universität wien

DISSERTATION

Titel der Dissertation

„RNA Secondary Structure Thermodynamics and Kinetics“

verfasst von

Dipl.-Inform. Ronny Lorenz

angestrebter akademischer Grad

Doktor der Naturwissenschaften (Dr.rer.nat.)

Wien, 2014

Studienkennzahl lt. Studienblatt:	A 091 490
Dissertationsgebiet lt. Studienblatt:	Molekulare Biologie
Betreut von:	Univ.-Prof Dipl.-Phys. Dr. Ivo L. Hofacker

RNA SECONDARY STRUCTURE THERMODYNAMICS AND KINETICS

RONNY LORENZ

A journey to Vienna with RNA structures and Schnitzel

August 2014

ABSTRACT

RNA molecules fold back onto themselves through interaction of complementary segments of the linear sequence. This process which leads to the biologically functional shape of an RNA molecule is called *"folding"*. The formed secondary structure captures most of the folding free energy and can be formally described by a set of base pairs.

Computational methods exist that allow one to predict ground states, base pairing probabilities, and other thermodynamic properties of RNA secondary structures with relatively high accuracy. These algorithms are based on experimentally determined enthalpy and entropy parameters.

Besides the ground state, an RNA can adopt a multitude of different structures that taken together constitute a conformational ensemble. By introducing a notion of structural neighborhood the ensemble and the free energies of its members form an energy landscape.

This sets the scene to investigate the folding and refolding dynamics of RNA molecules. Biological phenomena like regulation of gene expression by riboswitches, attenuators in prokaryotic cells, or structural changes of the genomic RNA of viroids and RNA-viruses require an understanding of refolding paths and their kinetics.

In my PhD thesis I present novel heuristic algorithms to predict optimal folding paths and folding behavior on top of an abstract view of the energy landscape using distance class partitioning. Great value is set on the applicability of these algorithms to RNA molecules of biologically significant sequence lengths. By exploration of variable and static energy landscapes with stochastic and deterministic techniques the new approaches can even be used to predict folding kinetics during transcription.

Furthermore, I developed an extension to existing RNA-RNA interaction prediction tools that allows their application to RNA/DNA heterodimers, which for instance occur in the transcription process. To provide an efficient and accurate framework for all my secondary structure based analyses, I implemented the latest thermodynamic energy models into the `ViennaRNA Package`. In addition, *G-Quadruplexes* were introduced into several programs of the `ViennaRNA Package` and constitute the first structural motif beyond classic secondary structures.

These new methods may help to better understand the relation between structural features of the energy landscape and their impact on the folding kinetics of RNA molecules. Therefore, this might open the path to *in silico* design of RNA molecules with prescribed folding behavior, knowledge that will facilitate research in highly interdisciplinary areas such as Synthetic Biology.

ZUSAMMENFASSUNG

RNA-Moleküle falten auf sich selbst zurück durch Interaktion komplementärer Segmente der linearen Sequenz. Dieser Prozess, der zu der biologisch funktionellen Form eines RNA-Moleküls führt, wird als *"Faltung"* bezeichnet. Die dabei gebildete Sekundärstruktur umfasst den größten Teil der freie Energie und kann formal durch eine Menge von Basenpaaren beschrieben werden.

Bestehende Berechnungsmethoden erlauben relativ genaue Vorhersagen von Grundzuständen, Basenpaarungswahrscheinlichkeiten und anderen thermodynamischen Eigenschaften von RNA Sekundärstrukturen. Diese Algorithmen basieren auf experimentell bestimmten Enthalpie- und Entropieparametern.

Neben den Grundzustand kann eine RNA eine Vielzahl verschiedener Strukturen annehmen, welche gemeinsam eine Ensemble verschiedenster Konformationen darstellen. Durch die Einführung des Begriffs der 'strukturellen Nachbarschaft' bildet dieses Ensemble, zusammen mit der freien Energien seiner Elemente, eine Energielandschaft.

Diese Konzepte erlauben nun, Faltungs- und Umfaltungsdynamiken von RNAs zu untersuchen. Biologische Phänomene, wie die Regulation der Genexpression durch RNA-Schalter, prokaryotische Attenuatoren oder strukturelle Veränderungen der genomischen RNA von Viroiden und RNA-Viren erfordern detailiertes Verständnis von Umfaltungspfaden und deren Kinetik.

In meiner Doktorarbeit präsentiere ich neue heuristische Algorithmen basierend auf einer abstrakten Darstellung der Energielandschaft mittels Distanzklassenpartitionierung, um optimale Faltungspfade und Faltungsverhalten vorherzusagen. Großer Wert wird hierbei darauf gelegt, diese Algorithmen auf RNA-Moleküle von biologisch signifikanter Sequenzlänge anwenden zu können. Durch Untersuchung variabler und statischer Energielandschaften mittels stochastischer und deterministischer Methoden können diese neuen Ansätze auch zur Vorhersage co-transkriptioneller Faltungskinetiken verwendet werden.

Darüber hinaus wurde eine Erweiterung vorhandener Methoden zur RNA-RNA-Interaktionsvorhersage entwickelt, sodaß diese auch auf RNA/DNA Heterodimere, die beispielsweise während des Transkriptionsprozesses auftreten, angewandt werden können. Um effiziente und genaue Rahmenbedingungen für alle meine Sekundärstrukturanalysen zu schaffen, implementierte ich die neuesten thermodynamischen Energiemodelle in das `ViennaRNA Package`. Weiters wurden *G*-Quadruplexe in mehrere Programme des `ViennaRNA Package` eingeführt und bilden somit das erste Strukturmotiv, das über die klassischen Sekundärstrukturen hinausgeht.

Diese neuen Verfahren können helfen, die Beziehung zwischen Strukturmerkmalen der Energielandschaft und ihren Einfluß auf Faltungskinetik von RNA-Molekülen bes-

ser zu verstehen. Weiters könnten dadurch neue Wege in Richtung *in silico* Design von RNAs mit bestimmten Faltungsverhalten eröffnet werden und somit in stark interdisziplinären Bereichen wie der Synthetischen Biologie Anwendung finden.

ACKNOWLEDGMENTS

Sincere thanks to everybody who helped me to succeed in writing this thesis.

First and foremost I want to thank my supervisor Ivo L. Hofacker, who was always patient and helpful and guided me through the many years of my PhD study. His scientific competence, and that of his group, allowed me to thoroughly investigate new fields of research.

Very special thanks go to my beloved Andrea who constantly supported me, even in times when my life seemed to consist only of work. Without her help on real-life related problems and my research, I wouldn't have managed to finish this part of my study.

Furthermore, I want to thank all my colleagues and collaborators at the TBI, in Leipzig, and elsewhere who provided me many new viewpoints and insights into many problems of RNA bioinformatics. Among them are Peter Stadler, Xtof, Berni, MTW, Yann, Christian, Stef, Peter, Marcel, Egg, Fall, and so many more!

Last but not least I thank my family and friends, especially my mother, my brother, my grandmother, my father, Pete and Jette, and everybody who is not mentioned here, for perpetually asking me when I finally finish this study. I can tell you: with this work I have just begun.

CONTENTS

LIST OF FIGURES

ACRONYMS

NMP nucleoside mono-Phosphate

NDP nucleoside di-Phosphate

NTP nucleoside tri-Phosphate

DP Dynamic Programming

MFE minimum free energy

MEA Maximum Expected Accuracy

RNA ribonucleic acid

ncRNA non-coding RNA

miRNA micro RNA

rRNA ribosomal RNA

tRNA transfer RNA

sRNA small RNA

snRNA small nuclear RNA

siRNA small interfering RNA

snRNP small nuclear ribonucleic particles

snoRNA small nucleolar RNA

mRNA messenger RNA

DNA deoxy-ribonucleic acid

UTR untranslated region

Part I

RNA IN VIVO, IN VITRO, AND IN SILICO

1 INTRODUCTION

1.1 BIOLOGICAL BACKGROUND AND MOTIVATION

Amongst the most prominent macro molecules in living cells are the nucleic acids deoxy-ribonucleic acid (DNA) and ribonucleic acid (RNA). While DNA serves as information storage forming a template for the transcription of RNA, it was long assumed that RNA itself merely represents an intermediate messenger RNA (mRNA) on the way to the final gene product, a *protein*. Therefore, its only purpose is to provide information flow from the DNA to the protein production facilities of the cell, the *ribosomes*. This model is well known as the *central dogma of molecular biology* [45, 44], depicted in Figure 1.

However, as it turned out, RNA is capable to do much more. Already at the time when the central dogma was proposed in the 1970s, it was found that two types of RNA, namely transfer RNA (tRNA) and ribosomal RNA (rRNA), are rather exceptional, since their nucleotide sequence does code for any proteins. Instead, they interact with the ribosome and play an active role in the *translation* of the genetic code into proteins [163]. Later, many more so-called non-coding RNA (ncRNA)s were discovered that, although being transcribed, do not encode proteins.

In the early 1980s, several labs discovered RNAs with catalytic activity, so-called *ribozymes*, that able to cleave and ligate their own phosphate backbone (*autocatalytic ribozymes*) or use other RNAs as substrates. This has been shown for instance for the RNA component of *ribonuclease P* [90], the *hairpin ribozyme* [225], and the *hammerhead ribozyme* [229, 26]. A more detailed investigation of the ribosome revealed, that the RNA component, the rRNA, is the active part of the ribosome that catalyzes the peptide bond between two consecutive amino acids [171, 193].

