Structure Prediction. *Frontiers in Bioinformatics*, 2, July 2022.

[FXM+98] Andrew Fire, SiQun Xu, Mary K. Montgomery, Steven A. Kostas, Samuel E. Driver, and Craig C. Mello. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–811, February 1998. Publisher: Nature Publishing Group.

[GAL+17] Jonathan S. Gootenberg, Omar O. Abudayyeh, Jeong Wook Lee, Patrick Essletzbichler, Aaron J. Dy, Julia Joung, Vanessa Verdine, Nina Donghia, Nichole M. Daringer, Catherine A. Freije, Cameron Myhrvold, Roby P. Bhattacharyya, Jonathan Livny, Aviv Regev, Eugene V. Koonin, Deborah T. Hung, Pardis C. Sabeti, James J. Collins, and Feng Zhang. Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science (New York, N.Y.)*, 356(6336):438–442, April 2017.

[GBK+23] Vrinda Gote, Pradeep Kumar Bolla, Nagavendra Kommineni, Arun Butreddy, Pavan Kumar Nukala, Sushesh Srivatsa Palakurthi, and Wahid Khan. A Comprehensive Review of mRNA Vaccines. *International Journal of Molecular Sciences*, 24(3):2700, January 2023. Number: 3 Publisher: Multidisciplinary Digital Publishing Institute.

[GC09] David P. Giedroc and Peter V. Cornish. Frameshifting RNA pseudoknots: Structure and mechanism. *Virus Research*, 139(2):193–208, February 2009.

[GFW+08] Michael Geis, Christoph Flamm, Michael T. Wolfinger, Andrea Tanzer, Ivo L. Hofacker, Martin Middendorf, Christian Mandl, Peter F. Stadler, and Caroline Thurner. Folding Kinetics of Large {RNAs}. *Journal of Molecular Biology*, 379(1):160–173, 2008.

[GR12] Mitchell Guttman and John L. Rinn. Modular regulatory principles of large non-coding RNAs. *Nature*, 482(7385):339–346, February 2012. Publisher: Nature Publishing Group.

[GSCY14] Alexander A. Green, Pamela A. Silver, James J. Collins, and Peng Yin. Toehold Switches: De-Novo-Designed Regulators of Gene Expression. *Cell*, 159(4):925–939, November 2014. Publisher: Elsevier.

[GSX+18] Jing Gong, Di Shao, Kui Xu, Zhipeng Lu, Zhi John Lu, Yucheng T Yang, and Qiangfeng Cliff Zhang. RISE: a database of RNA interactome from sequencing experiments. *Nucleic Acids Research*, 46(D1):D194–D201, January 2018.

[GT04] Thomas A Geissmann and Danièle Touati. Hfq, a new chaperoning role:

binding to messenger RNA determines access for small RNA regulator. *The EMBO Journal*, 23(2):396–405, January 2004. Num Pages: 405 Publisher: John Wiley & Sons, Ltd.

[HF07] Lance M. Hellman and Michael G. Fried. Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. *Nature Protocols*, 2(8):1849–1861, August 2007. Publisher: Nature Publishing Group.

[HFS+94a] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly*, 125(2):167–188, February 1994.

[HFS+94b] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly*, 125(2):167–188, Feb 1994.

[HLC+22] Hsi-Yuan Huang, Yang-Chi-Dung Lin, Shidong Cui, Yixian Huang, Yun Tang, Jiatong Xu, Jiayang Bao, Yulin Li, Jia Wen, Huali Zuo, Weijuan Wang, Jing Li, Jie Ni, Yini Ruan, Liping Li, Yidan Chen, Yueyang Xie, Zihao Zhu, Xiaoxuan Cai, Xinyi Chen, Lantian Yao, Yigang Chen, Yijun Luo, Shupeng LuXu, Mengqi Luo, Chih-Min Chiu, Kun Ma, Lizhe Zhu, Gui-Juan Cheng, Chen Bai, Ying-Chih Chiang, Liping Wang, Fengxiang Wei, Tzong-Yi Lee, and Hsien-Da Huang. miRTarBase update 2022: an informative resource for experimentally validated miRNA–target interactions. *Nucleic Acids Research*, 50(D1):D222–D230, January 2022.

