universität wien

DISSERTATION / DOCTORAL THESIS

Titel der Dissertation /Title of the Doctoral Thesis

Control of RNA function by conformational design

verfasst von / submitted by

Mag. rer. nat. Stefan Badelt

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

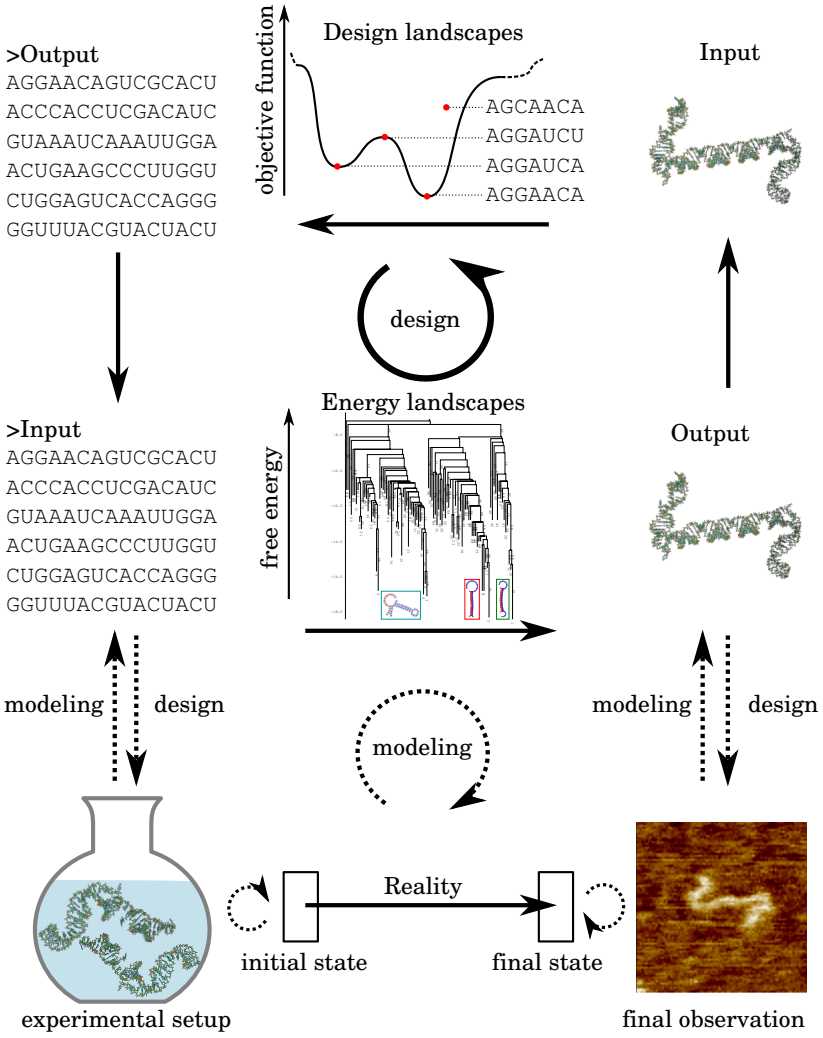
Doctor of Philosophy (PhD)

Wien, 2016 / Vienna, 2016

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

CONTROL OF RNA FUNCTION BY CONFORMATIONAL DESIGN

STEFAN BADELT



Cycling in energy and design landscapes

Stefan Badelt: *Control of RNA function by conformational design*: Cycling in energy and design landscapes

A thesis submitted in partial fulfillment of the requirements for the degree of:

Ph.D.

at the

Institut für Theoretische Chemie
Fakultät für Chemie
Universität Wien

ABSTRACT

RNAs play an essential role in the life cycle of every cell. RNA is not only the intermediate between the genetic blueprint (DNA) and the proteins produced, but also performs a variety of regulatory tasks. The function of an RNA molecule is determined by its structure, which can be reasonably well predicted following the biophysical rules implemented in the most popular RNA structure prediction programs. I present four projects, two of them in form of a peer-reviewed publication, the other two are unpublished work with preliminary results.

The first publication describes how RNAs with catalytic function, ribozymes, can be designed to concatenate multiple copies of themselves (i. e. they *self-polymerize*), into longer molecules. This is important since the RNA World hypothesis claims that RNA emerged before DNA and proteins. It has, however, been hard to imagine how sufficiently long RNA molecules could exist in the RNA world. Our results suggest how pre-biological RNA genomes may have been built up by concatenation of shorter sequences.

The second publication shows conformational switching of RNAs through the interaction of two copies. Such a *conformational self-replication* was so far known only from proteins, where it forms the molecular basis of prion diseases such as Creutzfeldt-Jakob. The artificial RNAs designed in this project could help to better understand the mechanism of such diseases, but might also be useful as molecular sensors and amplifiers in biotechnology.

A central challenge to both publications mentioned above are RNA-RNA interactions. While we have used thermodynamic criteria combined with existing algorithms for intramolecular folding kinetics to design sequences, we are now developing new algorithms to model folding kinetics of interacting RNAs. This problem is much more complicated by the fact that intermolecular base-pairing is concentration dependent. Preliminary results to model the kinetics of small interacting RNAs are promising and will serve as a basis for a separate, peer-reviewed publication.

The last major topic addressed in this thesis concerns the synthesis of RNA molecules in cells, i. e. cotranscriptional folding. We first show how small metabolites can be included into an existing algorithm to model intramolecular folding during transcription. We have published these results also in the context of a recent book chapter on the computational modeling of riboswitches. However, the approach is limited to short RNA transcripts and we have now developed a faster heuristic to model cotranscriptional folding for longer molecules. The results presented from this new program, `DrTransformer`, will also be used in a separate, peer-reviewed publication.

ZUSAMMENFASSUNG

RNAs (Ribonukleinsäuren) spielen eine tragende Rolle im Lebenszyklus jeder Zelle. Sie bilden das Bindeglied zwischen dem genetischen Bauplan (DNA) und den daraus erzeugten Proteinen und übernehmen eine Vielzahl von regulatorischen Aufgaben. Diese Aufgaben werden großteils von der Molekülstruktur bestimmt, welche wiederum auf Grund experimenteller Daten und der daraus abgeleiteten physikalischen Regeln vorhergesagt werden kann. Im Rahmen dieser Dissertation werde ich vier Projekte vorstellen, die sich mit RNA Design und der Faltungskinetik interagierender RNAs beschäftigen. Zwei der Projekte sind bereits in öffentlichen Journalen erschienen, für die beiden anderen werden vorläufige Resultate präsentiert die als Basis für eine separate Publikationen dienen.

Die erste Publikation beschäftigt sich mit dem Design von RNAs mit katalytischer Funktion, sogenannten Ribozymen. Das Augenmerk liegt dabei auf Sequenzen die sich selbst prozessieren können, sodass mehrere Kopien des selben Moleküls aneinandergehängt werden. Die Resultate zeigen das Potential von RNAs im Ursprung des Lebens, nämlich dass präbiologische RNA-Genome aus kürzeren Sequenzen gebaut werden konnten.

In der zweiten Publikation wurden RNAs designed, die die Möglichkeit haben Kopien von sich selbst von einer aktiven Struktur in eine andere umzufalten. Solche Mechanismen sind bisher nur für Proteine beschrieben worden. Die sogenannten Prionen sind Auslöser neurologischer Erkrankungen, z.B. von Creutzfeldt-Jakob. Biotechnologish, können diese RNAs als molekulare Sensoren eingesetzt werden, aber sie können auch dabei helfen die molekularen Mechanismen von Prionen-Krankheitserregern besser zu verstehen.

Eine zentrale Herausforderung der oben genannten Publikationen bestand darin RNA-RNA Interaktionen zu modellieren. Die Konzentrationsabhängikeit der beteiligten Moleküle macht eine genaue Vorhersage der kinetischen Prozesse komplizierter als für intramolekulare Faltung. Die Entwicklung eines Programs zur Modelierung von intermolekluarer Faltunskinetik, ermöglicht detailierte Simulationen von kurzen interagierenden RNAs.

Das letzte Kapitel betrifft die Synthese von RNA in der Zelle, i. e. co-transkriptionelle Faltung. Anhand eines Beispiels wird beschrieben, wie man den Einfluss kleiner Metaboliten in die Simulation von co-transcriptioneller Faltung einbeziehen kann. Das Beispiel wurde im Kontext eines Buchkapitels zur Modellierung von Riboswitches verwendet. Zusätzlich wird hier ein neues Programm präsentiert, `DrTransformer`, um co-transcriptionelle Faltung auch für längere RNAs mit höherer Genauigkeit vorherzusagen, als es bisher möglich war.

ACKNOWLEDGMENTS

At first I want to thank my supervisor *Ivo Hofacker* for guiding me through my PhD studies and always being helpful with good advice. He gave me the freedom to follow my own research interests and, at the same time, pointed out different directions when I got stuck. He has changed my perspective on describing biological problems by explaining processes in the context of chemical and physical laws.

During my PhD studies *Christoph Flamm* has been a constant source of inspiration and ideas. A lot of this work builds on his previous publications and he has impressed me with his playful approach to formulate scientific problems as small, exciting challenges, or vice versa.

I want to use this possibility to thank my love *Anela Tosevska*. Our adventures when we are traveling as well as our discussions on lazy days have given me the energy to focus on my work in rough times. Also, she has greatly improved this thesis with critical feedback.

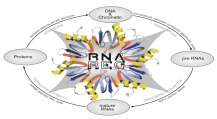
The RNA design projects build on theoretical models that have been developed during fruitful discussions with my PhD Committee: *Sabine Müller* and *Peter Stadler*. New ideas regarding RNA folding kinetics have mostly been discussed first with my colleague *Ronny Lorenz*. He has always been happy to share his detailed knowledge on RNA folding algorithms and their implementation in the `ViennaRNA package`. I want to thank *Andrea Tanzer* and *Michael Wolfinger* for involving me into a project about cotranscriptional design of large RNAs, which eventually led to the idea for the `DrTransformer` algorithm described in this thesis. My coworkers *Sonja Petkovic*, *Stephan Block*, *Stefan Hammer* and *Peter Kerpedjiev* have made essential contributions to the work presented here. Sonja Petkovic and Stephan Block have experimentally characterized the folding of self-processing ribozymes (Chapter 6), Stefan Hammer has written the front-end for an RNA design webservice (Chapter 5) and Peter Kerpedjiev has implemented visualization software for cotranscriptional folding (Chapter 4). I have had a great time with my colleagues at the TBI: *Florian Eggenhofer*, *Joerg Fallmann*, *Fabian Amman*, *Marcel Kucharik*, *Roman Ochsenreiter*, *Sven Findeiß*, *Craig Zirbel*, *Maria Waldl*, *Judith Ivansits*, *Richard Neuböck*, *Dominik Steininger*, *Cristina Wagner* and *Bernhard Thiel*. You guys are awesome.

I am blessed with a great family. My mother *Doris Badelt* and my father *Felix Badelt* have supported me (also financially) in my decision to study which now resulted in this thesis. I am also thankful to *Walter Karban*, *Evelyne Badelt*, *Christoph Badelt* and *Alexander Badelt* for their moral support, and I want to thank all my friends that I did not mention here.

Many thanks also to *Judith Ivansits*, *Nicola Wiskocil* and *Gerlinde Aschauer*, who have helped me with organizational challenges, travel refunding, and all sorts of bureaucratic pitfalls that otherwise would have dominated my life as a scientist. *Richard Neuböck* maintains an incredibly stable infrastructure at the TBI and has helped me to restore important data after I removed it `'--really-fast'`.

FUNDING

The research presented in this thesis was funded in parts by the FWF International Programme I670, the DK RNA program FG748004 and the FWF project "SFB F43 RNA regulation of the transcriptome".



CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

ACRONYMS

AFM	atomic force microscopy
DNA	deoxy-ribonucleic acid
DP	dynamic programming
EFE	ensemble free energy
MFE	minimum free energy
NN	nearest neighbor
ODE	ordinary differential equation
PCR	polymerase chain reaction
RNA	ribonucleic acid
gRNA	guide-RNA
mRNA	messenger-RNA
tRNA	transfer-RNA
rRNA	ribosomal-RNA
miRNA	micro-RNA

Part I

INTRODUCTION

1 FROM MOLECULAR TO SYNTHETIC BIOLOGY

Our understanding of molecular biological processes increases steadily with new experimental technologies and their combination with theoretical models. Experiments range from characterization of single and interacting molecules to large-scale analysis of whole metabolic networks. In combination with today's computational power, theoretical models can be tested *in silico* to analyze, formalize, and visualize the behavior of living systems.

This thesis contributes to the field of biological engineering, which is at the interface of molecular biology, biochemistry, physics and computer science. Biological processes are formulated as algorithms, which requires a well defined set of assumptions and rules in oder to predict behavior of molecules in artificial context. As it is common practice in computer science, biological engineering lives from the *debugging* and problem solving after an artificial biological process has been tested experimentally. This often reveals important little details that did not get enough attention during the initial design process.

As the modeling of living matter always contains uncertainties arising from stochastic (*random*) elements, errors do accumulate. However, if the details are modeled with reasonable accuracy, then theoretical models are applicable for bigger, more complex systems and they can be combined to explain observations that seem to be arbitrary in the first place.

This thesis is split into two parts. First, I combine algorithms for single ribonucleic acid (RNA) folding kinetics with thermodynamic modeling of interacting RNAs and I develop new algorithms for modeling RNA folding kinetics during RNA synthesis. Second, I demonstrate intrinsic RNA mechanisms that may have increased functional diversity in the origin of life and that can be used as building blocks during metabolic engineering. However, before providing details about the theoretical background, this chapter presents some basic concepts in molecular biology that put the work into a broader context.

1.1 A SNAPSHOT OF LIFE AND ITS MOLECULES

Three types of molecules are believed to be central for all known life forms: deoxyribonucleic acids (DNAs), RNAs and proteins. The genetic blueprint DNA encodes all the information to orchestrate a genetic program, from the first cell in a new living species to the (controlled) cell death when sufficient time has passed. The information to develop a human being, for instance, is encoded in a DNA double helix with roughly

three billion ($3 \cdot 10^9$) base-pairs. In order to pack this information into human cells, the DNA is compressed into 23 chromosomes, that have to be selectively decoded when new information is needed.

RNA is the first layer of decoding the genetic material. It is produced from a template DNA in a process called *transcription*: one or more proteins assemble at a particular chromosomal region, unpack it, unwind the double-helix structure, and then transcribe the nucleotide sequence into an RNA molecule. The produced RNA molecule can be of variable length (up to multiple thousand nucleotides), it can get chopped into smaller pieces, it can get chemically modified, and it can have a multitude of different functions which will be discussed below.

Proteins are synthesized from RNAs in a process called *translation*. Compared to RNAs, proteins are chemically more diverse, their production requires more energy, and their lifetime is generally longer. Proteins build complex structures for cellular communication, energy production, recycling, chromosome packing, spacial arrangements in cells, cellular transport and many more fundamental processes. Figuratively speaking, proteins are highly specialized machines that are produced and often operated by short-lived RNAs.

The functional repertoire of an RNA is determined by two factors, (1) the sequence can be exposed to form hybridization interactions with other molecules in the cell and (2) the structure, i. e. the three-dimensional arrangement, can form binding pockets for metabolites or build scaffolds to arrange other metabolites into reactive assemblies. For instance, the sequences of messenger-RNAs (mRNAs) encode the information to produce a protein, ribosomal-RNAs (rRNAs) and transfer-RNAs (tRNAs) together translate this information by synthesizing proteins. The rRNAs serve as a big scaffold structure arranging mRNA and tRNAs in order to catalyze the necessary chemical reactions.

Short RNA molecules, such as the 21 nucleotide long micro-RNAs (miRNAs) regulate transcription and translation via RNA-RNA interactions. In particular, they direct protein complexes to mRNAs in order to degrade the transcript before it is translated into a protein [Cai et al., 2009]. The recently discovered (in parts engineered) guide-RNAs (gRNAs) [Jinek et al., 2012] recruit protein complexes to genomic DNA. This has been especially popular during the last years, as it is a new strategy to modify genomic DNA and to transcribe selected regions in the genome [Zalatan et al., 2015]. Natural systems can employ nucleic acid hybridization as an immune defense against viruses. Short fragments of foreign nucleic acid are integrated into the host chromosome and serve as a library to guide degradation complexes to the invading viral genomes [Wiedenheft et al., 2012].

It is important to point out that the RNAs mentioned above are just a subset of those with well characterized functions. Taken together, RNAs are molecules that are necessary to translate the genetic material into a cellular program, but also to adapt the genetic material and the program to environmental influences. Additionally, RNA has

the functional repertoire to store genetic material (e. g. RNA viruses) and to catalyze reactions (e. g. protein synthesis).

1.2 SYNTHETIC BIOLOGY AND ARTIFICIAL LIFE

What are the essential properties of a living system? Does it require evolution, communication, reproduction or self-reflection? Does it require the molecules described above? From which point on can a life form be considered as artificial? Is it enough to remove or replace parts in an existing organism or does it need the assembly and understanding of a new system from scratch? Can we explain and reproduce how early life emerged on earth?

The studies of synthetic biology and artificial life raise many philosophical questions, but the current challenges lie in the details. Synthetic biology aims for the inverse problem of modeling biological systems: designing functional parts that can be embedded into molecular networks. This can be used to gain control over the output of natural systems, i. e. metabolic pathways are used as a function to translate input into output molecules. For artificial life, it is neither required to produce a living species in the first instance, nor to have a complete understanding of underlying mechanisms. The goal is to develop a system (from scratch) that has certain properties of living systems, e. g. a unit that evolves, replicates or just seems more complex than the sum of its parts.

Let us come back to the title of this thesis "Control of RNA function by conformational design". We have previously discussed that RNA function emerges from a combination of its sequence and structure and I will explain later in this thesis how the structure can efficiently be predicted from the sequence information. For example, it is possible to identify highly structured metabolite binding pockets, in combination with formally unstructured regions, that are available to bind other RNAs in the environment. Similarly, metabolite binding can induce a conformational change that alters RNA mediated information transfer.

RNA modeling primarily supports theories about RNA function by confirming experimental observations. It is important to stress the word *observation*, as measurements always influence the model system and even a computational model perfectly matching experimental data cannot capture the full complexity of a real world phenomenon.

RNA design is the inverse problem: to optimize RNAs with a particular function that has to be tested experimentally. Figure 1 shows a feedback loop to design RNA sequences. The process is essentially an *in silico* evolution of a sequence until it suffices the design criteria. The selection criteria are evaluated with RNA modeling, thus, build on the observation of experimental data. However, the artificial evolution process makes RNA sequence designs so robust, that also perturbations in model parameters may only have little effect on the designed RNA function [Dirks et al., 2004]. Along these lines, it is often more difficult to model (confirm) an experimentally determined

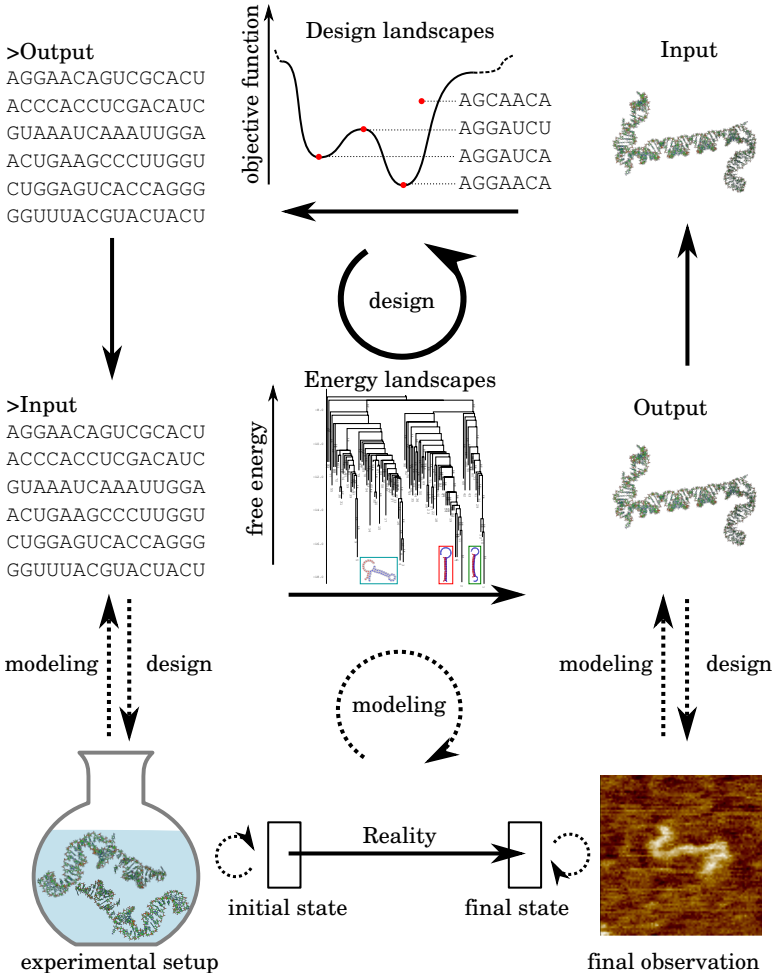


Figure 1: The process of designing a biological system. Arrows indicate an information transfer that can lead to a loss of accuracy. For solid arrows, this loss of information is avoidable or calculable, dashed lines connect theory with experiments and bear inherent uncertainties that are they key for understanding a real world phenomenon. Circular arrows depict feedback loops, i. e. the process of iterative model refinement. **Bottom row:** The experimental setup and its final observation lead to an initial theoretical model, which can be improved with new technologies. **Center:** Computational modeling translates an input into output. Both input and output are inferred from experimental findings with the prospect of simulating a real-world phenomenon. A well trained model can also detect flaws in experimental observations. **Top:** RNA design is the inverse problem of modeling. The final observation serves as input in an artificial evolution process to find an appropriate experimental setup. The design cycle uses modeling as selection criterion.

function of a natural RNA molecule, while the design of such a system can be surprisingly simple. In parts, because natural RNA molecules are evolved to fulfill a multitude of different (unknown) tasks that are influenced by stress response mechanisms, alternative expression levels, environmental context, different cell types, or from different stages in the cell cycle. Artificial RNAs, on the other hand, are simpler in the sense that their desired function is mostly independent of all these influences.

However, the design methods presented in this thesis optimize RNAs with multiple, combined functions and include properties that are not very well characterized in the model. In particular, RNAs switching between different conformations, forming catalytically active sites, or binding other metabolites. The results do not only give feedback about the functional repertoire of nucleic acid folding but are used to establish, debug, and refine theories for molecular modeling.

1.3 A PREVIEW OF THIS THESIS

Chapter 2 provides background on existing programs for RNA modeling. This includes the chemical decomposition of nucleic acids, the biophysical energy model and existing algorithms for modeling the thermodynamics and kinetics of single RNA folding.

Chapter 3 introduces background on thermodynamic properties of interacting RNA molecules and then shows my own contributions for the kinetic modeling of RNA-RNA interactions. The algorithmic work builds on coarse-grained RNA energy landscapes described in Chapter 2.

Chapter 4 introduces previous work on cotranscriptional folding and then continues with my own contributions. First, I present a strategy to include the binding of small metabolites, which has also been published as a separate book chapter [Badelt et al., 2015b]. Second, I present a new program `DrTransformer` to model the folding kinetics of larger RNAs during transcription.

Chapter 5 provides background about sequence design landscapes and their relationships to RNA energy landscapes and then continues describing the most effective, existing sequence design algorithms. In that context, I present a new RNA design library, which is part of the newest release of the `ViennaRNA package`.

Chapter 6 is a standalone publication with shared first authorship [Petkovic et al., 2015] that shows RNAs with self-processing activity. The ribozymes presented in the study catalyze their own circularization or elongation.

Chapter 7 is a standalone publication [Badelt et al., 2015a], following up a shorter conference paper [Badelt et al., 2014] that shows conformational self-replication. The

RNAs induce a conformational switching between other copies of the same molecule. A mechanism that has so far only been known from the protein world.

The structure of this thesis does not reflect the historical order of my research. In fact, the peer-reviewed publications in Chapters 6 and 7 have raised my attention for folding kinetics of interacting RNAs, as this is a central concept in both publications. While it is possible to design RNAs that interact in a specified way, modeling the kinetic processes of the interaction (as discussed in Chapter 3) is a much greater challenge. However, with the perspective of designing multi-functional, artificial networks, it will be necessary to design more complicated RNAs such as those interacting during transcription (Chapter 4).

Part II

RNA MODELING

2 MODELING OF SINGLE RNA MOLECULES

This chapter is written as an introduction into the basic chemistry that governs RNA folding and a summary of how these findings have been incorporated into RNA folding algorithms. Importantly, I want to emphasize that RNA molecules are dynamic polymers and there is never *one single* important conformation. Instead, to understand the function of an RNA molecule, one has to look at an ensemble of structures. This chapter will provide fundamental details that are necessary to understand my own contributions to the field of RNA interactions (Chapter 3) and RNA folding kinetics on dynamic energy landscapes (Chapter 4), as well as for the following chapters on RNA design in Part iii.

Also, this chapter introduces a notation to formulate RNA structure prediction as a mathematical problem. This notation will be mostly consistent with the notation in Lorenz [2014], a recent PhD thesis that focuses on single stranded RNA structure prediction. More detailed information about implementations of algorithms and properties of RNA energy models can be found, for example, in Lorenz [2014] and Andronescu [2008].

2.1 PROPERTIES OF RNA MOLECULES

RNA molecules can in principle form an infinite number of conformations, however, certain chemical properties of the molecule enabled researchers to (i) infer rules for RNA folding, (ii) define different levels of abstraction for the term *RNA structure* and (iii) incorporate thermodynamics to formulate RNA folding as a physical problem of minimizing free energy. This section is a brief summary of this process, before we pinpoint mathematical models and their implementation.

2.1.1 *Chemistry of RNA molecules*

An RNA molecule is an elastic ribose-phosphate polymer chain, with the ribose being covalently bound to one of four differently interacting nucleotides: Adenine (A), Cytosine (C), Guanine (G) or Uracil (U). The chemical composition of these nucleotides determines their possibility to interact with each other and with the RNA backbone. The RNA backbone determines the steric flexibility, i. e. which nucleotides of an RNA molecule can find each other for interactions.

Early observations were that single RNA molecules tend to form very characteristic helices and that these helices are formed mainly from six *canonical* base-pairs AU, UA,

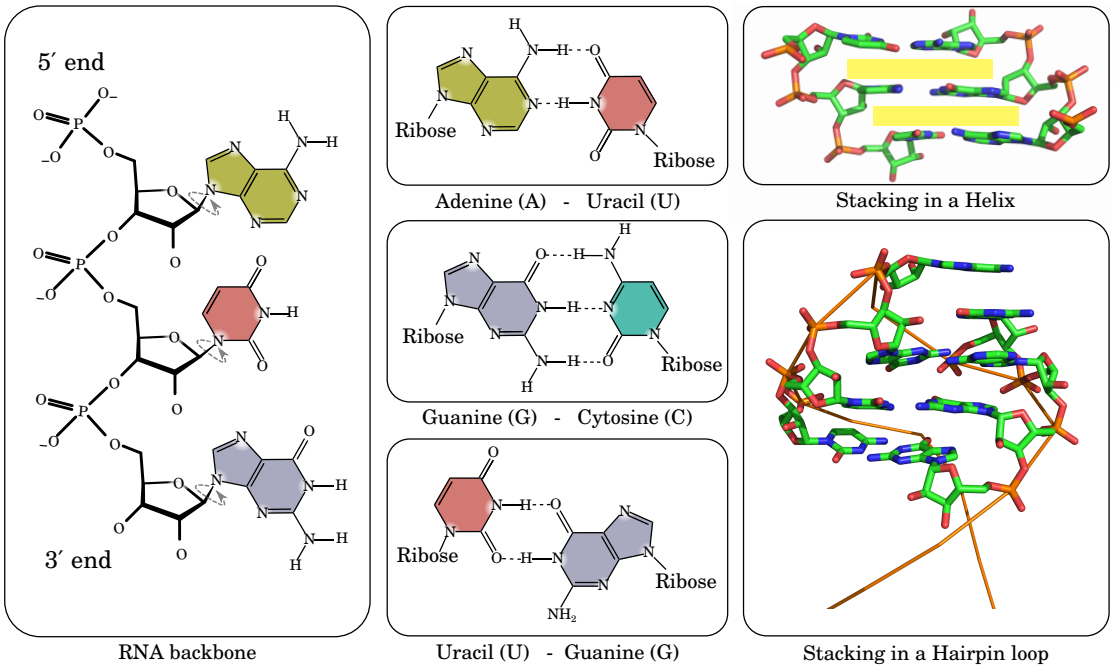


Figure 2: Canonical properties of RNA molecules. **Left panel:** chemical decomposition of a single stranded RNA Molecule with the sequence AUG in 5′ to 3′ direction. **Center:** hydrogen-bonding between the three canonical base-pairs (AU, GC, UG). **Right panel:** (top) yellow background indicates base-pair stacking in a helix. (bottom) base-pair stacking in a hairpin loop.

GC, CG, GU, UG. From more recent systematic searches in *reduced-redundancy* lists of experimentally determined RNA structures, these base-pairs were found to make about 76% of the total base-pairing pattern [Stombaugh et al., 2009].

It turned out that there are two fundamental reasons why canonical base-pairs are the driving force for RNA folding: first, these base-pairs are *isosteric*, i. e. they can substitute each other by mutation without disrupting the typical helical conformation. Second, the typical helical conformation enables nucleotides to share π-orbitales of their aromatic rings. This mechanism is known as *base-stacking* and turns out to be the main stabilizing factor for helices (see Figure 2).

In addition to the six canonical base-pairs, many other base-pairing interactions have been discovered and can be described following a systematic nomenclature suggested by Neocles Leontis and Eric Westhof Leontis and Westhof [2001]; Almakarem et al. [2011], see Figure 3 for a summary.

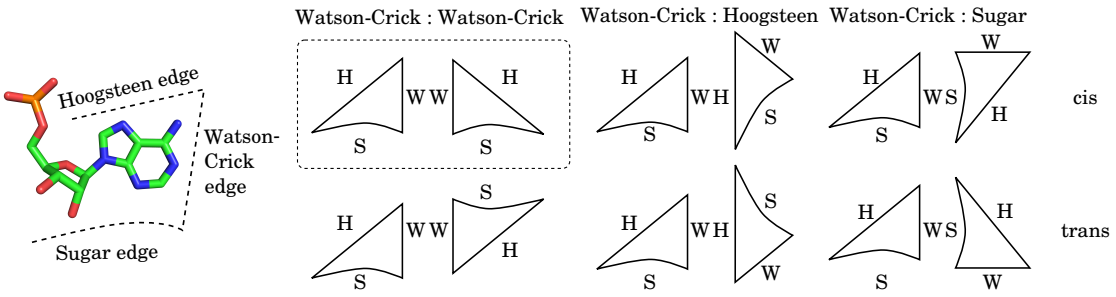


Figure 3: Canonical and non-canonical base-pairs. According to the Leontis/Westhof base-pairing classification, each nucleotide has three edges for interaciton. The Watson-Crick edge (W), Hoogsteen edge (H) and Sugar edge (S). All of these edges can interact in *cis* or *trans*, relative to the backbone orientation. Only a subset of possible pairwise interactions is shown. It is straightforward to extrapolate the remaining possibilities as well as more complex formations of base-triplets.

2.1.2 *RNA secondary structure*

Before going into details on energy models, the notion of RNA structure must be consistent. Researchers commonly distinguish three levels of abstraction:

PRIMARY STRUCTURE The *primary structure* is the sequence of nucleotides, without any steric information. This simple representation is often used to predict whether an RNA encodes the information to produce a protein, or has the potential to interfere with regulatory pathways by intermolecular base-pairing.

Definition 2.1 *The primary structure of an RNA molecule is an ordered sequence of letters* $\sigma = (N_1, \ldots, N_n)$ *where* $N \in \{A,C,G,U\}$ *and* n *is the length of the molecule.*

SECONDARY STRUCTURE The *secondary structure* contains the information of the primary structure and the helices formed, but no steric information how these helices are arranged in space. This representation of RNA structure is closely connected to the concept of hierarchical folding [Brion and Westhof, 1997]: RNA folding is a two step process in which first individual helices form and then they arrange relative to each other. The secondary structure thus determines the possible 3-dimensional arrangements and, thereby, the functional space of a particular RNA. A formal definition of a pseudoknot free, canonical secondary structure will be given below and I will refer to this definition in the remainder of this thesis.

Definition 2.2 *An RNA secondary structure* s *is a connected graph with vertices and edges* $G = \{V, E\}$*, where the set of vertices corresponds to nucleotides, and the edges correspond to (i) the covalently bound backbone and (ii) the hydrogen bonds forming base-pairs. The edges representing the backbone connect every pair of consecutive bases in the sequence interval* $[1, n]$*. The edges representing the base-pairs* $(i \cdot j)$ *must fulfill the following properties:*

GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA

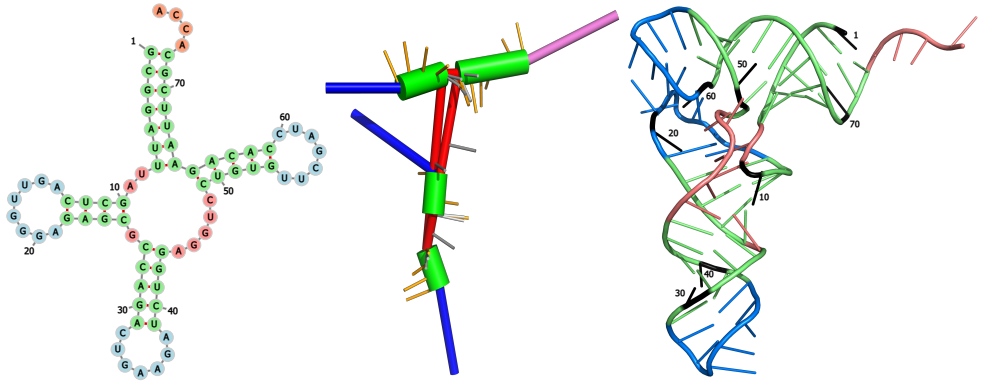


Figure 4: Different possibilities to display a tRNA molecule. **Top:** The primary structure is a string of nucleotides. **Left:** The secondary structure shows the base-pairing pattern of an RNA (visualized with `forna` [Kerpedjiev et al., 2015a]). **Center:** The tertiary structures shown with helices as stiff green cylinders, hairpin loops as blue sticks and the multiloop in red. **Right:** A more detailed representation of the tertiary structure. Black residues with labels correspond to the secondary structure visualization. The colors green, blue and magenta show the different loop-types seen in the secondary structure drawing. Two of the hairpin loops are engaged in a pseudoknot interaction. The structure was drawn using `Pymol` [Schrödinger, LLC, 2010] such that bases are depicted as single sticks. The tertiary structures were determined with electron microscopy, PDB-ID: `1Z01` [Allen et al., 2005].

1. $(i \cdot j) \in \{A \cdot U, U \cdot A, G \cdot C, C \cdot G, G \cdot U, U \cdot G\}$
2. *if* $(i \cdot j)$ *and* $(k \cdot l)$ *and* $i = k$ *then* $j = l$
3. *if* $(i \cdot j)$ *and* $(k \cdot l)$ *and* $i < k < j$ *then* $i < l < j$
4. *if* $(i \cdot j)$ *and* $i < j$ *then* $i + 3 < j$

Conditions 1–4 state that: (1) base-pairs have to be canonical, (2) bases may only form one base-pair at a time, (3) base-pairs have to be nested, such that no pseudoknots are possible, and (4) the minimal hairpin loop size has to span 3 unpaired nucleotides for steric reasons. Importantly, all secondary structures within that definition are sterically possible.

It is worth pointing out that researchers can have alternative definitions of secondary structure, e. g. above mentioned non-canonical base-pairs can be included, as well as base-triplets, different kinds of stacking interactions and pseudoknot structures. Figure 4 shows a secondary structure including a common pseudoknot interaction of tRNAs.

TERTIARY STRUCTURE The *tertiary structure* is the most complete form of RNA representation and its prediction relies heavily on the correct secondary structure. It commonly refers to a three-dimensional model, that either depicts the steric conformation of an RNA molecule with all details about the atomic positions, or various levels of abstractions. Figure 4 shows and example where bases are visualized with single sticks, and the backbone is a smooth line. Other visualizations, such as used by the coarse-grained tertiary structure prediction tool `Ernwin` [Kerpedjiev et al., 2015b] show the positioning of helices relative to each other. Tertiary structure prediction is an extremely challenging research area, often with the intention to predict RNA molecules that have been crystallized, down to Ångström resolution.

NOTATION Unless explicitly stated otherwise, I will refer to RNA primary structure (Definition 2.1) whenever I use the terms *RNA*, *RNA sequence* or *RNA molecule*, and to the definition of RNA secondary structure (Definition 2.2) whenever the terms *RNA conformation*, *RNA structure* or *RNA secondary structure* are used.

2.1.3 *The nearest neighbor energy model*

Thermodynamic modeling of RNA secondary structure quantifies the folding process by the change of free energy ΔG. This change is computed as the difference between an unfolded molecule (i. e. the open chain conformation) with a free energy of 0 kcal/mol and the structure of interest. The free energy of a reaction is computed as

$$\Delta G = \Delta H - T\Delta S \tag{1}$$

where ΔH describes the enthalpic contribution and ΔS the entropic, temperature dependent contribution of a chemical reaction. Note that the term $E(s)$ will be used to quantify the free energy of a structure s, which is conceptually equivalent to the free energy difference between the structure s and the open chain.

LOOP DECOMPOSITION The nearest neighbor (NN) energy model is based on the realization that stacking interactions are the driving force for RNA folding. This significantly increased the accuracy of RNA structure prediction compared to early attempts that maximized the number of canonical base-pairs in an RNA molecule [Nussinov and Jacobson, 1980]. The basic idea is to decompose RNA secondary structures into unique, additive elements called *loops*. The energy contributions of the most common loops are measured experimentally and tabulated in parameter-files, while energy contributions for the rest are extrapolated with mathematical models. The free energy of an RNA structure is computed by the sum of all loop free energies:

$$E(s) = \sum_{l \in s} e(l) \tag{2}$$

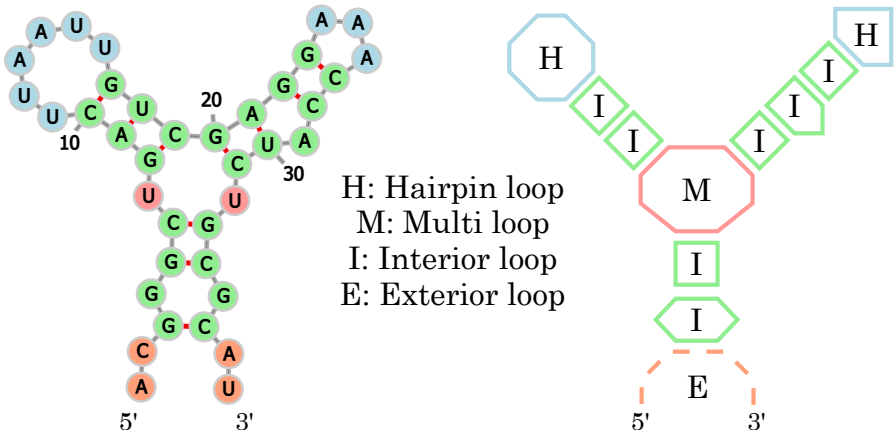


Figure 5: An RNA secondary structure is decomposed into the loop types from the nearest neighbor (NN) energy model. The sum over all loop energies yields the free energy of the molecule.

See Figure 5 for an RNA structure decomposed into three different loop types: *Hairpin loops* are enclosed by a single base-pair, *interior loops* are enclosed by two base-pairs and *multi loops* are enclosed by more than two base-pairs. Using this loop-decomposition, a stacking interaction between two adjacent Watson-Crick base-pairs is a special case of an interior loop that happens to be energetically favorable. All remaining single-stranded regions that are not enclosed by base-pairs (e. g. unpaired 3′ and 5′ ends) are denoted as *exterior loop* with the free energy contribution of 0 kcal/mol.

This loop-decomposition reveals abundant RNA structure elements that were measured by optical melting experiments and whose energies are publicly available in the nearest neighbor parameter database [Turner and Mathews, 2009]. For example, today's parameter files list the free energies for every interior loop of the form: 0-0, 0-1, 1-1, 1-2, 2-2, 2-3, where these numbers correspond to the unpaired bases between two closing base-pairs. The majority of loops have not been experimentally measured, but are extrapolated from mathematical models. Note that the 1-1 case describes a non-standard base-pair embedded in a canonical helix. This base-pair may slightly displace the RNA backbone, but can still contribute hydrogen bondings and favorable stacking interactions to the helix.

ENVIRONMENTAL AND MODEL PARAMETERS USED IN THIS THESIS

- Temperature: RNA folding is a temperature dependent process and parameter files list the free energy at 37°C together with the enthalpy. This allows the calculation of the free energy for every other temperature following equation 1. Unless explicitly stated otherwise, RNA folding is modeled at 37°C in this thesis.

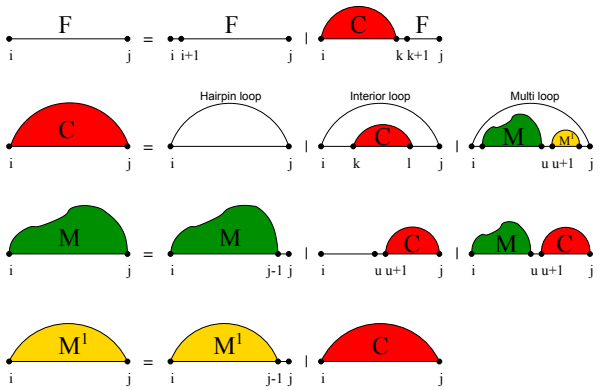
- Ion concentrations: The majority of NN energy parameters is measured at a ion concentration of 1M NaCl. The idea is that high concentrations of monovalent ions compensates for the lack of divalent ions such as Mg^{2+}. Recent experiments demonstrate the changes in helix stability for varying ion concentrations, and suggest to include varying ion concentrations into RNA folding e. g. Draper [2004]; Tan and Chen [2006, 2007]. However, the theory for an efficient algorithmic incorporation of these parameters is complicated and remains to be developed. Until then, the experimental results can be used to reassess the energy contribution of a particular structure, rather than predicting the structure in the first place. Hence, results in this thesis are computed using the standard environment of 1M NaCl.

- Dangling ends: Stacking interactions do also occur in single stranded regions. Especially engaged are those bases immediately adjacent to helix-closing base-pairs, so-called *dangling ends* [Neilson et al., 1979]. Their energy contributions have been included into RNA folding, however, the user is free to choose from different models of how these parameters are evaluated. The default model used in this thesis allows a single nucleotide to contribute with all its possible favorable interactions. Alternatively, an (algorithmically) more complex model has been suggested in which the unpaired base can stack with at most one base pair. Also stacking interactions between two helices emerging from the same loop region, *coaxial stacking* [Walter et al., 1994] can be included.

- DNA: While RNA favors an A-helix conformation, DNA favors the very characteristic B-helix. The set of canonical base-pairs is slightly different with Thymine replacing Uracil (i. e. AT, TA, GC, CG, GT, TG). However, also DNA strands can fold back on themselves and the principle of the NN energy model remains the same. In general, all folding algorithms described in this thesis are easily adaptable to DNA modeling by switching to the appropriate NN energy parameters, e. g. SantaLucia Jr and Hicks [2004].

2.2 RNA SECONDARY STRUCTURE PREDICTION

The number of RNA secondary structures (see Definition 2.2) grows exponentially with sequence length (approximately $1.4848n^{-3/2}(1.8488)^n$ [Schuster et al., 1994]), however, the minimum free energy (MFE) secondary structure can be computed in polynomial time $O(n^3)$. Here, I will give an overview of the dynamic programming (DP) algorithms to compute properties of RNA molecules, such as MFE, ensemble free energy, and base-pair probabilities. All of these algorithms are implemented in the `ViennaRNA package` [Hofacker et al., 1994; Lorenz et al., 2011], a comprehensive, well-maintained and runtime-optimized `C` library that enables programmers a "plug-and-play" devel-



$$F_{i,j} = \min \begin{cases} F_{i+1,j} \\ \min\limits_{i<k\leq j} C_{i,k} + F_{k+1,j} \end{cases}$$

$$C_{i,j} = \min \begin{cases} H_{(i,j)} \\ \min\limits_{i<k<l<j} I_{(i,j;k,l)} + C_{k,l} \\ \min\limits_{i+1<u<j-1} M_{i+1,u} + M^1_{u+1,j-1} + a \end{cases}$$

$$M_{i,j} = \min \begin{cases} M_{i,j-1} + c \\ \min\limits_{i<u<j} (u-i+1)c + C_{u+1,j} + b \\ \min\limits_{i<u<j} M_{i,u} + C_{u+1,j} + b \end{cases}$$

$$M^1_{i,j} = \min \begin{cases} M^1_{i,j-1} + c \\ C_{i,j} + b \end{cases}$$

Figure 6: The dynamic programming recursions for MFE secondary structure prediction as implemented in the `ViennaRNA package`. Image adapted from Bompfünewerer et al. [2008].

opment using the scripting languages `Perl` and `Python`. In later chapters, I will present new algorithms that are using this interface.

Note that there are multiple other programs for structure prediction, design and related problems. The probably most popular examples are `mfold` [Zuker, 2003], `UNAfold` [Markham and Zuker, 2008], `RNAstructure` [Reuter and Mathews, 2010], `NUPACK` [Dirks et al., 2007]. These programs can have small differences in the energy model parameters and also the implementations are often directed to address particular challenges related to RNA folding.

2.2.1 *Minimum free energy prediction*

There is an important feature about our definition of RNA secondary structures that makes calculation of equilibrium properties feasible in polynomial time: Every base-pair divides a structure into two separate parts and the optimal solution for the complete structure is exactly the sum of the optimal solutions for the two substructures. This feature is described by *Bellman's Principle of Optimality* and makes RNA structure prediction solvable with DP: the MFE of the smallest sub-sequences (length 5) is calculated, tabulated, and looked up for the calculation of the next bigger sub-sequences (length 6), and so forth, until the solution for the full-length molecule is found. Importantly, only optimal solutions are tabulated, but not the history of how these solutions were computed. Finding the secondary structure corresponding to the MFE then requires an additional backtracking routine in order to reconstruct the optimal path through the DP matrices.

Figure 6 shows recursions to calculate the MFE for any subsequence ($F_{i,j}$). Using DP, this problem can be solved in $O(n^4)$ time and $O(n^2)$ space, where the dominating

factor for computation time are interior loops $O(n^4)$ followed by multi loop recursions $O(n^3)$. The `ViennaRNA package` reduces time requirements by excluding interior loops that are larger than 30 unpaired nucleotides in total. This is reasonable, as such large interior loops are not observed in crystallized structures and there are no measured energy parameters available. Thus, the formal time-complexity to find optimal interior loops is reduced to $O(n^2)$, making multi loop recursions the now determining factor for RNA folding at $O(n^3)$ runtime.

2.2.2 *Ensemble properties and base-pair probabilities*

The MFE secondary structure alone provides little information about the RNA molecule. It is the most likely structure in thermodynamic equilibrium, yet it is unclear *how* likely this structure is. With the equilibrium partition function of a system, it is possible to calculate statistical properties considering the thermodynamic ensemble of *all* secondary structures. The equilibrium partition function Z is computed as the sum over all possible structures Ω (Definition 2.2), weighted by their energy in an Boltzmann-distributed ensemble:

$$Z = \sum_{s \in \Omega} e^{\frac{-E(s)}{RT}} \tag{3}$$

McCaskill [1990] introduced a DP algorithm to compute this partition function with the same time complexity as computing the MFE structure. A basic requisite is a unique decomposition of RNA secondary structures, ensuring that every structure is computed exactly once (e. g. Figure 6). If these criteria are fulfilled, one can use the same DP algorithms as described for MFE folding, but replace min with a $\sum$ operation and add up the Boltzmann distributed free energies of sub-sequences.

The partition function is of particular importance to compute properties of the secondary structure ensemble. The ensemble free energy G of an RNA molecule is given as the free energy of the partition function:

$$G = -RT \cdot \ln Z \tag{4}$$

The difference between MFE and ensemble free energy G will be used later in this thesis to determine the occupancy of the MFE secondary structure in thermodynamic equilibrium. In general, if the free energy of a structure is close to the free energy of the ensemble, then the ensemble is dominated by this structure.

The second important application for the equilibrium partition function Z is to compute the probability of forming a particular secondary structure $P(s)$

$$P(s) = \frac{e^{\frac{-E(s)}{RT}}}{Z} \tag{5}$$

Also, the efficient computation of the ensemble probability of a base-pair $P(i \cdot j)$ or likewise of a particular sub-structure is possible by computing a *constrained* partition function $Z_{(i \cdot j)} = \sum_{s \in \Omega_{(i \cdot j)}} e^{\frac{-E(s)}{RT}}$ where $\Omega_{(i \cdot j)}$ are all secondary structures forming the base-pair $(i \cdot j)$. Hence, the probability $P(i \cdot j)$ is given as

$$P(i \cdot j) = \frac{Z_{(i \cdot j)}}{Z} \tag{6}$$

2.2.3 *Suboptimal secondary structures*

Different types of suboptimal secondary structures have been introduced for RNA folding. *Zuker suboptimal structures* are defined as calculating for every possible base-pair the optimal structure given that base-pair [Zuker, 1989]. This set of structures grows only quadratically with sequence length and can be used for heuristic approaches to RNA folding. In practice, however, it follows that every base-pair present in the MFE structure will regenerate the MFE structure and it is impossible to find suboptimal structures that are composed of more than one suboptimal substructures.

In contrast, *Wuchty suboptimal structures* [Wuchty et al., 1999] capture the whole secondary structure space we have previously introduced as Ω and hence grow exponentially with sequence length. In practice, suboptimal structures with a low free energy are especially interesting for RNA folding. `RNAsubopt` [Wuchty et al., 1999] computes all conformations within a specified energy range above the MFE by altering the DP backtracking routine.

2.3 RNA ENERGY LANDSCAPES AND FOLDING KINETICS

While thermodynamic properties of RNA molecules provide a static, steady state image of RNA folding, this section addresses the dynamics of RNA molecules when being *out of equilibrium*. RNA folding kinetics often plays an important role in living systems, either because it prevents the MFE secondary structure from being formed during the life-time of a molecule, or because environmental changes induce a *re*-folding of an RNA molecule into a different conformation with an alternative function.

This section will introduce the concept of *energy landscapes* and their inherent properties. It will show two-dimensional projections of these landscapes and discuss biophysical models to simulate the stochastic process of RNA folding kinetics.

2.3.1 *RNA energy landscapes*

With the previously introduced energy model $E(s)$ and the set of suboptimal structures Ω, we have already introduced two out of three components to define an RNA

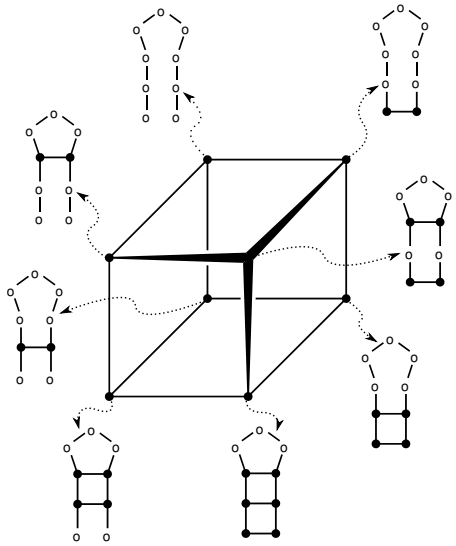


Figure 7: The move set (neighborhood relation) of an RNA molecule that can form three different base-pairs.

energy landscape. The last missing piece is a *move set* M to describe all possible folding reactions within the set of suboptimal structures, formally:

Definition 2.3 *Let $\mathcal{L} = (\Omega, M, E)$ be the energy landscape of an RNA molecule. Ω is the set of RNA secondary structures, M is a move-set that defines neighboring structures as those that differ by a single base-pair and E is the NN energy model assigning a fitness value in form of a free energy to each conformation.*

Single base-pair moves are always reversible and ensure ergodicity

Each of the three properties can be replaced for an alternative landscape definition. In order to make landscape analysis computationally tractable, Wuchty suboptimal structures can be used to compute only the energetically low parts. The move-set M, has to ensure *ergodicity*, i. e. that every conformation can be reached from every other conformation. The concept of single-base-pair moves (see Figure 7) satisfies this condition. Alternative move-sets insert and break whole helices in order to explore folding landscapes faster. This sacrifices the reversibility of moves, since the possibility of folding into particular secondary structures depends on the order of how helices have been inserted [Flamm and Hofacker, 2008]. E. g. insertion of two helices can result in one big helix and therefore the insertion of the second helix is not reversible. In section 2.3.5 models of well defined *coarse-grained* energy landscapes will be introduced that reduce the number of conformations while keeping reversibility.