Moreover, a small nuclear RNA (snRNA) is typically bound in complexes of specific proteins, the so called small nuclear ribonucleic particles (snRNP), and plays a significant role in enzymatic reactions of RNA intron splicing, maintenance of DNA telomeres, as well as the regulation of transcription in the nucleus [62, 194, 139]. In fact, RNA is able to catalyze a wide variety of reaction types, including phosphoryl group transfer, isomerisation of C-C bonds and hydrolysis [219, 26, 24].

In 1986, Thomas Cech proposed *"A model for the RNA-catalyzed replication of RNA"* [36]. This ground breaking work showed, for the first time, that RNA itself may act as a self-replicating system without the need for any auxiliary protein enzymes. Thus, RNA may have been among the first self-sustaining macro molecules in a *prebiotic world*, prior to the emergence of the last universal common ancestor (LUCA). Its capability

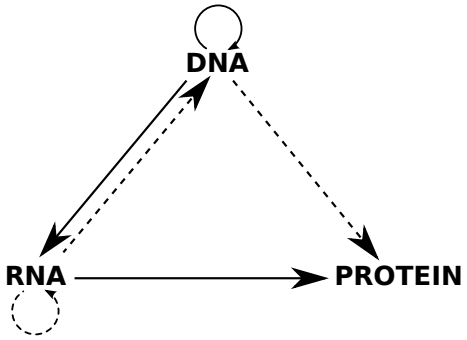


Figure 1: The *Central Dogma of Molecular Biology*. Solid and dashed arrows show general and special transfer of information, respectively. Information transfer that was not detected by 1970 [44] is indicated by missing arrows. Despite being outdated due to new insights in the role of RNA on cellular processes, the depicted dogma still is quite popular.

to store information, replicate itself, and act as a catalytic entity opens the possibility of a precellular evolution,where RNAs may have been one basic step at the origin of life, known as the *RNA World hypothesis* [23, 50, 83, 101, 120, 35]. This hypothesis is also supported by the findings of quasi-species effects in RNA *viruses*, subviral circular RNA plant-pathogens (*viroids*) [43, 98], and evolutionary effects of other short, replicating RNA sequences [144].

The last feature of RNAs that I want to mention here, is in my opinion also the most intriguing: their role in gene regulation. For many years it was widely believed that regulation and control over of gene expression is mainly determined by proteins and enzymes, both on transcriptional and translational level. However, within the last decade, more and more evidence has been provided which suggests, that RNA itself acts as the major driving force in gene regulation and thus servers as an additional RNA based layer of gene regulation [165]. Bacterial riboswitches, for instance, use transient structures allowing them to sense small ligand molecules or temperature changes in their environment, in order to activate or inactive gene expression [243]. This mode of gene regulation is widely used amongst bacteria and archaea, and can act on transcriptional, as well as translational level [235]. Many bacterial mRNAs contain regulatory elements in their 5′ untranslated region (UTR) that cotranscriptionally fold into structures which in turn tightly sequester the *Shine-Dalgarno* (SD) sequence, and thus inhibit translation [22, 89, 232]. A complex example for an RNA based control of gene expression provides the *Hok-Sok system* [222, 78] in *Escherichia coli*. This system constitutes a control mechanism to ensure correct cell division and plasmid stabilization. Both, the ncRNA Sok (suppression-of-killing) and the toxin encoding Hok gene (host-killing) reside on a plasmid, named R1. The principle of this system can be sketched as follows: Sok binds to hok mRNA and prevents translation of HOK pro-

tein. In the absence of Sok, Hok is expressed and the bacterium dies. The interesting part of this system, however, lies in the details. During transcription, the Hok mRNA becomes trapped in a translationally inactive form that masks the SD sequence. The fully transcribed Hok mRNA then refolds into a further inactive state that is subjected to $3'$ exonucleolytic cleavage. As degradation proceeds, it triggers several RNA structure refolding events that finally convert the mRNA into an active form, where the SD sequence becomes accessible to the ribosome. Hence, translation of the HOK protein is initiated leading the death of the cell through depolarization of the cell membrane. At this step Sok ncRNA antidote comes into play. In bacteria carrying this gene, Sok is constitutively expressed and can bind proximal to the SD sequence of the active form of Hok mRNA. This interaction finally triggers degradation of Hok by RNaseIII, thus keeping the cell viable.

1.2 THE STRUCTURE OF THIS THESIS

Within this thesis, four papers are presented, that set the foundation for my main focus of research: to detect, understand, predict, and possibly design *cis*-acting multistable RNA structures (see also Chapter 4). Of course, the methods presented here are applicable to a wide range of RNA secondary structure related analyses. On top of these methods and tools, the potential of applications will be discussed exemplarily for three important bioinformatics analysis, that have not yet been published [1]:

1. The detection of meta stable secondary structures, i.e. RNAs with the potential to switch between at least two low free energy conformations.

2. A method for coarse grained simulation if RNA folding dynamics that utilizes *ab initio* partitioning of the secondary structure space.

3. Cotranscriptional folding simulation and the detection of *kinetic traps*.

The most important toolset for *in silico* RNA secondary structure analysis is a framework that allows efficient and accurate predictions. Such a toolset is provided by the `ViennaRNA Package` [104, 137]. Although initially developed about 20 years ago, it still is one of the most accurate and fastest Dynamic Programming (DP) implementations for RNA secondary structure prediction [75, 137]. The programs included in the `ViennaRNA Package` can be used for a variety of secondary structure analyses, as for instance (i) simple minimum free energy (MFE) structure prediction for single sequences, sequence alignments, and interacting RNAs [104, 20, 19], (ii) genome-wide scans for interactions between a long target RNA and a (semi-) complementary short query sequence [213], and (iii) the generation of sets of suboptimal secondary structures by three different approaches[264, 255, 54]. The core of the `ViennaRNA Package` is a library of implemented algorithms written in the programming language `C`, that allows third party applications to easily interface with the very fast realizations of the corresponding methods. For rapid prototyping, this library is also exported to the scripting language `Perl`[2].

However, its a long and winding road from the development of new models to their later application. Much effort is necessary to include them into large software projects. This was especially true for the `ViennaRNA Package`. Over the past two decades, the group around David Mathews made an enormous effort to measure melting temperatures for a comprehensive set of small RNA structures [227, 152, 150, 228]. These constitute the source of the free energy parameters used in virtually all physics based secondary structure prediction algorithms to-date. Moreover, their enduring effort to revise previous experiments resulted in more accurate, yet more complex models. The

1 Manuscript in preparation

2 By the time of writing this thesis I also included an interface for the scripting language `Python`, which is now included in the distributed source code archive of the `ViennaRNA Package`

ViennaRNA Package as of the year 2008, i.e. in the beginning of my years of PhD study, implemented an energy model that got a bit long in the tooth. Already four years had passed since a new, more accurate model had been proposed [151].

To bring the `ViennaRNA Package` and its powerful set of tools back on track was therefore one of the top priorities on my to-do list, during my time at the TBI, the home of the `ViennaRNA Package`. But as it turned out, that over the years contributions by dozens of researchers introduced thousands of lines of redundant code. Implementing the latest energy model into every single application of the `ViennaRNA Package` would have required me to repeat the same task over and over again. Thus, I decided to reimplement major parts of the software package and unify the code base to simplify eventual future updates. The result of this work is presented in my publication *"ViennaRNA Package 2.0"* [137], in Part iv, Chapter 7.

The development of methods that allow one to investigate multi-stable RNA secondary structures requires many different techniques. Starting with a static view on the ensemble of secondary structure, my first steps were towards a more generalized view on the idea of *distance class partitioning* using a reference structure as introduced by Freyhult et al. [72]. This allowed me to develop an algorithm that not only predicts meta-stable low free energy structures, but also opens up new ways to tackle various problems requiring a deeper understanding of the entire secondary structure space. The `RNA2Dfold` algorithm partitions the structure space into distance classes with respect to two references, while exploiting the sparseness of the resulting DP matrices. The latter permits analysis of relatively long RNAs as found in many biological model systems.

My publication *"2D Projections of RNA folding Landscapes"* [136] sets the foundation for my work on distance classes, see Chapter 9. Potential applications for the concept of distance class partitioning are exemplarily discussed in much detail in Chapters 11, and 12. They include methods for the estimation of refolding energy barriers, and the prediction RNA folding dynamics, respectively. Moreover, efficient methods to simulate RNA folding dynamics offer the potential to investigate the folding behavior of RNAs under varying environmental settings. One such environmental setting is chain growth during transcription, which is why I discuss *cotranscriptional folding* and its dynamics in the context of distance class partitioning within the last section of Chapter 12.

As further outlined in Chapters 11, 12, and 13, the ideas of distance class partitioning are not restricted to only two reference structures. In fact, they can be applied to an arbitrary number of reference structures. However, the computational effort for this purpose would render most ideas practically impossible. Thus, I suggest an alternative, computationally cheaper method in the last part of this thesis (see *"Conclusions and Outlook"*, Chapter 13).

Since RNA molecules in living cells are produced by a process called *transcription*, where an enzyme, the *RNA polymerase*, assembles the nucleotides of the RNA accord-

ing the sequence of a DNA template, the interplay between RNA and DNA is of uttermost importance. This is due to the fact, that an already transcribed RNA segment of about 8 nt, that still resides within the *transcription bubble*, stays attached to its complementary DNA template. Depending on the strength of this interaction, the RNA/DNA hybrid is able to slow down the transcription process or even terminate it [174, 260]. Furthermore, RNA/DNA hybrids have been found to be involved in many more contexts such as gene silencing, genome stability, and DNA damage [10, 3, 113].