[HLD+25] Zhian Hu, Shen Ling, Jialin Duan, Zixiao Yu, Yanfei Che, Song Wang, Sichun Zhang, Xinrong Zhang, and Zhengping Li. Proximity-activated guide RNA of CRISPR–Cas12a for programmable diagnostic detection and gene regulation. *Nucleic Acids Research*, 53(3):gkaf017, February 2025.

[HSS98] Ivo L. Hofacker, Peter Schuster, and Peter F. Stadler. Combinatorics of RNA secondary structures. *Discrete Applied Mathematics*, 88(1):207–237, November 1998.

[HT14] Aleksandra Helwak and David Tollervey. Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). *Nature Protocols*, 9(3):711–728, March 2014. Publisher: Nature Publishing Group.

[HZW+20] Bo Hu, Liping Zhong, Yuhua Weng, Ling Peng, Yuanyu Huang, Yongxiang Zhao, and Xing-Jie Liang. Therapeutic siRNA: state of the art. *Signal Transduction and Targeted Therapy*, 5(1):1–25, June 2020. Publisher: Nature Publishing Group.

[JD08] Kirk B. Jensen and Robert B. Darnell. CLIP: Crosslinking and ImmunoPrecipitation of In Vivo RNA Targets of RNA-Binding Proteins. In Ren-Jang Lin, editor, *RNA-Protein Interaction Protocols*, pages 85–98. Humana Press, Totowa, NJ, 2008.

[JPK20] Mikkel Girke Jørgensen, Jens Sivkær Pettersen, and Birgitte H. Kallipolitis. sRNA-mediated control in bacteria: An increasing diversity of regulatory mechanisms. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1863(5):194504, May 2020.

[JSR23] Ryland W. Jackson, Claire M. Smathers, and Aaron R. Robart. General Strategies for RNA X-ray Crystallography. *Molecules*, 28(5):2111, January 2023. Number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[KAA+23] Dorian Kariuki, Kesava Asam, Bradley E Aouizerat, Kimberly A Lewis, Jose C Florez, and Elena Flowers. Review of databases for experimentally validated human microRNA–mRNA interactions. *Database*, 2023:baad014, January 2023.

[Kis01] Tamás Kiss. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *The EMBO Journal*, 20(14):3617–3622, July 2001. Num Pages: 3622 Publisher: John Wiley & Sons, Ltd.

[KTH+22a] Juanjuan Kang, Qiang Tang, Jun He, Le Li, Nianling Yang, Shuiyan Yu, Mengyao Wang, Yuchen Zhang, Jiahao Lin, Tianyu Cui, Yongfei Hu, Puwen Tan, Jun Cheng, Hailong Zheng, Dong Wang, Xi Su, Wei Chen, and Yan Huang. RNAInter v4.0: RNA interactome repository with redefined confidence scoring system and improved accessibility. *Nucleic Acids Research*, 50(D1):D326–D332, January 2022.

[KTH+22b] Juanjuan Kang, Qiang Tang, Jun He, Le Li, Nianling Yang, Shuiyan Yu, Mengyao Wang, Yuchen Zhang, Jiahao Lin, Tianyu Cui, Yongfei Hu, Puwen Tan, Jun Cheng, Hailong Zheng, Dong Wang, Xi Su, Wei Chen, and Yan Huang. RNAInter v4.0: RNA interactome repository with redefined confidence scoring system and improved accessibility. *Nucleic Acids Res.*, 50(D1):D326–D332, January 2022.

[LBH+11] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[LBHzS+11] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, November 2011.