2.3.2 *Walks and folding pathways*

Folding pathways may be direct or indirect, but they should be cycle-free for computational modeling.

RNA folding pathways describe structural rearrangements as a sequence of moves to

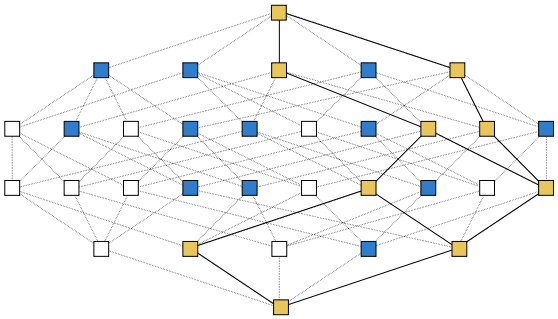


Figure 8: Direct paths connecting two structures with base-pair distance $d = 5$. When setting the upper bound $w = 2$, only the blue and yellow structures are evaluated and the best two (yellow) structures are selected to generate conformations in the next distance class. Figure adapted from Flamm et al. [2001].

transform one structure s_i into another structure s_j, we write $\mathcal{P} = (s_i, s_{i+1}, \ldots, s_{n-2}, s_j)$ for a path of length n. RNA folding is a stochastic process, thus, a typical RNA folding pathway visits the same conformations multiple times before a large structural rearrangement takes place. For computational modeling, we are mostly interested in *cycle-free* paths, i. e. a path is cycle-free if every structure s_k along the path is visited exactly once. Furthermore, we distinguish between *direct* and *indirect* paths: a path is called direct if there is no shorter path between the start and the end structure, and indirect otherwise. Hence, the length of a direct path is exactly the base-pair distance between the first and last structure.

If the energy function E is taken into account, then any folding path $\mathcal{P}$ has a saddle point with energy $E_{\mathcal{P}} = \max_{s \in \mathcal{P}} E(s)$, and the energy barrier $\Delta G^{\ddagger}$ between the start structure s_i and the stop structure s_j is determined by the lowest saddle point among all possible paths $\mathcal{P}_{s_i \to s_j}$

$$\Delta G^{\ddagger} = \min_{\mathcal{P}_{s_i \to s_j}} E_{\mathcal{P}} - E(s_i) = \min_{\mathcal{P}} \max_{s \in \mathcal{P}} E(s) - E(s_i) \tag{7}$$

Finding the best folding path has been shown to be a NP-hard problem [Maňuch et al., 2011]. However, there exist fast heuristics to compute near optimal direct paths between two structures, such as the `findpath` [Flamm et al., 2001] breadth-first search algorithm available in the `ViennaRNA package`. For every current structure s_k on the path $\mathcal{P}$, a list of neighbors $s_l \in N(s_k)$ is generated with base-pair distance $d(s_l, s_j) = d(s_k, s_j) - 1$. In practice, that means that s_n has a base-pair removed which is not present in s_j or a base-pair inserted which is required for the final conformation s_j. From this set of neighbors, `findpath` choses the energetically best w solutions for the next round of neighbor generation (see Figure 8). With the parameter $w = 1$ the method implements a greedy heuristic based on the NN energy model. For higher search widths, `findpath` starts an iterative procedure initialized with $w = 1$ to find an *upper bound* for subsequent iterations and then gradually doubles the search width

until the specified parameter w has been reached. Hence, the upper bound of the current iteration is determined from the previous result, at the cost of doubling the total runtime.

WALKING ON THE LANDSCAPE If the energy function E is taken into account, then some pathways become more likely than others and we speak of so-called *walks* on the energy landscape.

Definition 2.4 *Take an arbitrary secondary structure s_k as starting point for a walk in the landscape, than a* random walk *can take arbitrary steps, while an* adaptive walk *is biased to chose only from energetically better neighbors $E(s_{k+1}) < E(s_k)$. A* gradient walk *only choses from the neighbors where $E(s_{k+1})$ is minimal. Both the adaptive and the gradient walk therefore terminate once they reach a local minimum.*

Gradient walks always chose an energetically best neighboring conformation

2.3.3 *Rates of RNA folding reactions*

In terms of chemical kinetics, RNA folding is a network of isomerisation reactions ($s_i \rightleftharpoons s_j$) and, hence, a set of first order chemical reactions. The reaction constants are the *chemical reaction rates* k, and the occupancy of a particular structure $P_i(t)$ during RNA folding gives the probability of observing the structure s_i at time t. Following the first oder chemical master equation

$$\frac{dP_i(t)}{dt} = \sum_{i \neq j} (P_j(t)k_{ji} - P_i(t)k_{ij}) \tag{8}$$

the change of $P_i(t)$ is given by the sum of influx ($P_j(t)k_{ji}$) minus outflux ($P_i(t)k_{ij}$) of the state s_i. Thermodynamic equilibrium is reached when the occupancy change as a function of time $\frac{dP_i(t)}{dt} = 0 \,\forall\, i \in \Omega$. The chemical reaction rates k_{ij} have to be related to the change in free energy and they have to satisfy *detailed balance*

Ergodicity and detailed balance ensure that simulations end in thermodynamic equilibrium.

$$P_i k_{ij} = P_j k_{ji} \tag{9}$$

as implied by the chemical master equation 8.

For single-base-pair changes, reaction rates are commonly computed based on the Metropolis rule [Metropolis et al., 1953] from statistical physics or the Kawasaki rule [Kawasaki, 1966] originally developed to model spin diffusion constants for time-dependent Ising models. According to the Metropolis rule, the reaction rate is constant for all energetically favorable reactions, otherwise the rate is calculated from the free energy difference of the transition $\Delta G^{\ddagger} = E(s_j) - E(s_i)$

$$k_{ij} = \begin{cases} k_0 & \text{if } \Delta G^{\ddagger} \leqslant 0 \\ k_0 e^{-\frac{\Delta G^{\ddagger}}{RT}} & \text{otherwise} \end{cases} \tag{10}$$

where k_0 is a parameter to incorporate other physical processes, that are necessary to convert the time units to wall clock time. Kawasaki computes all rates evenly from the change in free energy:

$$k_{ij} = k_0 e^{-\frac{1}{2}\frac{\Delta G^\ddagger}{RT}} \tag{11}$$

ARRHENIUS MODEL Reaction rates for larger structural transitions are computed with the Arrhenius equation, which quantifies the "activation energy" for a reaction. This activation energy is computed as the change in free energy between the starting conformation s_i and the transition state $\Delta G^\ddagger = E(s_t) - E(s_i)$ where the transition state s_t is the state with the worst energy during the folding reaction

$$k = k_0 e^{-\frac{\Delta G^\ddagger}{RT}}, \text{ for } \Delta G^\ddagger > 0 \tag{12}$$

Arrhenius kinetics have also been suggested for a higher resolution modeling of single-base-pair changes [Schmitz and Steger, 1996; Zhang and Chen, 2006a]. The free energy is split into its entropic and enthalpic contributions, such that for base-pair formation, the entropic penalty has to be paid, before the enthalpic energy contribution is received and vise versa. Accordingly, Zhang and Chen [2006a] formulate the Arrhenius equation such that rates for formation of a helix (or base-pair) k_+ and the opening of a helix (or base-pair) k_- are treated separately

$$k_\pm = k_0 e^{-\frac{\Delta G^\ddagger_\pm}{RT}} \tag{13}$$

The nucleation of a helix involves an unfavorable entropy loss ΔS before enthalpic contributions from stacking interactions are received. The barrier is therefore entropic:

$$\Delta G^\ddagger_+ = T\Delta S \tag{14}$$

while the barrier for opening stacking base-pairs first involves the enthalpy increase ΔH when bonds and stacks have been disrupted, before relaxed torsional angles of the chain have favorable entropic effects:

$$\Delta G^\ddagger_- = \Delta H \tag{15}$$

The model satisfies the detailed balance condition (Equation 9) with

$$\frac{k_+}{k_-} = e^{\frac{(\Delta H - T\Delta S)}{kT}} \tag{16}$$

SOLVING THE CHEMICAL MASTER EQUATION An exact solution to the chemical master equation can be obtained by numerical integration. One way is to use matrix exponentials, as implemented in the program `treekin` [Wolfinger et al., 2004]. Given

a population vector $\vec{p}(t) = (P_1(t), \ldots, P_N(t))$ and a matrix of transition rates $R = \{k_{ij}\}$, the master equation can be rewritten as

$$\frac{d\vec{p}(t)}{dt} = \vec{p}(t)R \tag{17}$$

with the benefit that the formal solution can be computed from an initial population vector $\vec{p}(0)$ as

$$\vec{p}(t) = \vec{p}(0)e^{tR} \tag{18}$$

Using *matrix decomposition* methods [Moler and Van Loan, 2003], landscapes with roughly 10^4 structures can be solved explicitly. Unfortunately, already the conformation space of short molecules easily exceeds this number and thus there are generally two strategies to approximate folding kinetics: stochastic simulations of statistically correct trajectories and exact solutions on coarse-grained energy landscapes. Both strategies will be addressed in the remainder of this chapter.

2.3.4 *Stochastic modeling of RNA folding kinetics*

RNA folding can be modeled with a statistically correct random walk in the energy landscape. Assuming that every structural transition is independent of the previous transitions, the formally correct approach is based on sampling using a memoryless Markov Chain Monte Carlo method. Potential structural transitions are chosen from a random distribution of neighboring conformations and accepted or rejected according to the probability of the transition. The time needed for a structural transition is proportional to the trial and error process until a selected move has also been accepted.

The problem gets slightly more complicated for RNA energy landscapes, as the standard Markov Chain Monte Carlo assumes a constant neighborhood when sampling trajectories. The neighborhood of RNA secondary structures is variable, i. e. dependent on the current conformation, and has to be computed at every step in order to ensure detailed balance. Flamm et al. [2000] therefore implemented a rejection-less, continuous-time Markov Chain Monte Carlo method, also known as Gillespie-type simulation [Gillespie, 1977]. The neighborhood is computed to calculate the total flux out of the current stat s_i as

$$\Phi = \sum_{i \neq j} k(s_i \to s_j) \tag{19}$$

where k depends on the rate model chosen from the ones discussed above. Using the flux Φ, the rates can be corrected proportionally

$$k'(s_{i \to j}) = \frac{k(s_{i \to j})}{\Phi} \tag{20}$$

and the neighboring conformation can be chosen from the flux-corrected distribution using a random number $r_k \in [0, 1]$. In a standard Markov Chain Monte Carlo implementation, the waiting time for accepting a move decreases exponentially during the trial and error process. Hence, in the rejection-less method, the flux Φ can be used to compute the waiting time before moving to the neighboring conformation as

$$\tau = \frac{-\ln(r_\tau)}{\Phi} \tag{21}$$

where $r_\tau \in [0, 1]$ is a second random number. Using a rejection-less Markov Chain Monte Caro is also convenient, as RNA energy landscapes are *rough*, i. e. they contain lots of local minima that trap structures for a long time. The acceptance rate is therefore very low and standard Markov Chain Monte Carlo can get stuck in a current structure for a long time.

While `Kinfold` [Flamm et al., 2000] models stochastic trajectories using single base-pair moves, `Kinefold` [Isambert and Siggia, 2000] models RNA folding using whole helix transitions. More recently, Schaeffer et al. [2015] have implemented `Multistrand` for stochastic modeling of multiple interacting nucleic acid molecules.

2.3.5 *Reduced and coarse-grained energy landscapes*

RNA energy landscapes are rough and degenerate

RNA energy landscapes are high-dimensional with valleys, mountains, funnels, ridges and plateaus. In the last section we have called them *rough*, as structures easily get trapped in locally optimal conformations. Now, we note *degeneracy* as another unfortunate characteristic: multiple suboptimal structures exist on the exact same energy level, making it impossible to pick a best structure without introducing other artificial criteria.

We will now discuss how to chose a representative set of RNA secondary structures and transition rates between them such that the overall kinetics on that reduced set resembles the kinetics in the full energy landscape. Such a process is called *coarse-graining*: a given set of micro-states is partitioned with respect to selected features and the partitions form macro-states in a system. It is often convenient to have a representative for each macro-state, e. g. the secondary structure with the lowest free energy.

As an example, consider secondary structures without lonely base-pairs, also called *canonical secondary structures* [Bompfünewerer et al., 2008] as coarse-graining for the landscape of Wuchty suboptimal structures. The partitioning is easy: removing all lonely base-pairs from a given suboptimal conformation (micro-state) yields the respective canonical structure (macro-state). For such a moderate partitioning, it is easy to show that representative macro-states closely resemble the dominating micro-states and hence the kinetics on the coarse-grained energy landscape is a reasonable abstraction of the true micro-system. Canonical energy landscapes are only a small reduction,

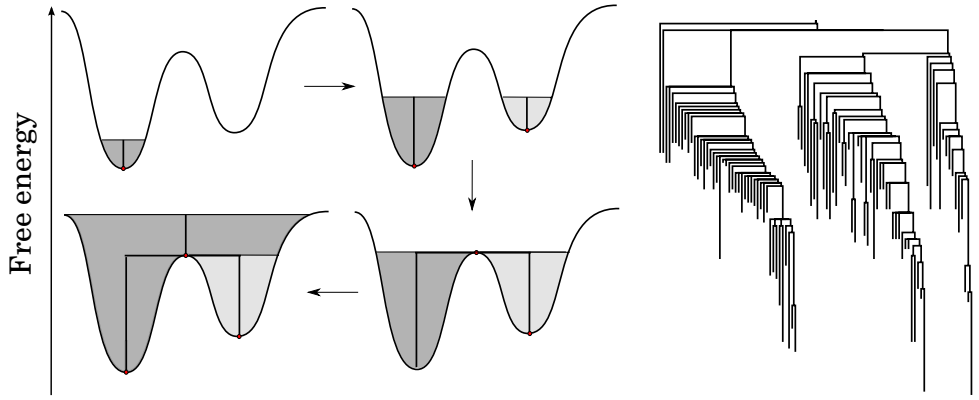


Figure 9: Projection of an energy landscape into a barrier tree using a flooding algorithm. A sorted list of Wuchty suboptimal structures is processed to identify basins in the energy landscape. A current structure is the transition state between two basins, if it has neighbors in both basins. The vertical lines in a barrier tree show the energy barrier separating local minimum conformations. The calculation of basin partition functions (not shown) can be computed with minimal extra effort using (deterministic) gradient walks after two basins have been merged by the flooding algorithm.

however, they are often used in practice as a top-level coarse-graining in order to reduce the number of suboptimal secondary structures in the first place.

GRADIENT BASINS A particularly elegant way of coarse-graining has been introduced with the program `barriers` [Flamm et al., 2000, 2002] and will serve as a basis to coarse-grain landscapes of interacting RNA molecules in Chapter 3 and cotranscriptional landscapes in Chapter 4. The partitioning works on two levels, the first level is a *flooding algorithm* to identify local minimum conformations and the lowest energy barriers separating them, the second level is a partitioning using gradient walks (see Section 2.3.2) to ensure a unique landscape decomposition and calculate the partition function for macro-states, so-called *gradient basins*.

Coarse-graining energy landscapes with a flooding algorithm works as follows: for every structure in an energetically sorted list of suboptimals $s \in \Omega$, the neighbors $N(s)$ are generated and compared with the set of previously processed structures. If none of the neighbors matches a previously processed structure, then s is a new local minimum and hence the representative structure in a new macro-state. If all matched neighbors $N^+(s)$ belong to the same local minimum α, then s is assigned to α and if matched neighbors are part of different local minima then s is a saddle point in the energy landscape connecting these local minima. From that point on, the connected local minima are merged into one macro-state represented by the former deepest minimum.

In order to compute the partition-function for macro-states, the basins of attraction comprise all structures that lead there with a gradient walk. For kinetic simulations on a gradient-basin landscape, the transition rates between macro states are calculated as the sum over all micro-rates between the basins

$$k(\alpha \to \beta) \approx \sum_{x \in \alpha} \sum_{y \in \beta} P(x|\alpha) k(x \to y) \tag{22}$$

where $P(x|\alpha)$ is computed assuming that the ensemble of structures in basin α is instantaneously in equilibrium. Thus, the coarse-graining works well if equilibrium within a macro-state is fast compared to transitions between macro-states.

In the program `barriers`, the size of gradient basins, as well as the number of suboptimal gradient basins is adjustable. Figure 9 shows the flooding of an energy landscape and an example barrier tree. Macro-state decomposition for the degenerate case can be dealt with correctly [Flamm et al., 2002], but is ignored in this thesis as it makes subsequent analysis of landscapes more complicated. For example, gradient walks must chose randomly from all energetically best neighbors, which destroys the deterministic computation of macro-state partition functions.

3 MODELING OF INTERACTING RNA MOLECULES

We have now studied the folding of single RNA molecules, assuming that they first arrange into their active conformation and then interact with the environment. In a cellular context RNA folding itself can already depend on environmental stimuli such as interacting proteins, (small) metabolites, or simply other nucleic acid molecules present in the vicinity. Most intermolecular binding reactions are not well characterized and we lack experimental energy parameters to systematically include them into nucleic acid folding. However, interactions of two or more nucleic acid molecules are particularly important and can be dealt with the same general methods as the folding of single RNA molecules, although many details become more complicated.

Employing hybridization reactions *in vitro* led to ground breaking technologies such as polymerase chain reaction (PCR) or reverse transcription and is a fundamental concept in nucleic acid computation. In natural systems, certain classes of RNA molecules are produced primarily to locate other nucleic acid molecules via hybridization reactions. Most prominent are micro-RNAs (miRNAs) that induce the degradation of messenger-RNAs (mRNAs), and guide-RNAs (gRNAs) hybridizing to DNA and recruiting enzymes or a translation complex [Zalatan et al., 2015]. Researchers more and more use these systems for metabolic engineering and, hence, the demand for more accurate prediction and design tools is steadily increasing.

THIS CHAPTER first provides background information about thermodynamic modeling of intermolecular folding in section 3.1 and then continues with my own contribution to intermolecular folding kinetics starting with section 3.2. The algorithmic work described here, as well as the results, will be adapted for publication together with Christoph Flamm and Ivo L. Hofacker. Sources for the corresponding programs `interkin` and `SundialsWrapper` will then be available at `http://www.tbi.univie.ac.at/software`

3.1 THERMODYNAMICS OF RNA-RNA INTERACTIONS

Several methods have been proposed to deal with nucleic acid interactions, among them Hofacker et al. [1994] showed a schematic example to compute the minimum free energy (MFE) secondary structure of two interacting RNAs. Mathews et al. [1999] presented `OligoWalk` to assess the quality of oligomer binding such as PCR primers, Rehmsmeier et al. [2004] published `RNAhybrid` to find miRNA target sites in mRNAs, and Andronescu et al. [2005] implemented `PairFold` to compute the MFE secondary

structure for two interacting RNA molecules using the dynamic programming (DP) recursions for intramolecular folding.

For molecular interactions concentration dependency becomes an issue, such that higher concentrations of strands are expected to increase the degree of intermolecular hybridization. This effect has been investigated for thermodynamic equilibrium using the partition function [Applequist and Damle, 1963] and refined and applied using the nearest neighbor (NN) energy model [Dimitrov and Zuker, 2004].

Most important for this work is `RNAcofold` [Bernhart et al., 2006], part of the `ViennaRNA package`, which combines MFE folding, the equilibrium partition function computation, and calculation of the monomer-dimer fractions in thermodynamic equilibrium.

It is worth mentioning at this point that Dirks et al. [2007] have extended previous work to efficiently compute thermodynamic properties for multiple interacting nucleic acid molecules. The algorithm is implemented in the `NUPACK` framework, and recently a Gillespie-type simulator `Multistrand` [Schaeffer et al., 2015] was developed to model folding kinetics of interacting nucleic acids using the same energy model [Schaeffer, 2013].

3.1.1 *Intermolecular nearest neighbor interactions*

In order to apply the same recursion schemes for intra- and intermolecular folding, the two RNAs are treated as one long molecule with a missing backbone edge between the 3′ end of the first molecule and the 5′ end of the second molecule. This missing backbone introduces an additional loop-type to the NN energy model: the *cut* loop (see Figure 10). If the cut lies in the external loop of a structure, then the two molecules do not interact, thus, there is no entropic penalty. In every other case, intermolecular base-pairs have formed and the (interior, hairpin, or multi-) loop energies have to be corrected to exterior loops. Note that our definition of valid secondary structures (Definition 2.2) excludes hairpin loops with less than three unpaired nucleotides for steric reasons. This requirement is sacrificed whenever the cut is in the hairpin loop. On the other hand, the restriction to at most 30 unpaired nucleotides in an interior loop remains for performance reasons. A side effect of this simplified interaction model is that some intermolecular motifs are now formally pseudoknots, such as the rather abundant base-pairing between two hairpin loops, the *kissing hairpin motif*. These motifs can be predicted using e. g. `RNAup` [Mückstein et al., 2008], however, `RNAup` predicts only one local interaction at a time, making `RNAcofold` the more general approach to model RNA-RNA interactions.

We assume a well-mixed, dilute, test-tube setting where the initiation of intermolecular RNA-RNA binding comes with a measurable free energy change E^{init}. This free energy change quantifies the entropy-loss when forming the initial RNA-RNA interaction and is assumed to be independent of sequence length and composition. It has been measured as $E^{init} = 4.1$ kcal/mol [Mathews et al., 1999; Turner and Mathews,

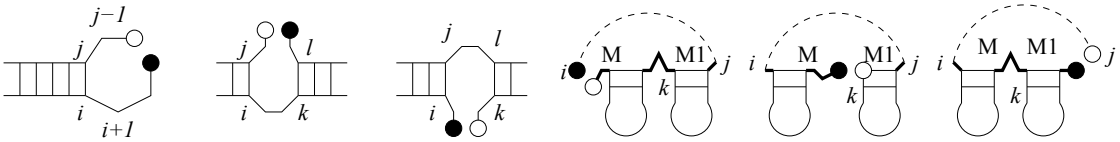


Figure 10: Hairpins, interior loops and multi-loops with cuts have to be scored as exterior loops. The cut is between two nucleotides (black ball and white ball). All possible cases that have to be considered in the DP recursions are shown. Figure adapted from Bernhart et al. [2006].

2009] at standard buffer conditions and is added to every conformation once, given at least one intermolecular base-pair has formed.

The computation of the MFE secondary structure and the partition function Z enables the calculation of all the thermodynamic properties we have previously discussed for intramolecular RNA folding (e. g. suboptimal structures, ensemble free energy, and base-pair probabilities) and, additionally, the concentration dependent equilibrium distributions of monomers and dimers.

For the computation of the partition function, the initiation energy E^{init} of the intermolecular contact has to be taken into account exactly once for every structure if and only if it contains an intermolecular base-pair. This additional bookkeeping can be avoided by introducing the term after the dynamic programming tables have been filled. `RNAcofold` calculates the partition function of the full structure ensemble Z^f_{AB} and also the partition functions of the sequence intervals $[1, c-1]$ and $[c, n]$, where n is the full sequence length and c is the index of the first nucleotide of the second molecule. This yields the partition functions for isolated monomers Z_A, Z_B and we observe the partition function for entropy-corrected dimers as

$$Z'_{AB} = (Z^f_{AB} - Z_A Z_B) e^{\frac{-E^{init}}{RT}} \tag{23}$$

An additional complication for homodimers is that they have to be corrected for a two-fold rotational symmetry that reduces their conformation space by a factor of 2. Since the partition function computations assumes two distinguishable molecules, Z'_{AA} is computed as

$$Z'_{AA} = \frac{(Z^f_{AA} - Z^2_A)}{2} e^{\frac{-E^{init}}{RT}} \tag{24}$$

For MFE computations and suboptimal structures, this symmetry correction is not implemented, hence, every symmetric homodimer should get an entropic penalty of $\delta = -RT \cdot \ln(2)$ [Dirks et al., 2007]. In order to guarantee finding the correct MFE structure, one has to compute suboptimal structures (`RNAsubopt`) with an energy range of at least δ and then add the penalty δ to all symmetric conformations.

NOTATION In the remainder of this chapter, we write Z'_{AB} for the dimer-only partition function and Z_{AB} for all species, with dimers corrected for the entropic interaction penalty

$$Z_{AB} = Z'_{AB} + Z_A Z_B \tag{25}$$

and equivalently for Z_{AA} and Z_{BB}.

3.1.2 *Concentrations at thermodynamic equilibrium*

Consider a dilute solution of two nucleic acid sequences A and B with concentrations $[A]$ and $[B]$. Hybridization yields the two homodimers AA, BB and the heterodimer AB. More complex oligomers are assumed to be disfavored by additional destabilizing initiation entropies and neglected in the following approach. The equilibrium constant of the chemical dimerization reaction $A + B \leftrightharpoons AB$ is expressed as

$$\frac{[AB]}{[A][B]} = K_{AB} = \frac{Z'_{AB}}{Z_A Z_B} \tag{26}$$

and, hence, can be directly computed from the partition functions Z'_{AB}, Z_A, Z_B of the respective molecules [Dimitrov and Zuker, 2004; Bernhart et al., 2006].

The distribution of monomers A, B and dimers AA, BB, AB at equilibrium can be calculated using the side condition that the number of particles in the system has to remain constant. Let $[A]_0$ and $[B]_0$ denote the initial concentrations of A and B in the system, then the side conditions are given as

$$\begin{aligned} [A]_0 &= [A] + 2[AA] + [AB] \\ [B]_0 &= [B] + 2[BB] + [AB] \end{aligned} \tag{27}$$

Taken together, equations 26 and 27 form a complete set of differential equations

$$\begin{aligned} 0 &= f([A],[B]) := [A] + K_{AB}[A][B] + 2K_{AA}[A]^2 - [A]_0 \\ 0 &= g([B],[A]) := [B] + K_{AB}[A][B] + 2K_{BB}[B]^2 - [B]_0 \end{aligned} \tag{28}$$

that can be solved by numeric integration with Newton's iteration method and yields the equilibrium concentrations $[A], [B], [AA], [BB], [AB]$.

3.2 MASS ACTION FOLDING KINETICS OF RNA-RNA INTERACTIONS

Thermodynamic design has recently led to a number of synthetic devices [Green et al., 2014; Chappell et al., 2015] that use RNA-RNA interactions to selectively perturb a metabolic state. While the programs `barriers` and `treekin` (see Section 2.3.5) are fast methods to model the folding kinetics of single RNA molecules, they cannot calculate

the kinetics of concentration dependent intermolecular reactions. However, concentration dependency is generally an important aspect for metabolic engineering. Future synthetic pathways have to be energy efficient, i. e. operate at lowest possible concentrations, and they have to be multi-functional. For example, a cellular stress response mechanism leads to a variation in molecule concentrations and, thereby, to an alternative target for a synthetic RNA.

A direct solution of the master equation using matrix decomposition methods, as implemented in `treekin`, cannot capture the chemical kinetics of bimolecular reactions. Therefore, we use a more complicated approach: first the `barriers` method is adapted to coarse-grain the landscapes of the interacting RNA molecules, then these landscapes are merged into a chemical reaction graph. Finally, this graph is translated into a system of ordinary differential equations (ODEs). Together with a vector of initial species concentrations, the ODE system can be solved using Newton's iteration method.

3.2.1 *Theory*

Coarse-graining

Let us assume the most simple scenario of RNA-RNA interactions: two RNA molecules A and B can form the heterodimer AB, then we can write a network of the following uni- and bimolecular chemical reactions:

$$\begin{aligned} A_i &\rightleftharpoons A_j && \text{unimolecular} \\ B_i &\rightleftharpoons B_j && \text{unimolecular} \\ A_k + B_l &\rightleftharpoons AB_m && \text{bimolecular} \\ AB_i &\rightleftharpoons AB_j && \text{unimolecular} \end{aligned}$$

where $A_i \rightleftharpoons A_j$ describes the set of intramolecular structural transitions of the molecule A between any two neighboring conformations i and j. Equivalently, k, l are any two monomeric secondary structures that can reversibly interact to form a dimer conformation m. The structure of such a chemical reaction network is known as a directed hypergraph, where molecules A, B, and (often) also AB form simple, connected subgraphs of isomerization reactions, while the bimolecular interactions introduce hyperedges connecting the three vertices $A + B + AB$ at once. A formal definition translating such hypergraphs into a bipartite chemical reaction graph follows below and is depicted in Figure 11.

Definition 3.1 *Consider a chemical reaction network $\mathcal{G}$ composed of a set of molecules $\mathcal{X}$ and a set of uni- and bimolecular chemical reactions $\mathcal{R}$. Such a structure composes a directed biparitite graph $\mathcal{G}(\mathcal{X}, \mathcal{R}, \mathcal{E})$, in which secondary structures $\mathcal{X}$ and reactions $\mathcal{R}$ are represented by two different types of vertices, and the set of edges $e \in \mathcal{E}$ corresponds to educts $e_{x \to r}$ or*

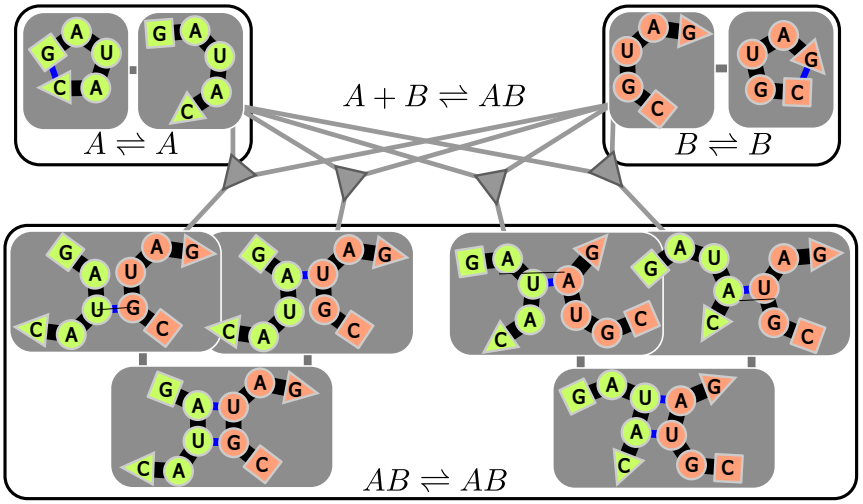


Figure 11: Graph representation of a fine-grained RNA-RNA interaction network involving the molecules A, B, AB. Gray boxes are nodes in the network, which can, alternatively, represent macro-states such as gradient basins of the energy landscape. Black borders separate sub-graphs of unimolecular isomerization reactions, the bimolecular reactions are depicted in the middle with edges connecting three species at once.

products $e_{r\to x}$ *with* $x \in \mathcal{X} \wedge r \in \mathcal{R}$*. For unimolecular reactions, there is only one educt and one product, such that we write the abbreviation* $e_{x\to y}$ *with* $x, y \in \mathcal{X}$ *for the formal bipartite path:* $e_{x\to r}, e_{r\to y}$.

The program `barriers` was adapted to process cofolded suboptimal structure output, and is used to construct the graph $\mathcal{G}$ in a stepwise process. First, the macro-states of monomer landscapes are added together with unimolecular isomerization reactions (A, B, AB). This yields (at least) three disconnected components of the monomer landscapes which are then connected by processing the energy landscape of cofolded secondary structures. Hence, only after this last step, the graph is fully connected and contains the bimolecular ($A + B \rightleftharpoons AB$) reaction vertices.

This approach is necessary, as `barriers` cannot distinguish whether a reaction between two RNA secondary structures is uni- or bimolecular and it is important to ensure that unimolecular reactions do not proceed via transition states of bimolecular reactions. Also, the approach is very useful in practice, since molecules A and B can be very different in length, leading to one molecule dominating the cofolded suboptimal structures. The separate computation of unimolecular reactions balances the number of conformations considered for each molecule, while the combined suboptimal structure space then identifies the lowest energy interaction pathways.

It is straightforward to add additional bimolecular reactions (e. g. $A + A \rightleftharpoons AA$) or a third species C. The currently limiting factors for more complicated systems are the re-

striction to cofolded structures using the `RNAcofold` energy model and, in general, the exponential structure space which limits the maximum complex size to approximately 70 nucleotides. Also, if the structure space of one of the monomers is to large, it is not feasible to connect the unimolecular subgraphs in the cofolded energy landscape.

NOTATION Following our previous notation of partition functions, we write the monomeric landscapes as $\mathcal{L}_A$ and $\mathcal{L}_B$, the landscape of cofolded monomers and dimers as $\mathcal{L}_{AB}$, and the dimer-only landscape as $\mathcal{L}'_{AB}$.

$\mathcal{L}_{AB}$ is exactly the union of unimolecular reaction landscapes $\mathcal{L}_{AB} = \mathcal{L}_A \cup \mathcal{L}_B \cup \mathcal{L}'_{AB}$, and it is the only landscape containing concentration dependent bimolecular reactions. Recall the flooding algorithm implemented in the program `barriers` (Section 2.3.5), then it is straightforward to partition the landscapes $\mathcal{L}_A$ and $\mathcal{L}_B$ into macrostates and compute the partition function for the basins of attraction. Using the same approach for $\mathcal{L}'_{AB}$, we note an important difference: secondary structures that do not interact, e. g. the open chain conformation of either molecule A or B, are not in the set of valid conformations. Ergo, the landscape $\mathcal{L}'_{AB}$ is not necessarily ergodic (e. g. Figure 11).

The suboptimal conformation space of $\mathcal{L}_{AB}$ is processed separately to add vertices and edges of bimolecular reactions. As pointed out previously, the `barriers` implementation does not distinguish between *true* dimers forming an intermolecular base-pair and two separate monomers. Thus, in some cases, gradient walks differ in unimolecular and bimolecular landscapes and we briefly address this problem to analyze the expected impact on kinetic simulations.

Let $x \in \mathcal{L}_A$ and $y \in \mathcal{L}_B$ be monomer conformations in the respective monomer landscapes, then the conformations can only form a dimer if there exits a neighboring conformation to both structures in $\mathcal{L}'_{AB}$. For a gradient walk to lead from two monomers into a dimer state, there must exist a dimer neighbor $z' \in \mathcal{L}'_{AB}$ with a free energy lower or equal than the sum of both monomer structures $E(z') \leqslant E(x) + E(y)$. We observe that this is impossible with current energy parameters, since the entropic dimer-initiation penalty cannot be overcome by a single-base-pair change. However, the reverse case *is* possible, as single-base-pair changes can be favorable if they come with an entropic bonus for breaking dimers into monomers. Thus, every gradient walk starting in a conformation with a single intermolecular base-pair is likely to lead into a gradient basin that formally describes two monomer states in the cofolded suboptimal structures. Figure 12 shows how we can count this effect when comparing unimolecular and bimolecular barrier trees. The gradient-basin partition functions of monomers are smaller in monomer landscapes, while the gradient basin partition functions of dimers are smaller in the landscape of cofolded monomers and dimers.

Occasionally, single-base-pair dimers are assigned to monomer gradient basins

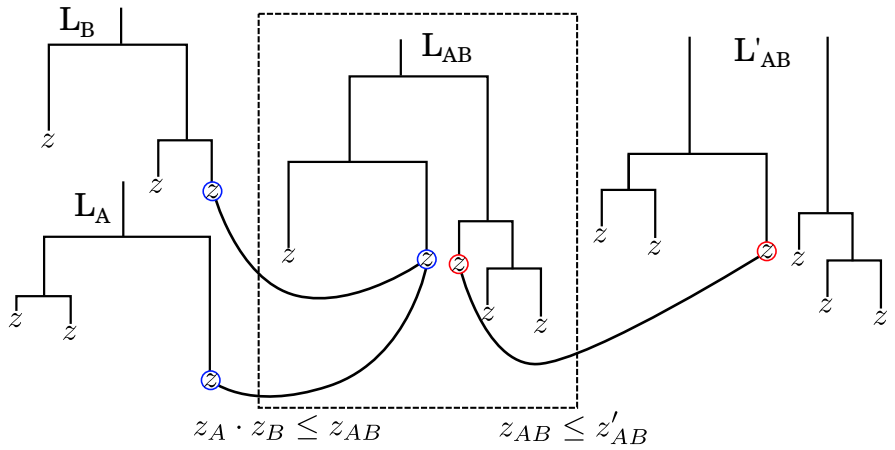


Figure 12: The barrier trees of four landscapes $\mathcal{L}_A, \mathcal{L}_B, \mathcal{L}_{AB}, \mathcal{L}'_{AB}$. $\mathcal{L}'_{AB}$ may be a forest, for the reasons described in the main text. The partition function of every molecule is split into gradient basin partition functions (z) for each leaf in the barrier tree. If the barrier trees are composed of the same set of suboptimal structures, then $Z_{AB} = Z_A Z_B + Z'_{AB}$, however, the gradient basin partition functions may still differ. Gradient walks assign dimer structures into monomer basins and hence lead to $z_A \cdot z_B \leqslant z_{AB}$ and $z_{AB} \leqslant z'_{AB}$.

Transition rates

The computation of effective transition rates between two gradient-basins α, β has been introduced in section 2.3.5, and is computed as

$$k(\alpha \to \beta) = \sum_{x \in \alpha} \sum_{y \in \beta} P(x|\alpha) k(x \to y) \tag{22}$$

where $P(x|\alpha)$ assumes that the ensemble of structures in basin α is instantaneously in equilibrium. Micro-rates $k_{x \to y}$ are computed using the Metropolis rule (Section 2.3.3) with free energy differences calculated from single base-pair transitions. The rates for uni- and bimolecular reactions are assigned to corresponding reaction vertices (or edges) in the previously defined reaction graph $\mathcal{G}$ and enable the formulation of a system of ODEs. In particular, the macro state A_α reacts with the macro-state A_β with forward and backward reaction rates k_f and k_b

$$A_\alpha \underset{k_b}{\overset{k_f}{\rightleftharpoons}} A_\beta \qquad \text{equilibrium:} \qquad k_f[A_\alpha] = k_b[A_\beta] \tag{29}$$

where $[A_\alpha]$ and $[A_\beta]$ denote molar concentrations. The first order ODEs derived from this reaction, calculate the change in concentration for $[A_\alpha]$ (and likewise $[A_\beta]$) within

time t, starting with an initial concentration and ending in thermodynamic equilibrium given sufficient time $t \to \infty$.

$$\frac{d[A_\alpha]}{dt} = k_b[A_\beta] - k_f[A_\alpha]$$
$$\frac{d[A_\beta]}{dt} = k_f[A_\alpha] - k_b[A_\beta]$$

The bimolecular reactions follow the same rules

$$A_\alpha + B_\beta \underset{k_b}{\overset{k_f}{\rightleftharpoons}} AB_\gamma \quad \text{equilibrium:} \quad k_f[A_\alpha][B_\beta] = k_b[AB_\gamma] \tag{30}$$

however, they lead to concentration-dependent second order ODEs

$$\frac{d[A_\alpha]}{dt} = k_b[AB_\gamma] - k_f[A_\alpha][B_\beta]$$
$$\frac{d[B_\beta]}{dt} = k_b[AB_\gamma] - k_f[A_\alpha][B_\beta]$$
$$\frac{d[AB_\gamma]}{dt} = k_f[A_\alpha][B_\beta] - k_b[AB_\gamma]$$

with $\frac{d[A_\alpha]}{dt} = \frac{d[B_\beta]}{dt} = \frac{d[AB_\gamma]}{dt} = 0$ at thermodynamic equilibrium.

3.2.2 *Implementation*

The above concepts are realized in the `Python/Perl` pipeline `interkin/SundialsWrapper` that writes and compiles `C` code for an intermolecular RNA folding kinetics simulation software (Algorithm 1). `interkin` first computes suboptimal structures with `RNAsubopt` [Wuchty et al., 1999], then coarse-grains the energy landscape using `barriers` [Flamm et al., 2002] and merges the output into the previously described bipartite graph $\mathcal{G}$ (see Definition 3.1). Upon coarse-graining from `RNAsubopt-barriers` runs for monomers, dimers and combined energy landscapes, the graph $\mathcal{G}$ is translated into a system of ODEs and written into ready-to-compile `C` code for numeric integration using the `Sundials CVODE` integrator [Hindmarsh et al., 2005; Cohen and Hindmarsh, 1996].

Computing the isomerization reactions within a dimer complex is done via filtering of the cofolded suboptimal structure output. All non-interacting structures are removed and only conformations with at least one intermolecular base-pair are processed with `barriers`. In order to support cofolded RNA structure input, the `barriers` neighbor generation routine had to be adapted to find hairpin loops with 0-2 unpaired nucleotides, which is the *only* case where the new intermolecular cut-loop violates the previous definition of single secondary structures 2.2. This new feature is activated automatically when reading cofolded `RNAsubopt` output in `barriers-v1.6`.

Suboptimal structures are sorted by energy *and* lexicographically, in order to ensure a consistent macro-state decomposition for degenerate landscapes. Hence, `barriers`

Algorithm 1 `interkin` pipeline

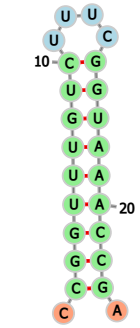
```
1:  S = set of single RNA molecules;                          ▷ Input
2:  G = ();                                                   ▷ Empty reaction graph
3:  for all x ∈ S do                                          ▷ Monomers
4:      G ← G ∪ coarse_grain(x, regular)
5:  end for
6:  for all {x, y} ∈ S do                                     ▷ Dimers for cofolding
7:      G ← G ∪ coarse_grain(x&y, filter)                     ▷ '&' connects two monomers
8:      G ← G ∪ coarse_grain(x&y, total)
9:  end for
10: print ODEs(G)
11:
12: procedure COARSE_GRAIN(x, mode)
13:     Ω ← RNAsubopt(x)                                      ▷ single or cofolded suboptimals
14:     if mode=filter then
15:         Ω ← filter_true_dimers(Ω)
16:     end if
17:     Ω ← sort(Ω)                                           ▷ 1: energetically, 2: lexicographically
18:     L ← barriers(Ω)                                       ▷ print and parse the rate matrix
19:     if mode=total then
20:         G ← add_bi_rates(L)                               ▷ connect subgraphs
21:     else
22:         G ← add_uni_rates(L)                              ▷ add subgraphs
23:     end if
24:     return G
25: end procedure
```

always selects the same conformation as a representative for the respective gradient basin, independently of whether the full landscape or the unimolecular subsets are coarse-grained. In order to assess the quality of kinetic folding simulations, the expected equilibrium distributions (including coarse-graining errors mentioned previously) can be directly computed from gradient basin partition functions of unimolecular and cofolded energy landscapes.

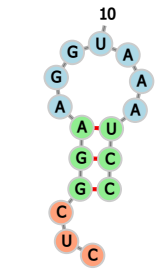
`barriers` returns rates in form of arbitrary time units^{-1}. `interkin` translates these rates into wall-clock time (seconds^{-1}) with the constant scaling factor $k_0 = 2 \cdot 10^5$. This value will be explained in detail in the context of cotranscriptional folding in section 4.1.2.

SundialsWrapper

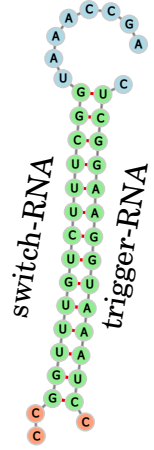
We have decoupled the process of translating the reaction-graph into a ready-to-compile `C` code as a standalone, `Perl` module: `Chemistry::SundialsWrapper`. `SundialsWrapper` comes with template `C` source-code files using the Sundials `CVODE` library [Hindmarsh et al., 2005; Cohen and Hindmarsh, 1996]. A reaction-graph, such as produced from `interkin`, is translated it into a system of ODEs and combined with the `CVODE`-templates to produce simulation software. For convenience, the programs have the same command line interface as the simulator for isomerization reactions `treekin`.



switch-RNA



trigger-RNA



3.2.3 *Results*

As a first example, we investigate the folding kinetics of a rationally designed pair of RNA molecules. Both sequences are short, 25 and 17 nucleotides, and they are designed to mimic the behavior of riboswitches. The sequence of 25 nucleotides is a *switch-RNA*, forming either a stable hairpin structure (off) or not (on). In the absence of the 17 nucleotide *trigger-RNA*, the equilibrium is dominated by the off state, otherwise, the trigger-RNA unfolds the switch-RNA from off to on conformation. This mechanism is generally called *on-switch*, while *off-switches* start in on-conformation and are turned off by a trigger molecule. The 17 nucleotide sequence is mostly unstructured but complementary to the riboswitch in order to favor the conformational switching.

Figure 13 shows the calculated equilibrium distributions of all possible monomer and dimer species as a function of molar concentrations. The switch RNA is in comparatively low concentrations, while the trigger RNA is supplied in excess. The distributions are computed from the partition functions using `RNAcofold`. As the sequences are rather short, high concentrations are needed to yield dimerization in the first place, e. g. 1 mM of the trigger-RNA converts about 50% of the switch monomer into a switch-trigger dimer.

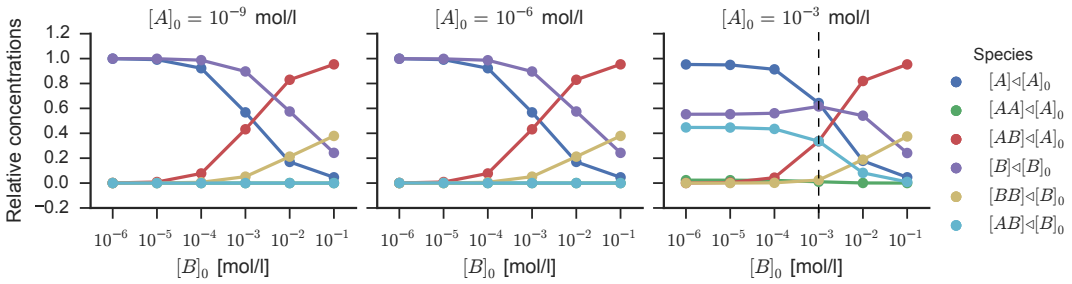
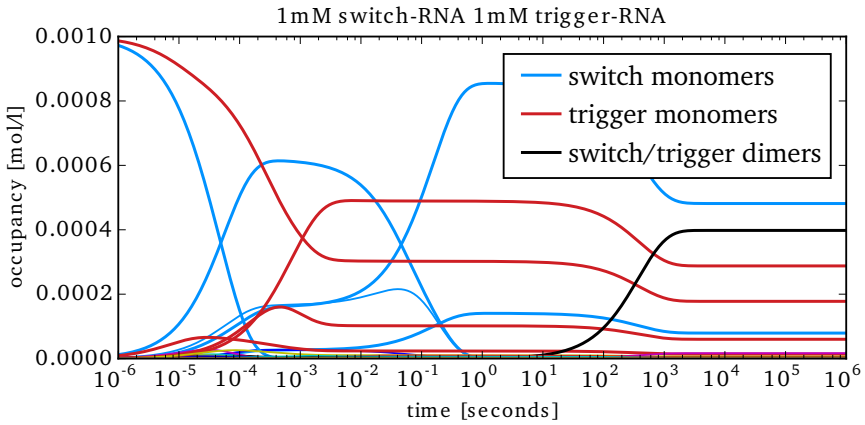


Figure 13: Equilibrium concentrations of a designed pair of RNA molecules. The switch-RNA A and the trigger-RNA B shown at their respective monomer concentrations, ([A], [B]) and dimer concentrations ([AA], [BB], [AB]). As molecules are mostly present in different concentrations, the plots show every species relative to the initial concentration of A, i. e. $[A]_0$ and the initial concentration of B, i. e. $[B]_0$. Increasing the molar concentrations leads to a higher yield of dimers in equilibrium, e. g. when 0.1 mol/l of B are present, the relative concentration of [BB] nearly approaches the maximum of 50%. Due to the high differences of concentrations shown in the first two plots, AB is generally not populated with respect to $[B]_0$, but reaches close to 100% relative to $[A]_0$. The vertical dashed line in the third plot marks the expected equilibrium of species in the kinetic simulation shown in Figure 14. Plots were drawn using `seaborn` [Waskom et al., 2014] and `matplotlib` [Hunter, 2007].

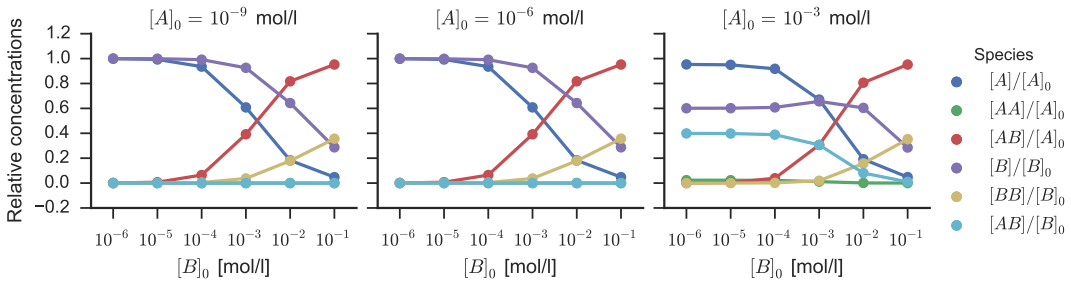
In Figure 14, we show the corresponding kinetic simulation, when both molecules are present in 1 mM concentration. The simulations start the folding process in the open chain conformation (i. e. with free energy of 0 kcal/mol). Intramolecular folding is fast, the molecules reach intramolecular thermodynamic equilibrium after less than one second. However, the formation of the heterodimer starts after only more than ten seconds and reaches the predicted 40 % occupancy (see Figure 13) after about 30 minutes.

`barriers` allows the user to reduce the number of macro-states by adjusting the minimum barrier height that separates a local minimum from its neighboring basins. Merging basins in bimolecular landscapes, however, can lead to a mixture of monomer and dimer basins, which are then assumed to be in instant equilibrium during kinetic simulations. The simulation in Figure 14 has been computed from landscapes with only 10^{-3} kcal/mol minimum basin height, in order to keep this effect as small as possible. Alternative values for basin heights and their influence on the equilibrium concentrations can be seen in Figure 15. The concentrations at thermodynamic equilibrium are calculated from the partition functions of all gradient basins in the respective barrier trees. The low barrier height of 10^{-3} kcal/mol only slightly increases the concentrations of monomers relative to the dimer formation, which is an expected side-effect of the `barriers` algorithm applied to cofolded secondary structures. Raising the barrier height to 1 kcal/mol does not make a larger difference, as the separation of

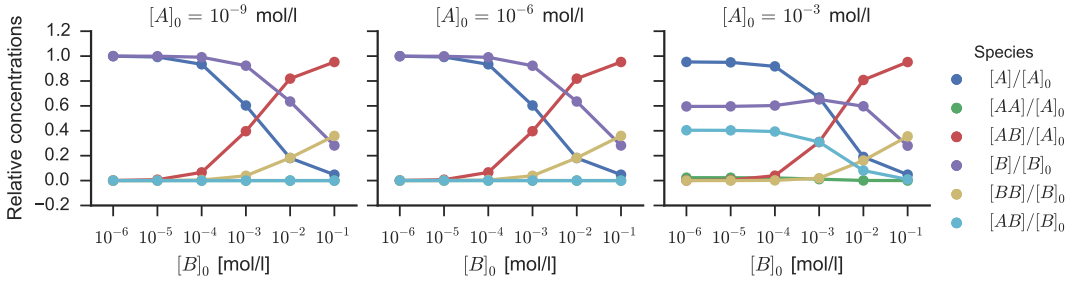


(a) basin height = 10^{-3} kcal/mol, 100 gradient basins

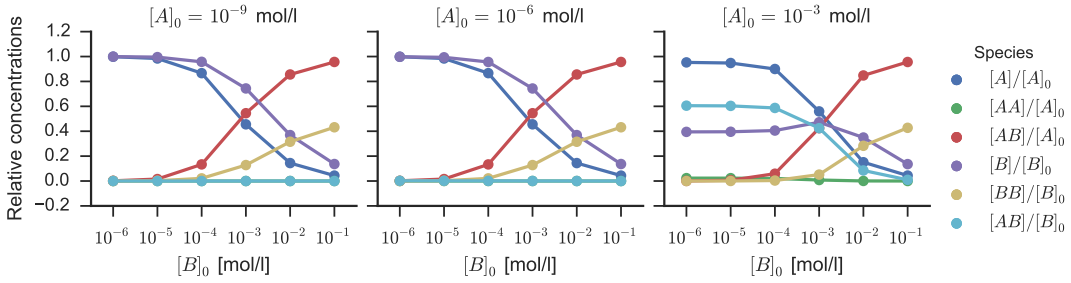
Figure 14: A folding simulation of two rationally designed RNA molecules (see main text). The expected equilibrium distribution can be seen in Figure 13. The switch-RNA (blue) and the trigger-RNA (red) are both present with a concentration of 10^{-3} mol/l. At this concentrations, no homo-dimers do form. Simulations start with both molecules being initially "unstructured", i. e. in their open chain conformation. Intramolecular equilibrium is reached in less than one second, with two populated structures for the switch-RNA and three populated conformations for the trigger-RNA. Note that the open chain conformation still has considerable occupancy when reaching equilibrium. The formation of the AB dimer MFE structure, black trajectory, starts after about 10 seconds and takes up to $2 \cdot 10^3$ seconds, i. e. roughly 30 minutes to reach the equilibrium occupancy of 40%.