Thus, efficient tools are required to predict the interactions, determine their stability, and investigate the nature of RNA/DNA hybrids. Such tools must take the concurrent *intramolecular* and *intermolecular* interactions between and within both nucleic acids into account [138] (See Chapter 8, *"Prediction of RNA/DNA hybrid structure"*). Consequently, the predicted thermodynamic properties allow for many applications, such as the prediction of *R-loops* which are known to be involved in genome instability and DNA damage [10, 3, 113]. Additionally, such a method also allows one to develop more sophisticated *cotranscriptional* folding algorithms. Here, the constant transcription speed used in virtually all available methods to-date may be replaced by a more elaborate approach that governs the transcription speed based on the interaction between RNA and DNA [259].

The restricted perspective of secondary structure prediction itself rules out many structural elements that are necessary to understand the functional relationship between an RNAs structure and its environment [256, 29]. However, the extension of the well established and efficient method of secondary structure prediction towards tertiary interactions of RNAs is difficult, both in terms of development of efficient models, and their accurate parametrization. A brief introduction to this topic is given in Chapter 3.8, followed by an outlook to future perspectives in the last chapter. In fact, many of the tertiary interactions can not be efficiently integrated into today's structure prediction algorithms. Nonetheless, some self-contained structural elements such as *G-Quadruplexes* [249], whose thermodynamic properties have been experimentally evaluated [156, 261] make it possible to do exactly that: incorporate RNA modules into fast and efficient secondary structure prediction algorithms.

Such an extension is provided with my work *"2D meets 4G: G-Quadruplexes in RNA Secondary Structure Prediction"*, as included in Chapter 10. Since G-Quadruplexes are extraordinary stable tertiary structure elements which control gene regulation on the mRNA level as parts of the 5′ UTR, act as recognition sites for RNA-protein interactions of transcripts involved in brain development [212], and maintain RNA stability by preventing degradation through *exonucleases* in the 3′ UTR, they constitute an integral part of the life-cycle of many RNAs [249]. The potential of the described method for G-Quadruplex prediction using secondary structure prediction algorithms is further outlined in the last part of this thesis, Chapter 13.

This thesis starts with an historical overview of the advances in RNA secondary structure prediction over the last four decades. Although within the last ten years or so

many probabilistic approaches on RNA secondary structure prediction emerged [124, 57, 96], I restrict myself here to an in-depth discussion of physics (thermodynamic) based DP approaches.Therefore, Chapter 2 sets the mathematical foundations that are necessary for the development of all structure prediction algorithms, which are explained in detail in Chapter 3. The subsequent Chapters 4, 5, and 6 then extend from the prospect of single secondary structures to the connections and dynamics between them. Part iv includes the most essential publications for my hitherto research, as outlined earlier in this section. The last part, Chapters 11, 12, and 13, includes a broad discussion about the potential and future perspectives of the methods that I developed for this PhD thesis.

2 RNA SECONDARY STRUCTURES

2.1 NUCLEIC ACID STRUCTURES

From the biochemical point of view, nucleic acids such as DNA or RNA are linear or circular biopolymers that consist of a sequence of nucleotides connected by a phosphodiester bond. The nucleotides themselves are composed of three organic molecules, a nitrogenous base, a five-carbon sugar, and a phosphate group. Monomers lacking the phosphate group are termed *nucleosides*. In the unpolymerized state, nucleotides may carry more than a single phosphate group. Accordingly, nucleic acid monomers carrying a single, two, or three phosphate groups are named nucleoside mono-Phosphate (NMP), nucleoside di-Phosphate (NDP), and nucleoside tri-Phosphate (NTP), respectively. The main difference between nucleic acids is in the ribose, precisely in its C2′ atom. While RNA has a hydroxyl group attached to this position, this group is substituted by a hydrogen atom in DNA, hence the name deoxyribonucleic acid. Furthermore, both biopolymers distinguish only a four letter alphabet in terms of the consecutive sequence of their nitrogenous bases, the purines *adenine* (A) and *guanine* (G), and the pyrimidines *cytosine* and *thymine* (T), where the latter is replaced by *uracil* (U) in RNA (see Figure 2). In nature, nucleotides are polymerized by an enzyme, the RNA polymerase, that catalyzes the phosphodiester bond between the C3′ atom of one, and the C5′ atom of another ribose part. Thus, a chain molecule of nucleotides has two free ends, 5′ and 3′, and it grows from its 5′ to 3′ end.

As a convention biochemistry discriminates four levels of abstraction for biopolymers that consist of a recurring, relatively small number of distinct monomers, the *primary*, *secondary*, *tertiary*, and *quaternary* structure. The first three of them are the most relevant in RNA biology, exemplified for the *hammerhead ribozyme* in Figure 3.

PRIMARY STRUCTURE The primary structure is just the sequence of monomers in a defined reading direction along the polymer. For RNA this is the sequence of ribonucleotides A, G, C, and U, read from the 5′ to 3′ end, since this is its most natural reading direction. Typical sequence lengths of RNA primary structures vary from about 22 nucleotides, for the short micro RNA (miRNA), to several thousand for some mRNA transcripts.

SECONDARY STRUCTURE Similar to the two complementary strands that form the double helix of DNA in all living organisms, RNA can form base pairs as well. However, RNA is typically found as a single molecule, where complementary parts of the

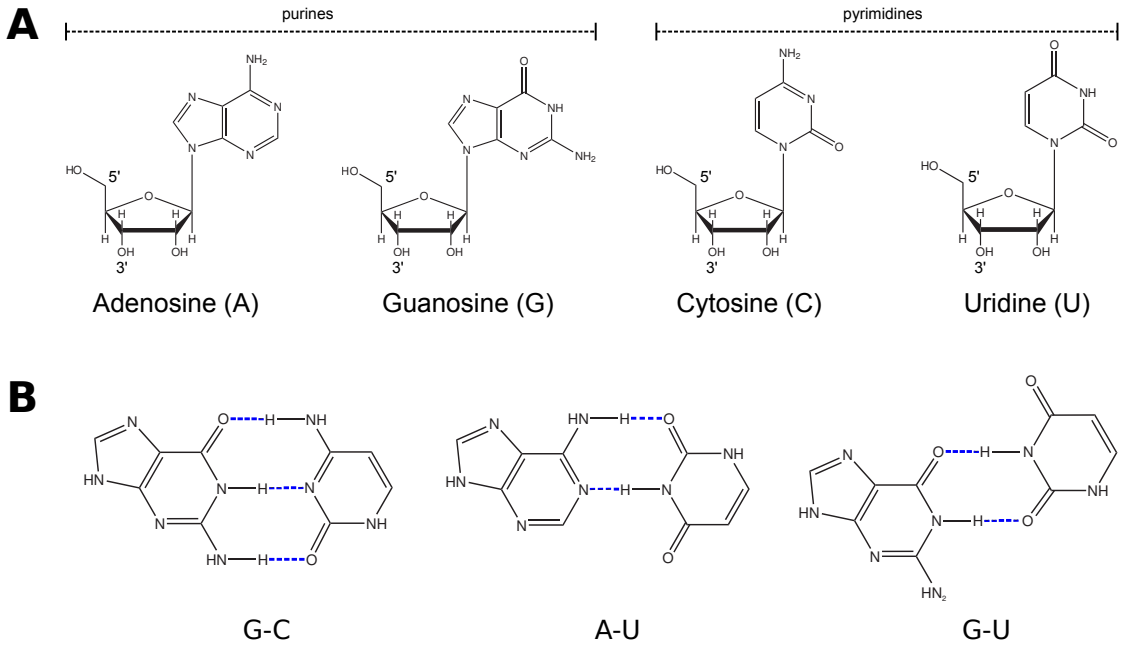


Figure 2: The building blocks of an RNA. **A** shows the nucleosides with the riboses in the lower left part, labeled at their C5′ and C3′ atom. Attached to the C1′ atom of the ribose are the nitrogenous bases adenine, guanine, cytosine, and uracil. **B** lists the three canonical base pairs, the Watson-Crick base pairs G-C, and A-U, as well as the Wobble pair G-U. Dashed lines between the bases indicate hydrogen bonds.

same sequence allow it to fold back on itself. As in DNA, the most abundant base pairs in RNAs are the so-called *Watson-Crick* base pairs, named after the heads of the team of scientists who discovered the DNA helix in 1953 [244]. They suggested that DNA consists of a helix of two complementary strands, bound together by hydrogen bonds between two opposing nucleotides. Particularly, a pyrimidine pairs with a purine, where A pairs with T, and G pairs with C, forming two and three hydrogen bonds, respectively. This finally answered the observation made by the chemist E. Chargaff, who found in the late 1940s, that the relative quantities of the DNA nucleotides in cells from diverse organisms are equally correlated [38]. In 1954, A. Rich and J.D. Watson proposed that RNA is able to form base pairs as well, having A-U pairs instead of A-T [189]. Furthermore, RNA frequently forms an additional, energetically weaker base pair between a G and a U, the so-called *Wobble pair* [233].

This set of three types of base pairs, the Watson-Crick base pairs *G-C* and *A-U*, and the Wobble pair *G-U*, as depicted in Figure 2, are assumed to largely determine an RNAs spatial arrangement, and may therefore be used to describe an RNAs conformation in the most conceptional way, the *secondary structure*. Still, this limited number of base pairs, the so-called *canonical base pairs*, are not the full story. In fact, there

are many more types of base pairs that belong to the so-called group of *non-canonical pairs* [247, 210]. However, they are generally treated as tertiary interactions and will be only briefly discussed in Chapter 3.8. A further restriction of the arrangement of these base pairs, conventionally used in bioinformatics dealing with RNA structure prediction, is introduced in the next section.