[Lev69] Cyrus Levinthal. How to fold graciously. In *Mössbaun Spectroscopy in Biological Systems Proceedings, Univ. of Illinois Bulletin*, volume 67, pages 22–24, 1969.

[LLH+16] Ronny Lorenz, Dominik Luntzer, Ivo L. Hofacker, Peter F. Stadler, and Michael T. Wolfinger. SHAPE directed RNA folding. *Bioinformatics*, 32(1):145–147, January 2016.

[LMT13] Ganqiang Liu, John S. Mattick, and Ryan J. Taft. A meta-analysis of

the genomic and transcriptomic composition of complex life. *Cell Cycle*, 12(13):2061–2072, July 2013.

[LMZ+15] Jianwei Li, Wei Ma, Pan Zeng, Junyi Wang, Bin Geng, Jichun Yang, and Qinghua Cui. LncTar: a tool for predicting the RNA targets of long noncoding RNAs. *Briefings in Bioinformatics*, 16(5):806–812, September 2015.

[LXL+19] Si Jie Lu, Juan Xie, Yang Li, Bin Yu, Qin Ma, and Bing Qiang Liu. Identification of lncRNAs-gene interactions in transcription regulation based on co-expression analysis of RNA-seq data. *Mathematical biosciences and engineering: MBE*, 16(6):7112–7125, August 2019.

[LZL+16] Zhipeng Lu, Qiangfeng Cliff Zhang, Byron Lee, Ryan A. Flynn, Martin A. Smith, James T. Robinson, Chen Davidovich, Anne R. Gooding, Karen J. Goodrich, John S. Mattick, Jill P. Mesirov, Thomas R. Cech, and Howard Y. Chang. RNA duplex map in living cells reveals higher order transcriptome structure. *Cell*, 165(5):1267–1279, May 2016.

[McC90a] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990. Publisher: Wiley Subscription Services, Inc., A Wiley Company.

[McC90b] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990.

[MJZS22] Haiyun Ma, Xinyu Jia, Kaiming Zhang, and Zhaoming Su. Cryo-EM advances in RNA structure determination. *Signal Transduction and Targeted Therapy*, 7(1):1–6, February 2022. Publisher: Nature Publishing Group.

[MLY22] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. TargetNet: functional microRNA target prediction with deep neural networks. *Bioinformatics*, 38(3):671–677, January 2022.

[MM06] John S. Mattick and Igor V. Makunin. Non-coding RNA. *Human Molecular Genetics*, 15(suppl_1):R17–R29, April 2006.

[MPW+22] Tyler Mrozowich, Sean M. Park, Maria Waldl, Amy Henrickson, Scott Tersteeg, Corey R. Nelson, Anneke Deklerk, Borries Demeler, Ivo L. Hofacker, Michael T. Wolfinger, and Trushar R. Patel. Investigating RNA-RNA interactions through computational and biophysical analysis, December 2022. Pages: 2022.02.01.478553 Section: New Results.

[MSP19] Maja Marušič, Judith Schlagnitweit, and Katja Petzold. RNA Dynamics by NMR Spectroscopy. *ChemBioChem*, 20(21):2685–2710, 2019. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cbic.201900072.

[MSZT99] David H. Mathews, Jeffrey Sabina, Michael Zuker, and Douglas H. Turner. Expanded sequence dependence of thermodynamic parameters improves

prediction of RNA secondary structure1. *Journal of Molecular Biology*, 288(5):911–940, May 1999.

[MTH+06] Ulrike Mückstein, Hakim Tafer, Jörg Hackermüller, Stephan H. Bernhart, Peter F. Stadler, and Ivo L. Hofacker. Thermodynamics of RNA–RNA binding. *Bioinformatics*, 22(10):1177–1182, May 2006.

[MTM24] Abhinav Mittal, Douglas H. Turner, and David H. Mathews. NNDB: An Expanded Database of Nearest Neighbor Parameters for Predicting Stability of Nucleic Acid Secondary Structures. *Journal of Molecular Biology*, 436(17):168549, September 2024.