(a) basin height = 10^{-3} kcal/mol, maximum number of basins = 500



(b) basin height = 1 kcal/mol, maximum number of basins = 500



(c) basin height = 2 kcal/mol, maximum number of basins = 500

Figure 15: Equilibrium concentrations of monomers [A], [B] and dimers [AA], [BB], [AB] at different concentrations. The sequences are the same as in Figure 13, but this time the equilibrium distributions are computed from the partition functions in cofolded barrier-trees. **(a)** A low minimum barrier height only leads to small differences compared to the previously shown calculations using `RNAcofold`. The findings fit well to the corresponding kinetic simulation shown in Figure 14. **(b)** Raising the barrier height to 1 kcal/mol has no visible effect on thermodynamic equilibrium, as the separation of monomers and dimers in the coarse-grained landscape is the same as for 0.001 kcal/mol. **(c)** A barrier of 2 kcal/mol substantially influences the accuracy of simulations. The formation of dimers is now favored, indicating that monomer basins of attraction have been merged into dimer basins while flooding the energy landscape.

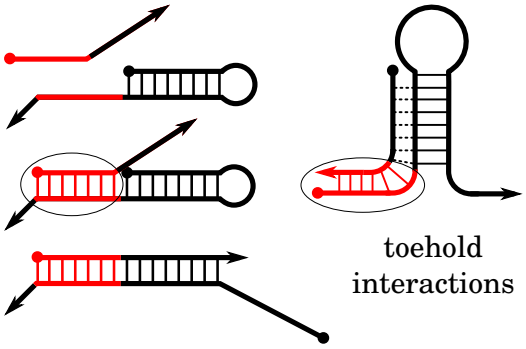


Figure 16: Schematic of the toehold mechanism used in nucleic acid design and computation. Arrows mark the direction from the 5′ to the 3′ end of the RNA. **Left:** The toehold (red) is an originally single-stranded region in both molecules that is available for intermolecular base-pairing. It is supposed to accelerate the rate for intermolecular conformational refolding. **Right:** the toehold mechanism as usually drawn for RNA switches to unwind a present helix.

monomers and dimers into different basins is still intact. In contrast, a further increase of the minimum barrier height to 2 kcal/mol destroys the separation of monomers and dimers, apparently leading to monomer basins being merged into dimer basins during the flooding of the energy landscape. As a consequence, dimers are overrepresented in thermodynamic equilibrium.

Toeholds to reduce refolding times

Kinetic analysis is a time consuming process and therefore, modern nucleic acid sequence design uses *ad-hoc* criteria as an alternative. With `interkin` it is now possible to analyze some of these strategies, in order to assess their efficiency. One popular design strategy is to make use of *toeholds* (see Figure 16). Toeholds are single stranded regions that lower the energy barriers for structural rearrangements. More detailed, different requirements for toehold design depend on the research question.

In many successful RNA sequence design studies [Isaacs et al., 2004; Green et al., 2014; Chappell et al., 2015], toeholds mediate conformational switching. Their primary function is to stabilize intermolecular RNA-RNA binding, before the actual refolding takes place. In this case, the toehold is designed to form a strong interaction that also contributes thermodynamically to the switching effect.

For nucleic acid computation, toehold-binding has to be reversible. The interaction must be strong enough to overcome entropic penalties for intermolecular base-pairing, but weak enough to dissociate again if there are no additional base-pairs formed. The efficiency of reversible toeholds has been studied experimentally [Yurke and Mills Jr, 2003; Zhang and Winfree, 2009; Srinivas et al., 2013], however, only recently compu-

tational methods have been used to enhance the understanding of the underlaying biophysical processes [Srinivas et al., 2013].

SEQUENCE DESIGN We now show comparisons of folding times for three slightly larger RNA switch/trigger pairs, with 32 and 17 nucleotides. In contrast to the previous example, two of the designs have toeholds to initiate molecular interactions. All of the following switch-trigger pairs are designed to be similar with respect to equilibrium properties (see Figure 17), which enables us to analyze kinetic effects only, rather than differences visible at thermodynamic equilibrium.

- Cliffhanger-design (Figure 17a): is a pair of sequences that uses a toehold for refolding.
- Conan-design (Figure 17b): is a pair of sequences not using a toehold for switching, i. e. the trigger-RNA needs to bind and unfold the hairpin of the switch-RNA at the same time. It is worth pointing out that in an experimental setting, the hairpin loop of the switch-RNA can serve as a toehold as well. This effect can be ignored in our models, as we only consider pseudoknot free transition states.
- Hulk-design (Figure 17c): is a pair of sequences that can chose between refolding with or without a toehold.

Note that Hulk-design had a rather complicated design objective, which required three competing stable conformations. Also sequences are short to make them applicable for analysis with low minimal basin heights. As a consequence, again, high concentrations of molecules are required to observe intermolecular interactions. For kinetic simulations, we specify a concentration of 1µM switch-RNA and 1mM trigger-RNA, to switch approximately 50% of the switch-RNA at thermodynamic equilibrium (see Figure 17). Relaxing the design requirements mentioned above allows the optimization of sequences with a higher tendency toward dimerization. All designs were made using a new nucleic acid sequence design library `RNA::Design`, that will be explained briefly in Chapter 5.

FOLDING KINETICS All sequence designs are expected to form only monomers and heterodimers at the specified concentrations and, therefore, we can assume that homodimers will not influence the intermolecular folding behavior. We have used this fact to reduce the size of the ODE system and compute folding simulations for heterodimer formation only. Figure 18 shows a comparison of folding times for Cliffhanger, Conan and Hulk designs. All simulations are shown in comparison to expected equilibrium distributions.

The results for Cliffhanger and Conan support the use of toeholds for riboswitch design, as Cliffhanger designs switch faster by an order of magnitude. The initial interaction in form of toehold binding forms long before the trigger/switch dimer

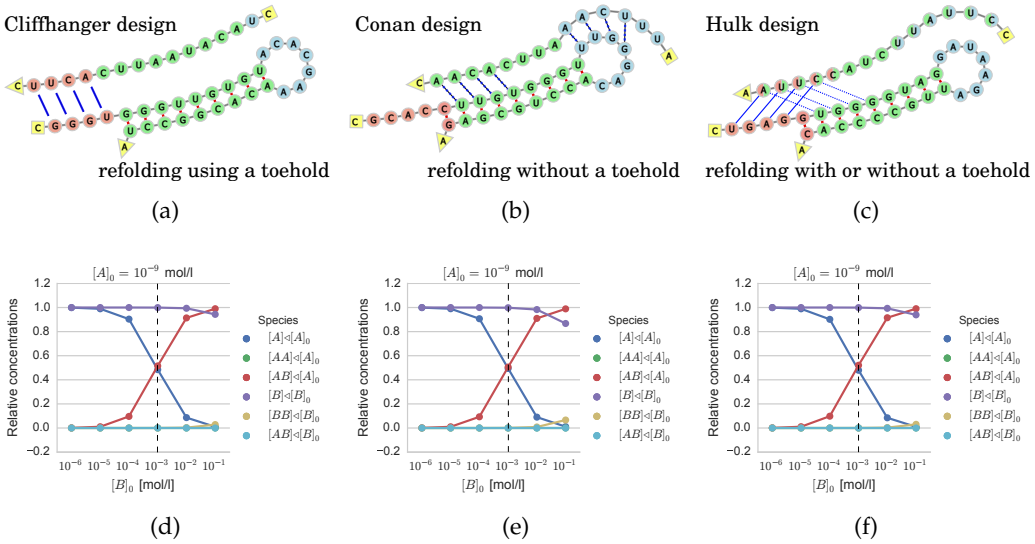
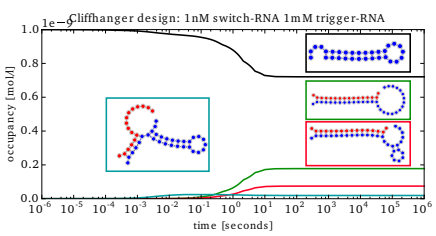
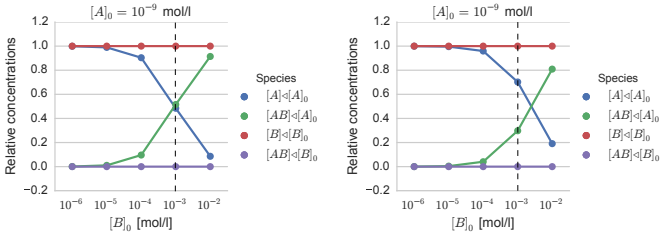


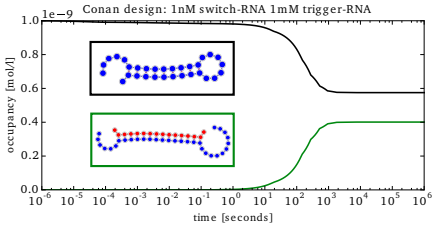
Figure 17: Three different pairs of riboswitches together with trigger molecules. The nucleotide sequences, as well as their intended mechanism of dimerization can be seen in the top row (a-c), while the expected yield of dimers at thermodynamic equilibrium as a function of concentration can be seen below (d-f). **(a)** Cliffhanger design uses a classical toehold as initial interaction. Two available single stranded regions (red) form base-pairs (blue) before the green intramolecular helix is opened. **(b)** Conan design does not have a toehold for refolding. In an experimental setting, one would have to consider that also the (blue) hairpin region may serve as a pseudoknot in the cofolding energy model and therefore cannot enhance refolding rates. **(c)** Hulk design can chose whether it refolds with or without the toehold. Both pathways lead to a helix of similar energy, forming 14 base-pairs. The final conformation is then either the same structure formed by Cliffhanger or Conan. **(d-f)** Concentrations at thermodynamic equilibrium are roughly the same for all three sequence designs. At 1 mM trigger-RNA (dashed line), 50% of the switch-RNA has folded into the AB dimer. Plots were drawn using `forna` [Kerpedjiev et al., 2015a], `seaborn` [Waskom et al., 2014] and `matplotlib` [Hunter, 2007].



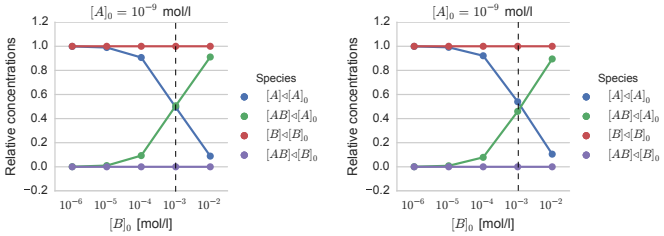
(a) Cliffhanger simulation



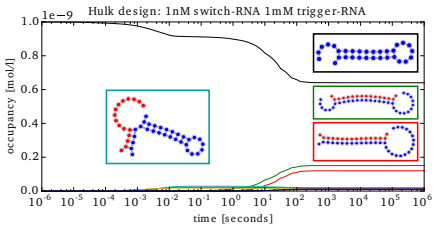
(b) Cliffhanger equilibrium



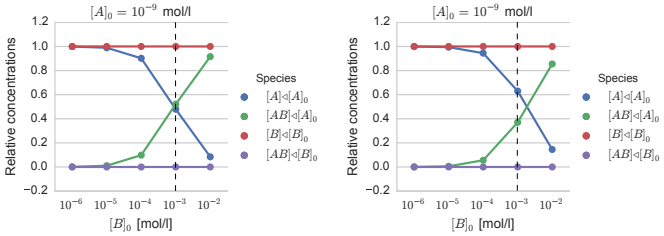
(c) Conan simulation



(d) Conan equilibrium



(e) Hulk simulation



(f) Hulk equilibrium

Figure 18: Differences in refolding times of Cliffhanger, Conan and Hulk designs and their expected equilibrium distributions. **Left column:** the kinetic simulations using `interkin`, **center:** exact equilibrium distributions calculated from `RNAcofold` partition functions, **right column:** expected equilibrium distributions calculated from the sum over all gradient basins in cofolded energy landscapes. **(a)** Cliffhanger: the toehold interaction forms fast, between 10^{-3} and 10^{-2} seconds, the unfolding of the switch-RNA from off into on conformation takes up to 10 seconds. **(b)** The equilibrium calculated from `RNAcofold` differs from the equilibrium observed in the kinetic simulation, while the equilibrium calculated from gradient basins in the cofolded energy landscapes (right) fits well with the observed equilibrium. **(c)** Conan: folding into the heterodimer is slower by 1-2 orders of magnitude, i. e. 10 to 10^3 seconds. **(d)** The simulation finishes roughly as expected from both the `RNAcofold` partition function and the sum-of-gradient basins partition function. **(e)** Hulk: similar to Cliffhanger design, Hulk forms the toehold after a few milliseconds, the folding into the MFE heterodimer takes longer than Cliffhanger, but is still faster than Conan. The heterodimer helix not using the toehold interaction (green) populates faster than the toehold-containing conformation (red). **(f)** The simulations differ from the exact equilibrium calculated from `RNAcofold`, but they end approximately in the equilibrium distribution calculated from gradient basins.

folds into the designed on state. The same can be observed for Hulk design, however, the trigger/switch ground states are reached at a slower rate compared to Cliffhanger. Surprisingly, from both populated conformations in equilibrium, the conformation *not* containing the toehold forms first, or at least at the same rate as the conformation containing the toehold.

This is in contrast to our initial intention: showing that the toehold containing structure populates first, and only then refolds into the equilibrium distribution. The toehold, however, enhanced the rate of both refolding pathways, although it was intentionally designed to favor only one of them.

3.2.4 *Discussion*

We have observed an additional important aspect during the sequence design and kinetic analysis process to produce the results shown. The main factor to determine reaction rates is the energy of the transition state, i. e. the energy barrier separating the conformations. Often this energy barrier occurs when forming the first intermolecular contact. In this case a toehold improves the rate of dimerization. However, we have initially planned to present switch-RNAs with hairpins known to terminate transcription, which will be discussed in more detail in the context of cotranscriptional folding in Chapter 4. The difference is that these conformations have stable tetra-loop motifs contributing a lot of folding free energy. In that case, the reaction rate is determined by unfolding the hairpin loop, not by the initial opening of the stem. This makes it possible to generate Conan designs (no toehold) that are faster than corresponding Cliffhanger designs (with toehold).

Our examples for toehold mediated interactions confirm these results. While we have demonstrated how toeholds are functional for increasing dimerization rates, the mode of action is more flexible than initially thought. Toehold regions do not need to be part of the final dimer structure, they are only required for establishing the initial contact. From there, intramolecular folding rates dominate the conformational rearrangements, and in the case of Hulk design, the transition state toward the non-toehold interaction is actually the same as that toward the toehold structure. The true-dimer barrier tree in Figure 19 visualizes this effect.

During the analysis of results, we realized that some of the simulations (e. g. Cliffhanger in Figure 18) do not reach the expected equilibrium distributions predicted from `RNAcofold`. In order to exclude accumulating errors from numerical instabilities we have reduced the problem to only heterodimer formation. However, the effect on the results was negligible, as homodimer species are not forming at the specified concentrations. We then calculated the equilibrium distributions directly from gradient basin partition functions in the coarse-grained cofolded energy landscapes. These new distribution fit better, as the macro-states are also used to calculate transition rates. A less well measurable factor inaccuracies is that transition rates between gradient

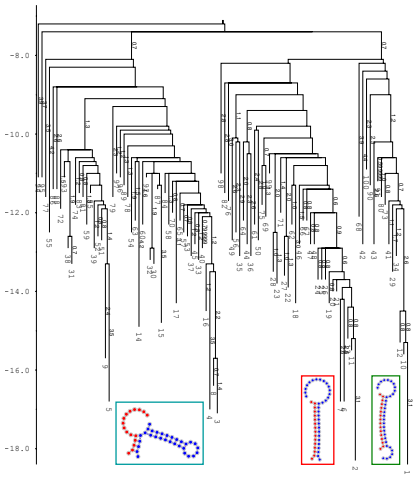


Figure 19: A dimer-only barrier tree from the Hulk sequence design. The red and green conformation correspond to local minimum 2 and 1 and differ by 0.10 kcal/mol. The depicted toehold interaction corresponds to local minimum 3. The three states are separated by the same transition states, thus have the same refolding barriers: $\Delta G^{\ddagger}_{3\to2} = \Delta G^{\ddagger}_{3\to1} = 9.70$ kcal/mol, $\Delta G^{\ddagger}_{2\to1} = 10.20$ kcal/mol.

basins can be prone to numerical errors, e. g. posterior evaluation of cliffhanger reaction rates showed that not all rates perfectly obeyed detailed balance, with the effect for the dimer-only landscape $\sum_{\alpha,\beta\in\mathcal{L}'_{AB}} P(\alpha)k_{\alpha\to\beta} - P(\beta)k_{\beta\to\alpha} = 0.59s^{-1}$. We hope to reduce this effect in the future by scaling the values appropriately.

It is also important to mention that dimers with a single intramolecular base-pair, i. e. the ones that are often assigned into monomer basins, are the transition states for bimolecular reactions. In Figure 20 we show this effect in the context of transiton rate computations: ideally, all bimolecular reactions proceed with a rate for forming or breaking a single intermolecular base-pair. However, if the gradient walk assigned a dimer structure into a monomer basin, then the probability $P(x|\alpha)$ becomes worse, and the reaction rate is spontaneous ($k_{x\to y} = k_0\, s^{-1}$, see Section 3.2.1). In the reverse reaction, $P(y|\beta)$ will be overestimated, but now the reaction rate quantifies the formation of a single intermolecular base-pair conformation rather than being spontaneous. A future implementation of `barriers` could circumvent this problem at the cost of increased runtime by keeping track of which cofolded structures correspond to monomers and dimers.

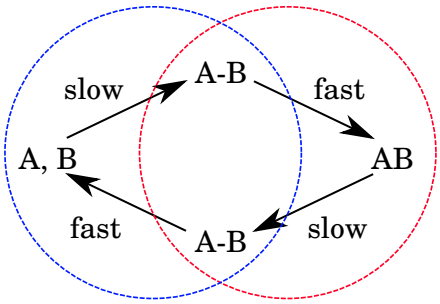


Figure 20: Schematic of a coarse-graining error and its impact on kinetic simulations. A and B are two cofolded RNA molecules coarse-grained using barriers. When the two molecules A, B react to form a dimer AB, the reaction has to proceed via the transition state forming a single base-pair A-B. The reaction into the transition state is always slow, while the subsequent reaction is spontaneous. The circles show macrostates that are modeled to be in instant equilibrium. The red circle (right) shows the proper way of separating monomers and dimers using gradient basins. In that case all AB dimers are in instant equilibrium. The blue circle (left) shows the problem when `barriers` assigns a transition state into monomer basins. Now the reaction reaction rate toward dimer formation is fast, while the backward reaction, forming monomers, is slow.

4

COTRANSCRIPTIONAL RNA FOLDING

RNA molecules are synthesized from DNA templates in a complex and tightly regulated process called *transcription*. The central molecule is the protein *RNA polymerase* that produces the RNA molecule while reading the DNA template. Additional transcription factors may stabilize the complex between RNA polymerase and DNA and they may assist in unwinding of the double stranded DNA helix structure [Larson et al., 2011; Dangkulwanich et al., 2014].

The rate for RNA synthesis has been found to vary between 10 to 100 nucleotides per second [Chamberlin and Ring, 1973; Bremer and Dennis, 1996; Larson et al., 2008], while the formation of single base-pairs in a helix can be on the order of 10^{-6} to 10^{-8} seconds [Pörschke and Eigen, 1971; Pörschke, 1974]. The RNA structures forming during transcription can alter the conformation found at the end of transcription, but they can also terminate the transcription process itself. For example, it has been shown that cotranscriptional folding can lead to significant changes in secondary structure [Kramer and Mills, 1981; Xayaphoummine et al., 2007], that it can speed up the folding into the MFE structure by sequential folding [Heilman-Miller and Woodson, 2003; Zhang et al., 2009] and that transcriptional pausing assists for folding large molecules [Wong et al., 2007]. Interactions with small metabolites can induce early transcription termination by mediating the formation of *terminator hairpin* structures [Wickiser et al., 2005; Mandal and Breaker, 2004; Lemay et al., 2011; Wachsmuth et al., 2013; Chappell et al., 2015].

IN THIS CHAPTER we model RNA folding kinetics during transcription. Section 4.1 introduces previous work on cotranscriptional folding, including a number of existing programs and their strategies to relate folding simulations to wall-clock time. Section 4.2 describes folding on dynamic energy landscapes as implemented in the original `BarMap` program [Hofacker et al., 2010] and my own contributions to improve and adapt `BarMap-v2.0` for metabolite binding riboswitches. Parts of this section have been published in a recent book chapter on computational modeling of riboswitches [Badelt et al., 2015b]. We predict the termintation of premature transcripts after about 70 nucleotides have been transcribed. Section 4.3 describes a new heuristic for cotranscriptional folding prediction implemented in the program `DrTransformer`. We compare different programs and show simulations of experimentally confirmed 200 and 660 nucleotide RNA molecules. The text will be adapted for publication together with Peter Kerpedjiev and Ivo L. Hofacker. Peter contributes visualization solutions for cotranscriptional folding [Kerpedjiev, 2016].

4.1 FOLDING ON DYNAMIC ENERGY LANDSCAPES

RNA molecules can sense and react to environmental stimuli, such as a change in temperature, a change in ion concentrations, or a change of present interaction partners. These environmental influences alter the energy landscape of the RNA molecule, e. g. ligands stabilize RNA binding pockets and lead to conformational rearrangements. Cotranscriptional folding is a quite drastic form of changing the energy landscape, since it grows exponentially with every newly transcribed nucleotide. The MFE secondary structure at a particular molecule length is often not a substructure of the full length MFE conformation. Hence, predictions of cotranscriptional folding dynamics may help to increase the accuracy of secondary structure prediction methods in terms of which conformations are actually formed during the lifetime of a molecule.

4.1.1 *Previous work on cotranscriptional folding*

Present algorithms fall into three categories: Stochastic simulations (`Kinfold` [Flamm et al., 2000], `Kinefold` [Xayaphoummine et al., 2005], `RNAkinetics` [Danilova et al., 2006]), master equation methods (`BarMap` [Hofacker et al., 2010], theoretical work from Zhao et al. [2011]) and deterministic prediction of a single folding trajectory (`Kinwalker` [Geis et al., 2008]).

Stochastic simulations model single folding trajectories through the energy landscape and provide detailed information about microscopic pathways. They are generally easy to adapt for cotranscriptional folding, as unpaired 3′ nucleotides are added whenever the simulation time exceeds the time-threshold for chain elongation. At the single-base-pair resolution, the Gillespie-type simulator `Kinfold` has been adapted to compute single statistically correct cotranscriptional folding trajectories. In practice, however, a single folding trajectory in the high dimensional energy landscape gives very little information, and while a set of e. g. 10^4 trajectories for RNA molecules of 30 nucleotides length can be considered as a *correct* result, for longer molecules the simulations become time intensive and potentially important parts of the energy landscape may still never have been observed.

Xayaphoummine et al. [2007] used the stochastic simulator `Kinefold` to design, model, and experimentally confirm RNA sequences with cotranscriptionally trapped folding pathways. `Kinefold` uses a move set based on whole-helix transitions, which lowers the chances of getting trapped in transient local minimum conformations. This makes their program applicable for longer RNA molecules, at the cost of ignoring parts of the energy landscape. `Kinefold` combines the stacking energies from the NN energy model with an analytic approximation of entropic loop penalties, enabling them to support pseuoknotted folding pathways [Isambert and Siggia, 2000].

`RNAKinetics` Danilova et al. [2006] is a stochastic Gillespie-type simulator using the standard NN free energy parameters. It uses a move set based on whole helix transi-

tions, combined with an alternative approach to compute transition rates. `Kinefold` and `RNAKinetics` are available as web-interfaces, however, unfortunately both programs are closed source.

Exact methods using the chemical master equation consider the complete ensemble of conformations and account for the average kinetic effect of each and every transition. In this case, the number of RNA secondary structures increases exponentially with chain length. However, clustering or coarse-graining methods have been developed to make the problem of RNA folding computationally tractable.

Zhang and Chen [2002] use the master equation in combination with a statistical mechanical energy model for RNA hairpin formation. Based on this energy model they developed a clustering approach to identify rate-limiting steps during hairpin formation [Zhang and Chen, 2006a,b] and simulate cotranscriptional folding using a helix-transition move set [Zhao et al., 2011].

`BarMap` [Hofacker et al., 2010] simulates cotranscriptional folding using the previously introduced coarse-graining of energy landscapes into barrier trees (Section 2.3.5). A mapping between the changing energy landscapes during transcription is used to transfer populations of dominant secondary structures. Algorithmic details will be explained in section 4.2. As every secondary structure in the low parts of the energy landscape is considered, `BarMap` is rarely applicable for sequences of more than 70 nucleotides length.

In contrast to previous methods, `Kinwalker` [Geis et al., 2008] is a deterministic approach to compute a single, best folding trajectory from the DP matrices filled during MFE folding. In particular, a series of metastable structures are constructed from a combination of thermodynamically optimal fragments. Selection of the next structure, as well as the time needed for a structural transition depends on the energy barrier. The estimation of barriers is done explicitly with either the Morgan-Higgs heuristic [Morgan and Higgs, 1998] or alternatively, the `findpath` algorithm [Flamm et al., 2001]. Both heuristics search for the lowest energy barrier within all shortest paths (see Section 2.3.2). Coarse-graining based on thermodynamic criteria makes `Kinwalker` applicable for sequences of more than 600 nucleotides.

`CoFold` [Proctor and Meyer, 2013] is an attempt to model cotranscriptional folding only with thermodynamic modeling and essentially predicts MFE secondary structures with a penalty on long-range base-pairs. While this reduces asymptotic complexity to that of standard MFE folding, it is hard to argue that the model actually addresses the dynamic aspects of cotranscriptional folding.

4.1.2 *Base-pair transitions at wall-clock time*

In order to adjust the time period for kinetic simulations during transcription, computational RNA folding speed has to be related to the real-time speed of the RNA polymerase. Transcription has mostly been reported on a timescale between 10 to 100

nucleotides per second [Chamberlin and Ring, 1973; Bremer and Dennis, 1996], while RNA folding is faster than 10^5 base-pairs per second during the formation of helices [Pörschke and Eigen, 1971; Pörschke, 1974]. This discrepancy allows small hairpin structures (about 25 bases) to fold even within the transcription bubble in order to stall the polymerase and terminate transcription, e. g. Larson et al. [2008].

A common approach for calculating folding rates is to find the (lowest) energy barrier for a structural transition ($\Delta G^{\ddagger}$) and compute the rate using the Arrhenius equation

$$k = k_0 e^{-\frac{\Delta G^{\ddagger}}{RT}}, \text{ for } \Delta G^{\ddagger} > 0 \tag{12}$$

where k_0 is a constant to adjust the free energy change to wallclock time. Detailed balance holds, if spontaneous reactions with $\Delta G^{\ddagger} \leqslant 0$ occur with the same rate k_0 according to Metropolis (see Section 2.3.3).

Early experiments from Poerschke [Pörschke and Eigen, 1971; Pörschke, 1974] showed that the formation of the nucleation site (i. e. the first 1-2 base-pair stacks) determines the rate for hairpin folding. The subsequent zipping of the adjacent stacks is comparatively fast. This effect is consistent with the parameters of the NN energy model, as it requires two or three base-pairs to compensate for entropic penalties of a hairpin loop closure. `Kinfold` uses the Metropolis or Kawasaki model, (see Section 2.3.3) to compute folding trajectories at unit time, i. e. $k_0 = 1$. Thus, the time for chain elongation is specified in arbitrary time units, and the user can freely adjust the conversion factor to seconds.

Methods using helix kinetics often approximate the rate for helix formation based on the energetically best nucleation point. `Kinefold` inserts a nucleus of length 3 and computes the rates using the Arrhenius law, Zhang and Chen [2006a] use their (kinetic) clustering approach to identify the rate-limiting entropic or enthalpic effects during hairpin formation.

The model used in `RNAKinetics` is based on the assumption that the formation of a helix is proportional to the probability of forming the first stacking base-pair and that every stack in a helix can serve as the nucleation point. Hence, the number of transition states is equal to the number of stacking pairs in the helix. The energy barrier is determined by the entropic penalty[1]

$$k_{form}^{eff} = \kappa_c N_h e^{\frac{T\Delta S}{kT}}$$

with κ_c between 10^{-6} to 10^{-8} (according to Poerschke's results) and N_h is the number of stacks in a helix. For spontaneous decay (k_{diss}^{eff}), the enthalpic free energy contribution has to be paid such that detailed balance holds.

`Kinwalker` also uses Poerschke's experiments to adjust the folding speed. The NN free energies are used to determine the energetically best transition state and barrier

1 http://bioinf.fbb.msu.ru/RNA/kinetics/theory.html

heights translate into first-passage times using the relation $t(\Delta G^{\ddagger}) = 10^{-7} e^{\frac{\Delta G^{\ddagger}}{kT}}$ seconds for $\Delta G > 0$ [Geis et al., 2008].

Sauerwine and Widom [2013] have concluded that (randomly distributed) `Kinfold` steps correspond to $5 \cdot 10^{-6}$ seconds and, hence, 4000 monte-carlo steps on average lead to the measured real-time RNA folding rate during transcription. We have used this as reference point in Badelt et al. [2015b] since `BarMap` and `Kinfold` are based on the same energy model. Also, their result is on the order of early work from Schmitz and Steger [1996], who observe an Arrhenius-prefactor $k_0 = 3.34 \cdot 10^6$ from calibration according to their previous work [Randles et al., 1982].

Hofacker et al. [2010] compared slow and fast transcription with `BarMap` using unit time (i. e. $k_0 = 1$). At slow transcription rate, 77 nucleotides were transcribed in 10^5 arbitrary time units, fast transcription finished in 10^4 arbitrary time units. This corresponds roughly to 100 and 1000 time units per nucleotide, i. e. a transcription speed of 200 and 2000 nucleotides per second when using Sauerwine's relation.

In conclusion, it is not easy to find a consistent factor for the Arrhenius model that converts unit time to wall-clock time, since there is little experimental data for microscopic rate models. Nevertheless, present experimental findings agree that folding speed is in the order of 10^{-6} to 10^{-8} s^{-1} and we use the conversion factor $k_0 = 2 \cdot 10^5$ to translate from arbitrary time units (atu) into seconds, for example:

$$
\begin{aligned}
1 \text{ atu/nuc} &= 5 \cdot 10^{-6} \text{ sec/nuc} \\
2 \cdot 10^5 \text{ atu/nuc} &= 1 \text{ sec/nuc} \\
4000 \text{ atu/nuc} &= 0.02 \text{ sec/nuc} && 50 \text{ nuc/sec} \\
1000 \text{ atu/nuc} &= 0.005 \text{ sec/nuc} && 200 \text{ nuc/sec}
\end{aligned}
$$

4.2 MASS-ACTION KINETICS OF METABOLITE-BINDING RIBOSWITCHES

Riboswitches are RNA molecules that translate environmental signals into a genetic program, in particular, the riboswitches discussed in this chapter can terminate their own transcription process when sensing a small metabolite. Such mechanisms are crucial for altered gene expression under stress conditions, e. g. high temperature, low nutrition, or immune defense. In the context of this work, design of riboswitches together with a trigger mechanism gives synthetic biologists a remote control for gene expression after a cell is successfully transfected. Figure 21 shows a cotranscriptional riboswitch inserted in the 5'-untranslated regions of a protein-coding transcript to sense the presence of a metabolite and toggle the formation of a transcription termination hairpin. See Serganov and Nudler [2013] for a review of other strategies to embed riboswitches into mRNA transcripts.

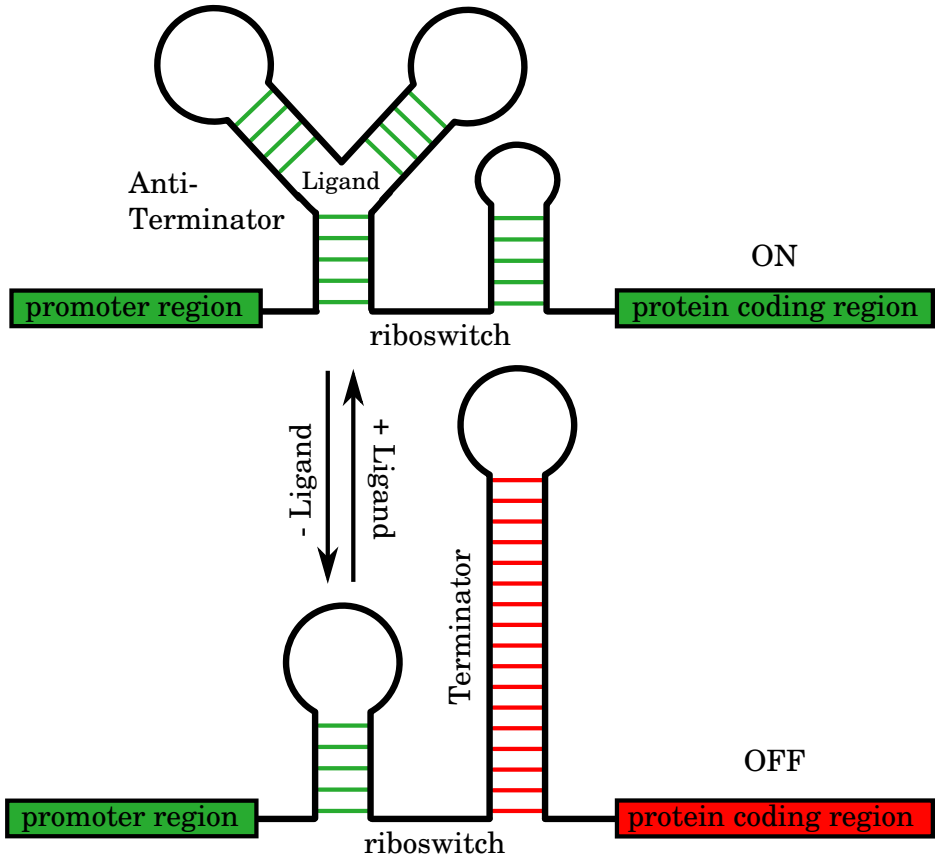


Figure 21: A blueprint to design a cotranscriptional riboswitch. The switch is inserted before the protein coding region in an RNA transcript. If the trigger molecule, e. g. a small ligand, is present, then an anti-terminator structure competes with the terminator hairpin and the polymerase can proceed to transcribe the subsequent protein coding region. Otherwise, the terminator stem blocks the polymerase and transcription terminates. The mechanism described here is an on-switch, as the ligand switches the RNA from off to on conformation, however, also off-switches use essentially the same principle.

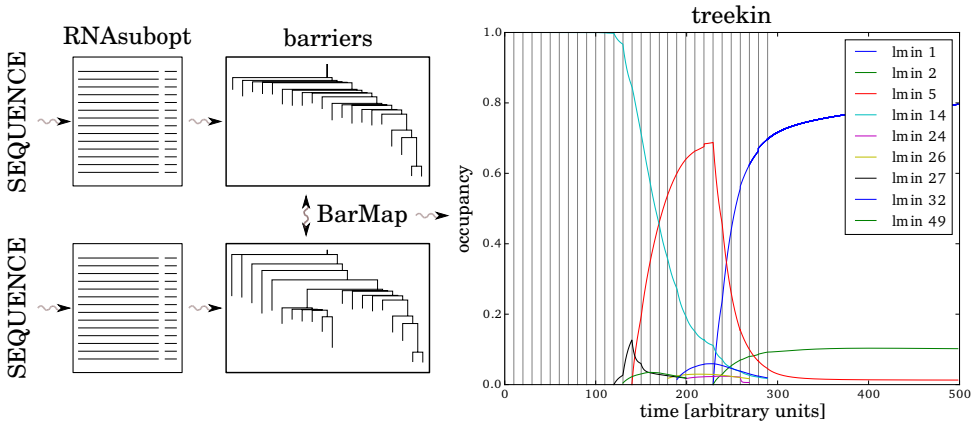


Figure 22: Schematic figure of the `BarMap` approach. The three programs `RNAsubopt`, `barriers` and `treekin` are used to simulate folding of a growing RNA transcript. `BarMap` finds a mapping between subsequent barrier trees and uses the information to transfer populations between `treekin` simulations. In this figure, 30 vertical gray lines in the `BarMap` simulation (right) show the time points of chain elongation, i. e. the time of mapping between landscapes. The following plots in this thesis only show the last vertical line to indicate the end of transcription. Simulations start in the MFE structure of the first energy landscape, usually the open chain conformation, the legend refers to the index of a local minimum in the last barrier tree.

4.2.1 *BarMap*

`BarMap` was published as a set of `Perl` scripts to build pipelines for simulating kinetics on various kinds of dynamic energy landscapes [Hofacker et al., 2010]. The program `barriers` [Flamm et al., 2002] is used to coarse-grain the energy landscapes into basins of attraction and saddle points connecting them (see Section 2.3.5), and the program `treekin` [Wolfinger et al., 2004] to compute chemical kinetics on the coarse-grained landscape using the master equation (see Equation 8, Section 2.3.3).

Applied to cotranscriptional folding, the algorithm can be decomposed into several steps: (i) coarse-grain every energy landscape for the growing RNA transcript (ii) find a mapping between local minima of consecutive landscapes (short: *barmaps*), and (iii) simulate folding kinetics starting in the first landscape and use barmaps to transfer populations between consecutive landscapes. The idea is that barrier trees and barmaps have to be constructed only once, and the kinetic analysis (e. g. with variable transcription speed) is independent of the computationally expensive coarse-graining. A correct mapping between two consecutive landscapes is crucial, and described below for the case of cotranscriptional folding:

NOTATION Recall that the three descriptors $(\Omega, \mathcal{M}, \mathrm{E})$ define an energy landscape (Definition 2.3). We write $\mathcal{L}$ for the fine-grained energy landscape with all subopti-

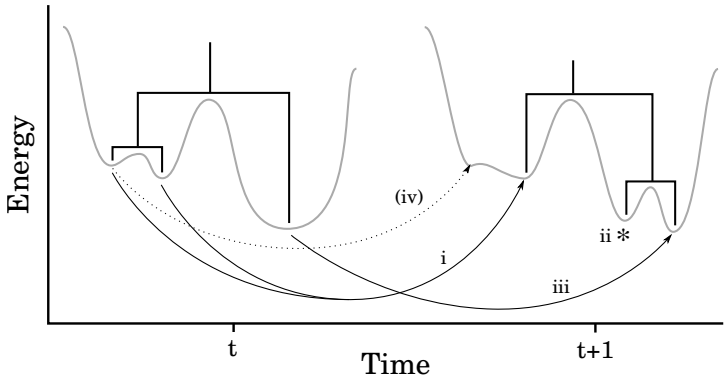


Figure 23: Mapping between local minima of consecutive landscapes. The events are possible, **(i)** two minima merge into one, **(ii)** a new minimum appears and **(iii)** a one to one mapping between landscapes. Case **(iv)** shows the mapping to a local minimum that has been merged in the new landscape, hence, `BarMap` has to ensure that the correct new gradient basin is found.

mal structures, single-base-pair transitions, and the NN energy model, and $\mathcal{B}$ for the gradient-basin landscape with transition rates between macro states and basin free energies according to the NN energy model.

Denote $\mathcal{L}_i$ as the energy landscape at transcript length i, then the mapping between consecutive landscapes $\mathcal{L}_{i-1} \rightarrow \mathcal{L}_i$ is trivial: Every structure in $\mathcal{L}_{i-1}$ is directly mapped to one structure in $\mathcal{L}_i$ by attaching an unpaired nucleotide at the 3′ end. In gradient basin landscapes, the mapping $\mathcal{B}_{i-1} \rightarrow \mathcal{B}_i$ is a two step process: First, an unpaired nucleotide is added to the local minimum conformations, then the gradient walk function $g(x)$ yields the (new) gradient basin conformation in $\mathcal{B}_i$.

Thus, a simple one-to-one mapping in gradient basin landscape $\mathcal{B}$ is not always possible. In fact, three events can happen (see Figure 23): (i) Two or more local minima merge into one, (ii) a new local minimum appears and (iii) a one-to-one correspondence between minima. In practice, there is also a forth case: (iv) a gradient walk does not result in the structure representing a local minimum. This forth case generally appears whenever multiple gradient basins are merged into the lowest one during landscape flooding. The original `BarMap` algorithm then maps this structure to the basin with the least base-pair distance. This enables a fast and simple population transfer between local minima that have been merged during barrier tree construction or between (negligible) disconnected valleys high up in the energy landscape.

`barmap-v2.0` It is worth pointing out that point (iv) has complicated the atomization of the mapping process. In particular, the user has to be aware of certain pitfalls, when analyzing landscapes at the limits of computational tractability (above 10^6 Wuchty suboptimal structures). In the worst case, populations are mapped across

large energy barriers. `BarMap-v2.0` now directly accesses the coarse-graining information computed by `barriers` and always maps to the correct gradient basin. The program terminates if populations get lost during subsequent simulations. Also, the reimplementation (1) combines features of the previous scripts in a `Python` library to automate the computation of a cotranscriptional folding simulation from single sequence input, and (2) updates the I/O interface to the latest releases of `barriers-v1.6` and `treekin-v0.4`. While the latest release of `treekin-v0.4` comes with important performance improvements, `barriers-v1.6` is essential to directly access the coarse-graining information for a given structure. The source code is available at `http://www.tbi.univie.ac.at/software`.

4.2.2 *RNA-ligand interactions*

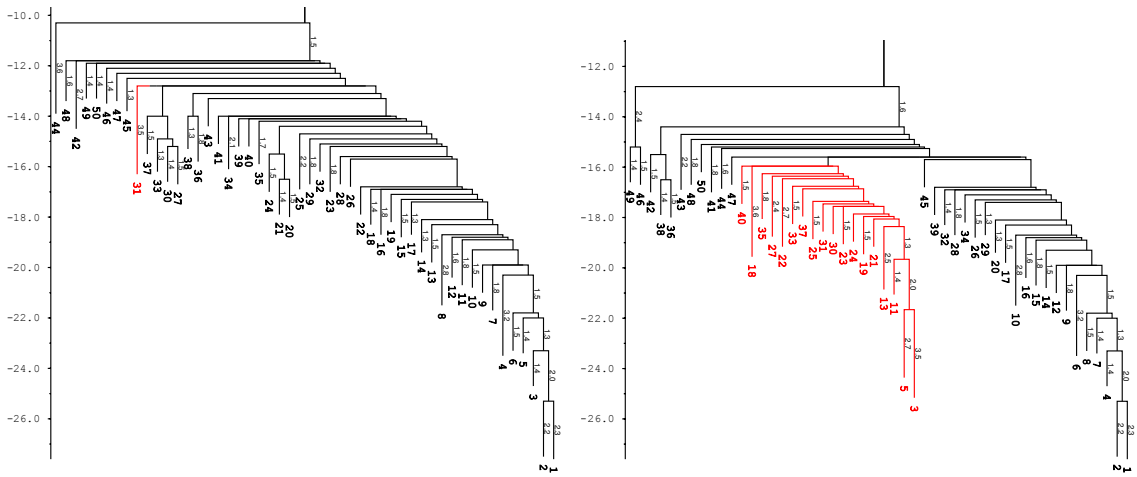
Binding of small metabolites to RNA molecules is one of the most direct reactions of a cell to environmental stimuli. Unfortunately, many metabolite binding pockets span regions of more than 200 nucleotides and involve pseudoknot conformations. Still, there are examples of small binding pockets (e. g. Theophylline [Jenison et al., 1994; Jucker et al., 2003; Gouda et al., 2003], Tetracycline [Berens et al., 2001; Müller et al., 2006], Adenine [Mandal and Breaker, 2004]) that only require small local binding pockets to form. The ligands stabilize these pockets with an experimentally measured dissociation constant

$$K_d = \frac{[L][R]}{[LR]} \tag{31}$$

where $[L]$ is the concentration of the ligand and $[R]$ is the concentration of the RNA. Toghether with the formula to convert an equilibrium constant into Gibbs free energy $\Delta G = -RT\ln(K)$ we can compute the free energy of *ligand binding* as

$$G_b = -RT \cdot \ln(K_b) = -RT \cdot \frac{1}{\ln(K_d)} \tag{32}$$

This binding free energy can then be included into the RNA energy evaluation as an additional stabilization term, i. e. added to every conformation that contains a ligand binding pocket. The effect of adding ligand binding free energies to the secondary structure ensemble can nicely be visualized with barrier trees, see Figure 24. Multiple conformations that have previously been merged into the basins of energetically better neighbors are now representative local minimum conformations. The best one, which has before been the suboptimal minimum number 31 is now the energetically third best local minimum conformation and part of whole subtree in the barriers representation. Unfortunately, the method assumes infinite concentration of the ligand and more realistic models such as described for nucleic acid only interactions (see Chapter 3) await to be developed.



(a) Barriertree w.o. Theophylline (b) Barriertree with Theophylline

Figure 24: A comparison of two energy landscapes for the same molecule. States marked in red form the theophylline binding pocket. The numbers correspond to the index of the local minimum in an energetically sorted list. **(a)** the best theophylline binding conformation is number 31 **(b)** the best theophylline binding conformation is number 3. We model a theophylline binding event to stabilize the respective pocket with −8.86 kcal/mol [Badelt et al., 2015b].

4.2.3 *Results*

In order to model termination of transcription we use the experimental results from Wachsmuth et al. [2013, 2015], Figure 25. Nine variants of cotranscriptional theophylline binding riboswitches have been analyzed in *Escherichia coli* by measuring the expression of β-Galactosidase reporter genes. The sequences are designed using the same theophylline binding pocket and differ only in their terminator hairpin loop. The switches, shown in their unbound, transcription terminating conformation, are turned on upon binding of theophylline. The coarse-grained energy landscapes from Figure 24 show a particularly effective riboswitch (Figure 25: RS10). The MFE secondary structure remains the same, independently of whether the ligand is bound, such that thermodynamic methods cannot explain the function of RS10. However, experimental findings show a 2-3 fold activation of the reporter gene product (in Miller Units) upon presence of the ligand.

The simulations of cotranscriptional folding kinetics shown in Figure 26 explain the experimental results. Without theophylline, two secondary structures are populated at the end of riboswitch transcription, i. e. before the β-Galactosidase gene is transcribed. Both of these structures correspond to a conformation where the terminator hairpin is completely formed. In presence of theophylline, again two structures are populated, now forming the *anti-terminator* structure including the ligand binding pocket. Forma-

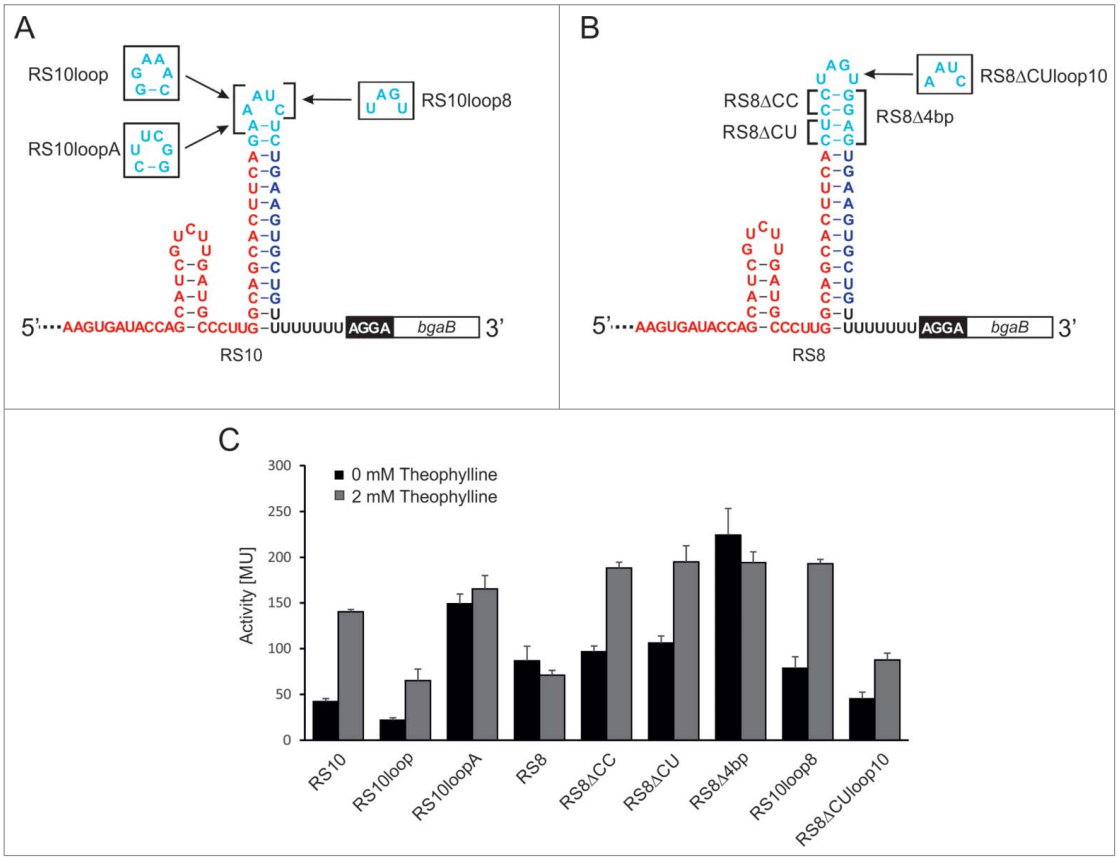


Figure 25: Experimental data on different variants of transcription terminating riboswitches. The red and blue regions are constant, only the hairpins vary in the tested examples. The red region contains the theophylline binding pocket (not shown), forming an anti-terminator structure upon a binding event. Although the ground states of the molecules do not change, see e. g. Figure 24, the experimental results show basically all possible variations of efficiency. Figure copied with permission from Wachsmuth et al. [2015].

tion of the terminator needs on the order of 10 seconds, hence a time period where the polymerase has already moved on transcribing about 500 more nucleotides.

While RS10 is in good agreement with experimental results, the behavior of riboswitches depends on many more factors than one particular ligand, and other present interaction partners are not included in the simulation. Hence, the simulations do not always confirm the experimental results. A particular extreme case is RS8, (also in Figure 25). The binding of theophylline does not improve the signal in an experimental setting.

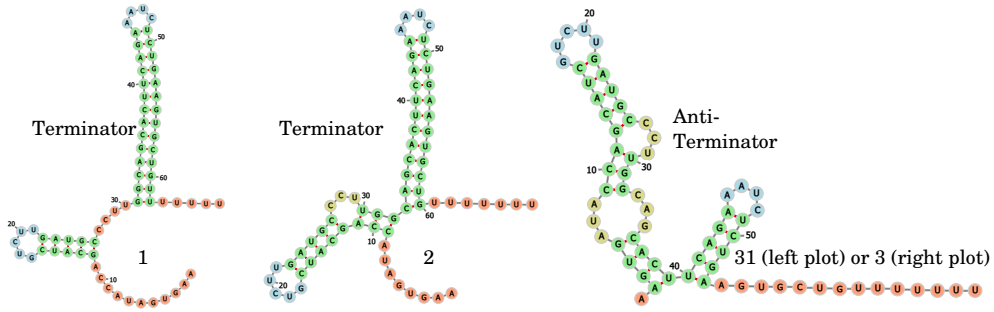
The simulations shown in Figure 27, however, suggest that RS8 is as effective as the previously shown RS10 riboswitch and also the barrier trees (not shown) are similar to RS10. It is worth pointing out that at much higher transcription speed and in the absence of theophylline an alternative structure to the terminator hairpin forms (as also observed with `Kinwalker` in Wachsmuth et al. [2015]). This might help to explain why there is no additional activation in presence of theophylline, but it cannot explain why the absolute activation of RS10 is better than for RS8.

In order to reduce the effects of coarse-graining on folding kinetics, we have produced all results with a low barrier height of 1 kcal/mol and allowing a maximum of 10^4 local minima. The latter is sufficiently large such that it does not limit the accuracy of simulations.

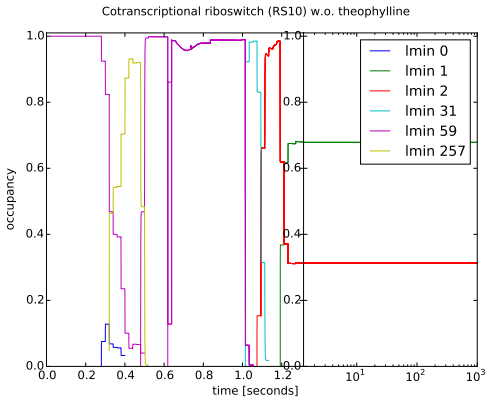
Discussion

We have shown that it is possible to model RNA-ligand interactions during cotranscriptional folding using coarse-grained energy landscapes. The results confirm and explain the trigger mechanism that had been assumed during the sequence design process in Wachsmuth et al. [2013]. As the method cannot include all environmental and experimental parameters, not all of the simulations exactly reproduce the tested riboswitch efficiency. However, `BarMap` can be used to select candidates for experimental fine-tuning, such as shown in Figure 25. Varying the transcription speed can help to further select for switches that are particularly stable in their folding behavior.