TERTIARY STRUCTURE The complete three dimensional conformation of an RNA contributes to the *tertiary structure*. It is usually given as the coordinates of its atoms in space, but also intramolecular interactions beyond pairs of nucleotides, such as *base triples*, are commonly accepted to be part of the tertiary, instead of the secondary structure. The same applies to *knotted* structures and *pseudo knots*, which are generally treated as tertiary interactions (see Chapters 2.2 and 3.8).

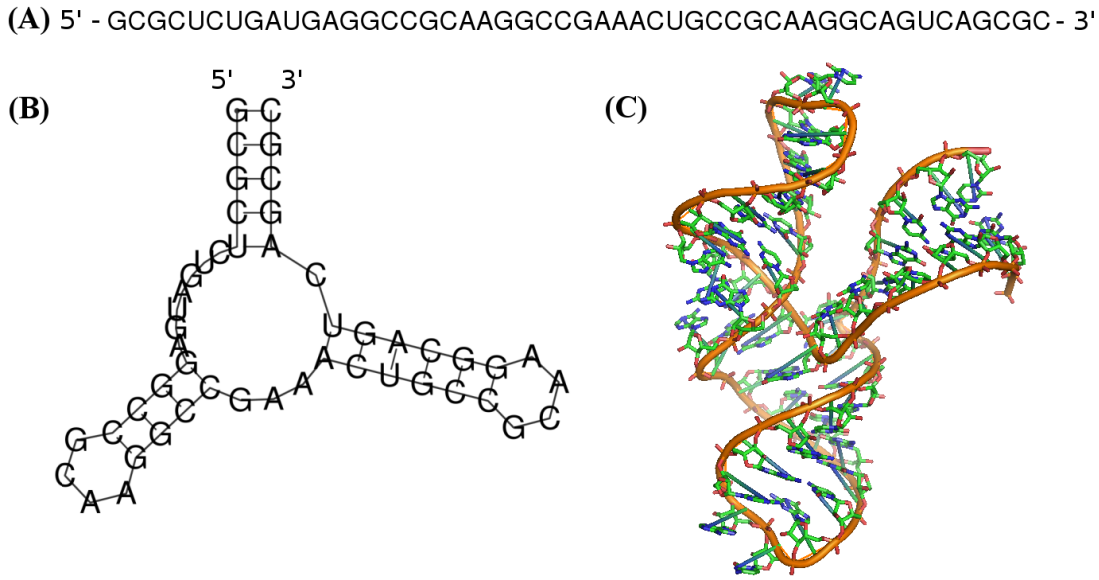


Figure 3: Structures of a 49 nucleotide RNA hammerhead ribozyme taken from [229]. **(A)** Primary structure as supplied in the article. **(B)** Secondary structure predicted with `RNAfold` of the `ViennaRNA Package` [104] **(C)** Tertiary structure displayed by `PyMOL`, 3-D structure based on PDB-ID 1RMN

2.2 ON RNA SECONDARY STRUCTURES AND GRAPHS

THE (FORMAL) DEFINITION OF A SECONDARY STRUCTURE As mentioned earlier, the primary structure of an RNA, the RNA sequence $\sigma = (N_1, \ldots, N_n)$, is an ordered set of letters $N_i \in \Sigma$ from $5'$ to $3'$ end, where the alphabet $\Sigma = \{A,G,C,U\}$, and n is the sequence length. The secondary structure s, is the set of base pairs, i.e. the set

of pairs (i, j) of sequence positions i and j that form hydrogen bonds between each other. Therefore, from a mathematical point of view, an RNA secondary structure can be considered a (labeled) graph $G = \{V, E\}$, where the nucleotides, particularly their positions in the sequence, constitute the vertex set $V = \{1, \ldots, n\}$. The edge set $E = \{(i, j)\}$, on the other hand, is derived from (i) the strong covalent bond between two consecutive nucleotides $(i, i+1)$, representing the *phosphate backbone* of the RNA, and (ii) its pairs of hydrogen bonds between two complementary nucleotides, the base pairs (i, j). As a convention, an RNA secondary structure is considered *valid*, if it complies with the following definition.

Definition 2.1 *A valid secondary structure graph* $G = \{V, E\}$ *can be described by an adjacency matrix* A *with entries* $a_{i,j} = 1$ *for each* $(i, j) \in E$*, and fulfills the following properties [240, 242]:*

1. $a_{i,i+1} = 1$ *for* $1 \leqslant i < n$.
2. *For each fixed* i, $1 \leqslant i \leqslant n$, *there is at most one* $j \neq i \pm 1$ *for which* $a_{i,j} = 1$.
3. *If* $a_{i,j} = a_{k,l} = 1$ *and* $i < k < j$ *and* $i \neq l \neq j$, *then* $i < l < j$.

Naturally speaking, the above definition ensures the following three conditions: 1. the phosphate backbone is part of the graph, 2. any nucleotide is involved in at most one base pair, and 3. base pairs do not cross each other. As a consequence, G is a *planar graph*, i.e. it can be drawn on a two-dimensional plane without crossing of its edges. Moreover, G is *outerplanar*, since drawn on the plane, no vertex is completely surrounded by edges. The last condition of Definition 2.1 basically excludes non-nested structures, so called *pseudo-knots*, which will be discussed in Chapter 3.8. This comes in handy for the prediction of secondary structures, as described in the following parts of this thesis.

Furthermore, the steric configuration of an RNA does not allow for base pairs (i, j) with $j - i \leqslant 4$, thus a minimal hairpin loop length of 3 will be assumed under all circumstances within this thesis. If not mentioned otherwise, the set of allowed base pairs that constitute a secondary structure is restricted to the six so-called *canonical base pairs*, i.e. the Watson-Crick pairs G-C,C-G,A-U, and U-A, as well as the Wobble pairs G-U, and U-G (see Figure 2). Therefore, a secondary structure s is considered *compatible* with a sequence σ if it is (i) valid, (ii) consists of canonical base pairs only, and (iii) no hairpin loop has a length shorter than 3.

NOTATION Throughout this entire thesis, I make use of some short, formal notations for sequences and secondary structures. This enables me to write down mathematical concepts and algorithms in a more compact and convenient way. While most of them are defined in the context of their appearance, the most frequently-used will be introduced here.

The i^{th} nucleotide N_i of an RNA sequence σ will most often be denoted as $\sigma[i]$, while entire subsequences from nucleotide N_i to N_j are written as $\sigma[i:j] = (N_i, \ldots, N_j)$. A pair of sequence positions i and j put in parentheses marks hydrogen bonding between the two nucleotides N_i and N_j, i.e. the base pair (i, j). For secondary structures, the lower case symbol $s = \{(i, j) \mid 1 \leqslant i < j \leqslant n\}$ will be used to identify the set of base pairs (i, j) they consist of. Analogous to subsequences, the notation $s[i:j]$ indicates the subset of base pairs $\{(k, l) \in s \mid i \leqslant k < l \leqslant j\} = s[i:j] \subseteq s$. For the ensemble of all secondary structures s compatible with a particular sequence σ, I will use the shorthand symbol Ω.

Additionally, several distance measures to assess the similarity of two secondary structures s_1 and s_2 have been proposed [214, 195, 200, 201, 27, 104]. The most frequently used among them is the *base pair distance* $d_{BP}(s_1, s_2)$, which is the number of base pairs the two structures do not have in common, i.e. the size of their *symmetric difference* $s_1 \triangle s_2$, hence

$$d_{BP}(s_1, s_2) = |s_1 \triangle s_2| = |s_1 \cup s_2| - |s_1 \cap s_2| \tag{1}$$

with

$$s_1 \triangle s_2 = \{(p, q) \mid ((p, q) \in s_1) \wedge ((p, q) \notin s_2)) \vee (((p, q) \in s_2) \wedge ((p, q) \notin s_1))\}$$

2.3 RNA SECONDARY STRUCTURE REPRESENTATIONS

Since RNA secondary structures are planar graphs, they can be more or less easily drawn on a plane. Indeed, they have been depicted in a multitude of publications for more than 50 years [71]. While most secondary structure drawings reveal common properties, for instance equidistantly drawn nucleotides and hydrogen bonds, the difference in the underlying layout algorithm can be large. Of course, hand-drawn images are still used in scientific literature, too, but generally (semi-)automated approaches are the preferred way for image generation. Here, I present some of the most commonly used forms of secondary structure representations.

CIRCULAR DRAWING Many structure plotting algorithms derive their layout from an initial circular representation of the molecule, that was already suggested in 1978 by Nussinov et al. [173]. Therefore, all nucleotides are placed equidistantly on a circle that depicts the phosphate backbone of the RNA. Base pairs are then highlighted by arcs between the corresponding nucleotides within the interior of the circle, see Figure 4 for an example. Alternatively, the circle can be replaced by a straight line with the 5′ and 3′ ends on its left and right, respectively. Such a representation is known as *linked graph representation*. Both kinds of this drawings can be easily generated in an automated manner.