[MTP23] Maja Marušič, Maria Toplishek, and Janez Plavec. NMR of RNA - Structure and interactions. *Current Opinion in Structural Biology*, 79:102532, April 2023.

[MW08] Stefanie A. Mortimer and Kevin M. Weeks. Time-Resolved RNA SHAPE Chemistry. *Journal of the American Chemical Society*, 130(48):16178–16180, December 2008. Publisher: American Chemical Society.

[MWB17] Martin Mann, Patrick R. Wright, and Rolf Backofen. IntaRNA 2.0: enhanced and customizable prediction of RNA–RNA interactions. *Nucleic Acids Research*, 45(W1):W435–W439, July 2017.

[NCY+16] Tri C. Nguyen, Xiaoyi Cao, Pengfei Yu, Shu Xiao, Jia Lu, Fernando H. Biase, Bharat Sridhar, Norman Huang, Kang Zhang, and Sheng Zhong. Mapping RNA–RNA interactome and RNA structure in vivo by MARIO. *Nature Communications*, 7:12023, June 2016.

[NDF+13] Olga A. Nikolaitchik, Kari A. Dilley, William Fu, Robert J. Gorelick, S.-H. Sheldon Tai, Ferri Soheilian, Roger G. Ptak, Kunio Nagashima, Vinay K. Pathak, and Wei-Shau Hu. Dimeric RNA Recognition Regulates HIV-1 Genome Packaging. *PLOS Pathogens*, 9(3):e1003249, March 2013. Publisher: Public Library of Science.

[Nil13] Timothy W. Nilsen. Toeprinting. *Cold Spring Harbor Protocols*, 2013(9):pdb.prot077180, September 2013. Publisher: Cold Spring Harbor Laboratory Press.

[NJ80] Nussinov R and Jacobson A B. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 77(11):6309–6313, nov 1980.

[NPGK78a] Ruth Nussinov, George Pieczenik, Jerrold R. Griggs, and Daniel J. Kleitman. Algorithms for Loop Matchings. *SIAM Journal on Applied Mathematics*, 35(1):68–82, 1978. _eprint: https://doi.org/10.1137/0135006.

[NPGK78b] Ruth Nussinov, George Pieczenik, Jerrold R. Griggs, and Daniel J. Kleitman. Algorithms for Loop Matchings. *SIAM Journal on Applied Mathematics*, 35(1):68–82, 1978.

[NSAB+17] Wy Ching Ng, Ruben Soto-Acosta, Shelton S. Bradrick, Mariano A. Garcia-Blanco, and Eng Eong Ooi. The 5' and 3' Untranslated Regions of the Flaviviral Genome. *Viruses*, 9(6):137, June 2017.

[OMMS22] Vaitea Opuu, Nono S. C. Merleau, Vincent Messow, and Matteo Smerlak. RAFFT: Efficient prediction of RNA folding pathways using the fast Fourier transform. *PLOS Computational Biology*, 18(8):e1010448, August 2022. Publisher: Public Library of Science.

[PHPW18] Norbert Pardi, Michael J. Hogan, Frederick W. Porter, and Drew Weissman. mRNA vaccines — a new era in vaccinology. *Nature Reviews Drug Discovery*, 17(4):261–279, April 2018. Publisher: Nature Publishing Group.

[PKB+12] Jordan Pischimarov, Carsten Kuenne, André Billion, Jüergen Hemberger, Franz Cemič, Trinad Chakraborty, and Torsten Hain. sRNAdb: A small non-coding RNA database for gram-positive bacteria. *BMC Genomics*, 13(1):384, August 2012.

[PMYK19] Alisha Parveen, Syed H. Mustafa, Pankaj Yadav, and Abhishek Kumar. Applications of Machine Learning in miRNA Discovery and Target Prediction. *Current Genomics*, 20(8):537–544, December 2019. Publisher: Bentham Science Publishers.