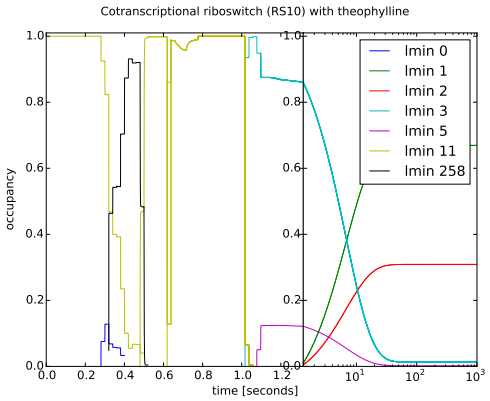
It is worth pointing out that the length of riboswitches presented here is close to the limit of `BarMap` modeling. The main problem is the high energy barrier separating a ligand-stabilized binding pocket from the unbound conformations. For this reason, we had to calculate a large number of suboptimal structures and reduce the volume by excluding conformations with lonely base-pairs. This introduces an artificial neighborhood relation, which is not properly compensated by the program `barriers`. In particular, two base-pairs are formed at a time during the nucleation of helices leading to a lower energy barrier than if the formation of a helix proceeds via a single-base-pair change. In order to assess the impact on `BarMap` simulations, we have calculated cotranscriptional folding from the full suboptimal structures (including lonely base-pairs) up to the length of 50 nucleotides. This includes the complete binding pocket



(a) RS10 local minumim conformations

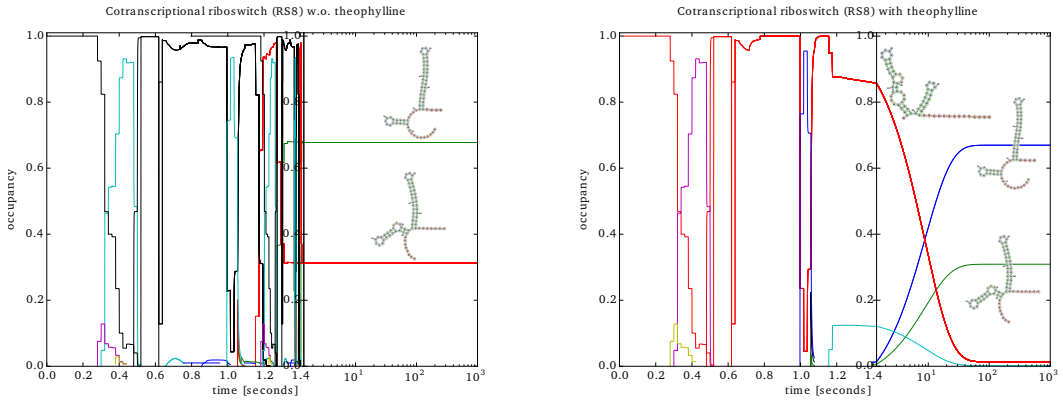


(b) `BarMap` without theophylline



(c) `BarMap` with theophylline

Figure 26: Cotranscriptional folding of the riboswitch RS10 in presence and absence of the trigger molecule theophylline. **(a)** Three structures that correspond to the local minima in the simulations (b,c). Local minimum 1 and 2 form a terminator hairpin, local minimum 31 (b) or 3 (a) forms an antiterminator structure stabilized by theophylline. **(b, c)** Simulations with a transcription speed of 50 nucleotides per second. The vertical line marks the end of transcription and starts a logarithmic timescale. Numbers in the legend correspond to indices in the barrier tree shown in Figure 24, the most important ones are also shown in (a). Without theophylline, the switch is in equilibrium at the end of transcription and both dominating secondary structures have formed the terminator hairpin. In presence of theophylline, the switch is out-of-equilibrium at the end of transcription, forming the anti-terminator with the theophylline binding pocket. At the time the system equilibrates, about 10 seconds, the polymerase has already moved on to transcribe the protein-coding region. lmin 0 indicates that a metastable structure could not be mapped to a local minimum in the final barrier tree. However, the population of this structure decreased below the threshold before it had to be transferred, otherwise `BarMap-v2.0` stops the calculation.



(a) `BarMap` without Theophylline

(b) `BarMap` with Theophylline

Figure 27: Cotranscriptional folding of the riboswitch RS8 in presence and absence of the trigger molecule theophylline. The results are very much comparable to RS10 (Figure 26). **(a)** In absence of theophylline, the switch forms only transcription termination structures (those with long terminator hairpins). **(b)** In presence of theophylline only anti-terminator structures have formed and the formation of terminators takes in the order of 10 seconds.

but not the terminator stem and we observe that the effects are negligible compared to the shown RS10 and RS8 examples without lonely base-pairs.

4.3 COTRANSCRIPTIONAL FOLDING OF LARGE RNAS

As introduced in section 4.1, present algorithms fall into three categories: stochastic simulations, master equation methods, and deterministic prediction of single trajectories. With the program `DrTransformer`, short for "DNA to RNA Transformer", we have implemented a heuristic that bridges the gap between master equation methods like `BarMap` that are rarely applicable for sequences longer than 70 bases and single trajectory prediction `Kinwalker`, which only selects one most populated secondary structure at each transcription step.

We will show that the results of `DrTransformer` compare well to statistically correct sampling of folding trajectories of short sequences using `Kinfold`, effectively making cotranscriptional folding of 200 nucleotide sequences easily tractable. The accuracy of simulations as well as the limits of sequence length in practice are heavily dependent on structural diversity, cotranscriptional folding traps, and on the parameters chosen. For example, it is possible to model artificially designed cotranscriptional RNA-origami sequences [Geary et al., 2014] with more than 1000 nucleotides.

Option	Default	Unit	Algorithm	Explanations
mfree	6	bases	3	minimum freed base-pairs during helix breathing
occut	1	%	5	minimal occupancy to keep structures after a kinetic simulation
effd	10^{-10}	mol/l	4	minimal rate to accept a new neighbor
fpwi	10		4	upper bound for `findpath` search
k_0	$2 \cdot 10^5$		4	convert arbitrary time units to seconds
t_0	$1 \cdot 10^{-6}$	sec	2	first output time of the simulation
t_8	$2 \cdot 10^{-2}$	sec	2	transcription speed, here 50 nuc/sec
t_X	$3.6 \cdot 10^3$	sec	2	simulation time after transcription stop (10 hours)

Table 1: Default parameters for `DrTransformer`. Detailed explanations are provided in the context of the individual procedures in the main text.

4.3.1 *Theory and Implementation*

Our model approximates an energy landscape with a conformation graph, such that structures (vertices) are connected with transition rates (edges). Whenever a nucleotide is added during transcription this graph is updated, i. e. every present structure gets an initially unpaired nucleotide attached and then its neighborhood is searched for new, energetically better conformations. A kinetic simulation redistributes occupancy according to the transition rates and structures with low occupancy are removed from the system.

This section will explain how our implementation, `DrTransformer`, finds favorable neighboring structures using constrained MFE folding, calculates transition rates between the currently dominant structures and their neighborhood, and computes the kinetics of isomerization reactions with `treekin` [Wolfinger et al., 2004] or `SundialsWrapper` (see Section 3.2.2). The output of `DrTransformer` can be fine-tuned by a number of parameters that are summarized in Table 1. The most important procedures are explaied in Algorithms 2, 3, 4, and 5.

NOTATION We formulate an energy landscape as a directed, strongly connected graph $\mathcal{G}(\mathcal{S}, \mathcal{K})$ with vertices $s \in \mathcal{S}$ representing the different structures and edges $k \in \mathcal{K}$ connecting neighboring conformations. Every vertex $s \in \mathcal{S}$ has some non-negative population assigned that can flow along the edges. The edges $k \in \mathcal{K}$ are directed and weighted by the rate of a folding reaction between two neighboring structures.

Algorithm 2 `DrTransformer` – core algorithm

```
1:  procedure DrTransformer (sequence Σ, options ...)
2:      G = ()                                              ▷ Empty reaction graph
3:      ViennaRNA.noLP ← True                               ▷ no lonely base-pairs
4:      for i = [1, ..., length(s)] do
5:          σ = Σ[1, i]
6:          m ← ViennaRNA_mfe(σ)
7:          S ← G.vertices
8:          for all s in S do
9:              AddTransitionEdges(G, σ, s, m, -)           ▷ effd, fpwi, k0, only direct path
10:             for all n ∈ BreathingNeighbors(σ, s) do     ▷ mfree
11:                 AddTransitionEdges(G, σ, s, n, -)       ▷ effd, fpwi, k0, only direct path
12:             end for
13:         end for
14:         if i < n then
15:             simulate(G, t0, t8)                         ▷ treekin
16:             G ← GraphPruning(G, σ)                      ▷ occut, effd, fpwi, k0
17:         else
18:             simulate(G, t0, tX)                         ▷ treekin
19:         end if
20:     end for
21: end procedure
```

The reaction graph is expanded whenever a nucleotide is transcribed, followed by a simulation to transfer the populations within the graph, and finally the graph is pruned to remove secondary structures with to little population (see Algorithm 2). At the start of RNA synthesis, the reaction graph $\mathcal{G}$ is small and the transition rates are high such that the majority of the population transfers quickly into the current MFE structure(s). At a later stage, populated structures define the active hubs within the network and serve as starting points to explore new areas in the growing energy landscape.

To ensure that simulations reach thermodynamic equilibrium eventually, the graph has to be ergodic, i. e. every state has to be reachable from every other state and detailed balance has to hold. Dependent on cotranscriptional folding dynamics, populated secondary structures can occupy parts of the energy landscape that evolved to be separated by high energy barriers. Hence, the transition rates become exponentially smaller leading to numeric instabilities in subsequent simulations, i. e. we say the landscape is *effectively disconnected*. Also, the MFE secondary structure might be effectively disconnected from the populated landscape at some point during transcription. As these effects violate the laws of thermodynamics they are highly unsatisfactory and `DrTransformer` employs a number of tricks to keep the landscape ergodic, if possible.

4.3.1.1 *Neighbor generation*

Exhaustive exploration of the full neighborhood in a conformation graph is time consuming. However, folding of secondary structures during transcription changes the conformation graph only in comparatively small areas, while the majority remains constant. A newly transcribed nucleotide can only interact with the exterior loop of a present conformation, otherwise the secondary structure would become pseudoknotted. The neighbor generation described here uses this property for finding potentially better secondary structures and searches only in exterior loops and in immediately adjacent helices.

A definition of helices able to *breath* (Figure 28) follows below and serves as a basis to explain the move set for finding favorable structural transitions:

Definition 4.1 *We define helices that only consist of one or more consecutive base-pair stacks as* stacked helices. *Such helices are flanked by bulges, mismatched interior loops, hairpin, multi, or exterior loops. A secondary structure* $s \in S$ *is called a* canonical *secondary structure [Bompfünewerer et al., 2008] if and only if every base-pair is part of a stacked helix. Stacked helices with a base-pair toward the exterior loop of a secondary structure are said to be able to* breathe.

In order to reduce conformations to canonical secondary structures, the `ViennaRNA package` option *no-lonely-base-pairs* excludes base-pairs that are not involved in stacking interactions. We may argue that these base-pairs are highly reactive and opened

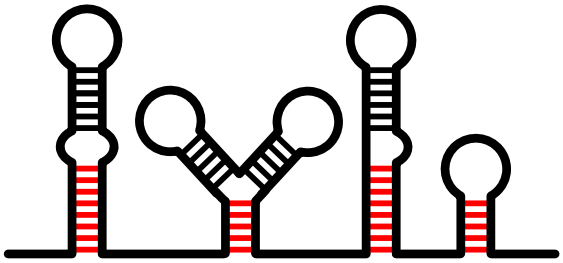


Figure 28: Helices able to breathe. Breathing helices are marked with red base-pairs, they share a base-pair with the exterior loop.

as a consequence. Then, one breathing helix is opened at a time and the remaining structure becomes a constraint for MFE folding. If that constrained folding results in a different conformation than the starting structure, then we have found an energetically better or equal conformation.

Stacked helices can vary greatly in length, starting with at least two base-pairs. Our method uses a parameter to choose the minimum amount of bases freed by helix breathing. As a default, we choose 6, which, for instance, corresponds to a stack of two base-pairs and a loop region of 2 nucleotides. If less bases are freed and there exists a nested stacked helix, this helix is considered to breathe as well. We further assume that breathing helices can compete with each other. Thus, one additional neighbor is generated by opening all breathing helices at once for constrained folding (see Algorithm 3). It is worth noting that also constrained MFE folding takes $O(n^3)$ time, with n being the sequence length. In practice, we only fold the exterior loop of the remaining conformations and merge the result with the constraint. Rates are computed in both directions to ensure ergodicity and detailed balance.

4.3.1.2 *Connecting conformations with folding pathways*

Reaction rates between two structures are computed using the Arrhenius model. Hence, the rate of a transition is directly proportional to the energy barrier separating two conformations. In order to find the lowest possible energy barrier among the set of all shortest refolding paths, the `findpath` algorithm [Flamm et al., 2001] from the `ViennaRNA` library is applied (see Section 2.3.2).

Depending on the settings of neighbor generation, the move set can allow structural transitions with high energy barriers. While spontaneous reactions take $5 \cdot 10^{-6}$ seconds, passing an energy barrier of 2 kcal/mol takes 10^{-4} seconds and a barrier of say 20 kcal/mol takes $6 \cdot 10^{8}$ seconds (i. e. 19 years) to be passed. Such differences can lead to numeric instabilities and to effectively disconnected components. In order to avoid this effect, transition rates (and the corresponding neighborhood relation) are rejected below a certain threshold (Default $10^{-10} k_0$ per second). The determination of refolding barriers to compute a transition rate is the most time consuming part of

Algorithm 3 `DrTransformer` – neighbor generation

```
1: procedure BreathingNeighbors (sequence σ, structure s)
2:     N ← {}                                              ▷ set of neighbors
3:     m ← s
4:     OpenHelices(N, σ, s, m, mfree)
5:     N ← N ∪ m
6:     for all n ∈ N do
7:         n ← fold_exterior_loop(σ, n)                    ▷ see main text
8:     end for
9:     return N
10: end procedure
11: procedure OpenHelices(neighbors N, sequence σ, structure s, neighbor m, mfree
    mfree )
12:     b ← {breathable(s)}                    ▷ set of base-pairs at exterior loop
13:     for all (x, y) ∈ b do
14:         n ← s
15:         (p, q) ← (x, y)
16:         loop ← 0
17:         open ← 0
18:         add ← True
19:         while p < q and (loop = 0 or open < mfree) do
20:             if (p, q) ∉ s then                          ▷ Multiloops
21:                 OpenHelices(N, σ[p : q], n[p : q], m, mfree − open)
22:                 add ← False
23:                 break                                   ▷ exit while loop
24:             end if
25:             n ← n \ (p, q)                              ▷ remove base-pair
26:             m ← m \ (p, q)                              ▷ remove base-pair
27:             open = open + 2
28:             loop ← 0
29:             while unpaired(p ← p + 1) and p < q do
30:                 loop ← loop + 1
31:             end while
32:             while unpaired(q ← q − 1) and p < q do
33:                 loop ← loop + 1
34:             end while
35:             open ← loop + open
36:         end while
37:         N ← N ∪ n if add
38:     end for
39: end procedure
```

`DrTransformer`. Hence, every computed energy barrier is stored for future calculations. Looking at Algorithm 4, we note that there is a second mode for transition rate computation, which is important for graph pruning. In particular, during pruning, transition rates are never rejected, but computed as the minimum over the currently best energy barrier via the transition state and the direct path barrier.

Algorithm 4 `DrTransformer` – computing transition rates

procedure AddTransitionEdges (Graph $\mathcal{G}$, sequence σ, structure s_2, structure s_1, structure s_t)
  $sE \leftarrow \text{findpath}(\sigma, s_2, s_1, w)$
  if s_t **then** ▷ transition state specified
    $tsE \leftarrow \text{findpath}(\sigma, s_2, s_t, w)$
    $tsE2 \leftarrow \text{findpath}(\sigma, s_t, s_1, w)$
    $tsE \leftarrow \max(tsE2, tsE)$
    $sE \leftarrow \min(tsE, sE)$
  end if
  $\Delta G_T \leftarrow sE - \text{get\_energy}(\sigma, s_2)$
  $k_{s_2 \to s_1} \leftarrow \text{Arrhenius}(k_0, \Delta G_T)$ ▷ see Equation 12
  $\Delta G_T \leftarrow sE - \text{get\_energy}(\sigma, s_1)$
  $k_{s_1 \to s_2} \leftarrow \text{Arrhenius}(k_0, \Delta G_T)$ ▷ see Equation 12
  if s_t **or** $k_{s_2 \to s_1} > k_0 \cdot c$ **then**
    $\mathcal{G}$.add_edge(s_2, s_1, k_{21})
    $\mathcal{G}$.add_edge(s_1, s_2, k_{12})
    return 1
  else
    return 0
  end if
end procedure

4.3.1.3 *Simulating folding kinetics*

Selecting neighbors of populated structures and connecting them with `findpath` yields a connected directed graph. In order to compute the population flow, the graph can be translated into a system of first order ODEs which are solved by numeric integration. The population of every vertex remains constant when thermodynamic equilibrium is reached, i. e. the populations of structures satisfy the previously discussed chemical master equation:

$$\frac{dP_i(t)}{dt} = \sum_{i \neq j} (P_j(t)k_{ji} - P_i(t)k_{ij}) = 0 \tag{8}$$

It is worth noting at this point that computing the folding kinetics using `treekin` is notably faster than `SundialsWrapper`. However, the latter turned out to be more stable for simulations on effectively disconnected landscapes and is used whenever `treekin` simulations fail.

4.3.1.4 *Discarding conformations*

After the simulation, the populations of individual secondary structures have changed and many of these structures have to be discarded in order to keep the system computationally tractable. This is known as a graph-pruning step, where all vertices with a population smaller than a particular cutoff are discarded. As long as ergodicity is of no concern, graph pruning is simple. A higher cutoff leads to smaller graphs and increasing population loss from the discarded vertices. The population loss can be distributed among remaining vertices, proportional to their present population.

In our approach (see Algorihtm 5), ensuring ergodicity during graph pruning is necessary, since identification and reconnection of disconnected subgraphs is computationally expensive and important intermediate states can get lost otherwise. We process an energetically descending list of unpopulated vertices and reconnect their energetically best neighbor to all the remaining neighbors. In order to avoid that the information of an important transition state is lost, the rate for new transitions is computed as the minimum over the direct path barrier and the indirect path barrier via the removed vertex. Denote k_{ab} as the rate from a to b, and $k_{a\to b}$ the rate from the direct path $a \to b$, then

$$k_{a\to b} = \min\{k_{a\to b}, k_{a\to i\to b}\} \tag{33}$$

where $k_{a\to i\to b}$ is the rate computed from the energy barrier of the two direct paths $k_{a\to i}$ and $k_{i\to b}$. Vertices that only have energetically worse neighbors are considered as *still reachable* and kept for the next round. However, they are excluded from neighbor generation as long as they remain populated below the threshold.

Algorithm 5 `DrTransformer` – graph pruning

```
1:  procedure GraphPruning (Graph G, sequence σ)
2:      n_e ← n.energy
3:      for all s in sort(G.vertices) do                ▷ decreasing energy
4:          p ← s.population
5:          continue if (p > occut)
6:          e ← s.energy
7:          N ← sort(s.neighbors)                       ▷ increasing energy
8:          m ← N[0]
9:          m_e ← m.energy
10:         continue if (n_e > e)                       ▷ still reachable
11:         for all n ∈ N \ N[0] do
12:             AddTransitionEdges(G, σ, n, m, s)       ▷ see Algorithm 4
13:         end for
14:         G.delete(s)
15:     end for
16: end procedure
```

4.3.2 *Results*

We start with testing the quality of cotranscriptional folding predictions comparing `DrTransformer` with `Kinfold`, `BarMap`, and `Kinwalker` for random sequences. In the second part we use `DrTransformer` to model experimentally tested cotranscriptional folding-traps presented in Xayaphoummine et al. [2007] and long RNAs that are expected to fold into a particular RNA Origami shape [Geary et al., 2014].

Random sequences

The data set consists of 200 random sequences with the length of 30, 40, 50, 60 and 70 nucleotides, i. e. 1000 sequences in total. In order to find sequences with potential cotranscriptional folding traps all sequences were sorted by the maximum barrier height between any two structures in the full-length energy landscape. Only the top 20 of each length were used for comparisons. Although we cannot guarantee that the maximum barrier is also a cotranscriptional folding trap, this way we were able to remove predominantly unstructured molecules where all programs have similar results, i. e. the MFE structure of the molecule.

We compare three of the methods discussed before, all using the same energy model: `Kinfold` does statistically correct sampling of trajectories and can be considered as the gold standard for short sequences. `BarMap` calculates cotranscriptional folding kinetics of the complete, coarse-grained secondary structure ensemble and therefore should perform well for RNAs up to a length of roughly 70 nucleotides. `Kinwalker` is a heuristic applicable to long RNA molecules, but only returns a single trajectory. All of the programs start transcription with the first nucleotide. Additional parameters were chosen to be similar to standard parameters of `BarMap` and `DrTransformer`.

- `Kinfold`: Time was set to unit time ($k_0 = 1$). The chain grows every 4000 time units, which corresponds to 50 nuc/sec. The total simulation time is $4000n$, where n is the sequence length. Simulations use the Metropolis rule and standard single-base-pair moves. The logarithmic multi-loop evaluation was disabled to ensure the same energy model as for the other methods. Every sequence was evaluated with 10^4 trajectories.

- `BarMap`: Time was set to unit time ($k_0 = 1$). The time for chain elongation and the last simulation $t_8 = t_X = 4000$ time units, which corresponds to 50 nuc/sec. Simulations were done using an energy range resulting in roughly $9 \cdot 10^6$ suboptimal secondary structures, a minimal barrier height of 1 kcal/mol, at most 9999 lowest minima per energy landscape (a value that is usually not reached at this length). The threshold to transfer occupancy between the landscapes was set to 0.01. Simulations are based on the Metropolis rule and standard single-base-pair moves.

- `DrTransformer`: Time was set to unit time ($k_0 = 1$). Transcription starts at the first nucleotide. The time for chain elongation and the last simulation $t_8 = t_X = 4000$ time units, which corresponds to 50 nuc/sec. The occupancy threshold to keep secondary structures during graph pruning was set to 0.01. Simulations are based on Arrhenius kinetics, the `findpath` routine to find best direct paths has an upper bound of 10.
- `Kinwalker`: The transcription time was set to 50 nuc/sec. We used the `findpath` routine with an upper bound of 10 for finding energy barriers.

Comparing the results of the different programs is complicated due to the fact that each program uses a different coarse-graining for the secondary structure ensemble. Multiple conformations returned by one method can be represented by a single secondary structure of another method, and vice versa. While `Kinfold` can in principle return every secondary structure, `BarMap` returns the populated leaves of the final barrier tree. Less well defined are the results of `DrTransformer` and `Kinwalker`. The sequences are assembled from thermodynamically optimal fragments, depending on the history of the simulation.

As a consequence, only barrier trees seem suitable to compare simulations, because every output structure can be mapped into its local minimum basin. This yields the same coarse-graining for all methods and enables the comparison of final population vectors $\vec{p}_n^A, \vec{p}_n^B$, where A and B denote the respective method and n is the index of the local minimum in the barrier tree. The similarity s is calculated as

$$s = \sum_n \sqrt{\vec{p}_n^A \vec{p}_n^B} \tag{34}$$

For example, `Kinwalker` will always have a population vector with only one local minium populated at the maximum of 1, while the remaining entries are 0.

Figure 29 shows the similarity scores of final population vectors using the barrier tree coarse-graining. The results show that `DrTransformer` compares well to the statistically exact method `Kinfold` and, surprisingly, often compares better than `Kinfold` to `BarMap`. However, it is important to point out that all methods are merged into the `BarMap` coarse-graining, but not the other way around. Hence, `BarMap` has generally the broadest range of conformations populated which systematically worsens the similarity score. Also, we have introduced a threshold of 1% population at the end of transcription in order to be considered for comparison. All structures exceeding this population threshold were normalized such that the total population is 100%. While in practice this has no effect on short sequences (30, 40 nucleotides) it can be problematic for the longer cases. Both `Kinfold` and `DrTransformer` occasionally predict structures that exceed the suboptimal structure range used in `BarMap`. Also, `Kinfold` returns a high diversity of conformations, such that for longer sequences only a few are populated above the 1% threshold. Subsequent normalization to 100% then overestimates their population and results in examples with no similarity to other methods.

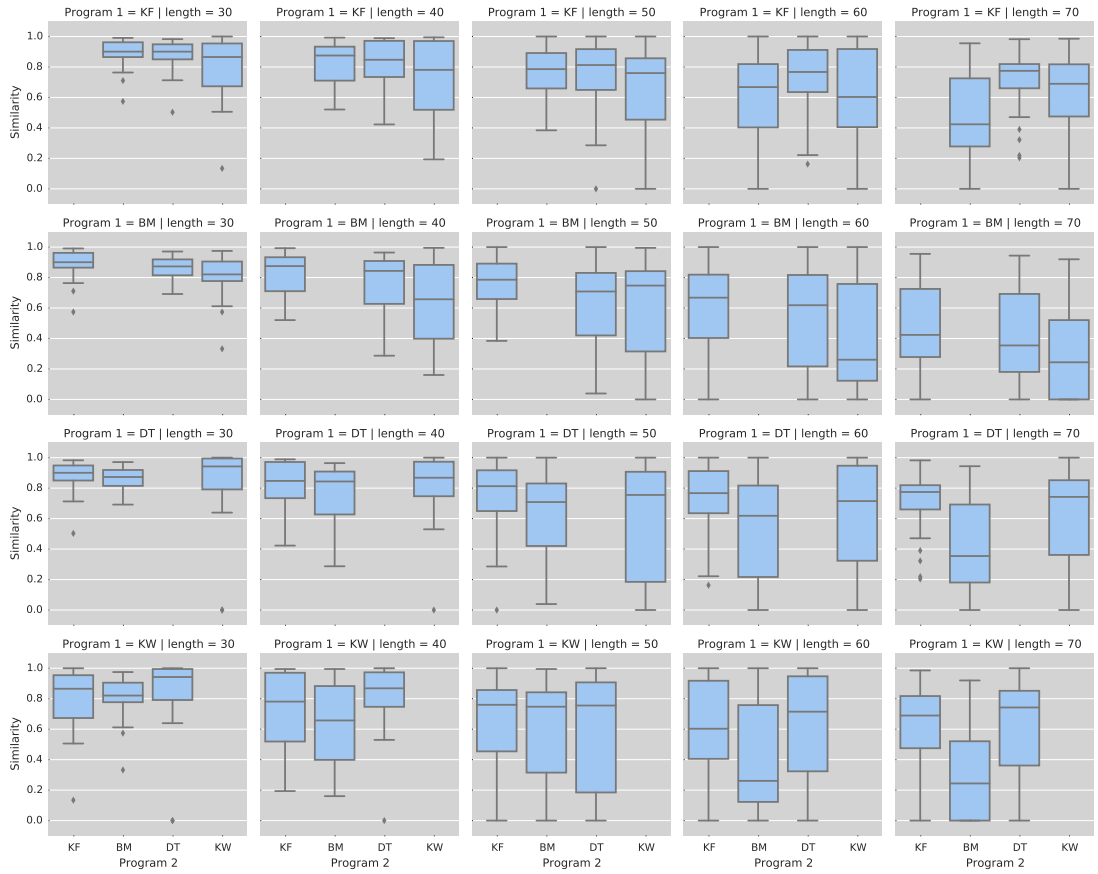


Figure 29: Four algorithms in a pairwise comparison of the results. The programs are in order of their expected prediction accuracy (`Kinfold` (KF), `BarMap`, (BM), `DrTransformer` (DT), `Kinwalker` (KW)) and the compared sequences are grouped into their length (20-70 nucleotides). Each box-plot contains the top 20 sequences with respect to the maximum barrier in the energy landscape, i. e. a potential cotranscriptional folding trap. The similarity is measured by comparing population vectors using coarse-graining of barrier trees (see Equation 34). **Top row:** all programs are compared to the gold standard `Kinfold`. The results of `DrTransformer` are more similar to `Kinfold` than both other programs `BarMap` and `Kinwalker`. It is important to point out that the results of `Kinfold` get highly diverse for sequences of length 60 and 70, such that even upon simulation of 10^4 trajectories only very few structures have been populated with more than 1%. For that reason, especially structures with co-transcriptional folding traps show sometimes no similarity between modeling with `BarMap`. **Third row:** `DrTransformer` is compared to all other programs. Independent of the sequence length, only the most accurate modeling with `Kinfold` compares well with `DrTransformer`, while both `BarMap` and `Kinwalker` have only little similarity for longer sequences. **Second row/Last row:** For the sake of completeness, also `BarMap` and `Kinwalker` are compared to all other programs. Interestingly, `DrTransformer` also has the highest similarity compared to `Kinwalker`, suggesting that those results where `DrTransformer` differs from `Kinfold` are close to the results of `Kinwalker`.

Modeling cotranscriptional folding traps

A very beautiful experimental test of cotranscriptional folding traps was shown by Xayaphoummine et al. [2007]. Three RNA sequences have been designed, two are composed of the exact same (palindromic) sub-sequences (A,B,C,D) in forward and reverse order (ABCD and DCBA). A third sequence (DCB'A) differs from DCBA only by a single point mutation in B. The sequences demonstrate how the order of helix formation determines which structure is formed at the end of transcription. In particular, ABCD has been shown to fold almost exclusively into the MFE structure, while the reverse sequence DCBA is cotranscriptionally trapped in a metastable state. The single-base-mutation in DCB'A decreases the effect of the cotranscriptional folding trap and yields roughly 50% of the MFE structure at the end of transcription.

Figure 30 shows a comparison of the three sequences using `DrTransformer`. The results exactly reproduce the experimental findings, and are largely independent of the transcription speed. ABCD folds almost exclusively into a two-helix conformation at the end of transcription, while BCDA folds into the metastable state with only one long hairpin. It then requires on the order of 10^7 seconds (115 days) to fold into the MFE conformation. DCB'A, on the other hand, folds with around 50% into both conformations, where the exact ratio depends on the transcription speed.

Modeling of long sequences

A particularly challenging example for RNA design is the synthesis of cotranscriptional RNA origami [Geary et al., 2014]. Sequences of multiple hundred nucleotides length have to be designed to fold directly into a particular shape. While the designs presented in Geary et al. [2014] are first successful demonstrations of polygon shaped RNAs, it is only a matter of time until experimenters attempt to design more complex structures, such as long-non-coding RNA scaffolds or artificial ribosomes.

The design of sequences has to either avoid cotranscriptional folding traps in the first place, or use them in an intentional way to guide the folding of the full-length transcript. The crucial question for RNA Origami is therefore not whether the sequence folds thermodynamically into the MFE structure, but to identify candidates that efficiently fold into the MFE structure during transcription. Using `DrTransformer`, we chose to model their largest Origami sequence (6H-AO, Geary et al. [2014]), that folds correctly using a mica annealing protocol, but does not form the correct structure with the cotransriptional assembly protocol.

The Origami consists of 3 tiles, each 661 nucleotides with a high sequence similarity. It is important to note that the designs involve kissing-loop interactions, which are not captured by the standard secondary structure energy model. Hence, `DrTransformer` can only predict the desired folding trajectory if it forms even without the stabilization energies from kissing loops. We may argue that kissing loops predominantly orient

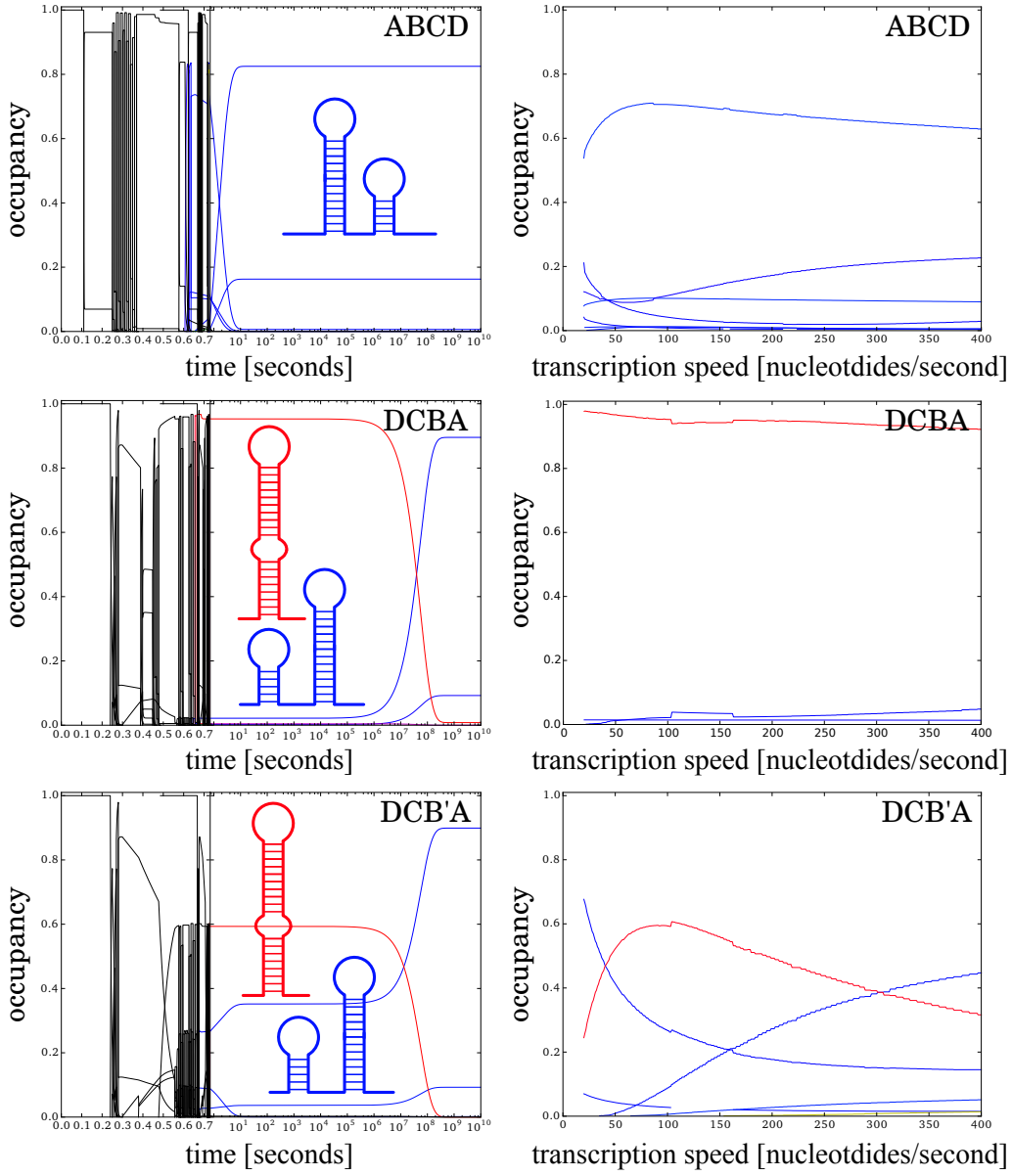


Figure 30: Cotranscriptional folding of three different sequences. **Left column:** simulations at a transcription speed of 100 nuc/sec are shown. A second y-axis divides the plot into a linear scale (during transcription) and a logarithmic scale (after transcription). Black trajectories correspond to intermediates during transcription, blue and red trajectories correspond to the shape shown in the respective RNA secondary structure representations. Multiple lines of the same color indicate small differences such as single base-pair variations. **Right column:** Change in occupancy as a function of transcription speed. The data ranges from from 20 to 400 nuc/sec and shows that only DCB'A shows variations dependent on the transcription speed. For experimental results confirming these observations see Xayaphoummine et al. [2007], Figures 2 and 3.

the origami into its well defined tertiary shape, rather than being essential for the formation of the secondary structure.

Figure 31 shows the results from `DrTransformer` modeling. Simulations at a transcription speed of 50 nucleotides per second suggest that two of the three tiles are highly effective, forming predominantly the MFE secondary structure after only few seconds. Only 6AO-C needs more than an hour to reach 40 % of MFE structure occupancy. At 250 nucleotides/second transcription speed, two out of three tiles need more than an hour to fold into their desired shape.

It is important to point out that the simulations for the Origami tiles are parameter dependent and prone to numeric instabilities. The simulations shown in Figure 31 use standard parameters, most importantly, a `findpath` upper bound of 10 and a minimal rate of $k_0 \cdot 10^{-10}$ s^{-1}. Both parameters influence the results, e. g. raising the `findpath` bound to 50 suggests that all tiles form within 10 seconds after transcription independently of transcription speed. Lowering the minimal rate to $k_0 \cdot 10^{-5}$ seconds then again slows down the formation of intended Origami shapes. Hence, drawing final conclusions why the cotranscriptional folding does not work in the experimental setting is impossible without additional experimental data.

4.3.3 *Discussion*

We have presented `DrTransformer`, a new approach to model cotranscriptional folding using heuristic folding kinetics. The program is a hybrid between master equation methods on the full landscape (`BarMap`) and single trajectory computation as done by `Kinwalker`. Conceptually, the method is similar to a `Kinwalker` implementation that allows suboptimal transition states at each chain length. The comparison of different methods for cotranscriptional folding is not easy to interpret, but we have shown that `DrTransformer` is more accurate than single trajectory prediction from `Kinwalker`, and it performs better then `BarMap` in a direct comparison with `Kinfold`.

The method is easily applicable to sequences of up to 200 nucleotides. For the modeling of longer sequences, the cotranscriptional folding traps may limit the accuracy, and greatly reduce the time for simulations. However, the program can be used to determine the folding efficiency for molecules that have been optimized for folding along specific trajectories. In that case, also simulations of 660 nucleotide Origami takes only about 10 minutes computation time on a single core of a personal computer (Intel i5-3570K).

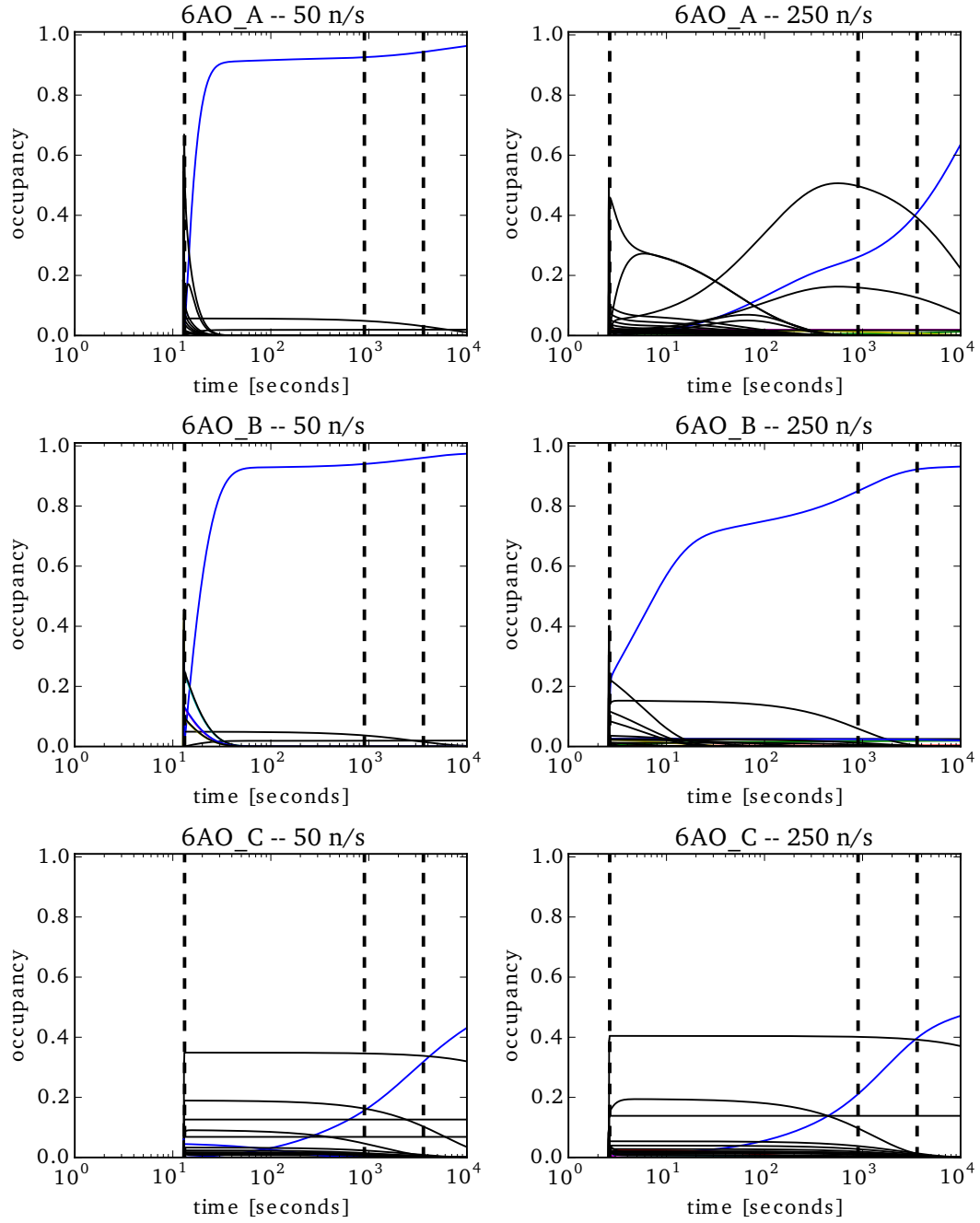


Figure 31: Cotranscriptional folding of RNA Origami at a transcription speed of 50 nucleotides/second (left) and 250 nucleotides/second (right). Three tiles are investigated (6AO-A, 6AO-B, 6AO-C from Geary et al. [2014]). The vertical dashed lines mark three time points, (1) the end of transcription, (2) 15 minutes and (3) one hour after transcription started. Only the blue trajectory corresponds to the desired RNA Origami shape, while black trajectories are not further discussed misfolded variants. **Left column:** at a transcription speed of 50 nucletides/second, the sequences 6AO-A and 6AO-B populate the desired Origami shape after about 20 seconds, while 6AO-C needs more than an hour to populate the conformation at 40 %. **Right column:** at a transcription speed of 250 nucleotides/second and 15 minutes total time, 6AO-A is populated at 20%, 6AO-B is populated at 80% and 6AO-C is populated at 20%.

Part III

RNA DESIGN

5 DESIGN OF RNA MOLECULES

The design of sequences is the inverse of the previously discussed RNA folding problem. Multiple designed riboswitches or miRNAs can form a logic regulatory networks and be used to control the information of which RNA or protein is expressed at what time point in a cell. For example, Xie et al. [2011] have used miRNAs to sense diverse disease-related phenotypes in order to trigger the expression of tumor-suppressor genes.

The remainder of this thesis, however, is about methods to design RNAs with a more diverse functional repertoire. In Chapter 6, RNAs are considered as a central molecule in the origins of life. In such a setting of extreme environmental conditions, RNA may have had a very different behavior than in todays sterile test-tubes. The method to design RNA prions presented in Chapter 7 introduces a novel self-switching mechanism. The design method optimizes a molecule to switch conformations if a particular concentration is reached.

THIS CHAPTER explains why it is possible to design RNAs that adopt one or more *common* secondary structure motifs, although the problem is formally NP-hard [Schnall-Levin, 2011]. We continue with algorithmic approaches toward RNA design and introduce a new `Perl` library shipped with the current `ViennaRNA package-v2.2`. The main advantage over existing methods is that it allows to formulate different design problems as simple scripts and it is available through the ViennaRNA web-services[1].

5.1 PROPERTIES OF RNA DESIGN LANDSCAPES

In order to get a complete picture of the RNA design problem, we come back to the concept of landscapes as previously introduced for folding kinetics in Section 2.3.

Definition 5.1 *Let $\mathcal{D} = (\Sigma, \mathcal{M}, \mathcal{O})$ be the design landscape for an RNA molecule. $\sigma \in \Sigma$ is a set of RNA sequences, $\mathcal{M}$ is a set of sequence mutations that defines neighboring sequences and $\mathcal{O}(\sigma)$ is the design objective function assigning a fitness value to each sequence.*

The size of the design landscape is 4^n where n is the sequence length, but there are many sequences which are not compatible with a given secondary structure constraint. Specifically, every specified base-pair dictates that the involved nucleotides are chosen from the set of canonical base-pairs. Hence, a sequence constraint reduces the solution space and therefore also design algorithms mutate only among these valid solutions.

1 http://rna.tbi.univie.ac.at/rnadesign

This has the interesting effect that a random sequence *given a structure constraint* has a higher probability of containing Guanine and Uracil in helical regions, than Adenine and Cytosine. The remainder of this section compares the sequence space with the structure space of RNA molecules, while Section 5.2 will discuss the inverse folding problem for uni-, bi-, and multistable sequences as well as present common objective functions.

In the previous century, a number of intrepid researchers have characterized the RNA folding problem in the context of evolution, using a map between landscapes of sequences (genotypes) and secondary structures (phenotypes) [Schuster et al., 1994; Huynen et al., 1996; Reidys et al., 1997]. Three findings are especially important for sequence design and seem to be insensitive to whether the folding algorithm is thermodynamic, kinetic, maximum matching or whether one considers one MFE structure or the entire Boltzmann ensemble [Huynen et al., 1996]. First, the mapping is often *redundant*, as there are many more sequences than structures. Second, some structures are realized much more frequently than others. As a consequence of these two properties, many mutations in the sequence space are *neutral* in a sense that they preserve the secondary structure. Third, the sequence to structure mapping is *sensitive*, such that already small changes of a sequence can lead to large changes of the structure.

For RNA design (or RNA evolution), this means that every random sequence is located in the proximity of a *common* secondary structure, and, on the other hand, it is possible to change (evolve) the entire sequence following *neutral networks*, i. e. following mutation pathways without disrupting the secondary structure [Reidys et al., 1997]. These characteristics make computational RNA design of sequences folding into a particular structure surprisingly easy, even though finding the formally *best* solution is NP-hard [Schnall-Levin, 2011].

5.2 INVERSE RNA FOLDING

In its most simple form, the inverse folding problem computes a sequence such that a target conformation is the MFE structure of this sequence. Based on the previous findings on sequence-structure maps, inverse folding can be modeled with simple adaptive walks in the design landscape. This basic idea is implemented in `RNAinverse` [Hofacker et al., 1994]. A decomposition of the target structure into smaller substructures reduces the computation time compared to a naive approach where the full sequence has to be evaluated in every step. A more efficient decomposition scheme has been implemented in `RNA-SSD` [Andronescu et al., 2004; Aguirre-Hernández et al., 2007] and eventually enabled Zadeh et al. [2011] to reduce the runtime to formally 4/3 of the RNA folding problem as implemented in `NUPACK`. An alternative design strategy is used by `INFO-RNA` [Busch and Backofen, 2006]: first the sequence with minimal free energy for the target structure is computed exactly, and then this sequence serves as

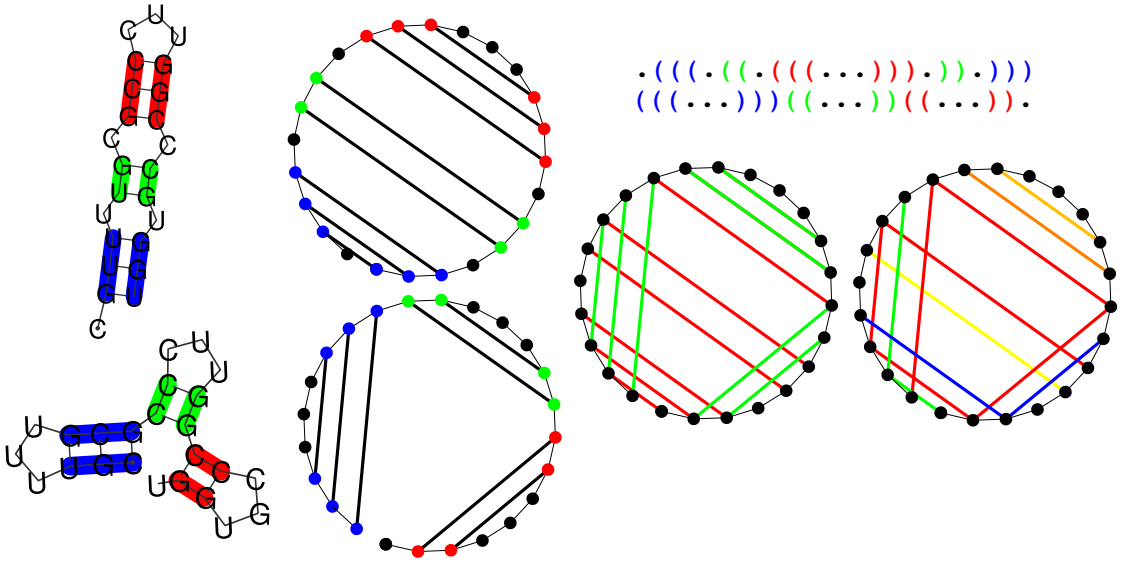


Figure 32: Visualization of dependency pathways of bistable sequences. Both structure constraints are shown in three different representations. **Left:** A common secondary structure visualization. **Top right:** The corresponding dot-bracket-strings used as standard `ViennaRNA` input format. **Center:** The corresponding circular secondary structure diagrams. **Bottom right:** The superposition of both circular diagrams reveals the dependency pathways and hence the move-set for sequence mutations.

starting point for adaptive walks to minimize the distance between the target structure and the MFE of the sequence.

Design of multistable sequences using dependency graphs

More complicated designs, such as presented in Chapters 6 and 7, require two or more target structures in order to toggle between states with different functions. In this case, every structure constraint reduces the number of feasible solutions. A visualization of this effect using dependency graphs can be seen in Figure 32. The enforced base-pairs form connected components in the dependency graph, i. e. a mutation of one base (usually) affects all other bases in the connected component.

Reidys et al. [1997] have proven that it is always possible to find a sequence compatible with two structural constraints. However, for more than two structures this is not the case. In order to determine whether a sequence is compatible with structure input, all connected components must be replaceable with a sequence of alternating purines and pyrimidines ("RYRYRYRY", e. g. "AUGCGUAU"). Formally, if the connected components are bipartite, then there exists a sequence compatible with the structural constraints and the number of possible solutions can be calculated explicitly [Höner zu Siederdissen et al., 2013].

Flamm et al. [2001] have shown an efficient dynamic programming (DP) algorithm to enumerate the sequence space of bistable molecules and enable statistically correct sampling of sequences in the reduced design landscape. This is important, as only statistically correct adaptive walks ensure a sufficient diversity of solutions at a reasonable sample size and non-structure constraints may significantly reduce the number of satisfactory solutions. A generalization of this algorithm for multistable sequence design problem has been implemented in `RNAdesign` [Höner zu Siederdissen et al., 2013] or `MODENA` [Taneda, 2015].

It is important to point out that also sequence constraints reduce the design space, such as constrained promoter regions, metabolite binding pockets, catalytically active sites (Chapter 6) and pseudoknot motifs (Chapter 7). Other design strategies involve reduced alphabets in order to avoid local interactions, e. g. 'AUC' rich regions are unlikely to form stable structures by themselves. Later in this chapter we will introduce a design library that properly addresses sequence contsraints within dependency-graphs.

Objective functions for thermodynamic optimizations

Formulation of an objective function is the most important step for the success of designing RNA molecules. On the one hand, calculations have to be fast, as they will typically be repeated hundreds or thousands of times before a solution is returned. On the other hand, the objective function can include multiple additional constraints and weight them according to their importance. In this section, a few popular objective functions will be presented. They are also often referred to as *cost functions*, because they are minimized during the sequence design process.

NOTATION We write $\sigma \in \Sigma$ for a sequence in the set of all sequences, $s \in \Omega$ for a structure in the ensemble of secondary structures s_T for the target secondary structure and $s_M = f(\sigma)$ is the MFE secondary structure of the sequence σ. The functions $e(\sigma, s_T)$, $p(\sigma, s_T)$, and $g(\sigma)$ compute the free energy of a structure, the probability of a structure, and the ensemble free energy of a sequence.