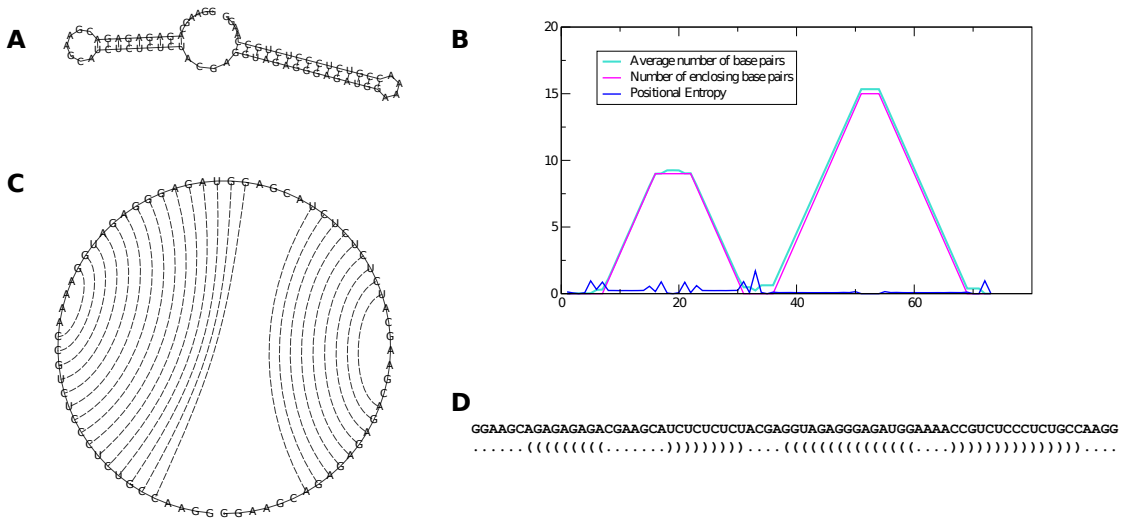


Figure 4: Different representations of a the secondary structure of an artificially designed RNA [258]. All programs used are part of the `ViennaRNA Package`[104, 135]. **A** "Classical plot" (radial drawing) as obtained from the `RNAfold` program **B** Mountain plot depicting the auxiliary information of average number of enclosing base pairs, i.e. $y(x) = \sum_{i<x<j} p_{ij}$ with base pair probabilities p_{ij}. Furthermore, the positional entropy $s_i = -\sum_j p_{ij} \cdot \ln(p_{ij})$ is shown. Base pair probabilities were computed with `RNAfold -p`. See Chapters 3.4 and 3.5 for details. **C** Circular drawing as obtained from the program `RNAplot -t2`. **D** Sequence and Dot-bracket annotation taken from the output of `RNAfold`.

RADIAL DRAWINGS The first more realistic drawings generated by computer algorithms appeared in the early 1980s [126]. However, theses early attempts produced many crossing helices in larger structures. Meanwhile, many more sophisticated algorithms were developed [202, 30, 12]. Still, the problem of generating overlap-free secondary structure plots that allow to easily compare structural features while being visually pleasing is not yet solved and leaves further space for improvements. The more recent approaches are usually implemented as standalone applications which sometimes also allow for user interaction, e.g. rotation and translation of helices and nucleotides [48, 47, 97]. These programs take sequence/structure pairs for input to produce a drawing. However, most RNA secondary structure prediction programs implement the *NAVIEW* algorithm developed by Bruccoleri and Heinrich [30], which constitutes the more or less *classical plot* of an RNA secondary structure. It extends the circular drawing by not placing the whole sequence on a circle but the central loop, which usually is the loop with the most branches. It then extrudes the branching helices toward the outside in optimal angles to allow enough space for the substructures, whose layout is determined iteratively. Still, this method is not perfect and may pro-

duce overlapping secondary structure parts. A radial drawing of an exemplary RNA secondary structure is given in Figure 4.

MOUNTAIN PLOTS A totally different kind of secondary structure representation is the so-called *mountain plot*, shown in Figure 4. This representation was suggested by Hogeweg and Hesper [107] and it allows to visually compare RNA secondary structures in a most simplistic way. The construction of this 2D plot is straight-forward. The x-axis denotes the nucleotide positions from the $5'$ to $3'$ end. Every base pair (i, j) with $i < j$ introduces a move in y-direction, where position i and j are associated with the moves 'upwards', and 'downwards', respectively. Any unpaired nucleotide does not change the value of y. Therefore, the y-coordinate at any nucleotide position, the height of the mountain, reflects the number of enclosing base pairs. A *peak* indicates a *hairpin*, where the unpaired nucleotides form a *plateau* enclosed by symmetric slopes representing the enclosing stem. A plateau interrupting a sloped region indicates a *bulge* if there is no other plateau of the same height on the other side of the mountain. Otherwise it depicts an *interior loop*. *Valleys* represent the unpaired nucleotides of a *multi loop*, and the *exterior loop* depending on whether they are above zero, i.e. $y > 0$, or not, respectively (See also Chapter 3.2).

DOT-BRACKET NOTATION A much more compact visualization that is both human- and machine-readable is provided by the *dot-bracket* notation. Here, a string of length n composed of the characters `'.'`, `'('`, and `')'` represents the pairing pattern of an RNA sequence with n nucleotides. Base pairs (i, j) are represented by a matching pair of *parentheses* at position i and j while unpaired nucleotides are denoted as *dots*. An example is shown in Figure 4.

3

RNA SECONDARY STRUCTURE PREDICTION

3.1 DECOMPOSITION OF SECONDARY STRUCTURES AND ITS IMPLICATIONS

The definition of secondary structures, as introduced in the previous chapter 2.1, reveals an intriguing property. Considering a sub structure $s[i:j]$ within the sequence interval $\sigma[i:j]$, there are only two alternatives how position i may contribute to $s[i:j]$. Either i has no pairing partner, or i pairs with another nucleotide k with $i < k \leqslant j$. In the first case, where i is unpaired, $s[i:j]$ consists of the base pairs in $s[i+1:j]$ only, i.e. $s[i:j] = s[i+1:j]$. Formation of a base pair (i,k), however, subdivides the structure into two parts, one enclosed by (i,k), namely $s[i+1:k-1]$, and the other adjacent to it, $s[k+1:j]$. Thus $s[i:j] = s[i+1:k-1] \cup s[k+1:j] \cup \{(i,k)\}$. Since condition (3) of 2.1 ensures that $s[i:j]$ can not contain base pairs that 'cross' (i,k), the two shorter substructures $s[i+1:k-1]$ and $s[k+1:j]$ can be treated independently for a large variety of purposes.

These observations lead to a recursive decomposition scheme for RNA secondary structures that is the basic step for a large variety of DP approaches that solve RNA secondary structure related problems. A graphical representation of the idea can be found in Figure 5.

NUMBER OF SECONDARY STRUCTURES The first application that made use of this recursive decomposition scheme targeted the number of secondary structures $N_{1,n}$ compatible with a given sequence $\sigma[1:n]$ [240, 242]. While structural alternatives add up to $N_{i,j}$, the combinatorial possibilities to form structures out of shorter and independent substructures demands for a multiplicative construction. This leads to the recursive relationship

$$N_{i,j} = N_{i+1,j} + \sum_{\substack{i<k\leqslant j \\ (\sigma[i],\sigma[k])\in\mathcal{B}}} \left(N_{i+1,k-1} \cdot N_{k+1,j}\right) \tag{2}$$

with

$$N_{i,i} = 1 \text{ and } \mathcal{B} = \{(G,C),(C,G),(A,U),(U,A),(G,U),(U,G)\},$$

i.e. only *canonical* Watson-Crick and Wobble base pairs are allowed to contribute to the set of solutions.

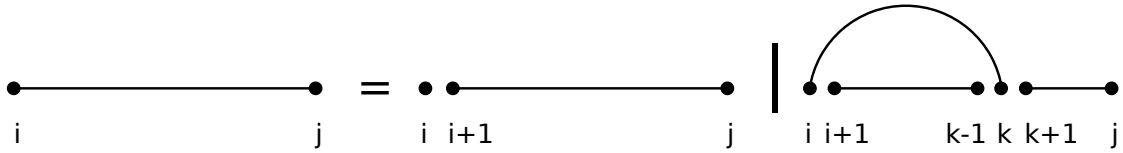


Figure 5: Any secondary (sub)structure $s[i:j]$ can be recursively decomposed into smaller parts by considering only one particular nucleotide, e.g. N_i. The vertical bar in the right-hand side of the figure separates the possibilities that emerge. Either this nucleotide position i is unpaired and the decomposition yields the substructure $s[i+1:j]$ (left part of the right-hand side), or i is paired with some other position k (right part of the right-hand side). For the latter, the $s[i:j]$ decomposes to the base pair (s,k), the substructure $s[i+1:k-1]$ it encloses, and the remaining part next to (i,k), namely $s[k+1:j]$.

Rewriting recursion (2) in terms of the number of structures S_n compatible with a sequence of length n, and minimal hairpin loop length m leads to

$$S_{n+1} = S_n + \sum_{i=m}^{n-1} S_k S_{n-k-1} \tag{3}$$

with

$$n \geqslant m+1 \text{ and } S_0 = S_1 = \ldots = S_{m+1} = 1.$$

This is most useful, since it allows to derive an asymptotic value $\psi(S_n) \approx 1.8^n$ for the number of secondary structures compatible with an RNA sequence of length n and equally probable nucleotides [265, 105].

STRUCTURES WITH MAXIMIZED NUMBER OF BASE PAIRS In the same year that Waterman et al. presented their recursive decomposition scheme for RNA secondary structures, Ruth Nussinov released the first efficient DP algorithm capable of predicting optimal secondary structures [173]. Here, the optimality criterion is a *maximization of base pairs*, hence, the algorithm solves the so called *maximum matching* problem. The decomposition scheme introduced in the beginning of this section recursively divides the space of possible structures within an interval $[i:j]$ into combinations of independent substructures. Now, a (sub)structure $s[i:j]$ maximizes the number of base pairs, if its decomposition leads to substructures with a maximum number of base pairs again. Therefore, *Bellman's principle of optimality* [15] holds, and a DP algorithm can be developed to compute the maximum number of base pairs E in a secondary structure s compatible with a sequence σ.