[RB12] Andreas S. Richter and Rolf Backofen. Accessibility and conservation: general features of bacterial small RNA-mRNA interactions? *RNA Biol.*, 9(7):954–965, July 2012.

[RC12] John L. Rinn and Howard Y. Chang. Genome Regulation by Long Noncoding RNAs. *Annual Review of Biochemistry*, 81(Volume 81, 2012):145–166, July 2012. Publisher: Annual Reviews.

[RCE17] Elena Rivas, Jody Clements, and Sean R. Eddy. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature Methods*, 14(1):45–48, January 2017.

[RGB09] Francis E. Reyes, Andrew D. Garst, and Robert T. Batey. Chapter 6 - Strategies in RNA Crystallography. In *Methods in Enzymology*, volume 469 of *Biophysical, Chemical, and Functional Probes of RNA Structure, Interactions and Folding: Part B*, pages 119–139. Academic Press, January 2009.

[RM10] Jessica S. Reuter and David H. Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11(1):129, March 2010.

[RSHG04] Marc Rehmsmeier, Peter Steffen, Matthias Höchsmann, and Robert Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–1517, October 2004. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold

Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[RW54] Alexander Rich and J. D. Watson. Some Relations between DNA and RNA. *Proceedings of the National Academy of Sciences of the United States of America*, 40(8):759–764, 1954. Publisher: National Academy of Sciences.

[RWC+24] Tom Rappol, Maria Waldl, Anastasia Chugunova, Ivo L Hofacker, Andrea Pauli, and Elisa Vilardo. tRNA expression and modification landscapes, and their dynamics during zebrafish embryo development. *Nucleic Acids Research*, 52(17):10575–10594, September 2024.

[SCW15] Matthew J. Smola, J. Mauro Calabrese, and Kevin M. Weeks. Detection of RNA–Protein Interactions in Living Cells with SHAPE. *Biochemistry*, 54(46):6867–6875, November 2015. Publisher: American Chemical Society.

[SDH+14] Felix Spenkuch, Olwen Domingo, Gerald Hinze, Thomas Basché, and Mark Helm. Studying RNA Using Single Molecule Fluorescence Resonance Energy Transfer. In *Handbook of RNA Biochemistry*, pages 499–526. John Wiley & Sons, Ltd, 2014. Section: 24 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/9783527647064.ch24.

[SGCL+23] Heladia Salgado, Socorro Gama-Castro, Paloma Lara, Citlalli Mejia-Almonte, Gabriel Alarcón-Carranza, Andrés G. López-Almazo, Felipe Betancourt-Figueroa, Pablo Peña-Loredo, Shirley Alquicira-Hernández, Daniela Ledezma-Tejeida, Lizeth Arizmendi-Zagal, Francisco Mendez-Hernandez, Ana K. Diaz-Gomez, Elizabeth Ochoa-Praxedis, Luis J. Muñiz-Rascado, Jair S. García-Sotelo, Fanny A. Flores-Gallegos, Laura Gómez, César Bonavides-Martínez, Víctor M. del Moral-Chávez, Alfredo J. Hernández-Alvarez, Alberto Santos-Zavaleta, Salvador Capella-Gutierrez, Josep Lluis Gelpi, and Julio Collado-Vides. RegulonDB v12.0: a comprehensive resource of transcriptional regulation in E. coli K-12. *Nucleic Acids Res.*, 52(D1):D255, November 2023.

[SLSL23] Renfu Shang, Seungjae Lee, Gayan Senavirathne, and Eric C. Lai. microRNAs in action: biogenesis, function and regulation. *Nature Reviews Genetics*, 24(12):816–833, December 2023. Publisher: Nature Publishing Group.