Let us have a look at typical objective functions for designing a sequence folding into a single target conformation s_T. The first variant requires the computation of the MFE conformation and the calculation of the base-pair distance between two structures. The base-pair distance is calculated as $d(s_i, s_j) = |(s_i \cup s_j) \setminus (s_i \cap s_j)|$, where s_i and s_j denote a set of base-pairs forming a secondary structure. The objective function minimizes the base-pair distance between the current MFE structure s_M and the target conformation s_T

$$\mathcal{O}(\sigma) = d(s_M, s_T) = d(f(\sigma), s_T) \tag{35}$$

Alternatively, the objective function can be formulated to maximize the probability of forming the target structure in the secondary structure ensemble.

$$\mathcal{O}(\sigma) = e(\sigma, s_T) - g(\sigma) = -RT \cdot \ln p(\sigma, s_T) \tag{36}$$

While equation 35 cannot guarantee that s_M will dominate the ensemble of structures Ω, equation 36 assumes that any structure $s_i \neq s_T$ is equally distant from the target structure s_T. The two objective function above have therefore been combined by Dirks et al. [2004] and are now known as the *ensemble defect* of a sequence implemented in the NUPACK sequence design framework [Zadeh et al., 2011].

$$\mathcal{O}(\sigma) = \sum_{s_i \in \Omega} p(\sigma, s_i) \cdot d(s_i, s_T) \tag{37}$$

Thus, the ensemble defect calculates the distance between every conformation in the ensemble $s_i \in \Omega$ and the target conformation s_T, weighted by the probability of observing s_i. The ensemble defect can alternatively be formulated as the mean base-pair distance between a random structure in the ensemble and the target structure.

$$\mathcal{O}(\sigma) = \sum_{i,j \in s_T} (1 - p_{ij}) + \sum_{i,j \notin s_T} (p_{ij}) \tag{38}$$

where p_{ij} is a matrix of base-pair probabilities, which can be computed with the same asymptotic time complexity as MFE folding.

A comparison of objective functions 35, 36 and 37 shows that both maximizing the probability of a structure, as well as the ensemble defect computation are similar in efficiency and better than only the minimization of base-pair distance [Dirks et al., 2004].

Multistable designs follow the same principle, but require additional terms to balance the free energy for the specified structures, e. g. Equation 36 for the bistable sequence optimization yields the objective function presented by Flamm et al. [2001] for bistable sequence design

$$\mathcal{O}(\sigma) = e(\sigma, s_{T1}) + e(\sigma, s_{T2}) - 2g(\sigma) + \alpha(e(\sigma, s_{T1}) - e(\sigma, s_{T2}) + \delta)^2 \tag{39}$$

where α is an optional weighting factor for the second term and δ adjusts the free energy difference between the two target conformations s_{T1} and s_{T2}.

Objective functions can also be used to include other constraints that are not strictly enforced by dependency graphs. For example, deviations from a desired GC-content, as well as particular sequence motifs can be penalized when evaluating the design objective function. Also kinetic properties, such as direct-path barriers between two conformations are applicable for RNA design [Flamm et al., 2001].

Taken together, optimization functions are composed of $O(n)$ operations such as computing the free energy of a given structure $e(\sigma, s)$ or the base-pair distance between two structures $d(s_i, s_j)$ and $O(n^3)$ methods such as the calculation of the ensemble free energy $g(\sigma)$, the minimum free energy $f(\sigma)$, the probability $p(\sigma, s_T)$.

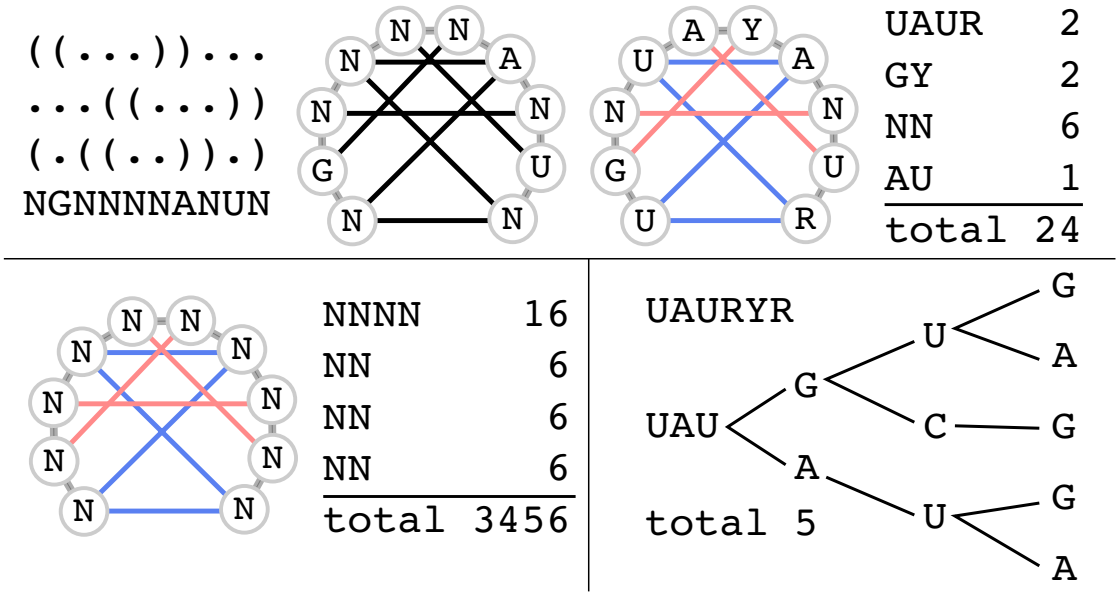


Figure 33: The impact of sequence constraints on the sequence space. **Top row:** Enumeration of the constrained sequence space as implemented in `RNA::Design`. An input for sequence design contains secondary structures in dot-bracket notation and an (optional) sequence constraint in IUPAC nomenclature. The Input is first translated into a dependency graph, then the sequence constraint is updated according to the dependencies (R: purine, Y: pyrimidine). The number of solutions for each dependency path (and cycle) are shown on the very right. There are only 24 sequences compatible with the input. **Bottom left:** If there is no sequence constrained specified, the base-pairs alone reduce the sequence space from 1048576 to 3456 possible solutions. **Bottom right:** Exhaustive representation of the dependency paths for the constraint "UAURYR". The number of leaves yields the number of possible sequences.

5.3 PLUG AND PLAY RNA DESIGN

The sequence designs presented in Chapters 7 and 6, were optimized using adapted versions of the `switch.pl` method discussed before. However, a number of ongoing side-projects showed that every sequence design task requires different numbers of sequence and structure constraints, as well as different types of objective functions. We have therefore written a `Perl` library `RNA::Design` to write different design tasks as simple scripts. The main features and performance improvements compared to `switch.pl` are described below. The library is distributed with the current release of the `ViennaRNA package-v2.2` and used in the new ViennaRNA sequence design web-interface at `http://rna.tbi.univie.ac.at/rnadesign`.

RNA design problems optimize a sequence from the following input: (1) One or more secondary structures, (2) an optional sequence constraint (3) an objective func-

tion and (4) additional parameters. Each of these inputs can be flexible and will be addressed below:

(1) SECONDARY STRUCTURES The secondary structure input can be used in two different ways. Either the conformations determine dependency paths for sequence mutations, or they are addressed in the objective function. Besides the standard dot-bracket notation, secondary structures may contain the following special characters:

- '&' connects two sequences to design a pair of interacting RNAs. We have used this feature to produce the sequences for Chapter 4.
- 'x' is only supported for structures that are used in the objective function and allows the computation of the accessibility of nucleotides (i. e. the probability of being unpaired).

The number of secondary structures is not restricted which enables the design of multistable RNAs such as shown in Figure 33. However, `RNA::Design` avoids difficulties of multstable designs where a single nucleotide has more than two dependencies. In that case, base-pair constraints are not added to the dependency graph, however, the constraints are still evaluated in the cost function.

(2) SEQUENCE CONSTRAINT A sequence constraint may be specified in IUPAC nomenclature and enforced during the optimization process. While unconstrained dependency pathways are enumerated and sampled exactly using the DP algorithm of `switch.pl`, the constrained pathways are sampled by an exhaustive approach shown in Figure 33. We have realized recently that this exhaustive approach can also be solved exactly using DP, which will implemented in future versions of the design library.

(3) OBJECTIVE FUNCTION The objective function can be customized using a simple interface to the functions of the `ViennaRNA package`. In particular, every input secondary structure can serve as full target conformation or structure constraint. The objective function currently supports to compute the free energy of a target structure, the (constrained) ensemble free energy, the (conditional) probabilities of secondary structure elements, the accessibility of subsequences and the direct-path barriers between two structures. All of these terms exist for linear, circular, and cofolded molecules, as well as for custom specified temperatures. A more detailed documentation can be found at `http://rna.tbi.univie.ac.at/rnadesign`.

(4) ADDITIONAL PARAMETERS The user may select a desired GC-content of target sequences and choose to avoid specific sequence motifs such as repetitions of the same nucleotide. These terms are then included into the optimization with the objective function.

5.4 DESIGN OF KINETIC PROPERTIES

The incorporation of kinetic simulations into RNA design is complicated by the fact that they are to time consuming. However, it is possible to infer ad-hoc rules to get reasonably accurate fast predictions. The probably simplest form are previously mentioned direct-path barriers.

For example, in Chapter 7 we compute the direct path barriers using the `findpath` method during the optimization, and then rank the best candidate molecules again using exact solutions from `barriers`. Similarly, one can specify dedicated intermediate states to optimize indirect folding pathways composed from multiple direct pathways.

As we have shown in Chapter 3, also structure constraints in form of toeholds are commonly used to lower kinetic folding barriers. Whenever applicable, this is an even faster alternative to heuristic evaluations with the cost-function, but should be used in combination with a thorough post-evaluation.

Computing accessibilities of unpaired regions can be used for cotranscriptional sequence designs. The sequence forming the nucleation point of a helix can be optimized to be unstructured before the binding partner is transcribed.

6

SEQUENCE-CONTROLLED RNA SELF-PROCESSING: COMPUTATIONAL DESIGN, BIOCHEMICAL ANALYSIS, AND VISUALIZATION BY AFM

ARTICLE

Sonja Petkovic*, Stefan Badelt*, Stephan Block, Christoph Flamm, Mihaela Delcea, Ivo L. Hofacker and Sabine Müller.
Sequence-controlled RNA self-processing: computational design, biochemical analysis, and visualization by AFM.
in *RNA* (2015), **Volume** 21, **pages** 1249–1260
doi: 10.1261/rna.047670.114

* Shared first authorship

AUTHOR CONTRIBUTIONS

Sonja Petkovic established the experimental setup under supervision of Sabine Müller and conducted the PAGE exeriments to assess the behavior of self-processing ribozymes. Stefan Badelt has developed the method to design self processing ribozymes under the supervision of Ivo L. Hofacker and Christoph Flamm, Stefan Badelt has designed the sequences and computationally analysed the experimental results. Stephan Block made the atomic force microscopy (AFM) images under supervision of Mihaela Delcea, measured the length-distribution of observed RNA species and prepared the corresponding statistics. Sonja Petkovic, Stephan Block and Stefan Badelt have interpreted the experimental results and written the sections of our own work. Sabine Müller and Stefan Badelt have written the final version of the manuscript including Abstract, Introduction and Conclusion.

LICENCE

Sequence-controlled RNA self-processing: computational design, biochemical analysis, and visualization by AFM

SONJA PETKOVIC,[1,5] STEFAN BADELT,[2,5] STEPHAN BLOCK,[3] CHRISTOPH FLAMM,[2] MIHAELA DELCEA,[3] IVO HOFACKER,[2,4] and SABINE MÜLLER[1]

[1]Institute for Biochemistry, Ernst-Moritz-Arndt University Greifswald, 17487 Greifswald, Germany
[2]Institute for Theoretical Chemistry, University of Vienna, A-1090 Vienna, Austria
[3]ZIK HIKE—Center for Innovation Competence, Humoral Immune Reactions in Cardiovascular Diseases, Ernst-Moritz-Arndt University Greifswald, 17489 Greifswald, Germany
[4]Research Group Bioinformatics and Computational Biology, University of Vienna, A-1090 Vienna, Austria

ABSTRACT

Reversible chemistry allowing for assembly and disassembly of molecular entities is important for biological self-organization. Thus, ribozymes that support both cleavage and formation of phosphodiester bonds may have contributed to the emergence of functional diversity and increasing complexity of regulatory RNAs in early life. We have previously engineered a variant of the hairpin ribozyme that shows how ribozymes may have circularized or extended their own length by forming concatemers. Using the Vienna RNA package, we now optimized this hairpin ribozyme variant and selected four different RNA sequences that were expected to circularize more efficiently or form longer concatemers upon transcription. (Two-dimensional) PAGE analysis confirms that (i) all four selected ribozymes are catalytically active and (ii) high yields of cyclic species are obtained. AFM imaging in combination with RNA structure prediction enabled us to calculate the distributions of monomers and self-concatenated dimers and trimers. Our results show that computationally optimized molecules do form reasonable amounts of trimers, which has not been observed for the original system so far, and we demonstrate that the combination of theoretical prediction, biochemical and physical analysis is a promising approach toward accurate prediction of ribozyme behavior and design of ribozymes with predefined functions.

Keywords: AFM; circularization; computational design; hairpin ribozyme; RNA; self-processing

INTRODUCTION

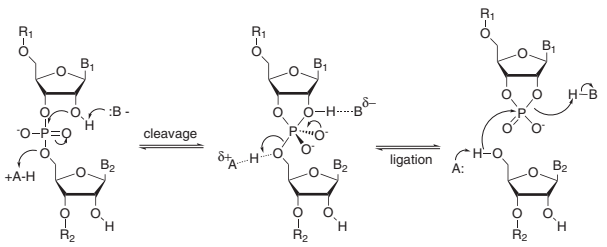
RNA processing plays a fundamental role in the cellular life cycle. RNA molecules are permanently synthesized, modified, edited, truncated, or abolished. In viruses, viroids, and satellite RNAs with circular RNA genomes, replication follows a rolling circle mechanism, thus initially producing linear concatemeric versions of the RNA genome (Flores et al. 2011). Further processing is required to convert the concatemers back to monomers that subsequently are cyclized to yield the final replication product: a cyclic RNA complementary to the template. This processing is dependent on specific RNA structural motifs that support reaction at the site of cleavage and ligation (Hampel and Tritz 1989; DeYoung et al. 1995). Among those, the hammerhead and the hairpin ribozyme are probably the best studied small RNAs with catalytic activity (Hammann et al. 2012; Müller et al. 2012).

Hairpin ribozyme catalyzed RNA cleavage and ligation reactions follow a transesterification mechanism (Cochrane and Strobel 2008). Cleavage occurs by nucleophilic attack of the 2′-oxygen on the neighboring phosphorous resulting in a trigonal-bipyramidal intermediate. Upon release of the 5′-OH-group, a 2′,3′-cyclic phosphate is formed. Ligation follows the same reaction path in opposite direction and proceeds via ring opening of the cyclic phosphate, exclusively delivering the natural 3′,5′-phosphodiester (Scheme 1). Ligation is enthalpically favored over cleavage, because ring strain energy is released when opening the cyclic phosphate. Entropically, ligation is disfavored, owing to the decrease in degrees of conformational freedom. However, the entropic cost of ligation is rather small and can be compensated by the favorable enthalpic contribution (Hegg and Fedor 1995, Nahas et al. 2004). In addition, ligation is about two times faster than cleavage (Liu et al. 2007). Thus, the internal equilibrium of the hairpin ribozyme is shifted toward ligation. Translated into practical use this means that the two activities

[5]**Shared first authorship.**

Corresponding authors: smueller@uni-greifswald.de, ivo@tbi.univie.ac.at, stephan.block@chalmers.se

Article published online ahead of print. Article and publication date are at http://www.rnajournal.org/cgi/doi/10.1261/rna.047670.114. Freely available online through the *RNA* Open Access option.



SCHEME 1. Hairpin ribozyme mechanism (from *left* to *right*) nucleophilic attack followed by intermediate formation and release of newly formed 5′- and 3′-termini (see text for details); (R_1, R_2) oligonucleotide chain, (B_1, B_2) nucleobase, (A) acid, (B^-) base.

can be controlled by structural modulation. Hairpin ribozymes that form a stable structure, such that fragments remain bound, favor ligation, whereas hairpin ribozymes that are less stable, such that cleavage fragments can easily dissociate, favor cleavage (Fedor 1999, Welz et al. 2003). These characteristic features distinguish the hairpin ribozyme from other small ribozymes, and we have shown in previous work that structural manipulation of hairpin ribozyme variants allows tuning of cleavage and ligation activity (Welz et al. 2003; Ivanov et al. 2005; Vauleon et al. 2005; Drude et al. 2007, 2011; Pieper et al. 2007; Petkovic and Müller 2013, Balke et al. 2014). Among these variants is a hairpin ribozyme that can cleave off its 5′- and 3′-end (Pieper et al. 2007). The cleaved product has two reactive ends that can ligate to circular species or concatemers of two or more molecules.

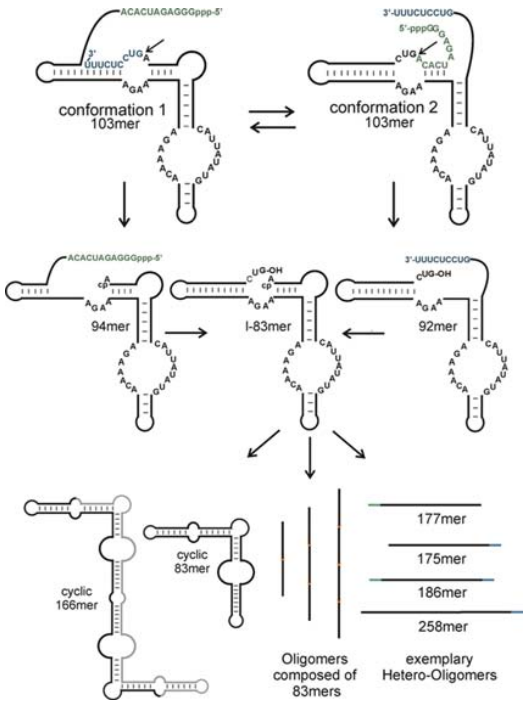
Herein we address the question whether it is possible to design entirely self-reactive RNAs to efficiently circularize OR polymerize to large RNA entities. In contrast to previous work, our purpose was to optimize RNA sequences for particular conformations favoring monomeric variants or multimerization, rather than tuning the cleavage and ligation rate itself. Self-reactive RNA molecules changing their properties by circularization or increasing their length by polymerization provide a good case-study to exploit the repertoire of state of the art computational design algorithms and to improve them by experimental verification. Good heuristics to embed catalytic activity into RNA molecules with desired functions are highly amendable for synthetic biology, since RNA cleaving or ligating ribozymes constitute an additional layer of gene regulation. Additionally, successful designs would have direct implications on the RNA world theory to explain the emergence of RNA genomes in an early RNA world.

Using the ribozyme CRZ-2 (Scheme 2), which was developed previously (Pieper et al. 2007) and recently analyzed in detail (Petkovic and Müller 2013), as template, we computationally optimized sequences using the program switch.pl (Flamm et al. 2001) of the Vienna RNA package (Lorenz et al. 2011). Four variants with different behavior according to our scoring functions were selected and analyzed in detail by polyacrylamide gel electrophoresis (PAGE) and atomic force microscopy (AFM). We present high resolution AFM images of the reaction mixtures, visualizing even the rather short 83mer RNA fragment.

RESULTS

Computer-aided sequence design

Compared with manual design that we had applied in previous work (Pieper et al. 2007; Petkovic and Müller 2013), computer-aided design is a more sophisticated way toward control of self-processing activity of RNA species. Therefore, we have started a bioinformatics approach to evolve hairpin ribozyme derived RNAs with self-processing activity. We have designed two classes of ribozyme species: Members of the first class should process themselves efficiently into circular monomers, whereas members of the second class would maximize the yield of ligation competent dimers. The design



SCHEME 2. RNA self-processing pathway of CRZ-2. The pathway also applies to designed sequence variants PBD1 to PBD4 (see below). Note that fragment lengths differ for PBD3 and PBD4. RNAs are programmed to fold in two distinct conformations (*top*). Both conformations favor cleavage, such that either the 5′-terminus (green) or the 3′-terminus (blue) can be cleaved off, resulting in a 94- or 92mer (*middle*). These intermediates can refold in the conformation required for cleavage of the remaining 3′- or 5′-end, respectively. The final cleavage product is always an 83mer, which can undergo intramolecular ligation to a circular species (*bottom*, *left*) or self-concatemerize by intermolecular ligation (*bottom*, *middle*). In addition, the fragments resulting from the first cleavage contain either the 2′,3′-cyclic phosphate or the 5′-OH group required for ligation, such that they can also oligomerize with each other or with one or more 83mers (*bottom*, *right*).

process is complicated by the fact that multiple constraints exist on both sequence and structure level. On the sequence level we included two well-conserved interior loop regions from the hairpin ribozyme (Berzal-Herranz et al. 1993), as well as a T7 RNA promotor sequence at the 5′-end for experimental implementation. On the structural level, the constructs have to be bistable, forming two distinct catalytic centers to cleave off both the 5′- and 3′-ends as depicted in Figure 1 and Supplemental Figure S1. Our approach is a two-step process that first computes a large set of RNA sequences with catalytic properties, and second scores these sequences to select for ribozymes with the desired behavior. Previously, we have shown that the efficient design of bistable molecules is surprisingly easy (Flamm et al. 2001). The algorithm, implemented in the program switch.pl of the Vienna RNA package, mutates initially random sequences into bistable switches *via* consistent mutations guided by a dependency graph. The mutations are meant to increase the probability of forming catalytically active structures and influence the conformations formed upon dimerization of the individual species. Using switch.pl, ~10,000 bistable RNA molecules conforming to the above design objective were designed and ranked according to two scoring functions κ_1 and κ_2 (Equations 2 and 4 in Materials and Methods) that evaluate the probabilities of forming reactive structures and the fraction of circular species in equilibrium. Results of the scoring function for all four sequences and the reference system CRZ-2 can be seen in Table 1. A lower κ_1-value indicates high catalytic activity of all monomeric variants; a lower κ_2 indicates a high probability of forming catalytically active homo-dimers. Hence, κ_2 was used to discriminate between ribozymes that are meant to favor formation of cyclic dimers (lower κ_2) and those that do not (higher κ_2). Detailed explanation and formulas can be found in Materials and Methods and Supplemental Figure S2a,b. Figure 1 shows four bistable ribozyme sequences (PBD1 to PBD4) that were selected for experimental validation in comparison to the reference system CRZ-2. Table 1 summarizes their expected properties and the results from the scoring functions (rounded to two decimal figures).

TABLE 1. Summarized properties of the designed sequences

RNA	Full length	Fragment length	κ_1	κ_2
CRZ-2	103	11 + 83 + 9	19.24	37.58
PBD1	103	11 + 83 + 9	9.38	13.71
PBD2	103	11 + 83 + 9	9.84	28.71
PBD3	105	8 + 83 + 14	12.14	18.77
PBD4	105	8 + 83 + 14	12.13	31.23

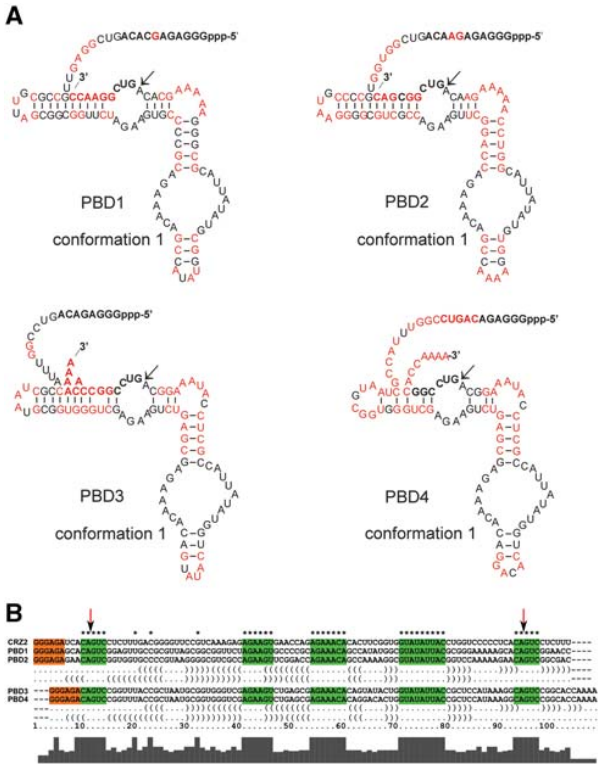


FIGURE 1. (*A*) Secondary structures of CRZ-2 and PBD 1–4 shown in one of the two possible conformations (cf. Scheme 2). Sequence changes in comparison to the reference RNA CRZ-2 are shown in red. (*B*) Sequence alignment of the four designed RNAs PBD1–4 with the reference system CRZ-2. Green interior loop areas are reported to be essential for cleavage/ligation activity and were therefore fixed during the design process. The orange colored T7 RNA promoter sequence was needed for experimental implementation. The secondary structure in dot-bracket notation *below* shows the constraints on a structural level.

Compared with the reference RNA CRZ-2, PBD1–4 differ by various base replacements in the nonconserved regions (Fig. 1). However, the new designed ribozymes were meant to undergo the same cleavage cascade reaction as described for CRZ-2 previously (Petkovic and Müller 2013) and depicted in Scheme 2. Dimerization of the hairpin ribozyme has been demonstrated previously (Butcher and Burke 1994), and is an essential prerequisite for the formation of concatemers by CRZ-2. We therefore assumed that sequences forming catalytically active, intermolecular ligation competent dimers favor concatemerization (PBD1 and PBD3), while sequences that have lower tendency to form these structures, are assumed to predominantly form cyclic monomers (PBD2 and PBD4). Our scoring functions furthermore indicate that PBD1 and PBD2 show increased efficiency to form cyclic monomers (Table 1, κ_1) compared with PBD3 and PBD4.

Biochemical analysis of the self-processing behavior of the designed sequences

The five RNAs, CRZ-2 and PBD1 to PBD4 were prepared by in vitro transcription with T7 RNA polymerase (see Supplemental Material) and incubated at conditions favoring self-cleavage followed by ligation (Petkovic and Müller 2013).

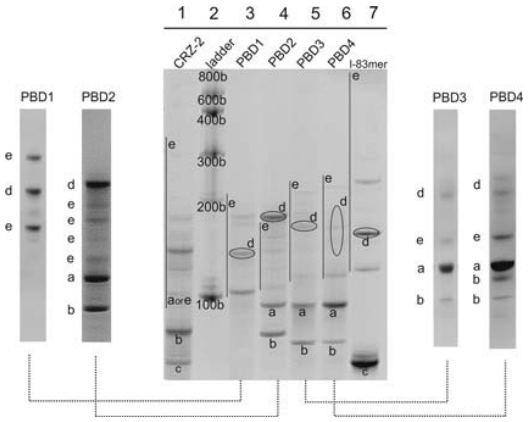


FIGURE 2. Analysis of self-processing reactions of sequences PBD1-4, CRZ-2, and the linear 83mer (l-83mer) in a 15% denaturing polyacrylamide gel, lane *2*: RNA size standard. For better visualization of individual bands, self-processing reactions of PBD1–PBD4 were analyzed separately with a higher amount of sample loaded onto the gel (separate lanes *left* and *right* to the gel. Note that large scale analysis was carried out separately for each of the designed RNAs PBD1–PBD4. Therefore, the relative positions of bands are not directly comparable between individual gels. Compare also Supplemental Figure S3. Bands were assigned as follows: full-length transcripts (a); cleavage intermediates (5′- or 3′-truncated transcripts) (b); final cleavage product (5′- and 3′-truncated transcripts) (c); cyclic RNA resulting from intramolecular ligation of species c (d); concatemers resulting from intermolecular ligation of species b and c (e).

First, reaction products were analyzed using denaturing polyacrylamide gels (Figs. 2, 3, 4). Bands in the gel were visualized by ethidium bromide staining. Table 2 shows the lengths (in number of bases) of products that theoretically can be formed; Figure 2 shows an overview of reactions of all self-processing ribozymes (CRZ-2, PBD1–PBD4). For comparison, the linear 83mer (l-83mer) resulting from two cleavage events in CRZ-2 and being incapable of further cleavage was isolated and incubated at identical conditions (Fig. 2, lane 7). The behavior of this 83mer was analyzed in detail recently (Petkovic and Müller 2013), such that the band pattern produced by the l-83mer could be used as guideline to navigate through the PAA gel and, with the aid of the 2D-gel electrophoresis results (see below), to assign the obtained bands to individual RNA species. This becomes especially important, since chemical modifications at RNA ends (such as OH, phosphate or cyclic phosphate), RNA sequence itself, and RNA structures formed in spite of denaturation can affect the migration behavior of RNA molecules. The standard length marker (lane 2) can only serve as an approximate guideline for higher ligation products.

*Full-length transcripts (**a**—103/105mers)*

The 103 (CRZ-2, PBD1 and PBD2) or 105mers (PBD3 and PBD4) are typically located below the 100 nt size standard mainly due to the triphosphate at the 5′-end resulting from in vitro transcription of the ribozymes (Fig. 2; Supplemental Fig. S3). The 103mer of CRZ-2 and PBD1 is barely detectable after ribozyme reaction (Fig. 2, lanes 1 and 3), whereas full length transcripts of PBD2 to PBD4 (lanes 4, 5, and 6) are still visible. This implies that CRZ-2 and PBD1 exhibit higher activity in cleaving off the 5′- or the 3′-end or both, and producing the shortened fragments denoted with ***b*** and ***c*** (Table 2; Fig. 2).

Cleavage products (b—97, 94, 92, 91mer and c—linear 83mer)

In CRZ-2, PBD1 and PBD2, a 92mer and a 94mer are produced as intermediates upon the first cleavage. These two intermediates occur as one band, since the 94mer carries additional charges from the triphosphate at the 5′-end. For CRZ-2 and PBD2 (Fig. 2, lanes 1, 4), a prominent 92/94mer band is visible, PBD1 shows none of these species. PBD3 and PBD4 produce a 91mer and a 97mer, with the 91mer carrying the triphosphate. Both systems show 91mers, whereas the 97mer is only detectable for PBD4 (lanes 5, 6, and gel pieces shown on the right). The final cleavage product of all test systems is a linear 83mer. Lane 7 shows the 83mer from CRZ-2 used as an additional size standard (Petkovic and Müller 2013). Interestingly, only CRZ-2 shows a band corresponding to the final cleavage product, while linear 83mers of PBD1-PBD4 are not detectable, suggesting an immediate consumption in ligation reactions.

Intramolecular monomeric ligation (d—cyclic 83mer)

From all produced monomers (l-83mer, 91mer, 92mer, 94mer, 97mer, 103mer, and 105mers) the linear 83mer is the only RNA that may perform cyclization due to the chemical constitution at its 3′- and 5′-end. However, the migration behavior of an unknown cyclic species in a PAA gel is impossible to predict by common size markers, since the overall

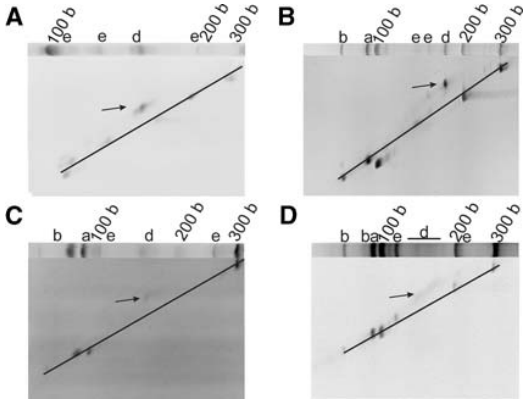


FIGURE 3. Two-dimensional gel electrophoretic analysis of PBD1 (*A*), PBD2 (*B*), PBD3 (*C*), and PBD4 (*D*). All samples were mixed with a linear RNA size standard prior to subjecting onto the gel. The first dimension gel of the respective system is implemented in each panel. The diagonal marks the linear RNAs; circular species are denoted by an arrow.

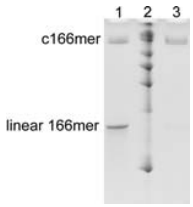


FIGURE 4. Enzymatic ligation of the inactive 166mer and treatment with exonuclease RNase R. Lane *1*: ligation mixture composed of ligation product and remaining nonligated linear transcript; lane *2*: RNA size standard, 100 bases, 200 bases, 300 bases, 400 bases, 600 bases, 800 bases, and 1000 bases; lane *3*: ligation product mixture after treatment with RNase R (for details see Supplemental Material).

shape and the migration behavior of the cyclic RNA strongly depend on the sequence (Grabowski et al. 1984; Sigurdsson and Eckstein 1996). Previously, we have set up a two-dimensional-PAA gel electrophoresis assay (see Experimental section of the Supplemental Material, and Petkovic and Müller 2013) to identify cyclic species by means of their nonlinear movement at different PAA concentrations. While linear species move on a diagonal in the second dimension, cyclic species are expected to show irregular movement. Full-length CRZ-2 appears to form, if at all, only traces of a circular 83mer (Fig. 2, lane 1; cf. also Supplemental Fig. S4), while incubation of the isolated linear 83mer of CRZ-2 alone (lane 7) clearly produces the cyclic species. The cyclic 83mer is located approximately at the 150 nucleotide size standard. For PBD1-PBD4, identification of cyclic species was possible by 2D gel electrophoresis (Fig. 3). Interestingly, all newly designed RNAs (PBD1-PBD4, lanes 3–6) do show circular 83mers, while not showing any linear 83mers. In case of PBD4, the cyclic species is not represented by a discrete band, but rather appears as a smear.

Higher noncyclic ligation products (e—dimers, trimers, concatemers)

Intermolecular backbone ligation can only occur upon dimerization of the 83mer and/or the intermediate cleavage products, carrying the required termini (5′-OH and 2′,3′-cyclic phosphate). A summary of these species can be seen in Table 2.

Identification of monomeric cleavage products was straight forward since they move roughly according to their size. Identification of higher ligation products is challenging, because their movement can be irregular (Cruz-Reyes et al. 1998). However, by means of the l-83mer marker (Fig. 2, lane 7) we know that the species moving ~150 nt length is actually the circular 83mer, while the covalently linked linear 166mer (83 + 83) is located roughly at 120 nt length. Bands above the 200 bases ladder correspond most likely to 249mers (83 + 83 + 83) and even longer molecules. In comparison, we do see multiple species between 100 and 200 nt in the full-length CRZ-2 lane (Fig. 2, lane 1). We can clearly identify the 166mer at the same height as the 166mer in the lane of the l-83mer reference marker (lane 7). Shortly above is a stronger band indicating intermolecular ligation of the 83mer with a 92/94mer, or of the 92mer with the 94mer, respectively. The ratio between linear 166mer and 175/177mer would also be similar to the observed ratio between 83mer (***c*** in lane 1) and 92/94mer (***b*** in lane 1). The bands further up are hard to interpret and might show a little of c83mer and 186mer (92 + 94), as well as a 258/260mer (83 + 83 + 92/94) next to the 300 bases ladder.

Assignment of bands becomes more difficult for PBD1 to PBD4. PBD1 (lane 3), our most efficient ribozyme concerning 5′- and 3′-end cleavage, shows two bands in addition to the c83mer, which most likely correspond to the linear 166mer (83 + 83) and 249mer (83 + 83 + 83), respectively. PBD2 (lane 4) shows four species between the 105mer (***a***) and the c83mer (***d***). Since we can see a clear band for 92/94mers (***b***) we suggest that these species took part in intermolecular ligation reactions with 83mers resulting in a diverse set of dimers (***e***). However, we cannot exclude that a low running trimer is present as well. PBD3 and PBD4 (lane 5, 6) show mostly the same species with different intensities. Analogous to PBD1 they show noncircular species that most likely represent the homo-dimer (83 + 83) and possibly hetero dimers and/or a trimer.

Cyclic dimer formation

With the purpose of identifying cyclic dimers in the reaction mixture we designed and synthesized an inactive dimer (CRZ*) which should mimic the behavior of its CRZ-2 equivalent (Supplemental Material). Figure 4 shows two versions of nonreactive CRZ*, the linear species at ~166 nt length and the enzymatically ligated circular version at a height of ~800 bases (lanes 1 and 3). By comparison with the results shown in Figure 2, a band at comparable height (~800 bases), is detected only in the l-83mer marker (lane

TABLE 2. Lengths of possible cleavage and ligation products upon RNA self-processing

RNA	Full length (a)[a]	Cleavage products (b, c)[a]	Ligation products composed of only 83mers (d + e)[a,b]	Ligation products of mixed composition (e)[a,b]
PBD1	103	83 (c), 92, 94 (b)	c83 (d), 166, c166, 249 (e)	175, 177, 186
PBD2	103	83 (c), 92, 94 (b)	c83 (d), 166, c166, 249 (e)	175, 177, 186
PBD3	105	83 (c), 91, 97 (b)	c83 (d), 166, c166, 249 (e)	174, 180, 188
PBD4	105	83 (c), 91, 97 (b)	c83 (d), 166, c166, 249 (e)	174, 180, 188
CRZ-2	103	83 (c), 92, 94 (b)	c83 (d), 166, c166, 249 (e)	175, 177, 186

[a](a) Full-length transcript, (b) cleavage intermediates, (c) final cleavage product, (d) cyclic monomer, (e) concatemers; compare also legend of Figure 2.
[b]Note that in addition to dimers and trimers also longer concatemers can be formed.

7), although being rather weak. All other ribozymes do not exhibit measurable amounts of circular dimers in PAA gels.

AFM measurements

To obtain deeper insight into the self-processing behavior of the designed RNAs and in particular into chain lengths distribution, we supplemented the biochemical analysis by Atomic Force Microscopy (AFM). AFM imaging is able to visualize RNA chains on single molecule level (Henn et al. 2001), allowing to characterize even rarely produced RNA species that are difficult (if not impossible) to observe in gel electrophoretic experiments. We analyzed four candidates out of the investigated self-processing RNAs with AFM imaging under semidenaturing conditions: CRZ-2, PBD1, PBD4, and the isolated linear 83mer of CRZ-2. These reaction mixtures showed high diversity upon biochemical analysis (Fig. 2, lanes 1, 3, 6, 7), with PBD1 forming predominantly cyclic 83mers, linear 166mers, and 249mers, and PBD4 expressing a plethora of dimeric and of multimeric species.

Figures 5 and 6 show representative examples for tapping mode (TM) AFM images of ribozymes (recorded in air after RNA immobilization on mica and drying). The observed RNA chains adopt either a coiled (see white arrow in Fig. 5A), or uncoiled conformation which consists of rod-like segments, connected by kinks. Hence, for the uncoiled conformation it is possible to measure the lengths of the constituting segments as well as the contour length of the whole chain. That allows a comparison of the observed molecules with secondary structure prediction of the species listed in Table 2. Immobilized under native conditions, all observed molecules had a coiled conformation (data not shown), while semidenaturing conditions resulted mostly in uncoiled conformations having the rod-kink-motif. Hence, the majority of the AFM measurements were done on RNA chains prepared under semidenaturing conditions (see Figs. 5, 6 for a representative overview). Histograms showing both the contour lengths and the segment lengths for all four analyzed ribozymes can be seen in Supplemental Figures S5 and S6, contour-length results are summarized in Table 3. These histograms are in agreement with the expected values from secondary and tertiary structure prediction: All single-molecule ribozyme species (83mer–105mer) are expected to form a reactive structure with two stiff helical regions (segments) connected with a flexible kink. If we assume a typical pitch of 0.3 nm per base pair (Arnott et al. 1973; Henn et al. 2001), the 83mer consists of two stiff regions with 5.4 and 6.3 nm length plus a kink of about five bases. The contour length would therefore be ~11.7 nm plus the kink region. Monomers that have noncleaved ends would form the same helices but have additional single stranded regions in the kink region or sticking out from one of the helices. Based on these single stranded regions, different monomer variants would be hardly distinguishable with AFM imaging. Accordingly, different dimer species are expected to fold into a conformation where ~166 bases are involved in successive helical regions (166×0.15 nm = 24.9 nm), trimers with 249 bases resulting in 37.35 nm and so on.

The contour-length histogram of the linear 83mer shows three contour-length peaks at 13.4, 24.5, and 36.5 nm, as well as very few molecules with even higher lengths (Supplemental Fig. S6a). These peaks closely map to expected values for regularly folded monomers, dimers and trimers, respectively, making an interpretation straightforward. The segment lengths, representing individual helical regions, showed peaks at 5.9, 8.5, and 13.8 nm (Supplemental Fig. S5a), which most likely correspond to the monomer helices and dimer-variants of these helices. AFM imaging reveals that the monomer of the l-83mer typically consists of

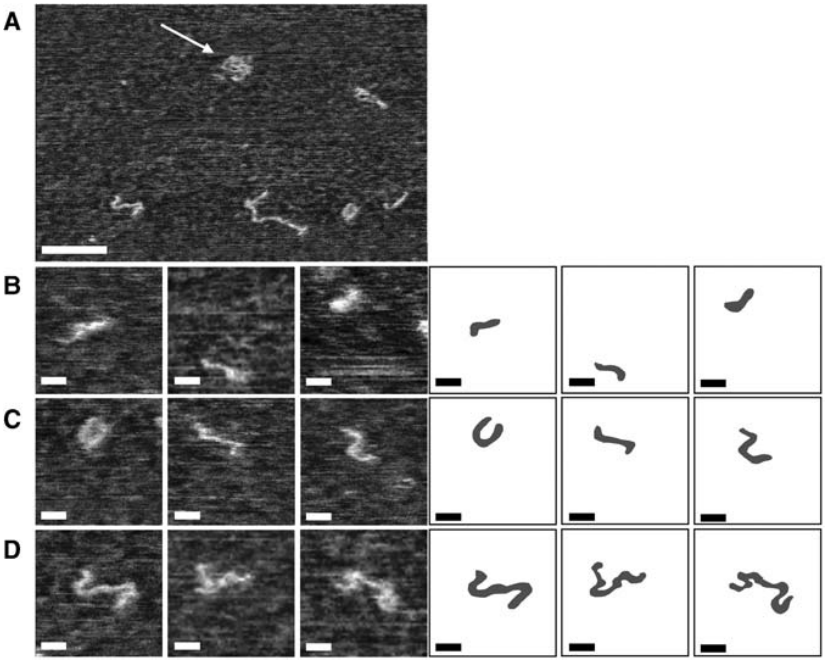


FIGURE 5. Tapping mode (TM) AFM phase images (range: 0°–30°) of the reaction products resulting from incubation of the l-83mer (isolated from CRZ-2 system) in cleavage/ligation buffer. For AFM analysis, samples were precipitated and resolved in 25 mM EDTA and 3.5 M urea (semidenaturing conditions). Scale bars: 50 nm (*A*), 10 nm (*B–D*). The overview scan (*A*) shows RNA chains in coiled (white arrow) and unwrapped conformation. High-resolution TM images (*B–D*) allow investigation of the internal structure of 83mers (*B*), dimers (*C*), and trimers (*D*). For convenience, schematics have been included on the *right* side to help with the interpretation of the AFM images. The corresponding height images are given in Supplemental Figure S7.

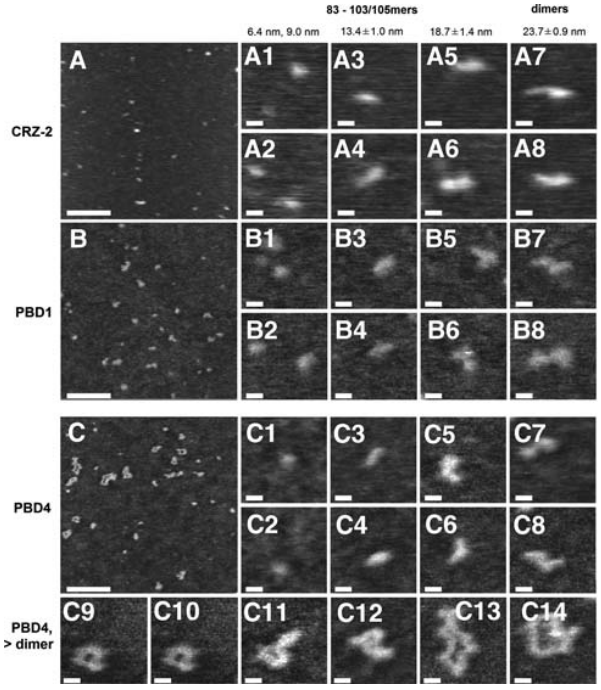


FIGURE 6. AFM images of RNA: CRZ-2 (*A*, a1–a8), PBD1 (*B*, b1–b8), and PBD4 (*C*, c1–c14). Scale bars: 100 nm (*A*–*C*), 10 nm (a1–c14); height scale 1 nm in all images. RNA chains have typically a height of ~0.4 nm in the AFM images. Overview scans (*A*–*C*) show a mixture of RNA chains of different contour length L_C for all three sequences investigated. Analysis of contour-length histograms (see Supplemental Fig. S5 and S6) for CRZ-2, PBD1, and PBD4 allowed association of most of the observed RNA chains with the species listed in Table 2 (as indicated in the figure). This procedure failed for two shortest species (a–c 1,2) and the one having a contour length ~18.7 nm (a–c 3,4), as the calculated numbers of bases did not match any entry of this table (see text for a detailed discussion).

one "short" and one "long" segment, which enclose an angle of ~110° (Fig. 5B). The dimer may be composed of two, three or even four segments (Fig. 5C, left to right), while the trimer shows typically a very complicated internal structure (Fig. 5D). Hence, a large variety of possible conformations is observed for dimers and trimers in the AFM images.

Full-length CRZ-2, as well as PBD1 and PBD4 show an even wider spectrum of contour-length peaks and chain conformations (see Supplemental Figs. S5, S6), which is expected from the computational design and the biochemical analysis. In the AFM measurements, full-length CRZ-2 creates predominantly species being shorter than 24 nm (Supplemental Fig. S6b). This "cut-off" shifts to 36 nm for PBD1, while much longer RNA chains (up to ~80 nm) are observed for PBD4. Hence, the population shifts progressively to longer RNA products from CRZ-2 over PBD1 to PBD4, which is in agreement with the gel electrophoretic analysis. All structures show contour-length peaks at expected values close to those from the l-83mer, which makes an identification of monomers, dimers, and trimers straightforward. Monomers appear mostly in a rod-like conformation, and a kink (similar to the l-83mer monomers) is rarely resolvable (Fig. 6, a3, a4, b3, b4, c3, c4). Dimers adopt L- and Z-like conformations (Fig. 6, a7, a8, b7, b8, c7, c8). Higher ligation products (trimers, etc.) are currently only observed for PBD4, which can lead (similar to the l-83mer) to very complicated and irregularly shaped internal chain structures (c9–c14 in Fig. 6). However, besides these species, also additional peaks are found at contour lengths that are (i) shorter than the expected value for a regularly folded monomer (6.5 and 9.0 nm, observed for CRZ-2, PBD1 and PBD4), (ii) between the monomer and dimer length (17.2 nm for CRZ-2, 19.0 nm for PBD1, and 20.0 nm for PBD4), or (iii) between the dimer and trimer length (31.7 nm for PBD1, 29.6 nm for PBD4). We can exclude that cleaved ends from processed full-length ribozymes would have a length of 6.8 nm or 9.5 nm in the AFM images (for the measurement conditions used in the experiments). Instead, we can show that the smaller peaks match very well to segment length measurements (Supplemental Fig. S5), suggesting that only one of the two helices is resolved by AFM imaging. Accordingly, we know that the catalytically active structure involves tertiary interactions to closely orient both helices to each other. Uncleaved structures with single stranded regions in the kink region might favor the back folding of the helices despite semidenaturing conditions, which are meant to destroy tertiary base pairs. The AFM images further support this interpretation: Species having a contour length ~18.7 ± 1.4 nm (i.e., between the monomer and dimer length) typically show an L-like conformation. Complementing this chain structure with a third segment (which might be irresolvable in the images due to back folding of one helix) having a length of 6.6 ± 0.4 nm (first peak in the segment length histograms, see Supplemental

TABLE 3. AFM contour-length measurements and their implications on the ratio between monomers (M), dimers (D) and trimers (T)

Species length	AFM contour-length measurements (nm)			
Number of bases	CRZ-2 l-83mer	CRZ-2	PBD1	PBD4
M (83–103/105)	–	6.8 ± 0.7	6.2 ± 0.7	6.3 ± 0.7
	–	9.5 ± 0.4	9.3 ± 0.7	8.3 ± 0.9
	13.4 ± 1.9	**12.3 ± 1.1**	**13.8 ± 1.3**	**14.2 ± 1.0**
D (166–186/188)	–	17.2 ± 0.4	19.0 ± 1.7	20.0 ± 1.0
	24.5 ± 1.7	**22.7±0.6**	**24.5 ± 1.0**	**24.0 ± 0.9**
T (249–269/271)	–	–	31.7 ± 0.9	29.6 ± 0.7
	36.5 ± 3.2	–	**36.5 ± 0.7**	**34.8 ± 0.5**
Ratio (M:D:T)	2:1.5:1	7.4:1:0	10:2.8:1	4.5:1:1
Bold ratio only (M:D:T)	2:1.5:1	6:1:0	9.5:4:1	2.5:1:1

Bold values correspond to contour lengths exactly matching expected values. Given are average values ± standard deviation of the Gaussian fits in the contour-length histograms (see Supplemental Fig. S6).

Fig. S6) gives a *Z*-like conformation with a total contour length of 25.3 ± 1.8 nm, matching well the expected value for a regularly folded dimer (24.9 nm). Using the same reasoning, the peaks ~6.4 ± 0.3 and 9.0 ± 0.6 nm may be interpreted as "partially back folded" monomer and the one ~30.7 ± 1.5 nm as a "partially back folded" trimer.

To compare the AFM findings with the computational design and the results of the biochemical analysis, we calculated the number frequency of each species from the contour-length histograms. The ratio between observed monomers, dimers, and trimers are given in the last two lines of Table 3: One line calculates the ratios regarding all peaks and one line regards only those peaks in the contour-length histograms that exactly matched the expected contour lengths. However, both lines show very similar ratios, indicating that using the "backfolding hypothesis" does not affect the final conclusions of the AFM measurements.

DISCUSSION

Taken together, the results of the biochemical analysis in combination with AFM imaging confirm the predicted behavior of the self-processing RNAs CRZ-2, PBD1, PBD2, PBD3, and PBD4. All RNAs undergo two initial cleavage events that truncate the full length transcript at the 5′- and 3′-end to a linear 83mer with 5′-hydroxyl group and 2′,3′-cyclic phosphate required for ligation. The subsequent intramolecular ligation delivers exclusively cyclic versions of the 83mer, whereas intermolecular ligation produces dimers and longer concatemers, which apart from PBD4, have no or rather low tendency toward cyclization. This implies that formation of cyclic dimers is extremely unfavored, since it requires the ligation at two sites simultaneously. The same applies to longer concatemers that are rather rare anyway.

Comparing experimental results with theoretical predictions, we observed two major points, as discussed in detail below. (i) All designed species are highly efficient in circularization or ligation of cleavage products (optimized with the scoring function κ_1). Thermodynamic optimization, however, resulted in less efficient cleavage reactions compared to CRZ-2, since the cleaved ends remain tightly bound and equilibrium is shifted toward ligation. (ii) PBD1–PBD4 vary in the formation of monomers and multimers (as intended by scoring function κ_2), but interestingly, molecule optimization for stable dimers reduced the variety of dimer species and did not lead to a higher yield of trimers. Figure 7 shows a detailed analysis for each ribozyme and will serve as a guideline to discuss observed results from PAA gel electrophoresis and AFM. During the cleavage cascade, we can distinguish three types of reaction steps: (i) formation of reactive structures, (ii) dissociation of cleaved ends after ribozyme reaction, and (iii) refolding of an unbound reaction product into a new reactive structure. In Figure 7, each of these steps is characterized by an activation free energy (see Supplemental Material for details).

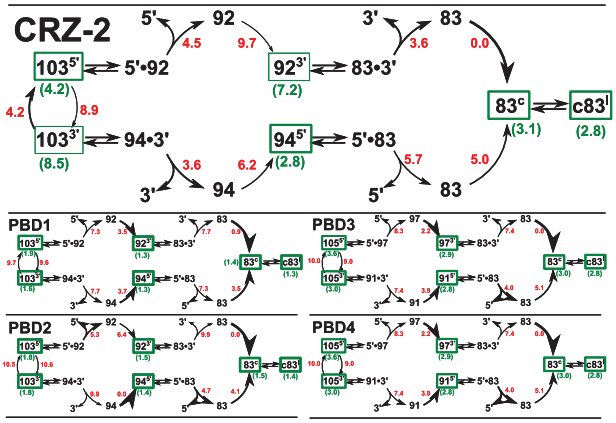


FIGURE 7. Cleavage cascades of molecules CRZ-2 and PBD1-4. Black numbers correspond to the length of the molecules or to the fragment to be cleaved (5′- and 3′-end). Superscripts 5′, 3′, c, or l mark molecules in a reactive conformation to cleave the 5′-end, the 3′-end, to circularize, or to linearize, respectively. Reversible cleavage reactions are indicated by double arrows, bent arcs denote refolding steps that are considered as irreversible. The line width of arcs is proportional to the refolding rate; red numbers denote the corresponding energy barriers (limiting the refolding rate). The line width of green boxes is proportional to the equilibrium probability of a reactive conformation; green numbers in parentheses denote the corresponding difference between the free energy of the reactive structure and the ensemble free energy.