Similar to the above structure counting algorithm, $E_{i,j}$ is selected from a set of combinations of maximal subsolutions. The maximum number of base pairs for a (sub)structure $s[i:j]$ is either the same as for $s[i+1:j]$ (i is unpaired), or if i pairs with some k, the score is the sum of $E_{i+1,k-1}$, $E_{k+1,j}$, and 1 for the pair (i,k). This results in an $\mathcal{O}(n^3)$ algorithm that requires $\mathcal{O}(n^2)$ memory for sequences of length n.

$$E_{i,j} = \max \left\{ \begin{array}{c} E_{i+1,j}, \\ \max_{\substack{i < k \leqslant j \\ (s[i], s[k]) \in \mathcal{B}}} \left\{ E_{i+1,k-1} + E_{k+1,j} + S(i,k) \right\} \end{array} \right\} \tag{4}$$

with

$$S(i,k) = 1$$

Once $E_{1.n}$ is computed, the corresponding secondary structure s_{max} can be determined by application of a recursive *backtracking* procedure. In principle, backtracking is the reverse operation of the above *forward recursion* (4) that, instead of constructing an optimal solution $E_{i,j}$, traces back the subsolutions that constitute $E_{i,j}$. For instance, if $E_{i,j} = E_{i+1,j}$ nucleotide i can be considered unpaired and backtracking continues in $E_{i+1,j}$. If the opposite is true, i.e. $E_{i,j} \neq E_{i,j}$, then i must be paired with some k. As soon as k is determined, the base pair (i,k) is put into s_{max} and backtracking continues for $E_{i+1,k-1}$ and $E_{k+1,j}$. Thus, to obtain the complete structure $s[1:n]$ with maximum number of base pairs, backtracking starts for $E_{1,n}$ and $s_{max} = \emptyset$.

A simple specialization of $S(i,k)$ into a base pair type dependent scoring function

$$S(i,k) = \begin{cases} 1 & \text{if } (s[i], s[k]) \in \{(G,U),(U,G)\}, \\ 2 & \text{if } (s[i], s[k]) \in \{(A,U),(U,A)\}, \\ 3 & \text{if } (s[i], s[k]) \in \{(G,C),(C,G)\} \end{cases} \tag{5}$$

transforms this algorithm into a more or less crude estimator for the most stable secondary structure, by taking the individual hydrogen bond strengths into account.

3.2 THERMODYNAMIC SECONDARY STRUCTURE PREDICTION

Although the maximum matching algorithm as presented in the previous section, was the first step towards optimal secondary structure prediction, its simplifications lead to poor prediction quality. This is mainly due to the fact that the major part in structure stability is contributed by stacking interactions of the π-electron systems of two adjacent bases, e. g. in a base pair stack, not from hydrogen bonds between two bound nucleotides. This observation was already made in the early 1970s when researchers performed the first RNA melting experiments to determine the free energy contributions of different structure motifs [64, 224, 87, 88, 11, 28, 6, 179]. In general, it turned out that the construction of an algorithm to compute a secondary structure with minimal free energy, i.e. the thermodynamically most stable one, is much more complicated and needs to distinguish base pairs according to the types of loop they enclose.

LOOP DECOMPOSITION Any secondary structure can be uniquely decomposed into basic components enclosed by base pairs, its loops L, see Fig. 6. Sparse energy contribution data for some of those primitive components, e.g. hairpin loops, bulge loops, and some interior loops, were already available in the beginning of the 1980s. This suggested that the total free energy $E(s)$ of relatively complex structures s can be adequately estimated by the sum of the free energy of its loops E_L [223, 224]

$$E(s) = \sum_{L \in s} E_L. \tag{6}$$

Here, the special case of the totally unfolded single stranded RNA serves as a reference for the relative free energy $E(s)$, hence it has a free energy of $0\ \mathrm{kcal/mol}$ by definition.

The decomposition of a secondary structure s into its constituent loops can be formally defined as follows. Each base pair (i,j) closes a loop L and thereby encloses further unpaired nucleotides $\{u \mid i < u < j\}$, and possibly other base pairs $\{(u,v) \mid i < u < v < j\}$ as well. Let u with $i < u < j$ being an unpaired nucleotide enclosed by the base pair (i,j). We denote u *immediately interior* to (i,j), if there is no other base pair (p,q) enclosing u with $i < p < u < q < j$. Similarly, a base pair (u,v) with $i < u < v < j$ is *immediately interior* to (i,j) if there is no other base pair (p,q) with $i < p < u < v < q < j$. Using this definition, a base pair (i,j) closes a cycle in the secondary structure graph which only consists of

1. an amount l of *immediately interior* unpaired nucleotides, and
2. a degree k of *immediately interior* base pairs (u,v).

The degree k of base pairs (u,v) on the path around such a cycle in turn determines a k–loop of size l, where the size may also be considered the length of the loop. The set of all k–loops constitute the faces of a planar drawing of a secondary structure

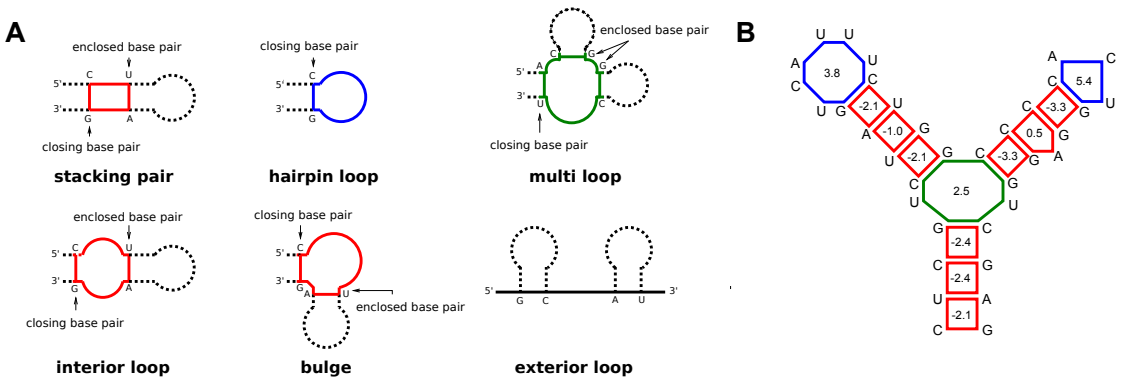


Figure 6: The different loop types distinguished by the *nearest neighbor energy model*. **A** Six classes of loops are distinguished, namely *hairpin loops* with exactly one closing base pair, *interior loops* with two base pairs involved, one enclosing the loop, the other being enclosed. Interior loops that do not enclose any unpaired nucleotides are further specialized to a *stacks of base pairs*, while completely asymmetric interior loops having unpaired nucleotides only in one side of the loop make up the *bulge* loops. Any other loop that consists of more than two base pairs is treated as a general *multi loop*, or *multibranch loop*. A special case is the *exterior loop*, since it does not have an enclosing pair, but only enclosed unpaired nucleotides and base pairs. **B** An example for the loop decomposition of a secondary structure with a central multi loop having three branching helices attached. The free energies of the individual loops according to the latest nearest neighbor energy parameters [228] are drawn on their inside, and add up to -6.5 kcal/mol in total.

(see Fig. 6). Since this set is unique for pseudo-knot-free secondary structures, it corresponds to the *unique minimal cycle basis* of the structure graph [132].

From now on, I denote k-loop with $k = 1$ *hairpin loop*, loops with $k = 2$ as a either a *stack*, if its size is 0, a *bulge* if there are unpaired nucleotides only on one side between the closing and the enclosed base pair, or simply an *interior loop* in all other cases. Loops with $k > 2$ are generally termed *multi loop*, or *multibranch loop*. All unpaired nucleotides or base pairs not enclosed by any other base pair (i, j) reside in the *exterior loop* which serves as a special case in this definition. Fig. 6 gives a graphical representation of the distinguished loop types.

NEAREST NEIGHBOR ENERGY PARAMETERS With the advent of reliable RNA ribonucleotide synthesis in the early 1980s, scientists were able to design new experiments to fill gaps and improve consistency in the knowledge of thermodynamic data on RNA structures [70, 116]. However, despite the large number of experiments, where for instance UV melting temperatures of a large variety of hairpin- and interior-loops were measured, they cannot cover all possible loop configurations[1]. The enormous

1 A hairpin loop of size n has 4^n possible arrangements of its loop sequence and may be closed by 6 different base pairs, resulting in $6 \cdot 4^n$ possibilities.

combinatoric variability simply does not allow one to measure all of them. As a consequence, the *nearest neighbor energy model* used in the algorithms discussed below distinguishes just three types of loops, namely hairpin loops, interior loops, and multibranch loops. It furthermore distinguishes between loops of small and large sizes in terms of the assessment of energy contributions. While tabulated, experimentally determined parameters are used for small loops only, mathematical models serve as estimates of the contributions for larger ones. In the past decades, the list of small, exhaustively tabulated loops grew a lot [9, 39, 228]. To date it accounts for certain short-sized hairpin loops that are exceptionally stable, and interior loops with very small loop length, such as $1-1$, $1-2$, $2-2$, and $2-3$ interior loops. For the latter, each pair of numbers indicates the length of the left and the right part of the loop, respectively. Some of these tabulated energy parameters can be considered to capture the stabilizing effect of *non-canonical* hydrogen bonds between unpaired nucleotides within these small loops, see also Chapter 3.8. In contrast to that, the mathematical models that predict the free energies of larger loops, do not take the loops entire sequence into account. They rather depend on the nucleotide composition of the base pairs that constitute the loop and their directly adjacent unpaired nucleotides, so called *terminal mismatches* and *dangling ends* [169, 238], hence the name *nearest neighbor energy model*. Furthermore, polymer theory predicts a logarithmic growth of a loop's free energy E_L in terms of its size l [115]. Accordingly, the model distinguishes three independent terms of energy contributions

$$E_L = E_{\text{mismatch}} + E_{\text{size}} + E_{\text{special}} \tag{7}$$

where E_{special} serves as a bonus energy for unusually stable loops, such as *tetra-loops*, which are certain hairpin motifs of size 4. Within the last three decades the set of energy parameters for this model was constantly upgraded and revised [227, 238, 152, 150, 40].