[SOS+13] Niranjan Srinivas, Thomas E. Ouldridge, Petr Sulc, Joseph M. Schaeffer, Bernard Yurke, Ard A. Louis, Jonathan P. K. Doye, and Erik Winfree. On the biophysics and kinetics of toehold-mediated DNA strand displacement. *Nucleic Acids Research*, 41(22):10641–10658, December 2013.

[SRB+15] Matthew J. Smola, Greggory M. Rice, Steven Busan, Nathan A. Siegfried, and Kevin M. Weeks. Selective 2'-hydroxyl acylation analyzed by primer extension and mutational profiling (SHAPE-MaP) for direct, versatile and accurate RNA structure analysis. *Nature Protocols*, 10(11):1643–1669, November 2015. Publisher: Nature Publishing Group.

[SS04] Svetlana A. Shabalina and Nikolay A. Spiridonov. The mammalian transcriptome and the function of non-coding DNA sequences. *Genome Biology*, 5(4):105, March 2004.

[SSB+23] Bohdan Schneider, Blake Alexander Sweeney, Alex Bateman, Jiri Cerny, Tomasz Zok, and Marta Szachniuk. When will RNA get its AlphaFold moment? *Nucleic Acids Research*, 51(18):9522–9532, October 2023.

[SSWOB16] Eesha Sharma, Tim Sterne-Weiler, Dave O'Hanlon, and Benjamin J. Blencowe. Global Mapping of Human RNA-RNA Interactions. *Molecular Cell*, 62(4):618–626, May 2016.

[SV21] Richard A Schäfer and Björn Voß. RNAnue: efficient data analysis for RNA–RNA interactomics. *Nucleic Acids Research*, 49(10):5493–5501, June 2021.

[SWD+22] Marcell Szikszai, Michael Wise, Amitava Datta, Max Ward, and David H. Mathews. Deep learning models for RNA secondary structure prediction (probably) do not generalize across families. *Bioinformatics (Oxford, England)*, 38(16):3892–3899, August 2022.

[SZWL09] Jesse Stombaugh, Craig L. Zirbel, Eric Westhof, and Neocles B. Leontis. Frequency and isostericity of RNA base pairs. *Nucleic Acids Research*, 37(7):2294–2312, April 2009.

[TBD+73] Ignacio Tinoco, Philip N. Borer, Barbara Dengler, Mark D. Levine, Olke C. Uhlenbeck, Donald M. Crothers, and Jay Gralla. Improved Estimation of Secondary Structure in Ribonucleic Acids. *Nature New Biology*, 246(150):40–41, November 1973. Publisher: Nature Publishing Group.

[TCKD14] Siqi Tian, Pablo Cordero, Wipapat Kladwang, and Rhiju Das. High-throughput mutate-map-rescue evaluates SHAPE-directed RNA structure and uncovers excited states. *RNA*, 20(11):1815–1826, November 2014. Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab.

[TS23] Jan-Niklas Tants and Andreas Schlundt. Advances, Applications, and Perspectives in Small-Angle X-ray Scattering of RNA. *ChemBioChem*, 24(17):e202300110, 2023. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cbic.202300110.

[TUL71] Ignacio Tinoco, Olke C. Uhlenbeck, and Mark D. Levine. Estimation of Secondary Structure in Ribonucleic Acids. *Nature*, 230(5293):362–367, April 1971. Publisher: Nature Publishing Group.

[UG17] Sinan Ugur Umu and Paul P. Gardner. A comprehensive benchmark of RNA-RNA interaction prediction tools for all domains of life. *Bioinformatics (Oxford, England)*, 33(7):988–996, April 2017.

[UP90] Eugen Uhlmann and Anusch Peyman. Antisense oligonucleotides: a new therapeutic principle. *Chemical Reviews*, 90(4):543–584, June 1990. Publisher: American Chemical Society.

[UZS16] Taylor B. Updegrove, Aixia Zhang, and Gisela Storz. Hfq: the flexible RNA matchmaker. *Current opinion in microbiology*, 30:133–138, April 2016.