Full-length transcripts (*a*—103/105mers)

Our theoretical analysis shows that dissociation of the cleaved ends from computationally optimized ribozymes (PBD1–4) has to overcome a higher energy barrier than in the case of manually designed CRZ-2 (Fig. 7). This is due to the fact that designed molecules are optimized to fold primarily into catalytically active conformations, and therefore also the cleaved conformations with tightly bound ends are very stable. It is known that tightly bound fragments shift equilibrium toward ligation (Fedor 1999), and this would explain why we see full-length product for three of our four designed ribozyme species (PBD2, PBD3, and PBD4), but not for CRZ-2. Here, the 5′-end is efficiently cleaved off, and the resulting transient fragments tend to accumulate as 92mers (Fig. 2, lane 1). The full-length transcript PBD1 (103mer) is completely consumed despite its high dissociation barriers, indicating that in agreement with theoretical analysis, the dissociation of cleaved ends is an irreversible step at experimental conditions.

Cleavage products (*b*—97, 94, 92, 91mer, and *c*—linear 83mer)

The cleavage cascade can start with either of two reactive conformations resulting in cleavage of the 5′- or 3′-end. In the case of CRZ-2, none of these conformations correspond to the ground state of the molecule, rather they are 4.2 and 8.5 kcal/mol above the ensemble free energy. Cleavage of the 5′-end is the favored reaction, but results in a structure that has to overcome a high barrier to fold into the reactive

92mer conformation for 3′-end cleavage. In equilibrium, the reactive structure is sparsely populated, since it is 7.2 kcal/mol above the ensemble free energy. We therefore expect that the prominent band ***b*** in the CRZ-2 lane in Figure 2 is mostly 92mer, since the 94mer is (i) the less favored cleavage product and (ii) more likely to undergo the following-up cleavage reaction. PBD3 and PBD4 enable a clear separation of cleavage products on the gel picture, due to their differently sized ends. Both molecules differ by only two point mutations in one hairpin loop of the reactive conformations (Fig. 1). Since this hairpin remains closed in all reactive species as well as on the most favorable refolding paths between the species, all important free energies differ by a constant factor (1.4 kcal/mol), and the barrier heights and structure ensemble probabilities are the same. The distribution of monomeric species should therefore be exactly the same for PBD3 and PBD4. Both molecules favor to cleave first the 3′-end, and second the 5′-end. This is in accordance with experimental results, showing mainly the 91mer as favored intermediate product. The 97mer is observed, particularly for PBD4, when higher amounts of RNA are subjected onto the gel (Fig. 2; Supplemental Fig. S3d).

Higher noncyclic ligation products (e—dimers, trimers, concatemers)

All hairpin-ribozyme variants can form two long helices, both of which have the possibility to form stable dimers that preserve the feature of catalytic activity. PBD1 and PBD3 have self-complementary hairpin loops, which results in a generally stronger tendency to dimerize, and a higher probability to retain catalytic activity upon dimerization. Since PBD1 is extremely efficient during the monomeric cleavage cascade, it shows only minor amounts of intermediate cleavage products (Fig. 2; Supplemental Fig. S3a) that could form dimers. Accordingly, the only dimer species we see is the 166mer (***e*** in lane 3). CRZ-2 and PBD2 show the greatest variety of concatemeric species. In the case of CRZ-2, we see stable 92-/94- and 83mer cleavage products (Fig. 2, bands ***b*** and ***c*** in lane 1). The probabilities to ligate their reactive ends are higher than the probabilities to cleave off the remaining terminal sequence patches, which corresponds to the fact that we observe a variety of concatemeric species. PBD2 shows only the 92-/94mer band ***b*** and no band for the 83mer ***c*** (Fig. 2, lane 4), but it has the highest probabilities (after PBD1) to ligate intermediate cleavage products, and, in contrast to PBD1, a low probability to cleave reactive ends upon dimerization.

PBD3 and PBD4 differ only in the self-complementarity of one hairpin loop and thus should be the best systems to study the influence of such mutations. Resulting from perfect self-complementarity, PBD3 has both higher probabilities to ligate intermediate products and higher probabilities to cleave ends from dimer species. However, the only detectable cleavage intermediate on gel pictures for PBD3 is the 91mer, which cannot ligate with itself to a higher species. Accordingly, we see exclusively the linear 166mer (dimer). For PBD4 also the 97mer is detectable (Fig. 2; Supplemental Fig. S3d). Interestingly, PBD4 shows more multimeric species, suggesting that design toward stable dimers (PBD3) leads to a lower diversity of multimers.

AFM visualization of RNA molecules

RNA species are identified using differences in their contour length, which however, can cause ambiguities if species differ only by a few nanometers. Hence, it was not possible to distinguish linear and cyclic species (same contour length), different monomeric cleavage products from the full length transcript (contour lengths differ by <2 nm), or to distinguish, which type of dimer, trimer, etc. is observed in the AFM image. However, AFM resolved structural features (helices, loop regions) and observed segment and contour lengths that match with secondary structure prediction for monomers, dimers, and trimers. In the case of the linear 83mer of CRZ-2 (which can only form multiples of itself) the typical pitch of 0.30 ± 0.02 nm per base pair in a helix (Arnott et al. 1973; Henn et al. 2001) matches exactly our observed segment and contour lengths. Contour lengths that do not match the values of regularly folded monomers, dimers, and trimers can be explained by spatial proximity of two adjacent segments, such that two segments appear as one. Supporting this hypothesis, adding a single segment from the segment length histogram to truncated RNA species would lead to expected contour lengths.

We furthermore observed that, although the samples were dried before imaging, the RNA chains kept most of the initial helical conformation. This was also observed in earlier studies (Bonin et al. 2000; Henn et al. 2001; Abels et al. 2005), and indicates that the RNA chain structure is sufficiently conserved to yield meaningful results using AFM imaging in air. Tip convolution, which may lead to a systematic overestimation of the contour lengths, introduces only minor disturbances. Using typical experimental parameters (tip radius <5 nm, RNA chain height <0.4 nm), tip convolution increases the lateral chain extension by <3 nm (measured as full-width half maximum/FWHM) (Ortinau et al. 2010), which is <20% of the monomeric contour length, <10% of the dimeric one, etc. However, the good quantitative agreement suggests that tip convolution effects (i.e., the effective tip radius) are smaller than expected for tip radii extracted from calibration measurements as described in Materials and Methods. Taken together, the results of AFM measurements confirmed and complemented the conclusions drawn from the gel electrophoretic analysis. While smaller fragments are dominating for CRZ-2, a tendency toward larger constructs is seen for PBD1, and for PBD4 a majority of rather complex structures is detected (Fig. 6). Comparing the outcome of the AFM analysis for PBD4 with the gel shown in Figure 2, these complex structures are either cyclic species of varying lengths or folded concatemers, since they most likely correspond to the smear of

bands in the area of ***d*** (Fig. 2, lane 4, cf. also Fig. 3D; Supplemental Fig. S4d), and thus show an anomalous migration behavior as typically observed for cyclic or folded RNA species (Grabowski et al. 1984; Sigurdsson and Eckstein 1996).

We obtain clear results from contour-length measurements counting *relative* amounts of monomers, dimers, and trimers (Table 3). Regardless of whether we compare ratios of theoretically expected peaks only, or include the peaks corresponding to partly unresolved molecules, we observe more dimeric and trimeric species for our newly designed species PBD1 and PBD4, which is in agreement with our design objective. Furthermore, PBD1 tends to form dimer species, again in agreement with our design goal, while PBD4, which was theoretically optimized to form cyclic monomers, shows the highest multimer variety both on PAA gel electrophoresis and AFM imaging.

To summarize, imaging ribozymes on the single-molecule level using AFM provides information that complements results obtained from the gel electrophoretic analysis and the computation analysis (and vice versa), making a combination of these techniques promising and very powerful. Our study revealed that (i) self-processing activity can be programmed into RNA structures, (ii) self-processing activity can be predicted and optimized by computer-aided design, and (iii) AFM turned out to be a powerful technique to image the reaction products at the single molecule level, even for short RNAs (<100mer). Dynamic processes like self-induced topology changes and oligomerization and their sensitivity upon sequence variation are essential for biological self-organization and evolution. Moreover, a large number of publications over the past two years have shown that biological processing of RNA into circular species with often still unknown function is widespread in nature. Thus, nowadays the emergence of circular RNAs and their cellular functionalities are actively investigated (Hansen et al. 2011; Abe et al. 2012; Danan et al. 2012; Jeck et al. 2013; Liu et al. 2013), making the development of in vitro techniques for RNA circularization and the study of models mimicking the processing into circular species even more important.

MATERIALS AND METHODS

Computational ribozyme design

To have a consistent, length-independent annotation for all possible RNA species that can emerge from a starting (full-length) ribozyme, we introduce the following notation: We denote the 5′- and 3′-ends of the full-length molecule as *L* (left) and *R* (right), respectively, and the linear core as *C* (center). An initial ribozyme species therefore consists of three parts and can be abbreviated as "*LCR*" molecule. Additionally we introduce the term *O* for the circular version of *C*. Despite the ability of *C* to form a circular *O*, multiple copies of *C* can ligate to one long strand that itself can form a maxi-cycle (e.g., $CCC \leftrightarrow C_3 \leftrightarrow O_3$).

The following two scoring functions (κ_1 and κ_2) were used to select for ribozymes which are able to process themselves efficiently into cyclic monomers (κ_1) and to differentiate between those, which predominantly form catalytically active or inactive dimers (κ_2). Both functions estimate rates for cleavage reactions by computing the probabilities of catalytic secondary structures, hence following two hypotheses: First, a cleavage/ligation rate is proportional to the equilibrium probability of a catalytically active secondary structure; second, the cleavage reaction leads to dissociation of the shorter fragment and is therefore irreversible. Equilibrium probabilities of RNA molecules can be calculated from the equilibrium partition function (Z); Z can be calculated using the McCaskill algorithm (McCaskill 1990) implemented in RNAfold of the Vienna RNA package (Lorenz et al. 2011). Let $Z(LCR)$ be the equilibrium partition function over all feasible secondary structures compatible with the molecule *LCR,* and $Z(LCR^L)$ be the constraint partition function over all reactive secondary structures in which *L* can be cleaved off, then the probability $P(LCR^L)$ can be computed as

$$P(LCR^L) = \frac{Z(LCR^L)}{Z(LCR)} \tag{1}$$

All computations were done using the Vienna RNA package Version 2.1.6 with standard energy parameters at 37°C. Supplemental Figure S2a shows our model of the cleavage cascade yielding cyclic monomers. It starts with a full-length molecule (*LCR*) that can process itself into the linear catalytic core in two parallel ways. Either the 5′-end (*L*) of the sequence is cleaved first and the resulting truncated version (*CR*) cleaves the 3′-end (*R*), or vice versa. For both of these parallel, two-step reaction pathways we are interested in the rate limiting step which determines the speed of the cascade. Since we approximate cleavage rates from probabilities of catalytic secondary structures, the rate limiting cleavage reaction is the minimum of both probabilities, and the total rate is the sum of both parallel cleavage pathways. The yield of circular reaction products is computed as the fraction of circular molecules in equilibrium ($Z(O)/(Z(O)+Z(C))$) resulting in the following scoring function:

$$\kappa_1 = -\ln\left(\left(\min\begin{Bmatrix} P(LCR^L) \\ P(CR^R) \end{Bmatrix} + \min\begin{Bmatrix} P(LCR^R) \\ P(LC^L) \end{Bmatrix}\right) \cdot \frac{Z(O)}{Z(O)+Z(C)}\right) \tag{2}$$

Our model of the cleavage/ligation cascade which forms circular dimers is shown in Supplemental Figure S2b. It follows the assumption that dimerization between full-length molecules happens first, then an intramolecular cleavage cascade is triggered, and finally the system equilibrates between all dimeric cleavage products. While monomers have one reactive ground state with two conserved interior loops to cleave one of their ends, dimers can form up to two reactive centers in three different ways to cleave one end (see Supplemental Fig. S2b). The two interior loops needed for a reaction are commonly called loop A (harboring the reactive site) and loop B. Our computations to score the dimer-cleavage cascade require at least one of these loop regions to be formed intermolecularly, since κ_1 already scores the molecules according to their intramolecular cleavage efficiency. The probability to cleave both 5′-ends (*L*) from a *LCR* dimer $P(LCR_d^{2L})$ can therefore be computed as

$$P(LCR_d^{2L}) = \frac{Z(LCR_d^{2L})}{Z(LCR_d)}, \tag{3}$$

where $Z(LCR_d^{2L})$ is the sum of two distinct sets of structures in which loop B is either formed intramolecularly or intermolecularly.

Similar to κ_1, the yield of circular dimers is computed as the fraction of circular dimers in equilibrium $(Z(O_2)/(Z(CC) + Z(C_2) + Z(O_2)))$ with CC and C_2 denoting noncovalently and covalently bound dimers, respectively. The second scoring function κ_2 is therefore computed as

$$\kappa_2 = -\ln\left(\frac{[LCR_d]_\theta}{[LCR]_\theta}\left(\min\begin{Bmatrix} P(LCR_d^{2L}) \\ P(CR_d^{2R}) \end{Bmatrix} + \min\begin{Bmatrix} P(LCR_d^{2R}) \\ P(LC_d^{2L}) \end{Bmatrix}\right)\cdot \frac{Z(O_2)}{Z(CC) + Z(C_2) + Z(O_2)}\right), \quad (4)$$

where the first term $[LCR_d]_\theta/[LCR]_\theta$ computes the equilibrium ratio between dimers and monomers at a given concentration θ (here 100 nM) for the *LCR* molecule following Bernhart et al. (2006) The scoring function only maximizes the probabilities for catalytically active homo-dimers; pathways that involve dehybridization of partially cleaved species are not included. This corresponds to the assumption that intramolecular cleavage reactions as well as intramolecular folding kinetics are faster than intermolecular structural rearrangements.

Self-processing reactions

RNAs (11.25 pmol) were solved in Tris–HCl buffer (10 mM, pH = 7.5). After denaturation for one minute at 90°C, RNA folding was allowed for 10 min at room temperature. To initiate the cleavage reaction, $MgCl_2$ hexahydrate to a final concentration of 10 mM was added and reaction was allowed to proceed for 2 h at 37°C. To favor ligation, Mg^{2+} concentration was increased up to 50 mM, and reaction proceeded for additional 2 h at 37°C. Reaction was stopped using an equal volume of stop mix composed of urea (7 M) and EDTA (50 mM) for the following PAGE analysis.

Two-dimensional electrophoresis

Identification of circular RNAs by 2D electrophoresis is based on the fact that the migration of linear and circular nucleic acids is distinctly dependent on the gel pore size (Sigurdsson and Eckstein 1996; Umekage and Kikuchi 2009). To enrich the samples with linear RNAs for better identification of the circular species, a commercially available RNA size standard (RiboRuler low-range RNA ladder; *Fermentas*) being composed exclusively of linear RNAs was added. Each individual mixture was separated in the first dimension gel. Then, the gel piece corresponding to the entire lane was cut out and used for electrophoresis in the second dimension, upon which linear RNAs are supposed to form a diagonal. Covalently closed cyclic RNAs possess reduced degrees of freedom, thus migrating in nonlinear dependence on the linear species and occurring beyond the diagonal (Pasman et al. 1996).

All ribozyme variants (11.25 pmol) were analyzed using 2D PAGE (for polyacrylamide gel composition, buffers and staining see "PAGE analysis"). First dimension: denaturing conditions (7 M urea) 15% polyacrylamide; second dimension: 17.5% denaturing polyacrylamide or 15% native polyacrylamide.

RNA preparation for AFM analysis

Ribozyme reactions were carried out as described above using 400 nM RNA. After reaction, the product mixture was diluted 1:10, and 5 µL of this solution were lyophilized. The pellet was taken up in 50 µL of semidenaturing buffer (25 mM EDTA, 3.5 M urea) to a final RNA concentration of 4 nM for imaging. Resolved RNA samples were frozen in liquid nitrogen until use.

Atomic force microscopy (AFM)

AFM imaging was performed in air using a Multimode atomic force microscope (Veeco/Digital Instruments) equipped with a Nanoscope IIIa controller. The AFM piezo scanner was calibrated using calibration gratings TGZ01 (rectangular 26 nm SiO_2 steps on silicon wafer; MicroMasch) and PG (chessboard-like pattern on silicon, 100 nm depth and 1 µm pitch; manufacturer: Digital Instruments).

RNA samples were prepared by placing a small droplet of RNA solution onto freshly cleaved mica (SPI Supplies). For the investigated RNA constructs, adsorption times of 30 sec to 2 min were sufficient to obtain a suitable RNA surface coverage on the mica substrate. After adsorption, the RNA samples were rinsed in Milli-Q water (Millipore) and dried in a laminar flow hood, followed by AFM imaging.

The images were recorded with conventional Tapping Mode in air using standard tapping mode cantilevers (OMCL-AC160TS, Olympus). Before usage the cantilevers were tested with a Nioprobe self-imaging sample (Aurora Nanodevices) and only cantilevers with tip radius <5 nm were used for imaging. To reduce tip contamination by RNA uptake during imaging process, cantilevers were functionalized with 3-aminopropyldimethyl-ethoxysilane (APDES) from *ABCR* (Karlsruhe) one day prior usage.

In contrast to DNA samples, whose structure often remains unchanged even after storage periods of several months (as judged by their spatial properties in AFM imaging), samples had to be imaged within few days after preparation. The highest resolutions were always obtained directly after preparation, while storing in air often led to post-preparational RNA chain degradation already after few weeks.

Images were analyzed using the software supplemented with the AFM. The shape of an RNA chain was "retraced" in terms of a sequence of connected straight segments, which allowed to calculate the contour length as sum of Euclidean distances (Rivetti et al. 1996). As shown by Rivetti and Codeluppi (2001) (who numerically assessed the accuracy of different methods for contour-length determination) this approach has an intrinsic error <1%. The main source of error in the contour-length determination is therefore given by the lateral resolution of the AFM, which is on the order of few nanometers (see Discussion for details).

SUPPLEMENTAL MATERIAL

Supplemental material is available for this article.

ACKNOWLEDGMENTS

S.Bl. and M.D. acknowledge support by the German Federal Ministry of Education and Research of Germany (BMBF) within the project FKZ 03Z2CN11. I.H. and S.M. acknowledge support by FWF (Austria) and Deutsche Forschungsgemeinschaft (DFG) (Germany) within the program ERA chemistry (MU1396/11). S.Ba. was

supported in part by the FWF International Program (Austrian Science Fund) I670 and the DK RNA program FG748004.

Received August 12, 2014; accepted March 7, 2015.

REFERENCES

Abe N, Abe H, Ito Y. 2012. Synthesis of dumbbell-shaped cyclic RNAs for RNA interference. *Curr Protoc Nucleic Acid Chem* Chapter 16: Unit 16.4.1–11.

Abels JA, Moreno-Herrero F, van der Heijden T, Dekker C, Dekker NH. 2005. Single-molecule measurements of the persistence length of double-stranded RNA. *Biophys J* **88:** 2737–2744.

Arnott S, Hukins DWL, Dover SD, Fuller W, Hodgson AR. 1973. Structures of synthetic polynucleotides in A-RNA and A′-RNA conformations: X-ray-diffraction analyses of molecular conformations of polyadenylic acid–polyuridylic acid and polyinosinic acid–polycytidylic acid. *J Mol Biol* **81:** 107–122.

Balke D, Zieten I, Strahl A, Müller O, Müller S. 2014. Design and characterization of a twin ribozyme for potential repair of a deletion mutation within the oncogenic *CTNNB1*-ΔS45-mRNA. *ChemMedChem* **9:** 2128–2137.

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.

Berzal-Herranz A, Joseph S, Chowrira BM, Butcher SE, Burke JM. 1993. Essential nucleotide sequences and secondary structure elements of the hairpin ribozyme. *EMBO J* **12:** 2567–2573.

Bonin M, Oberstrass J, Lukacs N, Ewert K, Oesterschulze E, Kassing R, Nellen W. 2000. Determination of preferential binding sites for anti-dsRNA antibodies on double-stranded RNA by scanning force microscopy. *RNA* **6:** 563–570.

Butcher SE, Burke JM. 1994. A photo-cross-linkable tertiary structure motif found in functionally distinct RNA molecules is essential for catalytic function of the hairpin ribozyme. *Biochemistry* **33:** 992–999.

Cochrane JC, Strobel SA. 2008. Catalytic strategies of self-cleaving ribozymes. *Acc Chem Res* **41:** 1027–1035.

Cruz-Reyes J, Piller KJ, Rusché LN, Mukherjee M, Sollner-Webb B. 1998. Unexpected electrophoretic migration of RNA with different 3′ termini causes a RNA sizing ambiguity that can be resolved using nuclease P1-generated sequencing ladders. *Biochemistry* **37:** 6059–6064.

Danan M, Schwartz S, Edelheit S, Sorek R. 2012. Transcriptome-wide discovery of circular RNAs in Archaea. *Nucleic Acids Res* **40:** 3131–3142.

DeYoung M, Siwkowski AM, Lian Y, Hampel A. 1995. Catalytic properties of hairpin ribozymes derived from Chicory yellow mottle virus and arabis mosaic virus satellite RNAs. *Biochemistry* **34:** 15785–15791.

Drude I, Vauléon S, Müller S. 2007. Twin ribozyme mediated removal of nucleotides from an internal RNA site. *Biochem Biophys Res Commun* **363:** 24–29.

Drude I, Strahl A, Galla D, Müller O, Müller S. 2011. Design of hairpin ribozyme variants with improved activity for poorly processed substrates. *FEBS J* **278:** 622–633.

Fedor MJ. 1999. Tertiary structure stabilization promotes hairpin ribozyme ligation. *Biochemistry* **38:** 11040–11050.

Flamm C, Hofacker IL, Maurer-Stroh S, Stadler PF, Zehl M. 2001. Design of multistable RNA molecules. *RNA* **7:** 254–265.

Flores R, Grubb D, Elleuch A, Nohales MÁ, Delgado S, Gago S. 2011. Rolling-circle replication of viroids, viroid-like satellite RNAs and hepatitis delta virus: variations on a theme. *RNA Biol* **8:** 200–206.

Grabowski PJ, Padgett RA, Sharp PA. 1984. Messenger RNA splicing in vitro: an excised intervening sequence and a potential intermediate. *Cell* **37:** 415–427.

Hammann C, Luptak A, Perreault J, de la Peña M. 2012. The ubiquitous hammerhead ribozyme. *RNA* **18:** 871–885.

Hampel A, Tritz R. 1989. RNA catalytic properties of the minimum (-) sTRSV sequence. *Biochemistry* **28:** 4929–4933.

Hansen TB, Wiklund ED, Bramsen JB, Villadsen SB, Statham AL, Clark SJ, Kjems J. 2011. miRNA-dependent gene silencing involving Ago2-mediated cleavage of a circular antisense RNA. *EMBO J* **30:** 4414–4422.

Hegg LA, Fedor MJ. 1995. Kinetics and thermodynamics of intermolecular catalysis by hairpin ribozymes. *Biochemistry* **34:** 15813–15828.

Henn A, Medalia O, Shi SP, Steinberg M, Franceschi F, Sagi I. 2001. Visualization of unwinding activity of duplex RNA by DbpA, a DEAD box helicase, at single-molecule resolution by atomic force microscopy. *Proc Natl Acad Sci* **98:** 5007–5012.

Ivanov SA, Vauleon S, Müller S. 2005. Efficient RNA ligation by reverse-joined hairpin ribozymes and engineering of twin ribozymes consisting of conventional and reverse-joined hairpin ribozyme units. *FEBS J* **272:** 4464–4474.

Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, Sharpless NE. 2013. Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* **19:** 141–157.

Liu S, Bokinsky G, Walter NG, Zhuang X. 2007. Dissecting the multistep reaction pathway of an RNA enzyme by single-molecule kinetic "fingerprinting". *Proc Natl Acad Sci* **104:** 12634–12639.

Liu Y, Cui H, Wang W, Li L, Wang Z, Yang S, Zhang X. 2013. Construction of circular miRNA sponges targeting miR-21 or miR-221 and demonstration of their excellent anticancer effects on malignant melanoma cells. *Int J Biochem Cell Biol* **45:** 2643–2650.

Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL. 2011. ViennaRNA Package 2.0. *Algorithms Mol Biol* **6:** 26.

McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29:** 1105–1119.

Müller S, Appel B, Krellenberg T, Petkovic S. 2012. The many faces of the hairpin ribozyme: structural and functional variants of a small catalytic RNA. *IUBMB Life* **64:** 36–47.

Nahas MK, Wilson TJ, Hohng S, Jarvie K, Lilley DM, Ha T. 2004. Observation of internal cleavage and ligation reactions of a ribozyme. *Nat Struct Biol* **11:** 1107–1113.

Ortinau S, Schmich J, Block S, Liedmann A, Jonas L, Weiss DG, Helm CA, Rolfs A, Frech MJ. 2010. Effect of 3D-scaffold formation on differentiation and survival in human neural progenitor cells. *Biomed Eng Online* **9:** 70.

Pasman Z, Been MD, Garcia-Blanco MA. 1996. Exon circularization in mammalian nuclear extracts. *RNA* **2:** 603–610.

Petkovic S, Müller S. 2013. RNA self-processing: formation of cyclic species and concatemers from a small engineered RNA. *FEBS Lett* **587:** 2435–2440.

Pieper S, Vauleon S, Muller S. 2007. RNA self-processing towards changed topology and sequence oligomerization. *Biol Chem* **388:** 743–746.

Rivetti C, Codeluppi S. 2001. Accurate length determination of DNA molecules visualized by atomic force microscopy: evidence for a partial B- to A-form transition on mica. *Ultramicroscopy* **87:** 55–66.

Rivetti C, Guthold M, Bustamante C. 1996. Scanning force microscopy of DNA deposited onto mica: equilibration versus kinetic trapping studied by statistical polymer chain analysis. *J Mol Biol* **264:** 919–932.

Sigurdsson ST, Eckstein F. 1996. Isolation of oligoribonucleotides containing intramolecular cross-links. *Anal Biochem* **235:** 241–242.

Umekage S, Kikuchi Y. 2009. *In vitro* and *in vivo* production and purification of circular RNA aptamer. *J Biotechnol* **139:** 265–272.

Vauleon S, Ivanov SA, Gwiazda S, Muller S. 2005. Site-specific fluorescent and affinity labelling of RNA by using a small engineered twin ribozyme. *ChemBiochem* **6:** 2158–2162.

Welz R, Bossmann K, Klug C, Schmidt C, Fritz HJ, Muller S. 2003. Site-directed alteration of RNA sequence mediated by an engineered twin ribozyme. *Angew Chem Int Ed Engl* **42:** 2424–2427.

7

DESIGN OF A CIRCULAR RNA WITH PRIONLIKE BEHAVIOR

ARTICLE

Stefan Badelt, Christoph Flamm, and Ivo L. Hofacker.
Computational design of a circular RNA with Prionlike behavior.
in press: *Artificial Life* (2015), **Volume** 22, **pages** 1–14
doi: 10.1162/ARTL_a_00197

AUTHOR CONTRIBUTIONS

Christoph Flamm, Ivo L Hofacker and Stefan Badelt have developed the theory for designing a prionlike RNA molecule. Stefan Badelt has established the method, designed the sequences, and written substantial parts of the manuscript.

LICENCE

Computational Design of a Circular RNA with Prionlike Behavior

Stefan Badelt**
Christoph Flamm**,†
Ivo L. Hofacker*,**,†,‡
University of Vienna

Keywords
RNA structure, sequence design, self-replication, folding kinetics

Abstract RNA molecules engineered to fold into predefined conformations have enabled the design of a multitude of functional RNA devices in the field of synthetic biology and nanotechnology. More complex designs require efficient computational methods, which need to consider not only equilibrium thermodynamics but also the kinetics of structure formation. Here we present a novel type of RNA design that mimics the behavior of prions, that is, sequences capable of interaction-triggered autocatalytic replication of conformations. Our design was computed with the ViennaRNA package and is based on circular RNA that embeds domains amenable to intermolecular kissing interactions.

1 Introduction

During the last decade, the field of synthetic biology has impressively illustrated that nucleic acids and in particular RNA molecules are reliable materials for the design and implementation of functional circuits as well as nano-scale devices and objects [18, 24, 1, 23]. The reasons for this success are grounded in the facts that for RNA (i) an experimentally measured energy model exists, (ii) regulation at the level of RNA molecules is faster than via the production of proteins, and (iii) design questions are more readily expressed in the discrete framework of binary base pairing than in continuous interactions between, say, the amino acids in proteins.

1.1 RNA Design

RNA molecules have been extensively engineered in the classical context of gene regulation. Successful designs include Boolean networks with miRNAs [27, 34, 47], synthetic RNA switches [22, 40, 17], and artificial ribozymes [43, 13]. In general, the RNA is optimized to switch conformations upon a hybridization interaction, in order to toggle between an ON and an OFF state. Turberfield et al. [38] showed for DNA that such hybridization reactions are reversible by a mechanism called *toehold exchange*, which was later used to design reversible logic circuits [14] as well as transcriptional oscillators [25] and transcription regulators [37]. Recently, multiple toehold switches have been integrated

* Contact author.
** Department of Theoretical Chemistry, University of Vienna, Währingerstraße 17/3, 1090 Vienna, Austria. E-mail: stef@tbi.univie.ac.at (S.B.); xtof@tbi.univie.ac.at (C.F.); ivo@tbi.univie.ac.at (I.L.H.)
† Forschungsverbund Chemistry meets Microbiology, University of Vienna, Althanstraße 14, 1090 Vienna, Austria.
‡ Research Group Bioinformatics and Computational Biology, University of Vienna, Währingerstraße 29, 1090 Vienna, Austria.

in vivo to activate gene expression in response to endogenous RNAs [17]. Cascading of strand-displacement reactions allows for the construction of multi-component chemical reaction networks, exhibiting complex computational and information-processing abilities [50, 41].

DNA and RNA hybridization effects are essential features for algorithmic self-assembly [36, 48] in nanotechnology. RNA nanoparticles have been constructed from smaller self-assembling units [6, 5, 16]. Small RNA motifs, such as complementary kissing hairpins, can be used to assemble complex 1D, 2D, and 3D shapes (for a recent review see [15]). Cayrol et al. [5] have found natural self-assemblies of the small RNA DsrA in *E. coli*, suggesting concentration-dependent RNA regulatory mechanisms via self-assembly.

The inherent complexity of nucleic-acid-based self-assembling systems makes it necessary to optimize hand-crafted designs computationally. Several energy-directed computational methods have been devised for the rational design of nucleic acid molecules that fold into single [20] or multiple [11, 21] predefined conformations or that form ensembles of interacting nucleic acid strands [49, 44]. While these methods carefully model equilibrium properties of the designed RNAs, they provide rudimentary or no support for designing kinetic properties, such as refolding times, which are much more expensive to compute and thus remain a challenge for computational design methods.

1.2 Prions

The *protein-only hypothesis* for the scrapie agent (for a review see [2]) proposes that a prion protein, with an altered (infectious) β-sheet-rich conformation, starts an autocatalytic cascade that uses the normally folded prion proteins as a substrate, converting them to the infectious form. This altered conformation then either self-assembles into fibers, which is the usual phenotype upon prion infection, or catalyzes the refolding of the remaining normally folded prions. A high activation energy between the normal and the infectious conformation prevents spontaneous conversion at detectable rates. The formation of a normal–infectious heteromeric complex lowers the activation energy barrier to convert the normally folded protein into a infectious species. This conversion leads to further recruitment of normally folded proteins in an autocatalytic process. In essence, a single infectious prion protein in a population of normally folded ones is enough to convert the whole population via autocatalytic structure replication into an all-infectious protein population, which self-assembles into long fibers.

1.3 Artificial Life

Prions represent a form of *conformational self-replication* that is so far not observed in the context of RNA biology. Minimal self-switching RNA prions can serve as a model for (i) a new class of riboswitches that provide exponential feedback, or (ii) self-induced nano-units that assemble or disassemble upon stimulation. Importantly, the computational design aspect allows for context-sensitive optimization, which is necessary for applications in synthetic biology and nanotechnology. Previous research in the field of bottom-up synthetic biology has already introduced designs of small self-assembling (for a recent review see [1]), self-replicating [35], and self-polymerizing [33] systems. While all of these results are valuable in showing that such designs are indeed possible, many designs are still done by intuition and hardly adjustable to a context-sensitive implementation such as prospective artificial life forms. Along those lines we recently submitted an experimentally verified computational design of multiple self-polymerizing ribozymes in order to study the dynamics of self-interactive systems [32].

This contribution was motivated by the question of whether RNA molecules can be designed in silico to exhibit the aforementioned prionlike behavior. We show that it is indeed possible to design such an *RNA prion*; whether the suggested sequence really shows the exponential refolding characteristics awaits experimental verification. The RNA prion presented here is a 49-nt-long, circular

RNA, designed as a bistable molecule. It thermodynamically favors one structure (S_1) if present as a monomer and the other structure (S_2) if present as a dimer.

2 Thermodynamics and Kinetics of RNA Prions

RNA secondary structure is a good approximation for the real RNA structure because, in contrast to proteins, RNA secondary structure captures the majority of the folding free energy. Furthermore, a well-established energy model for RNA secondary structures exists that is extensively parameterized via melting experiments [39]. On the computational side, algorithms have been developed for the thermodynamic and kinetic characterization of RNA secondary structures [28]. In particular, the energetically optimal structure as well as properties of the structural ensemble at thermodynamic equilibrium can be computed efficiently. The topology of the discrete folding landscape [12], which has a strong influence on the folding kinetics, can be analyzed in detail. Several approaches model the folding kinetics of RNA secondary structure as a stochastic process with different resolution of the folding landscape [9, 45, 26] (for a review see [10]). Recently these approaches have been extended to operate on folding landscapes that change with time, as in the case of folding during transcription [19].

2.1 Thermodynamics

Let the set Ω of RNA secondary structures be restricted to those that are formed from nested isosteric base pairs (GC, AU, GU), which have a minimal hairpin loop size of 3 nt and a maximum interior loop size of 30 nt. These restrictions are broad enough to include the vast majority of known pseudoknot-free RNA conformations, and they define a set of structures for which an experimentally determined energy model E exists [29]. Most importantly from the computational perspective, these definitions allow us to compute the structure of minimum free energy (MFE) and the equilibrium partition function (Z) in $O(n^3)$ time, where n is the sequence length.

For RNA design, a fast computation of the equilibrium partition function Z is of particular interest, since it allows for computing the probability $P(S)$ of forming a secondary structure S and the ensemble free energy G. Let Z be the sum of all Boltzmann-weighted energy contributions in the ensemble of RNA secondary structures Ω,

$$Z = \sum_{S \in \Omega} e^{-\frac{E(S)}{RT}} \tag{1}$$

Then we can compute the probability of any secondary structure as

$$P(S) = \frac{e^{-\frac{E(S)}{RT}}}{Z} \tag{2}$$

and the free energy of the ensemble as

$$G = -kT \cdot \ln(Z) \tag{3}$$

2.2 Kinetic Folding

Let us denote an RNA energy landscape as $\mathscr{L} = (\Omega, \mathscr{M}, E)$, where Ω is the previously introduced set of secondary structures, $\mathscr{M}$ is a move set comprising all possible transitions between structures, and E is an energy function assigning a fitness value for each secondary structure. For an ergodic move set $\mathscr{M}$, we choose the simplest reversible modification of an RNA structure, the opening and closing of a single base pair. If we have the full landscape $\mathscr{L}$, we can calculate folding kinetics as a

continuous-time Markov process where the influx and outflux rates, k_{ij} and k_{ji} respectively, between every two neighboring structures S_i and S_j determine the population $\pi_i(t)$ of a structure S_i at time point t. This transition process can be formulated with the master equation

$$\frac{d\pi_i(t)}{dt} = \sum_{j \neq i} \left(\pi_j(t)k_{ji} - \pi_i(t)k_{ij}\right) \tag{4}$$

An RNA energy landscape as defined above grows exponentially with sequence length, making exact folding simulations infeasible for molecules longer than 30 to 40 nucleotides. However, refolding times between two conformations correlate with the smallest energy barrier β separating them. Formally, any folding path $\mathcal{P}$ has a saddle point with energy $E_{\mathcal{P}} = \max_{s \in \mathcal{P}} E(s)$, and the energy barrier separating two minima S_1 from S_2 is determined by the lowest among all possible paths $\mathcal{P}_{S_1 \to S_2}$:

$$\beta_{S_1 \to S_2} = \min_{\mathcal{P}_{S_1 \to S_2}} E_{\mathcal{P}} - E(S_1) = \min_{\mathcal{P}} \max_{s \in \mathcal{P}} E(s) - E(S_1) \tag{5}$$

Finding the best folding path has been shown to be a NP-hard problem [30]. However, there exist fast heuristics to compute direct (shortest) paths between two structures, such as the `findpath` [11] method available in the ViennaRNA package, or `RNAtabupath` [7]. For sequences shorter than about 100 nt, paths with a minimal energy barrier can be computed exactly with `RNAsubopt` [46] and `barriers` [12]. The first program computes a list of all RNA secondary structures within a certain energy range; `barriers` processes a sorted list of these conformations to compute all local minima and the saddle points connecting them by a flooding algorithm.

2.3 Landscapes of RNA Prions

RNA prions are bistable molecules that favor one structure (S_1) if present as a monomer and the other structure (S_2) upon dimerization. To avoid spontaneous refolding between S_2 and S_1, the energy barrier $\beta_{S_2 \to S_1}$ has to be very high (see Figure 1).

In our design, S_2 forms two 10-nt hairpins that are prone to hybridize via a kissing interaction as reported for the HIV-DIS loop [8], a highly conserved stem–loop sequence found in many retroviruses. Importantly, S_2 should not only stabilize other molecules in S_2 conformation at high concentrations, but actively lower the energy barrier to refold S_1 into S_2 (see Figure 1).

If these landscape properties are fulfilled, it ensures that an initial population of *only* S_1 molecules will not refold into S_2 unless the refolding is triggered by an external mechanism. However, as soon as a small population of molecules in S_2 conformation is present, they can catalyze the refolding of S_1 molecules into S_2.

3 Design of an RNA Prion

RNA molecules with complex energy landscapes can be designed in a two-step process. First, a fast heuristic is used to generate and select promising candidates according to a cost function $\mathcal{C}$ that specifies thermodynamic aspects of the design objective. Second, minimal refolding barriers are computed for the set of candidate molecules and contribute to the final ranking of molecules (see Figure 2 for the prion design pipeline).

3.1 Cost Function

The cost function used for RNA prion design consists of two parts: $\mathcal{C} = \mathcal{C}_{\mathcal{M}} + \mathcal{C}_{\mathcal{D}}$. Here $\mathcal{C}_{\mathcal{M}}$ is the cost function to optimize bistable molecules (see Figure 1) without considering the refolding barrier:

$$\mathcal{C}_{\mathcal{M}} = E(S_1) - G + \alpha((E(S_1) - G_1) + (E(S_2) - G_2)) + \alpha(E(S_1) - E(S_2) + \epsilon)^2 \tag{6}$$

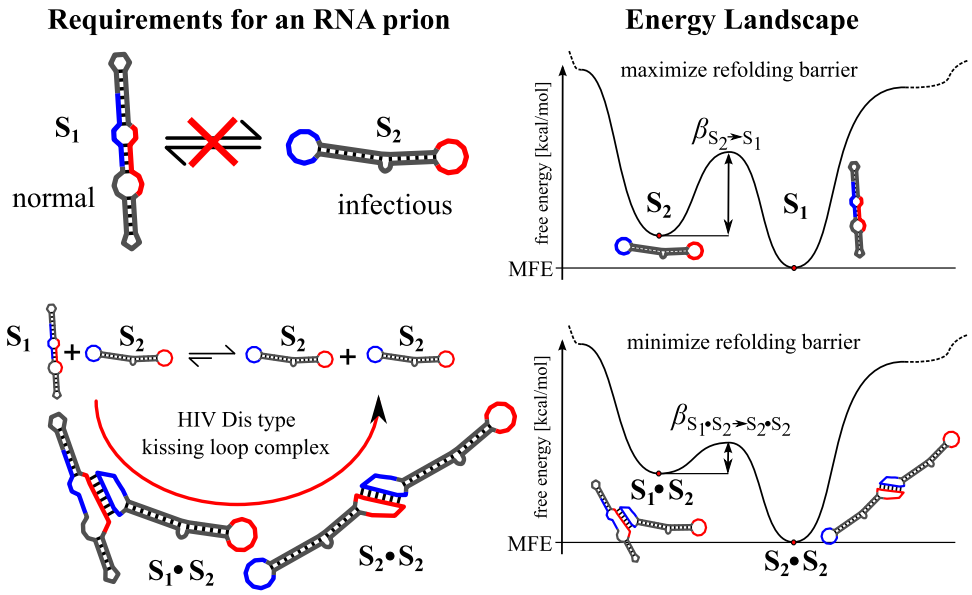


Figure 1. Schematic requirements for an RNA prion. Red and blue parts of the molecule are complementary and can form an intermolecular hybridization reaction. *Upper panel*: S_1 and S_2 are stable conformations that do not refold into each other, since the conformations are separated by a high energy barrier in the energy landscape. *Lower panel*: S_2 destabilizes S_1 with a HIV-Dis type kissing loop interaction. The energy barrier for the $S_1 \bullet S_2$ complex to refold into the $S_2 \bullet S_2$ complex is low and therefore allows spontaneous refolding.

where G is the ensemble free energy (see Equation 3), G_1 and G_2 are constrained ensemble free energies over all structures that form base pairs exclusively possible for S_1 and S_2 respectively (see Figure 3), ϵ specifies the desired energy difference between S_1 and S_2, and α is a constant to weight the terms. Thus, the first term ensures that S_1 is the ground state, while the next two terms optimze for sequences that have few alternative structures in the vicinity of S_1 and S_2.

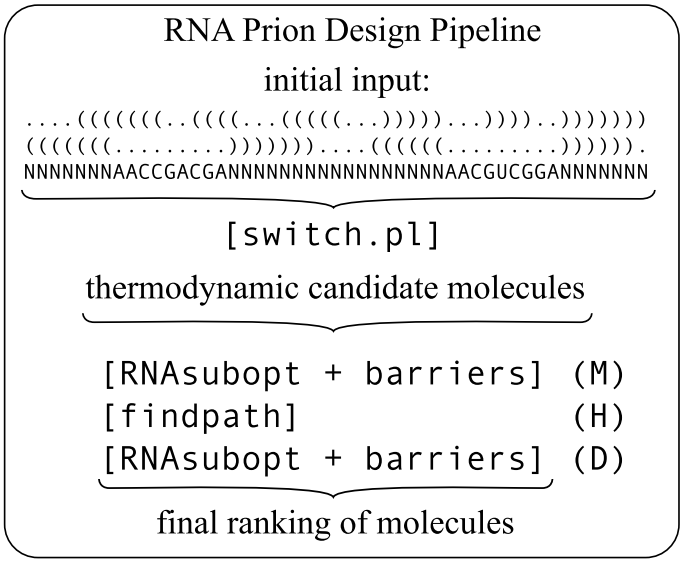


Figure 2. Programs used to design RNA prions. The initial input consists of two secondary structures in dot–bracket notation (i.e., every unpaired nucleotide is represented by a dot, and base pairs are shown as matching parentheses; see Figure 6 for the same structures shown in a graphical representation) and a sequence constraint for the kissing interaction. `switch.pl` returns a set of candidate molecules, which are analyzed with `RNAsubopt` and `barriers` to compute the minimum barrier height for monomer refolding (M) and dimer refolding (D). `findpath` is used to approximate the initial hybridization interaction of two monomers in structures S_1 and S_2 (H).

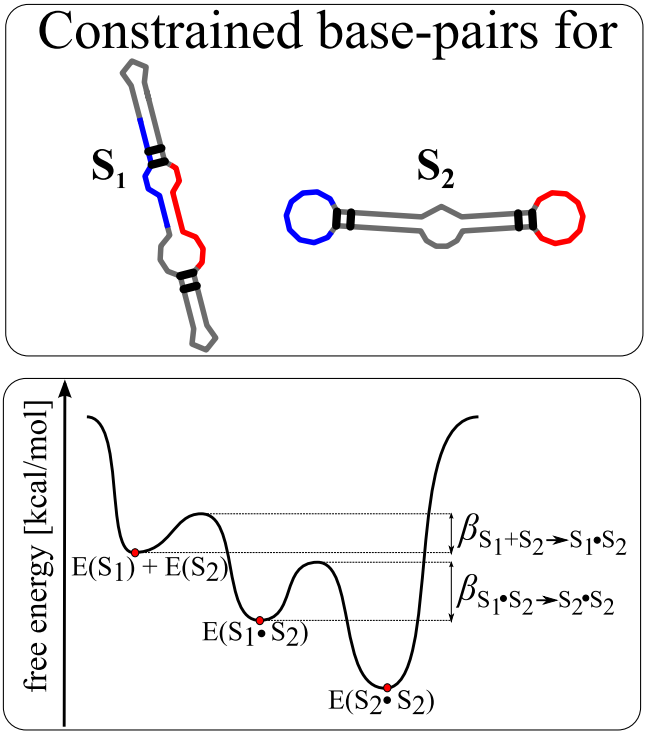


Figure 3. *Upper box*: Constraints for the monomer landscape. Among all structures that can form, the base pairs shown (S_1 or S_2) should be the best. The base pairs are chosen such that no structures can fulfill both constraints. *Lower box*: Optimization parameters for the dimer landscape. The energy of the hybridization complex $S_1 \bullet S_2$ should be half way between the sum of the monomer energies and the kissing complex $S_2 \bullet S_2$. The barrier heights β determine the ranking of final candidates.

$\mathcal{C}_\mathcal{D}$ optimizes the dimer landscape, so that the energy of the transition state formed by the initial intermolecular hybridization interaction ($S_1 \bullet S_2$) lies approximately half way between the energy of single molecules $E(S_1) + E(S_2)$ and the ground state of the kissing dimer interaction ($S_2 \bullet S_2$):

$$\mathcal{C}_\mathcal{D} = \left(\frac{E(S_2 \cdot S_2) + (E(S_1) + E(S_2))}{2} - E(S_1 \cdot S_2) \right)^2 \tag{7}$$

Finally, we rank the molecules by the difference of refolding barriers $\Delta\beta$ for monomers and dimers. Since refolding of dimers is a two-step process composed of (i) the initiation of the kissing action and (ii) the subsequent intramolecular refolding (see Figure 3), we only consider the rate-limiting step, that is, the one with the larger of the two barriers:

$$\Delta\beta = \beta_{S_2 \to S_1} - \max\left(\beta_{S_1+S_2 \to S_1 \cdot S_2}, \beta_{S_1 \cdot S_2 \to S_2 \cdot S_2}\right) \tag{8}$$

3.2 Energy Evaluation

RNA kissing interactions go beyond normal pseudoknot-free secondary structures as defined above, since they comprise non-nested base pairs, and their energies can therefore not be computed from the standard energy model. While we expect the energies of the intermolecular helix to be well described by standard energy parameters, it is less predictable how the (mostly entropic) contribution of the hairpin loops involved in the kiss will change. Thermodynamic stabilities of kissing interactions similar to the HIV–DIS loop, which features a 6-nt intermolecular interaction with a 9-nt hairpin loop, were studied in detail by Weixlbaumer et al. [42]. By varying the loop sequence

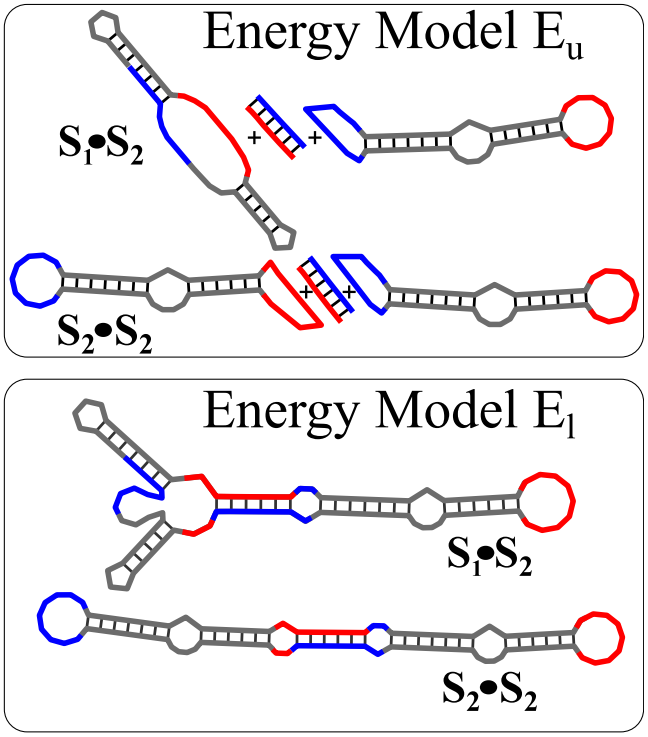


Figure 4. Two energy models to score the hybridization reactions also depicted in Figure 1. E_u is the sum of energies for S_1, S_2, and the hybridization reaction; E_l transforms the molecule into a hybrid that can be scored with standard energy parameters.

they demonstrated that normal stacking energies can indeed be used for the intermolecular helix. More importantly, they found kissing interactions to be surprisingly more stable than other hybridization structures. Based on the measurements in [42], we compute the energy of kissing hairpins as the energy of the intermolecular helix without energy bonuses for single-base stacking of adjacent unpaired bases (*dangling ends*) and add a loop energy of −4.2 kcal/mol. In contrast, other intermolecular hybridization structures are penalized by the usual intermolecular initiation energy of +4.1 kcal/mol, also confirmed by Weixlbaumer et al. [42].

These energy parameters enable us to accurately evaluate the energy of $S_2 \bullet S_2$, but, unfortunately, they are not sufficient for creating a consistent energy model for all intermediate conformations along the refolding paths $P_{S_1+S_2 \to S_1 \cdot S_2}$ and $P_{S_1 \cdot S_2 \to S_2 \cdot S_2}$. We therefore approximate the energy barrier $\beta_{S_1+S_2 \to S_1 \cdot S_2}$ with the `findpath` heuristic and compute the best energy barrier $\beta_{S_1 \cdot S_2 \to S_2 \cdot S_2}$, using the `RNAsubopt–barriers` approach with two different energy models that serve as an upper and a lower bound (see Figure 4).

The upper bound dimer energy model E_u to find the lowest energy barrier $\beta_{S_1 \cdot S_2 \to S_2 \cdot S_2}$ is computed from the energy of the structure formed from monomers 1 and 2—$E(M_1)$ and $E(M_2)$, respectively—and the energy of the duplex interaction is stabilized by a bonus of −4.2 kcal/mol:

$$E_u(D) = E(M_1) + E(M_2) + E(\mathrm{dup}_{12}) - 4.2 \tag{9}$$

This approach implicitly adds a penalty for *unpaired* loop regions, which are actually involved in the hybridization interaction, and therefore underestimates the actual energy of $S_2 \bullet S_2$.

Alternatively, the kissing interaction can be modeled as a regular intramolecular helix. If both molecules are cut after the 5′-AA of the interacting loops and connected to the beginning of the respective other strand, this results in a circular hybrid monomer that can be evaluated by standard energy parameters. Since the free-energy bonus for kissing interactions (−4.2 kcal/mol) approximately compensates for the entropic penalty of intermolecular interactions (+4.1 kcal/mol), $S_2 \bullet S_2$ is closer to the energy according to Weixlbaumer et al. [42]. Using this energy model, we increment the degree of loops involved in the kissing interaction as if a regular helix were formed, and thereby increase the entropic loop penalty. The interior loop of S_1 becomes a multi-loop, and the hairpin loop of S_2 becomes an interior loop. We have

$$E_I(D) = E\left(\text{Hybrid}_{M1\cdot\text{M2}}\right) \tag{10}$$

All optimization steps are based on the energy model E_u, for analysis of the best sequences both energy models were applied, and we await experimental feedback to decide which model is better.

The barrier for the initiation of the intermolecular hybridization $\beta_{S_1+S_2\rightarrow S_1\cdot S_2}$ is computed by first finding the best path for opening the competing helix in S_1 and second computing the energy barrier for the intermolecular hybridization reaction using the standard penalty of +4.1 kcal/mol. The bonus energy of −4.2 kcal/mol is added once the full kissing region is formed, to be consistent with the energy model E_u. The energy barrier for the initiation of the duplex formation is therefore usually either the last base pair in the process of unfolding the competing helix, or the first inserted base pair towards the kissing interaction.