For hairpin and interior loops, this model allows for a loop decomposition that enables the implementation of efficient DP algorithms to predict optimal secondary structures with respect to their free energy. However, the evaluation of multi branch loops does not allow for that, since a multiloop may consist of an arbitrary number of branching helices, and therefore an arbitrary number of differently sized segments of unpaired nucleotides connecting the branches. This renders a decomposition of a logarithmic evaluation of E_{size} infeasible [155, 150]. While the first secondary structure prediction algorithms that made use of the nearest neighbor model simply neglected the contributions of multi loops, a simple linear model is used in more recent algorithms [155, 104]. This ensures that multi loops can be efficiently and uniquely decomposed into their constituents. Nonetheless, most implementations allow one to re-evaluate multi loop contributions with respect to a more sophisticated energy model in a post-processing step [150]. The linear model only considers two properties of the multi branch loop, namely the total number of unpaired nucleotides C (the size),

and the number of branches B (the degree). Each of these features is assumed to be in a linear relationship with a free energy determining parameter c and b. Furthermore, since the closing pair of the loop introduces a restriction in the degrees of freedom of the enclosed part, a free energy penalty for closing the multibranch loop a is added, hence the simplistic model

$$\mathcal{M} = a + B \cdot b + C \cdot c. \tag{8}$$

Still, limited thermodynamic data for only a small subset of multibranch loops is available to even fit the parameters a, b, and c of this simple linear model [150]. Thus, the multi branch loop energy evaluation can be considered the weakest part of the nearest neighbor model. Furthermore, the nearest neighbor model neglects many stabilizing nucleotide-nucleotide, and nucleotide-backbone interactions. They are rather assumed to be parts of the so-called *tertiary structure*, since their incorporation into efficient DP algorithms introduces either too much complexity in computational terms, or parametrization is insufficient to date. Nevertheless, in Chapter 3.8 I briefly present the nature of some of these tertiary interactions together with the concepts how to include them in RNA structure prediction algorithms.

ZUKER'S ALGORITHM In 1981, Michael Zuker presented a modification to the maximum matching algorithm to determine the MFE and its corresponding structure for an RNA molecule [266]. It was the first time hairpin loops, interior loops and multibranch loops were distinguished in an efficient secondary structure prediction algorithm. However, this first approach did not explicitly take into account potential energy contributions of multibranch loops, but simply added the contributions of its enclosed substructures. It took another three years until secondary structure prediction using a full nearest neighbor model, as described above, was finally published [265].

COMPUTING THE MFE In principle, Zuker's MFE algorithm follows the same decomposition scheme that Ruth Nussinov already applied for the maximum matching problem, see Fig. 5. The major difference, however, is that due to the distinction of the three types of loops a single base pair (i, j) may enclose, some additional DP matrices had to be introduced. Below are the recursions as implemented in the program

`RNAfold` of the `ViennaRNA Package` [104, 137] that resemble the algorithm presented by Michael Zuker, graphically depicted in Figure Fig. 7.

$$
\begin{aligned}
F_{i,j} &= \min \begin{cases} F_{i,j-1}, \\ \min_{i<k<j} F_{i,k-1} + C_{k,j} \end{cases} \\
C_{i,j} &= \min \begin{cases} \mathcal{H}(i,j), \\ \min_{i<k<l<j} \mathcal{I}(i,j,k,l) + C_{k,l}, \\ a + b + \min_{i<k<j} M_{i+1,k-1} + M^1_{k,j-1} \end{cases} \\
M_{i,j} &= \min \begin{cases} c + M_{i,j-1}, \\ \sum_{i<k<j} (k-i)c + b + C_{k,j}, \\ \sum_{i<k<j} b + M_{i,k-1} + C_{k,j} \end{cases} \\
M^1_{i,j} &= \min \begin{cases} c + M^1_{i,j-1}, \\ b + C_{i,j} \end{cases}
\end{aligned}
\tag{9}
$$

While F_{ij} may be considered analogous to E_{ij} of Equation (4), C_{ij} is used to store the optimal free energy for any (sub-)structure given that i and j form a base pair. M_{ij} and M^1_{ij}, on the other hand, are necessary to decompose the structural parts of a multiloop. They are filled with the optimal energy for multiloop components with at least one branch, and exactly one branch, respectively. It should be noted that for MFE prediction the fourth DP matrix, M^1, is actually not essential, but may be replaced by M. Still, it is included here, since it allows for an unambiguous decomposition of multibranch loops which, for instance, is required for suboptimal structure prediction. Furthermore, this unique decomposition can be translated to the recursions necessary for computing the partition function as discussed in the upcoming sections. The asymptotic time complexity of the MFE algorithm is essentially determined by the evaluation of interior loops. Since for any enclosing pair (i,j) any possible enclosed pair (k,l) with $i < k < l < j$ needs to be considered, the amount of work required scales with $\mathcal{O}(n^4)$. However, it is generally assumed that large interior loops are unfavorable and may be ignored [155]. Therefore, most implementations restrict the maximal length of interior loops to a constant $l_{max} = 30$, which in turn renders the MFE algorithm an $\mathcal{O}(n^3)$ DP approach. Nonetheless, Lyngsø et al. presented an MFE algorithm that manages to compute interior loop contributions without restricting the loop length in $\mathcal{O}(n^3)$ [141].

Once the DP matrices F, C, M, and M^1 are filled, $F_{1,n}$ contains the minimum free energy of all structures compatible with the RNAs sequence. A corresponding secondary structure s_{MFE} can then be obtained via backtracking through the DP matrices, analogously to the backtracking procedure of the maximum matching algorithm (see Section 3.1).

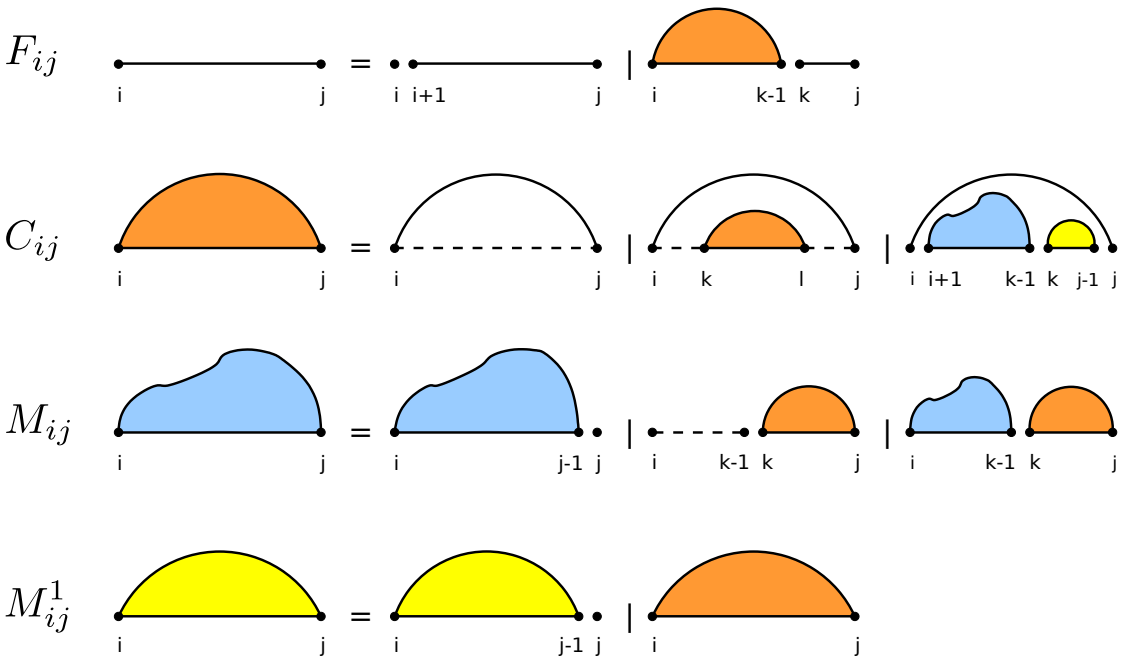


Figure 7: Decomposition scheme for the full Nearest Neighbor Energy Model. The first line shows the decomposition of exterior loops where either a position i is unpaired, or it is paired with some k. The second line highlights the decomposition of a base pair (i, j). Here, (i, j) may either form a hairpin loop, an interior loop with enclosed base pair (k, l), or a multibranch loop. For the latter, a unique decomposition into a part with at lease one branching stem within interval $[i + 1 : k - 1]$, and exactly one branch within interval $[k : j - 1]$ is applied. Unpaired nucleotides are represented by either an individual unconnected dot or a dashed line.

An implementation of Zuker's algorithm can be found in the software packages *RNAstructure* [188], *UNAfold* [147], and the *ViennaRNA Package* [104, 137]. The latter will be introduced in more detail in my publication *ViennaRNA Package 2.0* (Part iv, Chapter 7 of this thesis).

3.3 SUBOPTIMAL SECONDARY STRUCTURES

Predicting the MFE structure as shown in the section before retrieves only one single representative structure. However, RNA molecules in equilibrium are in constant flux between different low free energy structures. Furthermore, structure predictions may be inaccurate with no chance to assess their reliability. To circumvent this limitation, suboptimal secondary structures, i.e. a set of several low free energy structures can be generated giving more robust insight into the structure space.