[VBC+25] Julie Schock Vaiani, Mans Broekgaarden, Jean-Luc Coll, Lucie Sancey, and Benoit Busser. In vivo vectorization and delivery systems for gene therapies and RNA-based therapeutics in oncology. *Nanoscale*, 17(10):5501–5525, March 2025. Publisher: The Royal Society of Chemistry.

[VG09] Sergio M. Villordo and Andrea V. Gamarnik. Genome cyclization as strategy for flavivirus RNA replication. *Virus Research*, 139(2):230–239, February 2009.

[VL11] Jörg Vogel and Ben F. Luisi. Hfq and its constellation of RNA. *Nature Reviews. Microbiology*, 9(8):578–589, August 2011.

[VM00] Gabriele Varani and William H McClain. The G·U wobble base pair. *EMBO reports*, 1(1):18–23, July 2000. Num Pages: 23 Publisher: John Wiley & Sons, Ltd.

[VTPM+18] Briana Van Treeck, David S. W. Protter, Tyler Matheny, Anthony Khong, Christopher D. Link, and Roy Parker. RNA self-assembly contributes to stress granule formation and defining the stress granule transcriptome. *Proceedings of the National Academy of Sciences*, 115(11):2734–2739, March 2018. Publisher: Proceedings of the National Academy of Sciences.

[Wal17] Maria Waldl. Kinetics of RNA Structure Formation - Folding Pathways and Evolution. Master's thesis, University of Vienna, 2017.

[Wat78] Michael S Waterman. Secondary structure of single-stranded nucleic acids. *Studies in Foundations and Combinatorics*, 1:167–212, 1978.

[WCZ+18] Ming Wen, Peisheng Cong, Zhimin Zhang, Hongmei Lu, and Tonghua Li. DeepMirTar: a deep-learning approach for predicting human miRNA targets. *Bioinformatics*, 34(22):3781–3787, November 2018.

[Wee10] Kevin M Weeks. Advances in RNA structure analysis by chemical probing. *Current opinion in structural biology*, 20(3):295–304, June 2010.

[WHB+18] Wei-Sheng Wu, Wei-Che Huang, Jordan S Brown, Donglei Zhang, Xiaoyan Song, Hao Chen, Shikui Tu, Zhiping Weng, and Heng-Chi Lee. pirScan: a webserver to predict piRNA targeting sites and to avoid transgene silencing in C. elegans. *Nucleic Acids Research*, 46(Web Server issue):W43–W48, July 2018.

[WLZ+16] Jiang Wang, Tao Liu, Bo Zhao, Qixuan Lu, Zheng Wang, Yuan Cao, and Wuju Li. sRNATarBase 3.0: an updated database for sRNA-target

interactions in bacteria. *Nucleic Acids Research*, 44(D1):D248–D253, January 2016.

[WML+07] Jin-Der Wen, Maria Manosas, Pan T. X. Li, Steven B. Smith, Carlos Bustamante, Felix Ritort, and Ignacio Tinoco. Force Unfolding Kinetics of RNA Using Optical Tweezers. I. Effects of Experimental Variables on Measured Results. *Biophysical Journal*, 92(9):2996–3009, May 2007. Publisher: Elsevier.

[WRP+13] Patrick R. Wright, Andreas S. Richter, Kai Papenfort, Martin Mann, Jörg Vogel, Wolfgang R. Hess, Rolf Backofen, and Jens Georg. Comparative genomics boosts target prediction for bacterial small RNAs. *Proceedings of the National Academy of Sciences*, 110(37):E3487–E3496, September 2013. Publisher: Proceedings of the National Academy of Sciences.

[WS78a] M. S. Waterman and T. F. Smith. RNA secondary structure: a complete mathematical analysis. *Mathematical Biosciences*, 42(3):257–266, December 1978.

[WS78b] M.S. Waterman and T.F. Smith. RNA secondary structure: a complete mathematical analysis. *Mathematical Biosciences*, 42(3):257–266, 1978.