3.3 `switch.pl`

`switch.pl` [11] of the ViennaRNA package [28] was used to design bistable molecules and was modified to support the novel cost function composed of Equations 6 and 7 as well as the folding of circular RNAs. The algorithm first builds a dependence graph in order to efficiently and fairly sample RNA sequences that are compatible with both structural constraints. Since `switch.pl` can only design bistable and not tristable sequences, the structural constraint for the kissing interaction was specified indirectly as a sequence constraint. This reduces the number of candidate molecules, but ensures that they always have a experimentally validated, stable kissing interaction.

The chosen sequences for the kissing interaction have a similar free hybridization energy to that of the best kissing interaction examples shown in Weixlbaumer et al. [42], but differ by point mutations to be compatible with structural constraints for S_1 and S_2. Importantly, (i) the kissing hairpins cannot form intramolecular base pairs that would compete with the formation of an intermolecular kissing interaction, and (ii) the two complementary regions forming the kissing interaction should not form an intramolecular helix in S_1, which would make them inaccessible for intramolecular interactions. The asymmetric design shown in Figure 1 allows S_2 to open a shorter helical region that has a less stable free energy than the subsequent formation of the kissing interaction.

Alternatively, one could use `RNAdesign` [21], which can build dependence graphs for multiple structural constraints and therefore increases the search space, allowing for novel kissing interactions. This more general design attempt will be implemented upon experimental feedback for the design presented here.

3.4 `RNAsubopt`, `barriers`, `findpath`

The candidate molecules computed by `switch.pl` are subsequently ranked by the difference of barrier heights for single-molecule refolding and kissing-dimer refolding. Computation for monomer refolding is straightforward, by computing the suboptimal structures for the monomer using `RNAsubopt` followed by evaluation of the minimal barrier height $\beta_{S_2\rightarrow S_1}$ with `barriers`. For the kissing-dimer interaction, suboptimal structures were computed for both energy models described above. When using the energy model E_u (see Equation 9), suboptimals were computed with the constraint that the region involved in the kissing interaction of S_1 (red in Figure 1) is unpaired (i.e.,

involved in the kissing interaction); according to the energy model E_l (see Equation 10), suboptimal structures were computed for the hybrid, with the constraint that the kissing region is paired, while the part corresponding to molecule 2 was kept constant in conformation S_2. The barrier for the initiation of the intramolecular hybridization is computed using `findpath` for opening the competing helix in S_1 and subsequently adding the duplex energies for formation of the kissing interaction.

4 Results

Out of all possible sequences that are compatible with the sequence and structure constraints shown in Figure 2, 158 different sequences were returned from `switch.pl` using 1000 individual runs with each 2×10^6 optimization steps. After postprocessing with `RNAsubopt`, `barriers`, and `findpath` to compute the final ranking of the generated designs, 69 sequences either turned out not to fold exactly into the ground state structures specified as input for `switch.pl`, or had a higher barrier for dimer refolding than for monomer refolding and were therefore excluded from further analysis. In the remaining pool of 89 sequences, 23 showed differences between monomer and dimer refolding barriers higher than 4 kcal/mol according to E_u (see Equation 9) for dimers. These sequences were visually inspected, and we selected a candidate (ACCUGGGAACCGGC-GACCCAGGUUUUCGGAACCAACGUCGGAGGUUCCU) for demonstration of prion behavior, that has (i) a very high refolding barrier for monomer refolding with +16.70 kcal/mol, and (ii) an energy landscape with as few as possible competing local minima to S_1 and S_2. The dimer refolding barriers are +11.60 and +8.60 kcal/mol for E_u and E_b, respectively. In comparison, the molecule presented in Badelt et al. [3] has the same barrier for monomer refolding (+16.70 kcal/mol), but +13.60 and +9.80 kcal/mol for prion-induced refolding.

Figure 5 shows the equilibrium between the two stable conformations S_1 and S_2 as a function of RNA concentration. The concentrations of monomers [M] and dimers [D] in equilibrium can be computed by the equilibrium partition function Z (see Equation 1):

$$K = \frac{[D]}{[M]^2} = \frac{Z_D}{Z_M^2} \tag{11}$$

The equilibrium partition function of the monomer (Z_M) can be computed directly using the McCaskill algorithm [31] implemented in `RNAfold`; Z_D can be approximated as

$$Z_D = Z_{c1} \cdot Z_{c2} \cdot Z_{\text{dup}} \tag{12}$$

with Z_{c1} and Z_{c2} denoting the partition functions of two monomers under the constraint that the blue (c1) or red (c2) interaction region (see Figure 3) is unpaired and thus available for forming an intermolecular (kissing) interaction. Z_{dup} is the partition function of the intermolecular duplex formed between the two molecules. This model follows the assumption that dimerization can only involve an interaction between the strands of the kissing interaction.

Since we are interested in the conformations formed upon monomerization and dimerization, we divided the total partition function Z_M into three parts: Z_{S_1}, Z_{S_2}, and Z_o. Here Z_{S_1} and Z_{S_2} contain all conformations constrained to form base pairs that can only be formed in structure S_1 or S_2, respectively, whereas Z_o contains all other conformations, that is, conformations that are not compatible with both constraints. Constraints are chosen such that (i) the helices formed by S_1 and S_2 are preserved and (ii) there are no structures fulfilling both constraints (see Figure 3). We computed the relative concentration of S_1 in monomers and dimers as

$$[S_1] = \frac{Z_{S_1}}{Z_M} \cdot [M] + \left(\frac{Z_{S_1+c1}}{Z_{c1}} + \frac{Z_{S_1+c2}}{Z_{c2}} \right) \cdot [D] \tag{13}$$

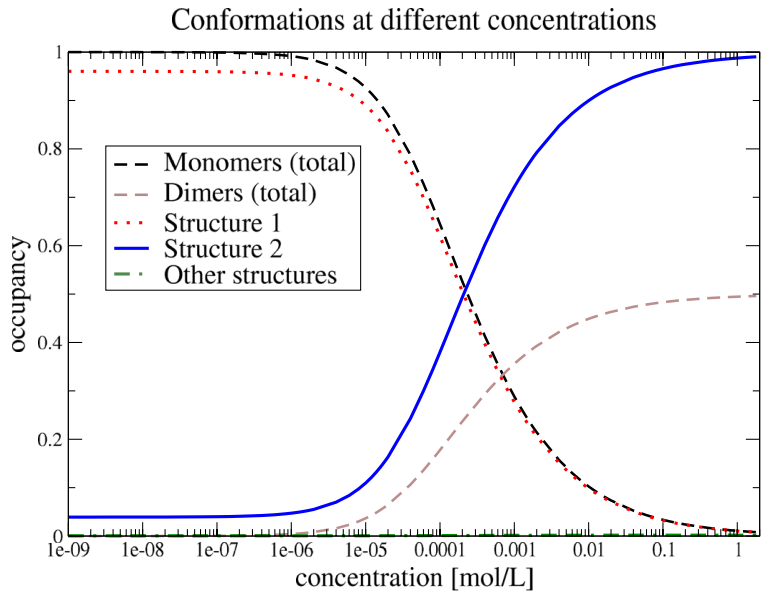


Figure 5. Conformational switching upon change of concentration. The transition from monomer to dimer conformations between 1 μM and 10 mM goes together with a switch from structure S_1 to S_2.

where $Z_{S_1+\bar{c}1}$ stands for a partition function that has both the constraint to fold into structure 1 (S_1) and the constraint to be unpaired in interaction region 1 ($\bar{c}1$). Relative concentrations of S_2 were computed accordingly and can be seen in Figure 5.

Figure 6 shows the energy profiles of the best refolding paths between S_1 and S_2, either for a single RNA monomer or for an RNA engaged in kissing interaction with another molecule. Since S_1 is the thermodynamically favored state in monomers, we show the refolding path from

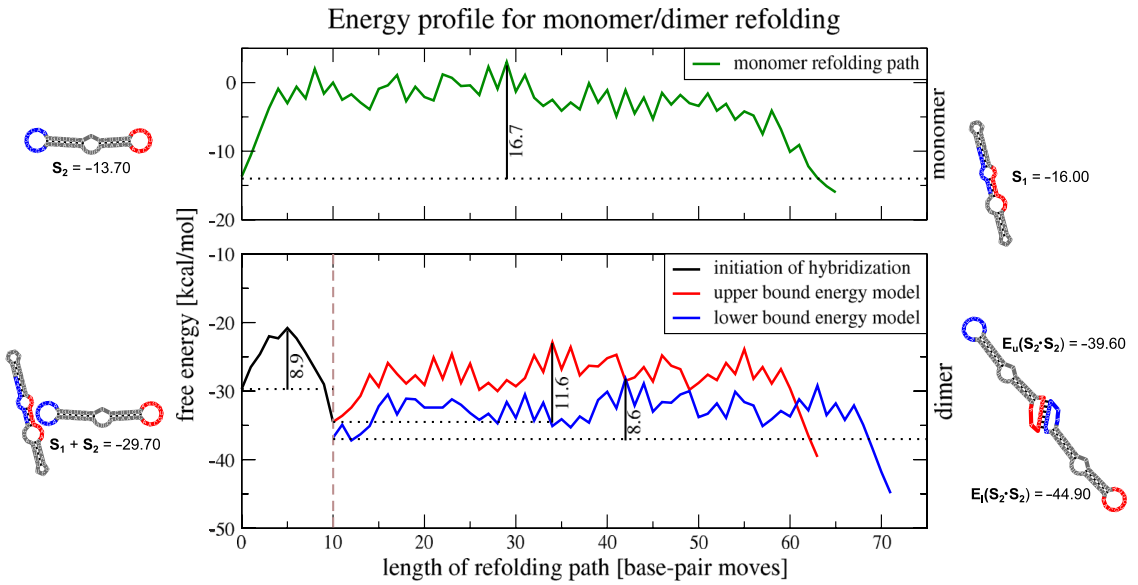


Figure 6. Energy profiles along the refolding path between structures S_1 and S_2. *Top panel*: refolding of a monomer; *bottom panel*: refolding while interacting with a second molecule. Blue- and red-colored regions are designed to form intermolecular base pairing. The lower panel shows a comparison between two energy models that differ in the energy contribution of the loop regions involved in the intermolecular pairing. In either case, the relative energy of refolding is lower than for the monomer.

S_2 (−13.70 kcal/mol) to S_1 (−16.00 kcal/mol) in the top panel. The barrier of this refolding path is 16.70 kcal/mol, making a non-induced switching of conformations unlikely.

The bottom panel of Figure 6 shows the energy profile for a scenario where an intermolecular interaction is first formed between one molecule in conformation S_1 and a second in S_2, followed by intramolecular refolding of the first molecule from S_1 into S_2. S_2 is now the favored conformation, since it is stabilized by the kissing interaction. In contrast, S_1 is destabilized, since one helix cannot be formed together with the intermolecular duplex. Theoretically, there would be a second possible duplex interaction that required S_1 to open eight base pairs in two helices, but since this interaction is not thermodynamically favored, it is not depicted in Figure 6.

For the initiation of the kissing interaction, all competing intramolecular base pairs of S_1 have to open first, and then the intermolecular base pairs can form. The energy barrier for this interaction is 8.90 kcal/mol and leads to a new local minimum conformation at −34.50 kcal/mol.

For the intramolecular refolding from S_1 to S_2, we compare the two energy models discussed above. The energy model E_u returns a barrier of 11.60 kcal/mol, while the energy model E_l results in a barrier of 8.60 kcal/mol. Note that also the folding path itself is different, due to the different modeling of involved loop regions.

5 Conclusion

In this contribution we have shown that the computational design of RNA molecules that exhibit prionlike behavior is feasible, and that the computational machinery is developed enough for a rigorous analysis of the behavior of the resulting sequences. As in the original prion system, the misfolded conformation forms, via a kissing interaction, a heterodimeric complex with the native conformation. This interaction destabilizes the native conformation and triggers refolding into the misfolded conformation. Hence, we demonstrated, at least in silico, that RNA molecules possess the necessary structural capabilities for conformational replication. The calculations show that the kissing interaction drastically lowers the activation energy for refolding. Furthermore, the misfolded conformation can oligomerize. In principle, the oligomerization could inhibit the exponential growth characteristics of the misfolded conformation. Experimental results from fiber-forming dynamic combinatorial libraries [4] show that mechanical forces lead to fiber breaking, restoring the exponential growth characteristics. One difficulty in our computational design is the lack of energy parameters for complex interaction structures that occur as folding intermediates. The design process is, however, flexible and can incorporate feedback from wet lab results. We therefore envision that practical RNA designs should be refined in a few rounds of wet lab testing and adaptation of the computational models.

Conformational replication constitutes a novel regulatory mechanism possessing highly nonlinear dynamic characteristics. This type of behavior is necessary for the construction of signal-enhancing molecular circuits. Such amplifying devices are usually hard to construct with RNA. Our design could easily be coupled with interaction sites for an external signal molecule that triggers the initial refolding event. Such a device could detect a single molecular event and translate it into a large, easily detected signal. In a prospective artificial metabolism, the infectious conformation may be used to trigger transcriptional or translational riboswitches, and, if it were regulating its own transcripition, the self-switching mechanism could be used for time-delayed feedback loops. The aspect of self-assembly to long fibers—which initially does not seem of particular interest, since it limits exponential growth—may provide a basic model towards the design of an RNA-based cytoskeleton. However, the current challenge is the synthesis of the RNA molecule in one particular conformation in vivo, which, for example, may be approached using techniques involving the tRNA ligase.

Acknowledgments

This work was supported in part by the FWF International Programme I670, the DK RNA program FG748004, the EU-FET grant RiboNets 323987, and the COST Action CM1304 "Emergence and Evolution of Complex Chemical Systems."

References

1. Afonin, K. A., Lindsay, B., & Shapiro, B. A. (2013). Engineered RNA nanodesigns for applications in RNA nanotechnology. *RNA Nanotechnology*, *1*, 1–15.
2. Aguzzi, A., Sigurdson, C., & Heikenwaelder, M. (2008). Molecular mechanisms of prion pathogenesis. *Annual Review of Pathology: Mechanisms of Disease*, *3*, 11–40.
3. Badelt, S., Flamm, C., & Hofacker, I. L. (2014). Computational design of a circular RNA with prion-like behaviour. In H. Sayama, J. Rieffel, S. Risi, R. Doursat, & H. Lipson (Eds.), *Artificial life 14: Proceedings of the Fourteenth International Conference on the Synthesis and Simulation of Living Systems* (pp. 565–568). Cambridge, MA: MIT Press.
4. Carnall, J. M. A., Waudby, C. A., Belenguer, A. M., Stuart, M. C. A., Peyralans, J. J.-P., & Otto, S. (2010). Mechanosensitive self-replication driven by self-organization. *Science*, *327*(5972), 1502–1506.
5. Cayrol, B., Nogues, C., Dawid, A., Sagi, I., Silberzan, P., & Isambert, H. (2009). A nanostructure made of a bacterial noncoding RNA. *Journal of the American Chemical Society*, *131*(47), 17270–17276.
6. Chen, C., Sheng, S., Shao, Z., & Guo, P. (2000). A dimer as a building block in assembling RNA. *Journal of Biological Chemistry*, *275*(23), 17510–17516.
7. Dotu, I., Lorenz, W. A., Van Hentenryck, P., & Clote, P. (2010). Computing folding pathways between RNA secondary structures. *Nucleic Acids Research*, *38*(5), 1711–1722.
8. Ennifar, E., Paillart, J.-C., Marquet, R., Ehresmann, B., Ehresmann, C., Dumas, P., & Walter, P. (2003). HIV-1 RNA dimerization initiation site is structurally similar to the ribosomal A site and binds aminoglycoside antibiotics. *Journal of Biological Chemistry*, *278*(4), 2723–2730.
9. Flamm, C., Fontana, W., Hofacker, I. L., & Schuster, P. (2000). RNA folding at elementary step resolution. *RNA*, *6*, 325–338.
10. Flamm, C., & Hofacker, I. L. (2008). Beyond energy minimization: Approaches to the kinetic folding of RNA. *Monatshefte für Chemie*, *139*(4), 447–457.
11. Flamm, C., Hofacker, I. L., Maurer-Stroh, S., Stadler, P. F., & Zehl, M. (2001). Design of multi-stable RNA molecules. *RNA*, *7*, 254–265.
12. Flamm, C., Hofacker, I. L., Stadler, P. F., & Wolfinger, M. T. (2002). Barrier trees of degenerate landscapes. *Zeitschrift für Physikalische Chemie*, *216*, 155–173.
13. Frommer, J., Appel, B., & Müller, S. (2015). Ribozymes that can be regulated by external stimuli. *Current Opinion in Biotechnology*, *31*, 35–41.
14. Genot, A. J., Bath, J., & Turberfield, A. J. (2011). Reversible logic circuits made of DNA. *Journal of the American Chemical Society*, *133*(50), 20080–20083.
15. Grabow, W. W., & Jaeger, L. (2014). RNA self-assembly and RNA nanotechnology. *Accounts of Chemical Research*, *47*(6), 1871–1880.
16. Grabow, W. W., Zakrevsky, P., Afonin, K. A., Chworos, A., Shapiro, B. A., & Jaeger, L. (2011). Self-assembling RNA nanorings based on RNAI/II inverse kissing complexes. *Nano Letters*, *11*(2), 878–887.
17. Green, A. A., Silver, P. A., Collins, J. J., & Yin, P. (2014). Toehold switches: De-novo-designed regulators of gene expression. *Cell*, *159*(4), 925–939.
18. Guo, P. (2010). The emerging field of RNA nanotechnology. *Nature Nanotechnology*, *5*, 833–842.
19. Hofacker, I. L., Flamm, C., Heine, C., Wolfinger, M. T., Scheuermann, G., & Stadler, P. F. (2010). BarMap: RNA folding on dynamic energy landscapes. *RNA*, *16*, 1308–1316.
20. Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., & Schuster, P. (1994). Fast folding and comparison of RNA secondary structures (the Vienna RNA Package). *Monatshefte für Chemie*, *125*(2), 167–188.
21. Höner zu Siederdissen, C., Hammer, S., Abfalter, I., Hofacker, I. L., Flamm, C., & Stadler, P. F. (2013). Computational design of RNAs with complex energy landscapes. *Biopolymers*, *99*(12), 1124–1136.
22. Isaacs, F. J., Dwyer, D. J., Ding, C., Pervouchine, D. D., Cantor, C. R., & Collins, J. J. (2004). Engineered riboregulators enable post-transcriptional control of gene expression. *Nature Biotechnology*, *22*(7), 841–847.
23. Ishikawa, J., Furuta, H., & Ikawa, Y. (2013). RNA tectonics (tectoRNA) for RNA nanostructure design and its application in synthetic biology. *WIREs RNA*, *4*, 651–664.

24. Khalil, A. S., & Collins, J. J. (2010). Synthetic biology: Applications come of age. *Nature Reviews Genetics*, *11*, 367379.

25. Kim, J., & Winfree, E. (2011). Synthetic in vitro transcriptional oscillators. *Molecular Systems Biology*, *7*(1), 465.

26. Kucharík, M., Hofacker, I. L., Stadler, P. F., & Qin, J. (2014). Basin hopping graph: A computational framework to characterize RNA folding landscapes. *Bioinformatics*, *30*(14), 2009–2017.

27. Leisner, M., Bleris, L., Lohmueller, J., Xie, Z., & Benenson, Y. (2010). Rationally designed logic integration of regulatory signals in mammalian cells. *Nature Nanotechnology*, *5*(9), 666–670.

28. Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA package 2.0. *Algorithms for Molecular Biology*, *6*(1), 26.

29. Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M., & Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences*, *101*(19), 7287–7292.

30. Maňuch, J., Thachuk, C., Stacho, L., & Condon, A. (2011). NP-completeness of the energy barrier problem without pseudoknots and temporary arcs. *Natural Computing*, *10*(1), 391–405.

31. McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, *29*, 1105–1119.

32. Petkovic, S., Badelt, S., Block, S., Flamm, C., Delcea, M., Hofacker, I. L., & Müller, S. (2015). Sequence-controlled RNA self-processing: Computational design, biochemical analysis and visualization by AFM. *RNA Biology*, *21*, 1249–1260.

33. Pieper, S., Vauleon, S., & Müller, S. (2007). RNA self-processing towards changed topology and sequence oligomerization. *Biological Chemistry*, *388*(7), 743–746.

34. Rinaudo, K., Bleris, L., Maddamsetti, R., Subramanian, S., Weiss, R., & Benenson, Y. (2006). A universal RNAi-based logic evaluator that operates in mammalian cells. *Nature Biotechnology*, *25*(7), 795–801.

35. Robertson, M. P., & Joyce, G. F. (2014). Highly efficient self-replicating RNA enzymes. *Chemistry & Biology*, *21*(2), 238–245.

36. Rothemund, P. W. K., Papadakis, N., & Winfree, E. (2004). Algorithmic self-assembly of DNA Sierpinski triangles. *PLoS Biology*, *2*(12), e424.

37. Subsoontorn, P., Kim, J., & Winfree, E. (2012). Ensemble Bayesian analysis of bistability in a synthetic transcriptional switch. *ACS Synthetic Biology*, *1*(8), 299–316.

38. Turberfield, A. J., Mitchell, J., Yurke, B., Mills Jr., A. P., Blakey, M., & Simmel, F. C. (2003). DNA fuel for free-running nanomachines. *Physical Review Letters*, *90*(11), 118102.

39. Turner, D. H., & Mathews, D. H. (2010). NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, *38*, D280–D282.

40. Wachsmuth, M., Sven Findeiß, S., Weissheimer, N., Stadler, P. F., & Mörl, M. (2013). *De novo* design of a synthetic riboswitch that regulates transcription termination. *Nucleic Acids Research*, *41*(4), 2541–2551.

41. Wang, F., Lu, C.-H., & Willner, I. (2014). From cascaded catalytic nucleic acids to enzyme–DNA nanostructures: Controlling reactivity, sensing, logic operations, and assembly of complex structures. *Chemical Reviews*, *114*, 2881–2941.

42. Weixlbaumer, A., Werner, A., Flamm, C., Westhof, E., & Schröder, R. (2004). Determination of thermodynamic parameters for HIV-1 DIS type loop–loop kissing complexes. *Nucleic Acids Research*, *32*, 5126–5133.

43. Wieland, M., Benz, A., Klauser, B., & Hartig, J. S. (2009). Artificial ribozyme switches containing natural riboswitch aptamer domains. *Angewandte Chemie*, *121*(15), 2753–2756.

44. Wolfe, B. R., & Pierce, N. A. (2014). Sequence design for a test tube of interacting nucleic acid strands. *ACS Synthetic Biology*, *4*(10), 1086–1100.

45. Wolfinger, M. T., Svrcek-Seiler, A. W., Flamm, C., Hofacker, I. L., & Stadler, P. F. (2004). Efficient computation of RNA folding dynamics. *Journal of Physics A: Mathematical and General*, *37*, 4731–4741.

46. Wuchty, S., Fontana, W., Hofacker, I. L., & Schuster, P. (1999). Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, *49*(2), 145–165.

47. Xie, Z., Wroblewska, L., Prochazka, L., Weiss, R., & Benenson, Y. (2011). Multi-input RNAi-based logic circuit for identification of specific cancer cells. *Science*, *333*(6047), 1307–1311.

48. Yin, P., Choi, H. M., Calvert, C. R., & Pierce, N. A. (2008). Programming biomolecular self-assembly pathways. *Nature*, *451*(7176), 318–322.

49. Zadeh, J. N., Steenberg, C. D., Bois, J. S., Wolfe, B. R., Pierce, M. B., Khan, A. R., Dirks, R. M., & Pierce, N. A. (2011). NUPACK: Analysis and design of nucleic acid systems. *Journal of Computational Chemistry*, *32*(1), 170–173.

50. Zhang, D. Y., & Seelig, G. (2011). Dynamic DNA nanotechnology using strand-displacement reactions. *Nature Chemistry*, *3*, 103–113.

Part IV

CONCLUDING REMARKS

8

FOLDING KINETICS AND RNA DESIGN

This thesis presented several individual projects that may be combined in order to address a diverse set of current challenges in bioengineering. We have (i) modeled the kinetics of pairwise RNA-RNA interactions in Chapter 3, (ii) included RNA-ligand interactions into RNA folding during transcription and modeled cotranscriptional folding of large RNAs in Chapter 4, (iii) designed ribozyme cleavage/ligation cascades in Chapter 6, and (iv) designed an (exponential) RNA self-switching mechanism triggered at high concentrations in Chapter 7.

From a bird eyes perspective, all projects relate to the prediction and design of interacting RNAs that are *out-of-equilibrium*. We have chosen different methods to coarse-grain energy landscapes and presented a new way to describe bimolecular reactions in the context of such energy landscapes. The models presented here are adaptable for experimental parameters. For example, one can and differ between uni- and bimolecular reactions using different scaling factors for transition rates, or one can model the free energy gain from a backbone formation when ribozymes ligate a substrate. In this chapter, I will focus on the central aspect of RNA-RNA interactions and its implications for all of the other projects.

Folding kinetics of interacting ribozymes (i) + (iii)

We have designed a ribozyme interaction network using the catalytic core of the hairpin ribozyme. A cascade of chemical reactions was formulated as objective function for sequence design, and self-circularization and self-polymerization was successfully confirmed by *in vitro* experiments. We found that simplified reaction schemes (Chapter 6, Figure 7) were able to qualitatively explain many of the differences between the five RNAs that were experimentally tested. However, simulations of larger reaction networks with more intermediates and including folding kinetics did not improve the quantitative match. In particular, our simulations required significantly higher RNA concentrations for concatemer formation than observed in experiments. Our method for computing RNA-RNA interactions now presents a more general way to model bimolecular reactions of nucleic acids. For future studies, the method can be trained with experimental data, e. g. derived from melting experiments, and then be applied to dynamic RNA energy landscapes that change due to cleavage/ligation events.

Figure 34 shows a two-dimensional pattern that in principle can be constructed from ribozymes, although steric considerations would require additional refinements. Parameters to improve the kinetic models of RNA-RNA interactions can be used to

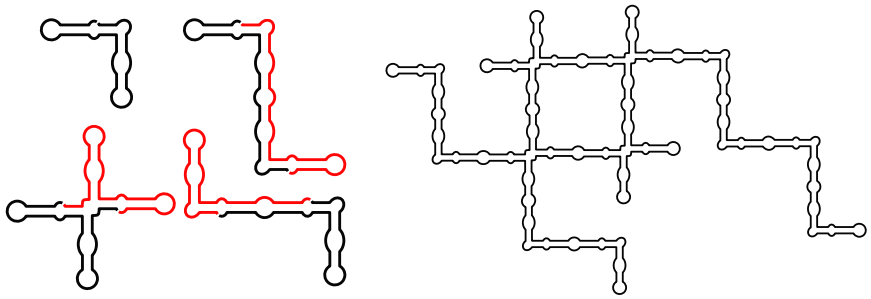


Figure 34: Theoretical networks of ribozyme self-processing. **Left:** Two monomers (black and red) can interact in three different ways to form a dimer. **Right:** Without paying respect to steric considerations, theses combinations yield networks of multiple interacting ribozymes forming regular shaped polygons composed of (multiple) circular ribozyme species.

simulate self-polymerizing RNAs, but also to simulate and design networks of multiple interacting catalytic RNA molecules. The regular shapes shown in Figure 34 could then be improved to dynamically assemble and disassemble more flexible scaffold structures in a cell. With respect to RNA functional diversity, one can model hairpin ribozymes that are capable of exponential self-replication, such as shown by Robertson and Joyce [2014]. Their approach used *in vitro* evolution, rather than *in silico* evolution presented in this work.

The main challenge for more robust design of self-*circularization* is the reversibility of the reaction. It is not clear whether the mechanism presented here can be used to produce stable circular RNA. Our experiments showed that the molecules are degraded by exonuclease RNase R, which indicates that the molecules equilibrate fast between their circular and linear version. However, overcoming this problem would allow the transcription of RNA molecules that circularize right after transcription and are then more stable against degradation.

Folding kinetics of interacting RNA-prions (i) + (iv)

Chapter 7 describes the thermodynamic design of a 49 nucleotide circular RNA molecule, that can propagate (self-replicate) a non-optimal, infectious conformation. A misfolded infectious RNA-prion can convert the other correctly folded RNA species into the infectious agent. We formalized the approach and showed that the design-space (i. e. the number of such prionlike RNAs) is much larger than expected. However, experimental feedback is essential to refine the method and utilize the mechanism for biological engineering. In combination with coarse-grained modeling of intermolecular folding kinetics, the response time of such switches could be adjusted. While fast response times result in an efficient signal amplification, delays in folding times can be impor-

tant to ensure proper timing in molecular response. The design method for circular RNA-prions may also be adapted to develop particularly stable, reversible triggers of riboswitches. At low concentrations, RNA-prions are enhancers of their own transcription, at high concentrations, RNA-prions trigger the response of a different transcript. Such basic feedback loops are used for molecular timing in natural cell-cycles. Alternatively, the kissing interactions used in RNA-prions allow the assembly of multiple copies. On the one hand, this limits the exponential nature of switching, on the other hand, it may serve as a primitive form of a dynamic cytoskeleton.

Nucleic acid interactions during transcription (i) + (ii)

We have modeled RNA-ligand interactions during transcription, assuming infinite RNA and ligand concentrations. However, our methods can easily be generalized to consider actual concentrations. RNA logic circuits composed of riboswitches and bacterial small RNAs can be used to identify whether a small RNA interaction site is accessible during transcription. At a slow transcription rate the transcription terminates due to limited substrate resources, while at fast transcription rate, the mRNA is translated into a protein. Simulations of RNA-RNA interactions allow the prediction of alternative targets, the response time, and how much miRNA is needed to degrade a substantial fraction of transcripts. The method may also be adapted for larger systems, e. g. in combination with the heuristics used in `DrTransformer`.

Design of multi-responsive riboswitches (i) + (ii) – inverse

Wachsmuth et al. [2015] showed that concatenation of multiple theophylline riboswitches yields a higher termination efficiency. The idea is appealing, as logic gates (AND, OR, NOT) can then be encoded using a single transcript instead of multiple interacting RNA. Our simulations of cotranscriptional ligand binding events now also enable the design of single riboswitches that have competing binding pockets. The presented `BarMap` approach is restricted to rather short molecules, however, `DrTransformer` combined with ligand binding would allow the design of larger logic gates in a single riboswitch.

Interactions of multiple RNAs (i)

Coarse-grained modeling of RNA-RNA interactions should also be applicable to simulating folding kinetics of multiple interacting RNAs, e. g. using the algorithm implemented in the NUPACK framework Dirks et al. [2007]. While investigating single RNA interactions in eukaryotes is crucial to develop RNA based medication [Andries et al., 2014], bigger logical circuits are expected to improve diagnostics. *In vitro* such logical

circuits can be used to enable a fast and cheap setup for diagnostic tests [Jung and Ellington, 2014] and are therefore also of economic interest. A profound understanding of molecular mechanisms will be crucial for this field and the methods developed in this project are adaptable and can incorporate experimental feedback.

The End

Part V

APPENDIX

A

SUPPLEMENTAL MATERIAL: SEQUENCE-CONTROLLED RNA SELF-PROCESSING: COMPUTATIONAL DESIGN, BIOCHEMICAL ANALYSIS, AND VISUALIZATION BY AFM

Supplemental Material

Sequence-controlled RNA self-processing: computational design, biochemical analysis and visualization by AFM

Sonja Petkovic, Stefan Badelt, Stephan Block, Christoph Flamm, Mihaela Delcea, Ivo Hofacker, Sabine Müller

Corresponding authors:

Prof. Dr. Sabine Müller; Email: smueller@uni-greifswald.de

Prof. Dr. Ivo Hofacker; Email: ivo@tbi.univie.ac.at

Dr. Stephan Block; Email: stephan.block@chalmers.se

Table of Contents

Experimental

General remarks and chemicals

Deoxynucleotide triphosphates (dNTPs), nucleotide triphosphates (NTPs), Klenow buffer, DNase I, T7 RNA polymerase, Klenow fragment exo-, RiboLock™, RiboRuler™ low range RNA ladder and polynucleotide kinase were purchased from *Fermentas Company* (Schwerte, Germany); T4 RNA Ligase 2 (T4 RnL2) and the appropriate buffer was obtained from *New England Biolabs* (Frankfurt am Main, Germany). DNA primers were provided by *Biomers.net* (Ulm, Germany). RNase R including the required buffer was obtained from *epicenter* (Oldendorf, Germany). All chemicals and reagents were of analytical grade and filtered through a 0.2 µm polyvinyl difluoride membrane before use. Upon electrophoresis, all gels were stained for 5 to 10 min with ethidium bromide. Final concentration of ethidium bromide in 1xTBE was 0.5 µg/ml. All UV spectra were recorded on a NanoDrop ND 1000 spectrophotometer. Stained gels (agarose or polyacrylamide) were visualized using Chemi-Smart 2000 WL/LC 26M or VWR GenoView.

RNA preparation

Klenow primers (for sequences see Supplemental Table S2) with 20 bp overlap (27 bp for the inactive dimer) were used in Klenow reactions with Klenow exo$^-$ polymerase following the manufacturer's protocol, and stopped by precipitation from ethanol at -20 °C overnight. DNA was isolated from native agarose gels (1.5%, EtBr stained). Product containing bands were cut out and DNA was isolated using QIA quick gel extraction kit (*Qiagen*, Venlo, The Netherlands). Since only one product was detectable after Klenow reaction, in later preparations the gel extraction step was skipped. Instead, after ethanol precipitation of the Klenow reaction product, the pellet was solved in 100 µl water, and 5 µl were used for subsequent *in vitro* transcription. RNAs were synthesized by *in vitro* transcription of double stranded DNA templates (1 µM concentration, or as mentioned above 5 µl of the Klenow product resolved in water after precipitation) with T7 RNA polymerase in the presence of the four ribonucleoside triphosphates (2 mM) and 1 U/µl RiboLock™ in 1x HEPES buffer (Na-HEPES 50 mM, $MgCl_2$*$6H_2O$ 12 mM, Spermidin 2 mM, pH = 7.5) in a total reaction volume of 50 µl for 3 h at 37 °C. DNA template was hydrolyzed adding 2 µl DNase I directly to the transcription mixture and left at 37 °C for additional 30-45 min. Final purification was achieved by electrophoresis on 15% denaturing polyacrylamide gels (for composition see subchapter *PAGE analysis* below), elution of the product-containing bands with sodium acetate (0.3 M, pH= 7, 3 times for at least two hours and overnight for the final elution step, shaking at approximately 500 rpm, at 10 °C) and precipitation with 250 vol.-% ethanol at -20 °C overnight.

PAGE analysis

For RNA species analysis or purification, denaturing (7 M urea) polyacrylamide gel electrophoresis (acrylamide: bisacrylamide 19:1 100 ml, ammonium persulfate 10 % w/v, 1 ml, *N,N,N',N'*-Tetramethylethane-1,2-diamine 50 µl) was applied, using 1xTBE buffer as running buffer and stop mix (7 M urea, 50 mM EDTA, bromophenol blue and xylene cyanol each 5 vol-%) for sample loading. After mixing samples and/or RNA size standard with buffer, RNAs were denatured at 90 °C for 2 min and directly loaded onto the gel. Loading buffer for the RNA size standard was provided by *Fermentas* (Schwerte, Germany).

Design, preparation and analysis of an inactive cyclic dimer

Design of the inactive cyclic dimer

Circular monomers have been successfully identified for CRZ-2 in our previous work (Petkovic and Muller 2013) however, the formation of circular dimers could not be shown. Since the size of cyclic RNAs is not assessable in PAA gels according to standard size markers, we designed an inactive circular dimer (CRZ*) as a reference. The cyclic reference dimer should be as similar as possible to the circular dimer of CRZ-2 and therefore help to identify this species in a PAA gel. However, due to its inactivity, the linear version of the reference dimer (produced by in vitro transcription) has to be ligated enzymatically to the desired cyclic product.

In particular, the design of an inactive circular dimer has to fulfill the following constraints with respect to the reference circular dimer (*CRZ*): *(i)* the secondary structure ensemble associated with the RNA has to be similar, *(ii)* the nucleotide content has to be equal, *(iii)* the sequence must not be symmetric, *(iv)* all conserved catalytic centers have to be destroyed and *(v)* a T7 promotor region is needed. Points *(i)* and *(ii)* shall insure a similar migration pattern on a polyacrylamide gel, whereas points *(iii)*, *(iv)* and *(v)* are necessary for experimental implementation. Asymmetry insures that only the defined 3'-terminal regions of Klenow primers overlap to obtain specific dsDNA of the desired length as template for RNA synthesis, and inactivity is necessary to avoid cleavage/ligation reactions after and during *in vitro* transcription. As a first step we preset the residual T7 RNA promotor sequence 5'-GGG AGA-3' as a non-mutable hexanucleotide at the 5'-end of the ribozyme. These bases will inevitably occur in the *in vitro* transcribed RNA due to usage of T7 RNA polymerase. Since this pre-processing step harms condition no. *(ii),* we mutated different helical regions to compensate for inequalities in the nucleotide content. Next, we randomly flipped base pairs within all helical regions (apart from the residual T7 RNA promotor region) and randomly shuffled the nucleotides from all loop regions to obtain loss of catalytic activity. With this approach we designed roughly 500 RNA species that fulfill conditions *(ii), (iii), (iv)* and *(v)*. To ensure similar folding behavior, the sequence should not only have the same ground state as the CRZ-2 dimer, also the whole structure ensemble should be similar. Therefore, we first selected for those sequences that have the smallest mean base pair distance within the equilibrium structure ensemble. This mean base pair distance (D) can be computed as

$$D(CRZ, CRZ^*) = \sum_{ij} P_{ij}^{CRZ}\left(1 - P_{ij}^{CRZ^*}\right) + P_{ij}^{CRZ^*}\left(1 - P_{ij}^{CRZ}\right) \quad [5]$$

With P_{ij}^{CRZ} denoting the probability of a single base pair between position i and j for the molecule *CRZ*. From the top 20 designed molecules, we selected a sequence (shown in Table S2, Supporting Information) that has a comparable minimal free energy (MFE).

Preparation of the inactive cyclic dimer by transcription priming with GMP

To obtain the inactive linear dimer (*in-l-166mer)*, with 5'-terminal monophosphate, GMP was added to the NTP mix following the protocol of Harris and Christian for incorporation of guanosine monophosphorothioate.(Harris and Christian 1999) A 4.8:1 ratio of GMP:GTP was used, and the double stranded Klenow DNA, buffer, RiboLock™ and polymerase were added as described above. *In vitro* transcription was stopped after 3 hours at 37 °C, and double stranded DNA template was hydrolyzed using DNase I following manufacturer's protocol. The reaction mixture was blended with 100 vol-% stop mix (7 M urea and 50 mM EDTA) and directly used for purification on a 15% denaturing polyacrylamide gel. After PAGE, elution of the desired RNA and ethanol precipitation as described above, RNAs were used for ligation. Enzymatic ligation in the double stranded region of the *in-l-166mer* to generate the cyclic species *in-c-166mer* was conducted using T4 RnL2 in a total reaction volume of 20 µl at 37 °C for 4 hrs following the suppliers protocol. RNA was purified using the RNA Clean & Concentrator™-5 kit (*ZymoResearch,* Freiburg, Germany*)* following the general protocol for total RNA purification. Elution of RNA was carried out with 50 µl desalted and purified millipore water. After addition of 50 µl stopmix, ligation products were analyzed on a 15% denaturing polyacrylamide gel.

Digestion with RNase R

10 µl of T4 RNA ligase 2 reaction mixture were used directly for hydrolysis using RNase R. $MgCl_2$ to a final concentration of 5 mM, RNase R buffer and water up to 17 µl were mixed and denatured at 90 °C for 5 min. The mixture was cooled down to 50 °C for 1 min before addition of 1 µl of an 1:1 freshly with water diluted RNase R solution. Hydrolysis occurred at 50 °C for 10 min. Reaction was stopped using an equal volume of stop mix, which is also used as loading buffer for electrophoresis, and the mixture was immediately frozen in liquid nitrogen. Reaction products were analyzed by electrophoresis through a 15% denaturing polyacrylamide gel.

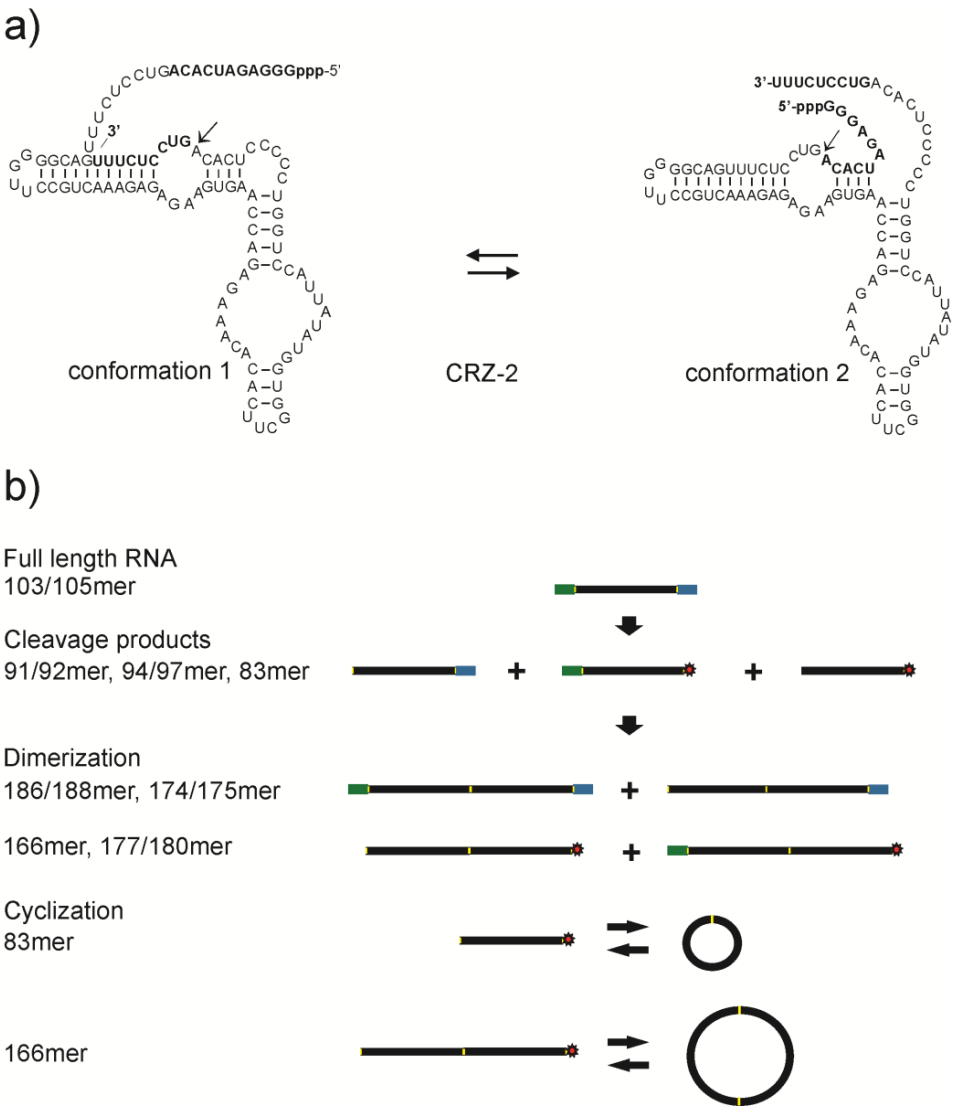


Figure S1: **a:** Two alternative cleavage-favoring conformations of the reference self-processing RNA CRZ-2, **b:** Schematic presentation of self-processing products.

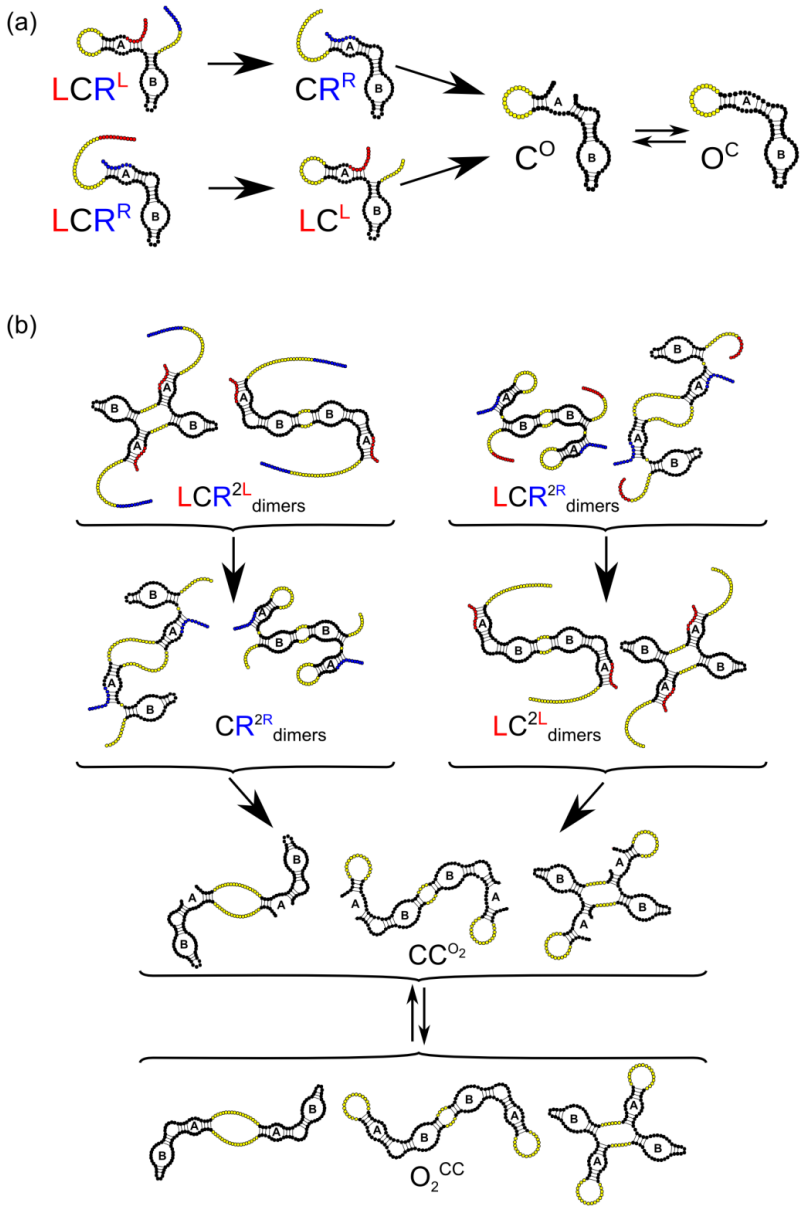


Figure S2: Models of (a) the monomer and (b) dimer cleavage cascade

Red and blue colored regions mark cleavable 5-’ and 3’-ends, respectively. A cleavage/ligation reaction can only occur when tertiary interactions between loop A and loop B are formed. In black we show structure constraints needed for such reactions, while yellow regions should be flexible without impairing catalytic activity. Importantly, every structure constraint defines a non-overlapping set of structures such that the probability of forming a reactive molecule can be computed from the sum of the constraint partition functions. RNA secondary structures were drawn using jViz (1).

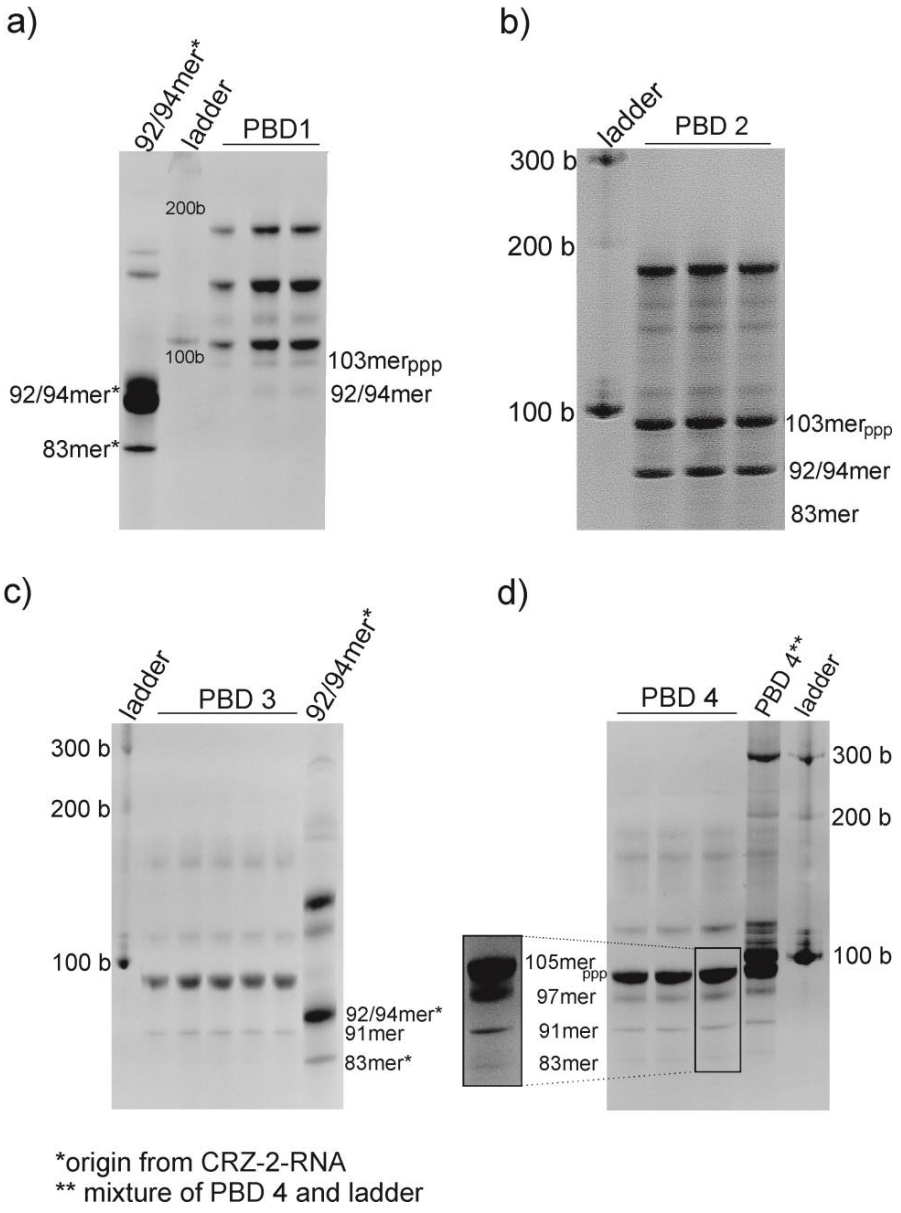


Figure S3: Self-processing products of the four designed RNAs PBD1 to 4 analyzed with a 15% denaturing polyacrylamide gel (preparative scale).

a: PBD1 self-processing products upon reaction at PBD1 starting concentration of 2.5 µM;
b: PBD2 self-processing products upon reaction at PBD2 starting concentration of 3 µM;
c: PBD3 self-processing products upon reaction at PBD3 starting concentration of 2 µM;
d: PBD4 self-processing products upon reaction at PBD4 starting concentration of 2 µM. For better visualization, the boxed area is shown slightly magnified and at higher contrast on the left.

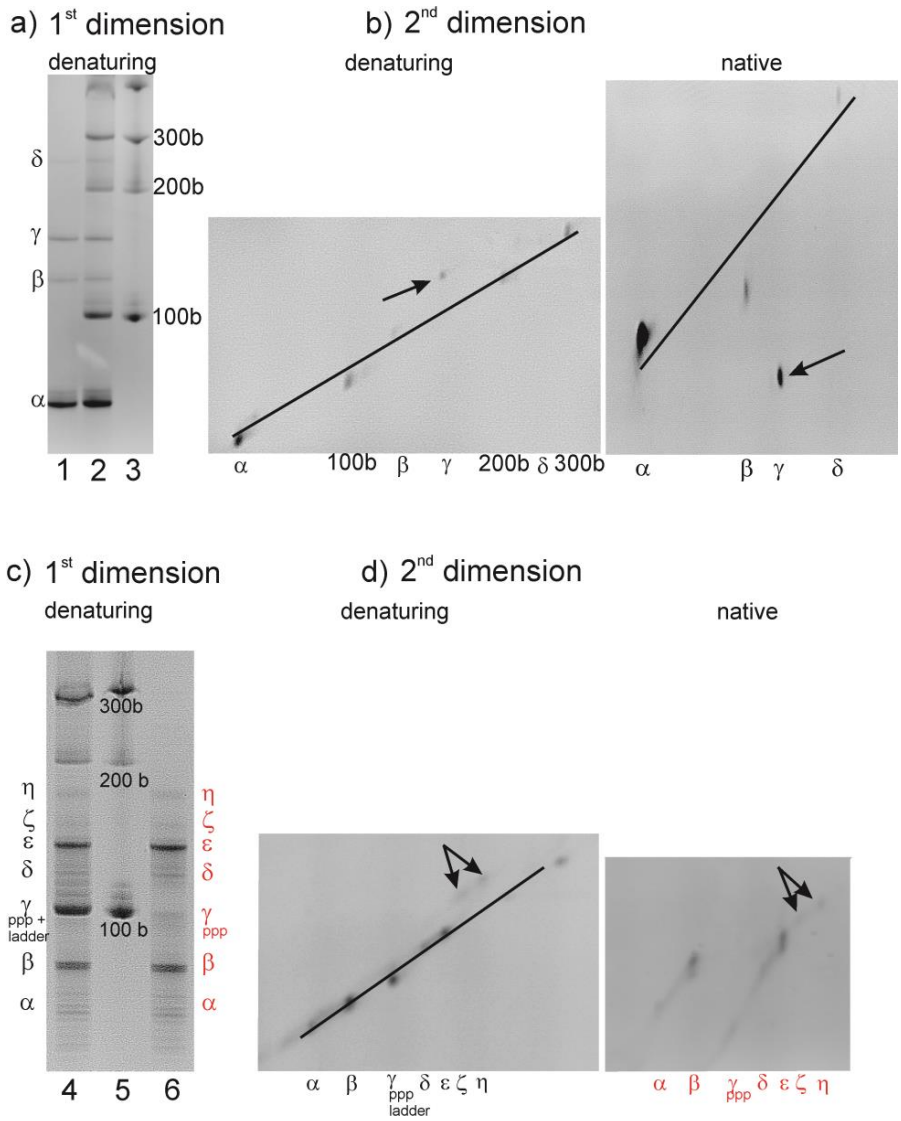


Figure S4: **Identification of cyclic RNA from l-83mer and from full-length CRZ-2 by 2D-PAGE.**

a: Self-processing of l-83mer analyzed on a 15% denaturing polyacrylamide gel. Lane 1: self-processing products denoted with Greek letters α to δ. Lane 2: self-processing products mixed with linear RNA size standard. Lane 3: linear RNA size standard. **b**: Second-dimension denaturing (left, to improve resolution polyacrylamide concentration was increased to 17.5%) and native (right, 15%) polyacrylamide gels. Lane 1 of the gel shown in panel (**a**) was cut off and used as "starting slot" for the native gel in second dimension (**b**, right), lane 2 of the gel shown in panel (**a**) was equally used for the denaturing gel in second dimension (**b**, left). Species γ (marked by an arrow) appears beyond the diagonal in both gels, implying its cyclic nature. **c**: Self-processing of full length CRZ-2 (103mer) analyzed on a 15% denaturing polyacrylamide gel. Lane 4: self-processing products denoted with Greek letters α to η, mixed with linear RNA size standard. Lane 5: linear RNA size standard. Lane 6: self-processing products denoted with Greek letters α to η. **d**: Second-dimension denaturing

(left, to improve resolution polyacrylamide concentration was increased to 17.5%) and native (right, 15%) polyacrylamide gels. Lane 4 of the gel shown in panel (**c**) was cut off and used as "starting slot" for the denaturing gel in second dimension (**d**, left), lane 6 of the gel shown in panel (**c**) was equally used for the native gel in second dimension (**d**, right). The two blurry spots ζ and η marked by arrows in the denaturing gel in panel (**d**) might correspond to cyclic RNAs. However, the corresponding native gel on the right reveals that both RNAs do not migrate as fast as would be expected for cyclic species according to the analysis of l-83mer shown in panel (**b**).

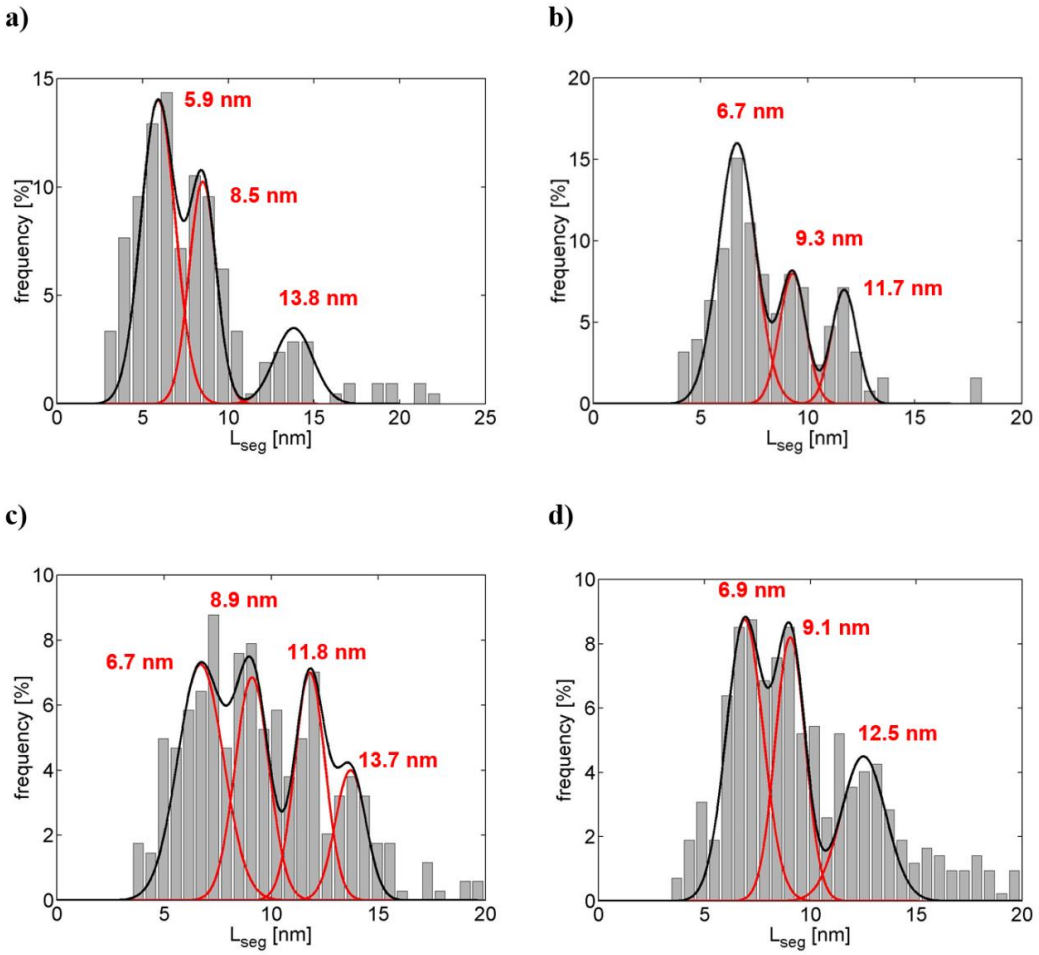


Figure S5: RNA segment length histograms of references: (a) l-83mer, and (b) CRZ-2and test sequences (c) PBD1 and (d) PBD4

In tapping mode AFM images the RNA chains typically consist of rod-like segments, which are connected via kinks (see Figs 5 and 6). From the AFM images, the length of these segments can be measured with nm accuracy, showing several well-resolvable peaks at (average ± standard deviation as averaged over the 4 different RNA sequences) 6.6±0.4 nm, 9.0±0.3 nm, 12.0±0.4 nm, and 13.8±0.1 nm. Assuming a double helical conformation of the rod-like segments and therefore a typical pitch of 0.3 nm per base pair, these peaks correspond to segments consisting of 22±1 bp, 30±1 bp, 40±1 bp, 46±0.3 bp. These histograms include the length of (**a**) 212, (**b**) 127, (**c**) 342 and (**d**) 423 segments.

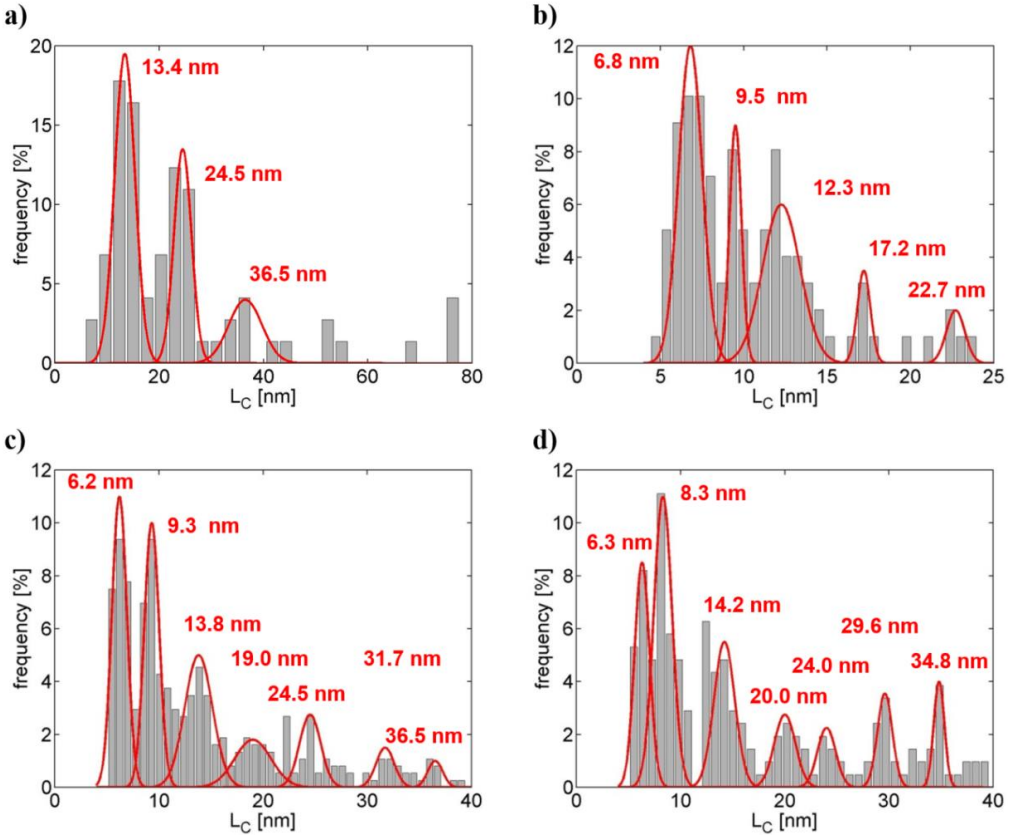


Figure S6: Contour length histograms of RNA chains for (a) I-83mer, (b) CRZ-2, (c) PBD1, and (d) PBD4.

For the linear 83mer, the histogram gives three peaks at 13.4, 24.5, and 36.5 nm. Assuming again a double helical conformation and therefore a typical pitch of 0.3 nm per base pair (= 0.15 nm per base), these peaks correspond to 89, 163 and 243 bases and can be identified as final cleavage products (83mer) and higher ligation products (dimer = 166mer; trimer = 249mer). (**b**-**d**) As the other RNA constructs additionally create intermediate cleavage products, their contour length histograms exhibit a more complicated peak structure. Generally, peaks are observed in all histograms at similar values allowing averaging the determined peak position over the three different RNA sequences. This yield the following values (average ± standard deviation of the respective peak position): 6.4±0.3 nm (first peak), 9.0±0.6 nm (second peak), 13.4±1.0 nm (third peak), 18.7±1.4 nm (fourth peak), 23.7±0.9 nm (fifth peak), and for PBD1 and PBD4 at 30.7±1.5 nm (sixth peak), and 35.7±1.2 nm (seventh peak). Note that the first two peaks coincide with the first two peaks of the segment length histograms in Figure S5. Hence, it is very likely that these peaks correspond to RNA chains, for which only one of the constituting segments was resolvable in the AFM image (*e.g.*, if two neighbouring segments enclose an angle of approximately 180° and therefore appear as a single segment in the measurement). These histograms include the contour length of (**a**) 73, (**b**) 100, (**c**) 218 and (**d**) 256 RNA chains.

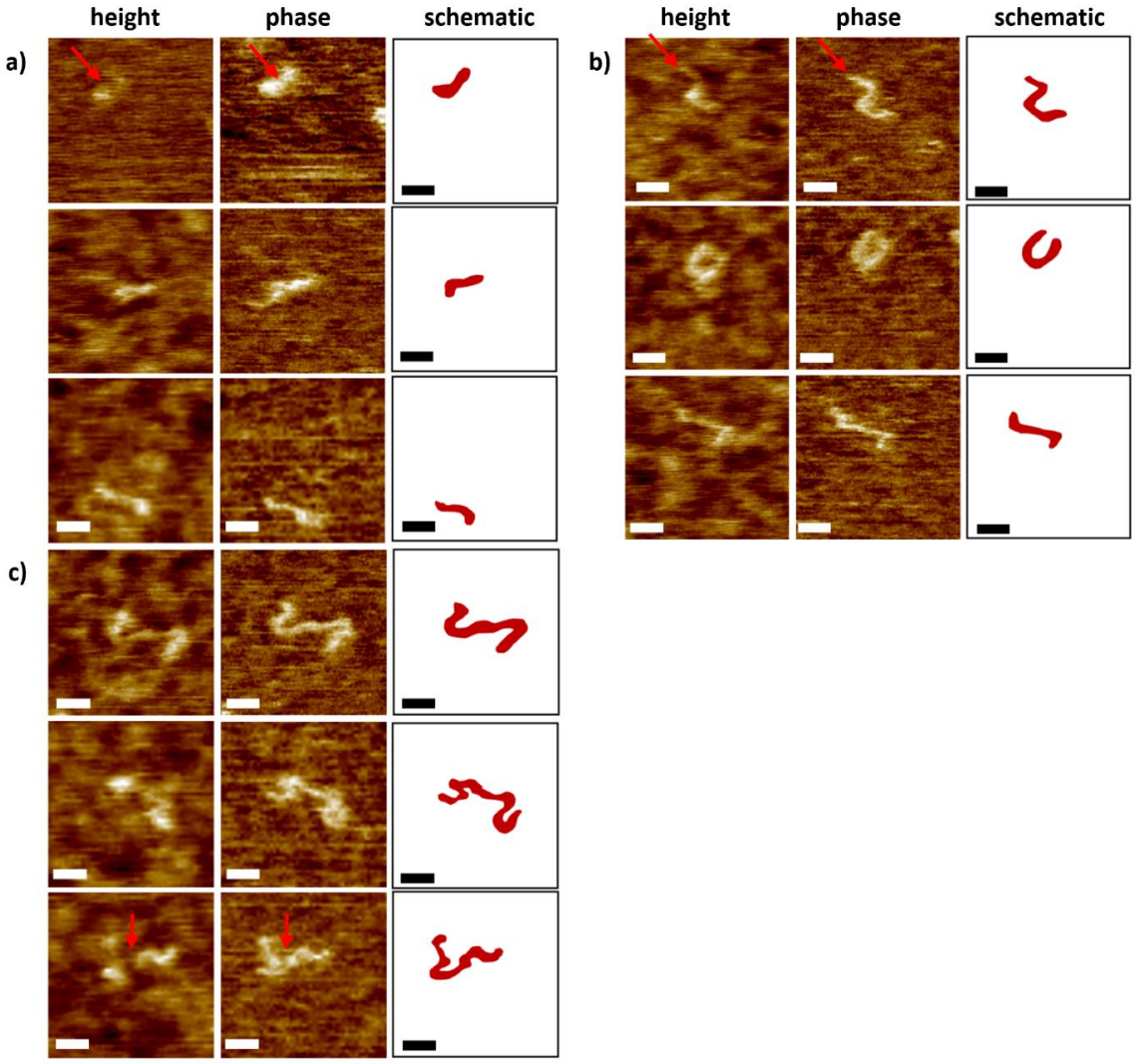


Figure S7: Comparison of AFM height and phase images of the l-83mer products.

Tapping mode (TM) AFM height images (height scale: 0.7 nm) and phase images (range: 0 – 30°) of the reaction products resulting from incubation of the l-83mer (isolated from CRZ-2 system) in cleavage/ligation buffer. Shown are 83mers (a), dimers (b) and trimers (c); scale bars correspond to 10 nm. For convenience, schematics have been included on the right side to help with the interpretation of the AFM images. For AFM analysis, samples were precipitated and resolved in 25 mM EDTA and 3.5 M urea (semi-denaturing conditions).

In most TM images, height and phase channels gave very similar results, i.e., they were in principle identical regarding RNA chain shape and regarding the obtained resolution. This holds for AFM tips with curvature radii down to approximately 4 nm (which was estimated from the resolution of the recorded AFM images) and applies for the majority of the measurements on CRZ-2, PDB-1, and PDB-1, so that for these ribozymes height images were given in the manuscript (Fig. 6). However, some AFM tips appeared to have a smaller

effective curvature radius, as indicated by a higher resolution in the phase channel and the tendency to deform/squeeze the ribozyme in the height image (compare the ribozyme structure at the positions indicated with red arrows in Fig. S7). The latter is understandable by the increased local pressure acting on the ribozyme if the interaction area is reduced (due to the reduction of the effective tip curvature radius). With these AFM tips the highest resolutions were achieved in this study, but with the sacrifice that sometimes parts of some ribozymes have been strongly deformed and are hard to see in the height channel, while the entire ribozyme is very well resolvable in the phase shift channel. As the l-83mer forms in principle only 3 products, such highly resolved images were obtained for all 3 products and therefore the phase images were given in the manuscript (Fig. 5), which enable the reader to get an impression of the ribozyme conformation with true nm resolution.

Estimation of activation energies (Figure 7, main text)

Figure 7 in the main paper shows a comprehensive view of the cleavage cascade for each of the experimentally tested ribozymes. We can distinguish three types of reaction steps *(i)* formation of reactive structures, *(ii)* dissociation of cleaved ends after ribozyme reaction, (iii) refolding of an unbound reaction product into a new reactive structure. Each of these steps is characterized by an activation free energy.

For *(i)* the Boltzmann probability of forming a reactive state is given by $\exp(-E_R/RT)/Z$, where E_R denotes the energy of the reactive state and Z is the partition function (see Equation [1], main text). Thus, the corresponding activation energy is the difference between the free energy of the reactive state and the ensemble free energy ($-RT\ln(Z)$). This activation energy is optimized through cost functionκ_1, energies to form reactive structures are therefore lower for all PBD molecules than for CRZ-2.

For *(ii)* and *(iii)* we approximate the best refolding path from the product conformation (reactive RNA dimer) to the next reactive species in the cascade. Finding the best refolding pathways is a computationally hard problem. The best direct refolding paths (i.e. paths of minimal length) can be estimated using the *findpath* heuristic (2). In order to get a better estimate of the energy barriers, we consider not only direct refolding paths but also detours via low-lying minima in the energy landscape. We computed low-lying minima of RNA landscapes with the program *barriers* (3) and selected the minimum free energy (MFE) conformation and up to three of the main alternative conformations. We then computed the direct refolding paths from the product conformation to each of these low-lying minima, from each low lying minimum the other and finally from each low-lying minimum to the reactive structure. The barriers along direct paths are computed as the difference between the worst energy along the refolding path the energy of the starting structure, the activation barriers *(ii)* and *(iii)* are selected such that the barrier of the total path is minimal. The dissociation barrier *(ii)* corresponds to the energy needed to dissociate the cleaved end, the refolding barrier *(iii)* describes the pathway from the unbound reaction product to the new reactive structure. The resulting values show that designed molecules often have to overcome higher dissociation barriers than CRZ-2.

Table S1: Sequences of self-processing RNAs CRZ2 and PBD1- 4. Sequenes are shown with separated fragments resulting according to the cleavage sites. Further processing occurs by intra- and intermolecular ligation of the central 83mers. In addition, intermolecular ligation of central 83mers still containing either the 5'- or the 3'-tail can take place (comp. Figure S1).

	5'-tail	central 83mer sequence	3'-tail
CRZ2	GGGAGAUCACA-cp	HO-GUCCUCUUUGACGGGGUUCCGUCAAAGAGAGAAGUGAACCAGAGAAACACACUUCGGUGGUAUAUUACCUGGUCCCCCUCACA-cp	HO-GUCCUCUUU
PBD1	GGGAGAGCACA-cp	HO-GUCGGAGUUGCCGCGUUAGCGGCGGUUCUAGAAGUGCCCCGCAGAAACAGCCAUAUGGCGUAUAUUACGCGGGAAAAAGCACA-cp	HO-GUCGGAACC
PBD2	GGGAGAGAACA-cp	HO-GUCGGUGGUGCCCCGUAAGGGGCGUCGCCAGAAGUUCGGACCAGAAACAGCCAAAAGGCGUAUAUUACGGUCCAAAAAGAACA-cp	HO-GUCGGCGAC
PBD3	GGGAGHACA-cp	HO-GUCCGGUUUACCGCUAAUGCGGUGGGUCGAGAAGUCUGAGCGAGAAACACAGUAUACUGGUAUAUUACCGCUCCAUAAAGGCA-cp	HO-GUCCGGCACCAAAA
PBD4	GGGAGACA-cp	HO-GUCCGGUUUACCGCUAAUGCGGUGGGUCGAGAAGUCUGAGCGAGAAACACAGGACACUGGUAUAUUACCGCUCCAUAAAGGCA-cp	HO-GUCCGGCACCAAAA

cp = 2', 3'-cyclic phosphate

Table S2: Klenow primer sequences for generation of double-stranded DNA templates to be used for enzymatic synthesis of RNAs and of an inactive dimer

	Klenow primer 1 including T7 RNA promoter sequence (in italics)	Klenow primer 2
PBD1	5'-*TAA TAC GAC TCA CTA TA* GGG AGA GCA CAG TCG GAG TTG CCG CGT TAG CGG CGG TTC TAG AAG TGC CCC GCA-3'	5'-GGT TGG CAC TGA GCT TTT TCC CGC GTA ATA TAC GCC ATA TGG CTG TTT CTG CGG GGC ACT TCT AGA ACC G-3'
PBD2	5'-*TAA TAC GAC TCA CTA TA* GGG AGA GAA CAG TCG GTG GTG CCC CGT AAG GGG CGT CGC CAG AAG TTC GGA CCA G-3'	5'-TCG CCG ACT GTT CTT TTT GGA CCG TAA TAT ACG CCT TTT GGC TGT TTC TGG TCC GAA CTT CTG GCG ACG-3'
PBD3	5'-*TAA TAC GAC TCA CTA TA* GGG AGA CAG TCC GGT TTA CCG CTA ATG CGG TGG GTC GAG AAG TCT GAG CGA GAA A-3'	5'-TTT TGG TGC CGG ACT GCC TTT ATG GAG CGG TAA TAT ACC AGT ATA CTG TGT TTC TCG CTC AGA CTT CTC GAC C-3'
PBD4	5'-*TAA TAC GAC TCA CTA TA* GGG AGA CAG TCC GGT TTA CCG CTA ATG CGG TGG GTC GAG AAG TCT GAG CGA GAA ACA-3'	5'-TTT TGG TGC CGG ACT GCC TTT ATG GAG CGG TAA TAT ACC AGT GTC CTG TGT TTC TCG CTC AGA CTT CTC G-3'
Inactive dimer	5'-*TAA TAC GAC TCA CTA TA* GGG AGA GGT GTT TCA GAC TCG AGA ACC AGA GAA TGA CAC GTA TGT GCA GGA TTA ACT GGT AAA ACT CTC ACA GCT GAA ACA CCT CTT TCG G-3'	5'-GGT CTA CGA GGA TGG TCA GGA TAA GGT CGC AAG GTT GGT GGC AGC ACG CAT TAG GAC CTT GAC TTC GCT CAC AGA CCG AAA GAG GTG TTT CAG CTG TGA GAG-3'
Full length RNA sequence of the inactive dimer	5'-GGG AGA GGU GUU UCA GAC UCG AGA ACC AGA GAA UGA CAC GUA UGU GCA GGA UUA ACU GGU AAA ACU CUC ACA GCU GAA ACA CCU CUU UCG GUC UGU GAG CGA AGU CAA GGU CCU AAU GCG UGC UGC CAC CAA CCU UGC GAC CUU AUC CUG ACC AUC CUC GUA GACC-3'	

References

1. Gan, H.H., Fera, D., Zorn, J., Shiffeldrim, N., Tang, M., Laserson, U., Kim, N. and Schlick, T. (2004) RAG: RNA-As-Graphs database--concepts, analysis, and features. *Bioinformatics*, **20**, 1285-1291.
2. Flamm, C., Hofacker, I.L., Maurer-Stroh, S., Stadler, P.F. and Zehl, M. (2001) Design of multistable RNA molecules. *RNA*, **7**, 254-265.
3. Flamm, C., Hofacker, I.L., Stadler, P.F. and Wolfinger, M.T. (2002) Barrier trees of degenerate landscapes. *Zeitschrift Fur Physikalische Chemie-International Journal of Research in Physical Chemistry & Chemical Physics*, **216**, 155-173.

BIBLIOGRAPHY

Rosalía Aguirre-Hernández, Holger H Hoos, and Anne Condon. Computational RNA secondary structure design: empirical complexity and improved methods. *BMC Bioinformatics*, 8(1):34, 2007.

Gregory S Allen, Andrey Zavialov, Richard Gursky, Måns Ehrenberg, and Joachim Frank. The cryo-EM structure of a translation initiation complex from escherichia coli. *Cell*, 121(5):703–712, 2005.

Amal S Abu Almakarem, Anton I Petrov, Jesse Stombaugh, Craig L Zirbel, and Neocles B Leontis. Comprehensive survey and geometric classification of base triples in RNA structures. *Nucleic Acids Research*, page gkr810, 2011.

Oliwia Andries, Tasuku Kitada, Katie Bodner, Niek N Sanders, and Ron Weiss. Synthetic biology devices and circuits for RNA-based'smart vaccines': a propositional review. *Expert Review of Vaccines*, 14(2):313–331, 2014.

Mirela Andronescu. *Computational approaches for RNA energy parameter estimation*. PhD thesis, University of British Columbia, 2008.

Mirela Andronescu, Anthony P Fejes, Frank Hutter, Holger H Hoos, and Anne Condon. A new algorithm for RNA secondary structure design. *Journal of Molecular Biology*, 336(3):607–624, 2004.

Mirela Andronescu, Zhi Chuan Zhang, and Anne Condon. Secondary structure prediction of interacting RNA molecules. *Journal of Molecular Biology*, 345(5):987–1001, 2005.

Jon Applequist and Vinayak Damle. Theory of the effects of concentration and chain length on helix-coil equilibria in two-stranded nucleic acids. *The Journal of Chemical Physics*, 39(10):2719–2721, 1963.

Stefan Badelt, Christoph Flamm, and Ivo L Hofacker. Computational design of a circular RNA with prion-like behaviour. In *ALIFE 14: The Fourteenth Conference on the Synthesis and Simulation of Living Systems*, volume 14, pages 565–568, 2014. doi: http://dx.doi.org/10.7551/978-0-262-32621-6-ch091.

Stefan Badelt, Christoph Flamm, and Ivo L Hofacker. Computational design of a circular RNA with prionlike behavior. *Artificial Life X*, pages 1–14, 2015a. doi: 10.1162/ARTL_a_00197.

Stefan Badelt, Stefan Hammer, Christoph Flamm, and Ivo L Hofacker. Thermodynamic and kinetic folding of riboswitches. In *Methods in Enzymology*, volume 553, pages 193–213. Elsevier, 2015b.

Christian Berens, Alison Thain, and Renée Schroeder. A tetracycline-binding RNA aptamer. *Bioorganic & Medicinal Chemistry*, 9(10):2549–2556, 2001.

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, 2006.

Athanasius F Bompfünewerer, Rolf Backofen, Stephan H Bernhart, Jana Hertel, Ivo L Hofacker, Peter F Stadler, and Sebastian Will. Variations on RNA folding and alignment: lessons from benasque. *Journal of Mathematical Biology*, 56(1-2):129–144, 2008.

Hans Bremer and Patrick P Dennis. Modulation of chemical composition and other parameters of the cell by growth rate. *Escherichia coli and Salmonella: cellular and molecular biology*, 2:1553–1569, 1996.

Philippe Brion and Eric Westhof. Hierarchy and dynamics of RNA folding. *Annual Review of Biophysics and Biomolecular Structure*, 26(1):113–137, 1997.

Anke Busch and Rolf Backofen. INFO-RNA–a fast approach to inverse RNA folding. *Bioinformatics*, 22(15):1823–1831, 2006.

Yimei Cai, Xiaomin Yu, Songnian Hu, and Jun Yu. A brief review on the mechanisms of miRNA regulation. *Genomics, Proteomics & Bioinformatics*, 7(4):147–154, 2009.

Michael Chamberlin and Janet Ring. Characterization of T7-specific ribonucleic acid polymerase I. general properties of the enzymatic reaction and the template specificity of the enzyme. *Journal of Biological Chemistry*, 248(6):2235–2244, 1973.

James Chappell, Melissa K Takahashi, and Julius B Lucks. Creating small transcription activating RNAs. *Nature Chemical Biology*, 2015.

Scott D Cohen and Alan C Hindmarsh. CVODE, a stiff/nonstiff ODE solver in C. *Computers in Physics*, 10(2):138–143, 1996.

Manchuta Dangkulwanich, Toyotaka Ishibashi, Lacramioara Bintu, and Carlos Bustamante. Molecular mechanisms of transcription through single-molecule experiments. *Chemical Reviews*, 114(6):3203–3223, 2014.

Ludmila V Danilova, Dmitri D Pervouchine, Alexander V Favorov, and Andrei A Mironov. RNAKinetics: a web server that models secondary structure kinetics of an elongating RNA. *Journal of Bioinformatics and Computational Biology*, 4(02):589–596, 2006.

Roumen A Dimitrov and Michael Zuker. Prediction of hybridization and melting for double-stranded nucleic acids. *Biophysical Journal*, 87(1):215–226, 2004.

Robert M Dirks, Milo Lin, Erik Winfree, and Niles A Pierce. Paradigms for computational nucleic acid design. *Nucleic Acids Research*, 32(4):1392–1403, 2004.

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, 2007.

David E Draper. A guide to ions and RNA structure. *RNA*, 10(3):335–343, 2004.

Christoph Flamm and Ivo L Hofacker. Beyond energy minimization: approaches to the kinetic folding of RNA. *Monatshefte für Chemie-Chemical Monthly*, 139(4):447–457, 2008.

Christoph Flamm, Walter Fontana, Ivo L Hofacker, and Peter Schuster. RNA folding at elementary step resolution. *RNA*, 6:325–338, 2000.

Christoph Flamm, Ivo L Hofacker, Sebastian Maurer-Stroh, Peter F Stadler, and Martin Zehl. Design of multi-stable RNA molecules. *RNA*, 7:254–265, 2001.

Christoph Flamm, Ivo L. Hofacker, Peter F. Stadler, and Michael T. Wolfinger. Barrier trees of degenerate landscapes. *Zeitschrift für Physikalische Chemie*, 216:155–173, 2002. doi: 10.1524/zpch.2002.216.2.155.

Cody Geary, Paul WK Rothemund, and Ebbe S Andersen. A single-stranded architecture for cotranscriptional folding of RNA nanostructures. *Science*, 345(6198):799–804, 2014.

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.

Daniel T Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81:2340–2361, 1977.

Hiroaki Gouda, Irwin D. Kuntz, David A. Case, and Peter A. Kollman. Free energy calculations for theophylline binding to an RNA aptamer: Comparison of MM-PBSA and thermodynamic integration methods. *Biopolymers*, 68(1):16–34, 2003.

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, 2014.

Susan L Heilman-Miller and Sarah A Woodson. Effect of transcription on folding of the tetrahymena ribozyme. *RNA*, 9(6):722–733, 2003.

Alan C Hindmarsh, Peter N Brown, Keith E Grant, Steven L Lee, Radu Serban, Dan E Shumaker, and Carol S Woodward. SUNDIALS: Suite of nonlinear and differential/algebraic equation solvers. *ACM Transactions on Mathematical Software (TOMS)*, 31(3):363–396, 2005.

Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie / Chemical Monthly*, 125:167–188, 1994. doi: doi: 10.1007/BF00818163.

Ivo L Hofacker, Christoph Flamm, Christian Heine, Michael T Wolfinger, Gerik Scheuermann, and Peter F Stadler. BarMap: RNA folding on dynamic energy landscapes. *RNA*, 16:1308–1316, 2010. doi: doi:10.1261/rna.2093310.

Christian Höner zu Siederdissen, Stefan Hammer, Ingrid Abfalter, Ivo L. Hofacker, Flamm Christoph, and Stadler Peter F. Computational design of RNAs with complex energy landscapes. *Biopolymers*, 99(12):1124–1136, 2013. doi: doi:10.1002/bip.22337.

J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing In Science & Engineering*, 9(3):90–95, 2007.

Martijn A Huynen, Peter F Stadler, and Walter Fontana. Smoothness within ruggedness: the role of neutrality in adaptation. *Proceedings of the National Academy of Sciences*, 93(1):397–401, 1996.

FJ Isaacs, DJ Dwyer, C Ding, DD Pervouchine, CR Cantor, and JJ Collins. Engineered riboregulators enable post-transcriptional control of gene expression. *Nature Biotechnology*, 22:841–7, Jul 2004.

Hervé Isambert and Eric D Siggia. Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proceedings of the National Academy of Sciences*, 97(12):6515–6520, 2000.

R. D. Jenison, S. C. Gill, A. Pardi, and B. Polisky. High-resolution molecular discrimination by RNA. *Science*, 263(5152):1425–1429, 1994.

Martin Jinek, Krzysztof Chylinski, Ines Fonfara, Michael Hauer, Jennifer A Doudna, and Emmanuelle Charpentier. A programmable dual-RNA–guided DNA endonuclease in adaptive bacterial immunity. *Science*, 337(6096):816–821, 2012.

Fiona M Jucker, Rebecca M Phillips, Scott A McCallum, and Arthur Pardi. Role of a heterogeneous free state in the formation of a specific RNA-theophylline complex. *Biochemistry*, 42(9):2560–2567, 2003.

Cheulhee Jung and Andrew D Ellington. Diagnostic applications of nucleic acid circuits. *Accounts of Chemical Research*, 47(6):1825–1835, 2014.

Kyozi Kawasaki. Diffusion constants near the critical point for time-dependent ising models. *Physical Review*, 145(1):224, 1966.

Peter Kerpedjiev. *Seeing Secondary, Sampling Tertiary: A parallel journey through the prediction and visualization of RNA tertiary and secondary structure*. PhD thesis, University of Vienna, 2016.

Peter Kerpedjiev, Stefan Hammer, and Ivo L. Hofacker. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. *Bioinformatics*, 31 (20):3377–3379, 2015a.

Peter Kerpedjiev, Christian Höner Zu Siederdissen, and Ivo L Hofacker. Predicting RNA 3D structure using a coarse-grain helix-centered model. *RNA*, 21(6):1110–1121, 2015b.

Fred Russell Kramer and Donald R Mills. Secondary structure formation during RNA synthesis. *Nucleic Acids Research*, 9(19):5109–5124, 1981.

Matthew H Larson, William J Greenleaf, Robert Landick, and Steven M Block. Applied force reveals mechanistic and energetic details of transcription termination. *Cell*, 132 (6):971–982, 2008.

Matthew H Larson, Robert Landick, and Steven M Block. Single-molecule studies of RNA polymerase: one singular sensation, every little step it takes. *Molecular Cell*, 41 (3):249–262, 2011.

Jean-François Lemay, Guillaume Desnoyers, Simon Blouin, Benoit Heppell, Laurène Bastet, Patrick St-Pierre, Eric Massé, and Daniel A Lafontaine. Comparative study between transcriptionally-and translationally-acting adenine riboswitches reveals key differences in riboswitch regulatory mechanisms. *PLoS Genetics*, 7(1):e1001278–e1001278, 2011.

Neocles B Leontis and Eric Westhof. Geometric nomenclature and classification of RNA base pairs. *RNA*, 7(04):499–512, 2001.

Ronny Lorenz. *RNA Secondary Structure Thermodynamics and Kinetics*. PhD thesis, University of Vienna, 2014.

Ronny Lorenz, Stefan H Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. ViennaRNA package 2.0. *Algorithms Mol Biol*, 6:26, 2011. doi: 10.1186/1748-7188-6-260.

Maumita Mandal and Ronald R Breaker. Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nature Structural & Molecular Biology*, 11(1): 29–35, 2004.

Nicholas R Markham and Michael Zuker. Unafold: Software for nucleic acid folding and hybridization. In *Bioinformatics*, pages 3–31. Springer, 2008.

David H Mathews, Mark E Burkard, Susan M Freier, Jacqueline R Wyatt, and Douglas H Turner. Predicting oligonucleotide affinity to nucleic acid targets. *RNA*, 5(11): 1458–1469, 1999.

Jan Maňuch, Chris Thachuk, Ladislav Stacho, and Anne Condon. NP-completeness of the energy barrier problem without pseudoknots and temporary arcs. *Natural Computing*, 10(1):391–405, 2011. doi: doi:10.1007/s11047-010-9239-4.

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

Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, 2003.

Steven R Morgan and Paul G Higgs. Barrier heights between ground states in a model of RNA secondary structure. *Journal of Physics A: Mathematical and General*, 31(14): 3153, 1998.

Ulrike Mückstein, Hakim Tafer, Stephan H Bernhart, Maribel Hernandez-Rosales, Jörg Vogel, Peter F Stadler, and Ivo L Hofacker. *Translational control by RNA-RNA interaction: Improved computation of RNA-RNA binding thermodynamics*. Springer, 2008.

Michael Müller, Julia E Weigand, Oliver Weichenrieder, and Beatrix Suess. Thermodynamic characterization of an engineered tetracycline-binding riboswitch. *Nucleic Acids Research*, 34(9):2607–2617, 2006.

T Neilson, PJ Romaniuk, D Alkema, DW Hughes, JR Everett, and RA Bell. The effects of base sequence and dangling bases on the stability of short ribonucleic acid duplexes. In *Nucleic Acids Symposium Series*, number 7, pages 293–311, 1979.

Ruth Nussinov and Ann B Jacobson. Fast algorithm for predicting the secondary structure of single-stranded RNA. *Proceedings of the National Academy of Sciences*, 77 (11):6309–6313, 1980.

Sonja Petkovic, Stefan Badelt, Stephan Block, Christoph Flamm, Mihaela Delcea, Ivo L Hofacker, and Sabine Müller. Sequence-controlled RNA self-processing: computational design, biochemical analysis and visualization by AFM. *RNA*, 21:1249–1260, 2015. doi: http://dx.doi.org/10.1261/rna.047670.114.

Dietmar Pörschke. Thermodynamic and kinetic parameters of an oligonucleotide hairpin helix. *Biophysical Chemistry*, 1(5):381–386, 1974.

Dietmar Pörschke and Manfred Eigen. Co-operative non-enzymatic base recognition iii. kinetics of the helix-coil transition of the oligoribouridylic – oligoriboadenylic acid system and of oligoriboadenylic acid alone at acidic ph. *Journal of Molecular Biology*, 62(2):361–381, 1971.

Jeff R Proctor and Irmtraud M Meyer. CoFold: an RNA secondary structure prediction method that takes co-transcriptional folding into account. *Nucleic Acids Research*, page gkt174, 2013.

John W Randles, Gerhard Steger, and Detiev Riesner. Structural transitions in viroid-like RNAs associated with cadang-cadang disease, velvet tobacco mottle virus, and solanum nodiflorum mottle virus. *Nucleic Acids Research*, 10(18):5569–5586, 1982.

Marc Rehmsmeier, Peter Steffen, Matthias Höchsmann, and Robert Giegerich. Fast and effective prediction of microrna/target duplexes. *RNA*, 10(10):1507–1517, 2004.

Christian Reidys, Peter F Stadler, and Peter Schuster. Generic properties of combinatory maps: neutral networks of RNA secondary structures. *Bulletin of Mathematical Biology*, 59(2):339–397, 1997.

Jessica S Reuter and David H Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC Bioinformatics*, 11(1):129, 2010.

Michael P Robertson and Gerald F Joyce. Highly efficient self-replicating RNA enzymes. *Chemistry & Biology*, 21(2):238–245, 2014.

John SantaLucia Jr and Donald Hicks. The thermodynamics of DNA structural motifs. *Annual Reviews of Biophysics and Biomolecular Structure*, 33:415–440, 2004.

Ben Sauerwine and Michael Widom. Folding kinetics of riboswitch transcriptional terminators and sequesterers. *Entropy*, 15(8):3088–3099, 2013.

Joseph M Schaeffer. *Stochastic simulation of the kinetics of multiple interacting nucleic acid strands*. PhD thesis, Caltech, 2013.

Joseph M Schaeffer, Chris Thachuk, and Erik Winfree. Stochastic simulation of the kinetics of multiple interacting nucleic acid strands. In Andrew Phillips and Peng Yin, editors, *DNA Computing and Molecular Programming*, volume 9211, pages 133–153. Springer International Publishing, 2015.

Michael Schmitz and Gerhard Steger. Description of RNA folding by "simulated annealing". *Journal of molecular biology*, 255(1):254–266, 1996.

Michael Schnall-Levin. *RNA: algorithms, evolution and design*. PhD thesis, Massachusetts Institute of Technology, 2011.

Schrödinger, LLC. The PyMOL molecular graphics system, version 1.3r1. August 2010.

Peter Schuster, Walter Fontana, Peter F Stadler, and Ivo L Hofacker. From sequences to shapes and back: a case study in RNA secondary structures. *Proceedings of the Royal Society of London B: Biological Sciences*, 255(1344):279–284, 1994.

Alexander Serganov and Evgeny Nudler. A decade of riboswitches. *Cell*, 152(1):17–24, 2013.

Niranjan Srinivas, Thomas E Ouldridge, Petr Šulc, Joseph M Schaeffer, Bernard Yurke, Ard A Louis, Jonathan PK Doye, and Erik Winfree. On the biophysics and kinetics of toehold-mediated DNA strand displacement. *Nucleic Acids Research*, 41(22):10641–10658, 2013.

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, 2009.

Zhi-Jie Tan and Shi-Jie Chen. Nucleic acid helix stability: effects of salt concentration, cation valence and size, and chain length. *Biophysical Journal*, 90(4):1175–1190, 2006.

Zhi-Jie Tan and Shi-Jie Chen. RNA helix stability in mixed Na^{+}/Mg^{2+} solution. *Biophysical Journal*, 92(10):3615–3632, 2007.

Akito Taneda. Multi-objective optimization for RNA design with multiple target secondary structures. *BMC Bioinformatics*, 16(1):1, 2015.

Douglas H Turner and David H Mathews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, 2009.

Manja Wachsmuth, Sven Sven Findeiß, Nadine Weissheimer, Peter F. Stadler, and Mario Mörl. *De novo* design of a synthetic riboswitch that regulates transcription termination. *Nucleic Acids Research*, 41(4):2541–2551, 2013. doi: doi:10.1093/nar/gks1330.

Manja Wachsmuth, Gesine Domin, Ronny Lorenz, Robert Serfling, Sven Findeiß, Peter F Stadler, and Mario Mörl. Design criteria for synthetic riboswitches acting on transcription. *RNA Biology*, 12(2):221–231, 2015.

Amy E Walter, Douglas H Turner, James Kim, Matthew H Lyttle, Peter Müller, David H Mathews, and Michael Zuker. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proceedings of the National Academy of Sciences*, 91(20):9218–9222, 1994.

Michael Waskom, Olga Botvinnik, Paul Hobson, John B. Cole, Yaroslav Halchenko, Stephan Hoyer, Alistair Miles, Tom Augspurger, Tal Yarkoni, Tobias Megies, Luis Pedro Coelho, Daniel Wehner, cynddl, Erik Ziegler, diegoo020, Yury V. Zaytsev, Travis Hoppe, Skipper Seabold, Phillip Cloud, Miikka Koskinen, Kyle Meyer, Adel Qalieh, and Dan Allan. seaborn: v0.5.0 (november 2014), November 2014. URL http://dx.doi.org/10.5281/zenodo.12710.

J Kenneth Wickiser, Ming T Cheah, Ronald R Breaker, and Donald M Crothers. The kinetics of ligand binding by an adenine-sensing riboswitch. *Biochemistry*, 44(40): 13404–13414, 2005.

Blake Wiedenheft, Samuel H Sternberg, and Jennifer A Doudna. RNA-guided genetic silencing systems in bacteria and archaea. *Nature*, 482(7385):331–338, 2012.

Michael T. Wolfinger, W. Andreas Svrcek-Seiler, Christoph Flamm, Ivo L. Hofacker, and Peter F. Stadler. Efficient computation of RNA folding dynamics. *Journal of Physics A: Mathematical and General*, 37:4731–4741, 2004. doi: 10.1088/0305-4470/37/17/005.

Terrence N Wong, Tobin R Sosnick, and Tao Pan. Folding of noncoding RNAs during transcription facilitated by pausing-induced nonnative structures. *Proceedings of the National Academy of Sciences*, 104(46):17995–18000, 2007.

Stefan Wuchty, Walter Fontana, Ivo L Hofacker, and Peter Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2): 145–165, 1999.

A Xayaphoummine, T Bucher, and H Isambert. Kinefold web server for RNA/DNA folding path and structure prediction including pseudoknots and knots. *Nucleic Acids Research*, 33(suppl 2):W605–W610, 2005.

A Xayaphoummine, V Viasnoff, S Harlepp, and H Isambert. Encoding folding paths of RNA switches. *Nucleic Acids Research*, 35(2):614–622, 2007.

Zhen Xie, Liliana Wroblewska, Laura Prochazka, Ron Weiss, and Yaakov Benenson. Multi-input RNAi-based logic circuit for identification of specific cancer cells. *Science*, 333(6047):1307–1311, 2011.

Bernard Yurke and Allen P Mills Jr. Using DNA to power nanostructures. *Genetic Programming and Evolvable Machines*, 4(2):111–122, 2003.

Joseph N Zadeh, Conrad D Steenberg, Justin S Bois, Brian R Wolfe, Marshall B Pierce, Asif R Khan, Robert M Dirks, and Niles A Pierce. NUPACK: analysis and design of nucleic acid systems. *Journal of Computational Chemistry*, 32(1):170–173, 2011.

Jesse G Zalatan, Michael E Lee, Ricardo Almeida, Luke A Gilbert, Evan H Whitehead, Marie La Russa, Jordan C Tsai, Jonathan S Weissman, John E Dueber, Lei S Qi, et al. Engineering complex synthetic transcriptional programs with CRISPR RNA scaffolds. *Cell*, 160(1):339–350, 2015.

David Yu Zhang and Erik Winfree. Control of DNA strand displacement kinetics using toehold exchange. *Journal of the American Chemical Society*, 131(47):17303–17314, 2009.

Libin Zhang, Penghui Bao, Michael J Leibowitz, and Yi Zhang. Slow formation of a pseudoknot structure is rate limiting in the productive co-transcriptional folding of the self-splicing candida intron. *RNA*, 15(11):1986–1992, 2009.

Wenbing Zhang and Shi-Jie Chen. RNA hairpin-folding kinetics. *Proceedings of the National Academy of Sciences*, 99(4):1931–1936, 2002.

Wenbing Zhang and Shi-Jie Chen. Exploring the complex folding kinetics of RNA hairpins: I. general folding kinetics analysis. *Biophysical Journal*, 90(3):765–777, 2006a.

Wenbing Zhang and Shi-Jie Chen. Exploring the complex folding kinetics of RNA hairpins: II. effect of sequence, length, and misfolded states. *Biophysical Journal*, 90 (3):778–787, 2006b.

Peinan Zhao, Wenbing Zhang, and Shi-Jie Chen. Cotranscriptional folding kinetics of ribonucleic acid secondary structures. *The Journal of Chemical Physics*, 135(24):245101, 2011.

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

Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406–3415, 2003.

Curriculum Vitae **Jan, 2016**

Stefan Badelt

http://www.tbi.univie.ac.at/~stef

stef@tbi.univie.ac.at

Educational Background

2011 – *present*	Graduate studies in Molecular Biology, University of Vienna Ph.D. thesis with Prof. Ivo Hofacker: *Control of RNA function by conformational design*
2004 – 2011	Undergraduate studies in Molecular Biology, University of Vienna Master's thesis with Prof. Ivo Hofacker: *RNA folding kinetics including pseudoknots* Graduation with distinction, Mag. rer. nat.

Professional Experience

2011/10 – *present*	**Ph.D. thesis** – Control of RNA function by conformational design with Ivo L. Hofacker – Theoretical Biochemistry *Institute for Theoretical Chemistry, Vienna, Austria*
WS2013 – SS2015	**Teaching** – Exercises for Foundations of Bioinformatics *University of Vienna, Austria*
2009/05 – 2011/09	**Master's thesis** – RNA folding kinetics including pseudoknots with Ivo L. Hofacker – Theoretical Biochemistry *Institute for Theoretical Chemistry, Vienna, Austria*
2008/07 – 2008/09	**Internship** – Chromosome degradation in apoptotic cells with Reinhard Ullmann – Molecular Cytogenetics *Max Planck Institute for Molecular Genetics, Berlin, Germany*
2008/03 – 2008/04	**Internship** – Interaction of Stat1-GRDBD-Stat1 and GRE with Pavel Kovarik – Infection Biology *Max F. Perutz Laboratories, Vienna, Austria*
2006/07 – 2009/03	**Technician** – Plasmid library administration, genotyping *Max F. Perutz Laboratories, Vienna, Austria* Group Kovarik – Infection Biology

Skills

- Languages: German (native), fluent English,
- Computer Skills: Perl, Python, Bash, Latex, R, C, C++
- Lab-Techniques: PCR, Real Time PCR, Tissue Culture work (including Nucleofection), Immunfluorescence, Immunoprecipitation, Nuclear Extract, Western Blot Analysis, Electrophoretic Mobility Shift Assay, DNA/RNA Extraction, DNA/RNA Gel Electrophoresis, Reverse Transcription, Array CGH, Oligoarray, BAC Array, ChIP on Chip.
- Snowboard and Windsurfing instructor

Publications

[1] S. Badelt, C. Flamm, and I. L. Hofacker, "Computational design of a circular RNA with prionlike behavior," *Artificial Life X*, pp. 1–14, 2015.

[2] S. Petkovic, S. Badelt, S. Block, C. Flamm, M. Delcea, I. L. Hofacker, and S. Müller, "Sequence-controlled RNA self-processing: computational design, biochemical analysis and visualization by AFM," *RNA*, vol. 21, pp. 1249–1260, 2015.

[3] S. Badelt, S. Hammer, C. Flamm, and I. L. Hofacker, "Thermodynamic and kinetic folding of riboswitches," in *Methods in Enzymology*, vol. 553, pp. 193–213, Elsevier, 2015.

[4] S. Badelt, C. Flamm, and I. L. Hofacker, "Computational design of a circular RNA with prion-like behaviour," in *ALIFE 14: The Fourteenth Conference on the Synthesis and Simulation of Living Systems*, vol. 14, pp. 565–568, 2014.

[5] G. Ebert, A. Steininger, R. Weißmann, V. Boldt, A. Lind-Thomsen, J. Grune, S. Badelt, M. Heßler, M. Peiser, M. Hitzler, *et al.*, "Distribution of segmental duplications in the context of higher order chromatin organisation of human chromosome 7," *BMC genomics*, vol. 15, no. 1, p. 537, 2014.

[6] M. Marz, A. R. Gruber, C. Höner zu Siederdissen, F. Amman, S. Badelt, S. Bartschat, S. H. Bernhart, W. Beyer, S. Kehr, R. Lorenz, A. Tanzer, D. Yusuf, H. Tafer, H. I. L, and P. F. Stadler, "Animal snoRNAs and scaRNAs with exceptional structures," *RNA*, vol. 8, no. 6, pp. 938–946, 2011.

Selected Talks & Posters

- *TBI Winterseminar* in Bled, Slovenia, Feb 15 - 22, 2015
  Talk: BarMap & SundialsWrapper – advanced RNA folding kinetics
- *Gordon Research Conference on RNA nanotechnology* in Ventura, USA, Feb 1 - 6, 2015
  Poster: Design of XOR riboswitches
- *Artificial Life Conference* in New York, USA, Jul 30 - Aug 2, 2014
  Talk: Design of a circular RNA with prion-like behavior
- *International Synthetic and Systems Biology Summer School* in Taormina, Italy, Jun 15 - 19, 2014
  Poster: Sequence-controlled RNA self-processing: computational design, biochemical analysis and visualization by AFM
- *TBI Winterseminar* in Bled, Slovenia, Feb 16 - 23, 2014
  Talk: Folding kinetics of self-polymerizing RNA
- *Herbstseminar Bioinformatik* in Decin, Czech Republic, Oct 2 - 7, 2013
  Talk: Circularization and multimerization of synthetic ribozymes
- *TBI Winterseminar* in Bled, Slovenia, Feb 13 - 20, 2011
  Talk: Energy barriers in pseudoknot conformation space
- *Herbstseminar Bioinformatik* in Vysoká Lípa, Czech Republic, Oct 5 - 10, 2010
  Talk: Design & future aspects of artificial RNA-switches in synthetic biology

COLOPHON

This document was typeset using the typographical look-and-feel `classicthesis` developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". `classicthesis` is available for both LaTeX and LyX:

http://code.google.com/p/classicthesis/

Happy users of `classicthesis` usually send a real postcard to the author, a collection of postcards received so far is featured here:

http://postcards.miede.de/

Final Version as of February 4, 2016 (`classicthesis` version 0.70).