[WS78c] M.S. Waterman and T.F. Smith. RNA secondary structure: a complete mathematical analysis. *Mathematical Biosciences*, 42(3):257–266, 1978.

[WTO+18] Maria Waldl, Bernhard C. Thiel, Roman Ochsenreiter, Alexander Holzenleiter, João Victor De Araujo Oliveira, Maria Emília M. T. Walter, Michael T. Wolfinger, and Peter F. Stadler. TERribly Difficult: Searching for Telomerase RNAs in Saccharomycetes. *Genes*, 9(8):372, August 2018. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute.

[WWL09] Markus C. Wahl, Cindy L. Will, and Reinhard Lührmann. The Spliceosome: Design Principles of a Dynamic RNP Machine. *Cell*, 136(4):701–718, February 2009. Publisher: Elsevier.

[XBTI03] A. Xayaphoummine, T. Bucher, F. Thalmann, and H. Isambert. Prediction and statistics of pseudoknots in RNA structures using exactly clustered stochastic simulations. *Proceedings of the National Academy of Sciences*, 100(26):15310–15315, 2003. _eprint: http://www.pnas.org/content/100/26/15310.full.pdf.

[YB22] Helen Yakhnin and Paul Babitzke. Toeprint Assays for Detecting RNA Structure and Protein–RNA Interactions. In Eveline Peeters and Indra Bervoets, editors, *Prokaryotic Gene Regulation: Methods and Protocols*, pages 305–316. Springer US, New York, NY, 2022.

[YMB+24] Hua-Ting Yao, Bertrand Marchand, Sarah J. Berkemer, Yann Ponty, and Sebastian Will. Infrared: a declarative tree decomposition-powered framework for bioinformatics. *Algorithms for Molecular Biology*, 19(1):13, March 2024.

[YMMS+21] Wencheng Yin, Luis Mendoza, Jimena Monzon-Sandoval, Araxi O. Urrutia, and Humberto Gutierrez. Emergence of co-expression in gene regulatory networks. *PLoS ONE*, 16(4):e0247671, April 2021.

[YNG13] Luke A. Yates, Chris J. Norbury, and Robert J. C. Gilbert. The Long and Short of MicroRNA. *Cell*, 153(3):516–519, April 2013. Publisher: Elsevier.

[YPFK06] Peter Yakovchuk, Ekaterina Protozanova, and Maxim D. Frank-Kamenetskii. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research*, 34(2):564–574, January 2006.

[YPW24] Hua-Ting Yao, Yann Ponty, and Sebastian Will. Developing Complex RNA Design Applications in the Infrared Framework. In Ronny Lorenz, editor, *RNA Folding: Methods and Protocols*, pages 285–313. Springer US, New York, NY, 2024.

[YZL+17] Ying Yi, Yue Zhao, Chunhua Li, Lin Zhang, Huiying Huang, Yana Li, Lanlan Liu, Ping Hou, Tianyu Cui, Puwen Tan, Yongfei Hu, Ting Zhang, Yan Huang, Xiaobo Li, Jia Yu, and Dong Wang. RAID v2.0: an updated resource of RNA-associated interactions across organisms. *Nucleic Acids Research*, 45(D1):D115–D118, January 2017.

[ZLT+23] Yu Zheng, Huaxia Luo, Xueyi Teng, Xinpei Hao, Xiaoyu Yan, Yiheng Tang, Wanyu Zhang, Yuanxin Wang, Peng Zhang, Yanyan Li, Yi Zhao, Runsheng Chen, and Shunmin He. NPInter v5.0: ncRNA interaction database in a new era. *Nucleic Acids Research*, 51(D1):D232–D239, January 2023.

[ZS81a] Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, January 1981.

[ZS81b] Zuker M and Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research*, 9(1):133–148, jan 1981.

[ZS84] Michael Zuker and David Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4):591–621, Jul 1984.