ZUKER'S SUBOPTIMALS The first attempt toward suboptimal structure was taken by Michael Zuker in 1989 [264]. He presented a modification in the backtracking procedure of his secondary structure prediction algorithm for circular RNAs that enabled him to produce for any fixed base pair (i, j) the optimal secondary structure that includes this pair. Ignoring the original implementation of the Zuker's suboptimal structure algorithm, the MFE under the constraint that i and j form a base pair can be formulated in terms of an *outside*-MFE O_{ij} with

$$O_{ij} = \min \begin{cases} F_{1,i-1} + C_{ij} + F_{j+1,n} \\ \min_{k<i<j<l} O_{kl} + C_{ij} + \mathcal{I}(k,l,i,j) \\ \min_{k<i<j<l} O_{kl} + C_{ij} + a + b + \min \begin{cases} (i-k-1)\cdot c + M_{j+1,l-1} \\ (l-j-1)\cdot c + M_{k+1,i-1} \\ M_{k+1,i-1} + M_{j+1,l-1} \end{cases} \end{cases} \tag{10}$$

Here, the first line corresponds to all cases where (i, j) is an exterior base pair, i.e. it is not surrounded by any other pair. The remaining terms cover the cases where (i, j) constitutes the enclosed pair of an interior loop, or the left, the right or some intermediate branch of a multi loop, respectively. The application of a backtracking procedure that starts with each possible base pair (i, j), i.e. backtracking from O_{ij}, then yields the optimal secondary structure under the constraint that i and j form a base pair (i, j).

However, although his paper had the title *"On finding all suboptimal foldings of an RNA molecule"*, the method he presented did not exactly do that. In fact, there are cases where this method might miss some possibly important low free energy structures [255]. Still, since Zuker's method of finding suboptimal structures generates only a relatively small amount of conformations that might be taken as representatives of the entire structure ensemble Ω structures[2]. Thus, this approach can be used quite conveniently to obtain a crude estimate for the diversity of low free energy structures.

2 On a sequence of length n, there are at most $(n \cdot (n-1))/2$ possible base pairs (i, j)

FINDING ALL SUBOPTIMAL STRUCTURES As stated above, the set of suboptimal structures generated by Zuker's suboptimal structure algorithm might miss some potentially important structures. This is due the design of backtracking a single structure only, for each fixed base pair taken into account. However, for some applications it is crucial to know about all the low free energy structures, even if they share large parts of substructures. When using the structure ensemble for simulation of RNA folding kinetics, for instance, very large numbers of closely related structures need to be generated, see Chapter 6.

A solution to this problem was provided in 1999 Wuchty et al. [255], who presented the first algorithm that exhaustively enumerates all secondary structures within a predefined energy range δ around the MFE. For this purpose, the authors of this method applied the ideas of Waterman and Byers [241] to RNA structure prediction. While using the MFE algorithm presented in Equation (9) to fill the corresponding matrices, its method modifies the backtracking procedure. Instead of tracing back the exact constituents of a particular solution in the DP matrices, it rather constructs a list $\mathfrak{S}$ of (partial) solutions of type $(E, \hat{s}, I)$, where E is the free energy gained so far from backtracking the substructure $\hat{s}$, and $I = \{[i:j]_X\}$ is a set of position intervals $[i:j]$ in matrix $X \in \{F, C, M, M^1\}$ that still need to be backtracked.

Let's first consider only a single partial solution in the list. The algorithm starts by removing an interval $[i,j]_X$ from I to find the subsolutions that constitute X_{ij}. The resulting subintervals $[k:l]_Y$ with $i \leqslant k \leqslant l \leqslant j$ are then inserted into I, any backtracked base pair is inserted into $\hat{s}$, and the corresponding loop energy is added to E. This process constitutes the backtracking procedure of Zuker's MFE algorithm, see Section 3.2.

However, in contrast to that, the suboptimal secondary structure algorithm by Wuchty et al. extends $\mathfrak{S}$ with alternative backtracking routes. In particular, for any partial solution $(E, \hat{s}, I)$, an alternative backtracking path produces some subintervals $[k:l]_Y$, possibly a base par (p,q) and a corresponding loop free energy E_L. Accordingly, the partial solution $(E + E_L, \hat{s} \cup (p,q), I \cup [k:l]_Y)$ is inserted into $\mathfrak{S}$, if the following condition is fulfilled

$$MFE + \delta \geqslant X_{ij} + E + E_L + \sum_{[k,l]_Y \in I} Y_{k,l}. \tag{11}$$

The exhaustive backtracking procedure starts with a single list entry $(0, \emptyset, \{[1:n]_F\})$, and terminates as soon as all partial solutions in $\mathfrak{S}$ have been processed, i.e. all structures s with $E(s) \leqslant MFE + \delta$ are backtracked.

3.4 THE THERMODYNAMIC ENSEMBLE OF SECONDARY STRUCTURES

Almost a decade after the first efficient algorithm for RNA secondary structure prediction, scientists were still limited to explore only a few representative structures, for instance the MFE structure, or some suboptimal structures. The only reliability measure of a single secondary structure prediction, e.g. the MFE structure, was to investigate a (possibly small) set of other structures with similar free energy and a sufficient amount of unshared base pairs. However, such an approach depends strongly on the choice of suboptimal structures to compare to and may therefore be rather artificial. A thorough, straight-forward solution to this dilemma would require taking the complete ensemble of secondary structures into account. But exhaustive enumeration of all suboptimal secondary structures is only possible for RNAs with very small sequence length. In 1990 the situation was about to change: John McCaskill presented an algorithm to compute the partition function Q of an RNA [155]. This finally enabled one to compute equilibrium properties of an RNA, e.g. various reliability measures, and base pairing probabilities.

MCCASKILL'S PARTITION FUNCTION ALGORITHM The algorithm presented by McCaskill applied the physical ideas of statistical mechanics and polymer science to RNA structure prediction. Each secondary structure $s \in \Omega$ compatible with a particular RNA sequence is a state within the system of structures, and is associated with a free energy $E(s)$. Considering the ensemble of structures in thermal equilibrium, the frequency or probability $p(s)$ of a particular state s follows a Boltzmann distribution, i.e.

$$p(s) \propto e^{-\beta E(s)} \text{ with } \beta = \frac{1}{kT} \tag{12}$$

where $k \approx 1.987 \cdot 10^{-3} \frac{kcal}{mol\,K}$ is the Boltzmann constant, and T the thermodynamic temperature. However, to obtain an actual probability, the above proportionality needs to be scaled. Such a scaling factor provides the *canonical partition function*

$$Q = \sum_{s \in \Omega} e^{-\beta E(s)} \tag{13}$$

which determines the free energy of the system

$$G = -\frac{1}{\beta} \ln Q. \tag{14}$$

McCaskill realized that a simple change in the algebraic structure of the MFE problem, as presented in Equation (9), yields an algorithm to compute Q. In detail, the introduced change of algebra is a substitution of each addition with a multiplication, each minimum becomes a sum, and the energy contributions are substituted with their Boltzmann factor counterparts.

$$
\begin{aligned}
Q_{i,j} &= Q_{i,j-1} + \sum_{i \leqslant k < j} Q_{i,k-1} Q^{B}_{k,j} \qquad (15)\\
Q^{B}_{i,j} &= e^{-\beta \mathcal{H}(i,j)} \\
&\quad + \sum_{i<k<l<j} e^{-\beta \mathcal{J}(i,j,k,l)} \cdot Q^{B}_{k,l} \\
&\quad + \sum_{i<k<j} e^{-\beta(a+b)} \cdot Q^{M}_{i+1,k} Q^{M1}_{k+1,j-1} \\
Q^{M}_{i,j} &= e^{-\beta c} \cdot Q^{M}_{i,j-1} \\
&\quad \sum_{i<k<j} e^{-\beta((i-k)c+b)} \cdot Q^{B}_{k,j} \\
&\quad \sum_{i<k<j} e^{-\beta b} \cdot Q^{M}_{i,k-1} Q^{B}_{k,j} \\
Q^{M1}_{i,j} &= e^{-\beta c} \cdot Q^{M}_{i,j-1} + e^{-\beta b} \cdot Q^{B}_{i,j}
\end{aligned}
$$

The result is an algorithm which is generally as fast as the MFE algorithm[3].

EQUILIBRIUM PROBABILITIES OF STRUCTURAL FEATURES Once the partition function Q of the ensemble of RNA secondary structures is known, it opens a great variety of applications. Not only can the (equilibrium) probability $p(s) = \frac{e^{-\beta E(s)}}{Q}$ of a particular secondary structure s be predicted, but also the probability $p(\mathcal{F})$ to observe any structural feature $\mathcal{F}$. Therefore, the only requirement is the partition function $Q_{\mathcal{F}} = \sum_{s_{\mathcal{F}}} e^{-\beta E(s_{\mathcal{F}})}$ of all structures $s_{\mathcal{F}}$ that exhibit $\mathcal{F}$, and

$$
p(\mathcal{F}) = \frac{Q_{\mathcal{F}}}{Q}. \qquad (16)
$$

This exact same relation was already applied by McCaskill in his 1990 article about the partition function algorithm. Using the restricted partition function matrices Q, Q^{B}, Q^{M}, and Q^{M1}, he presented an efficient DP algorithm to compute the equilibrium probabilities $p_{ij} = \frac{Q^{ij}}{Q}$ for all base pairs (i,j) in $\mathcal{O}(n^3)$. Since in this case, the structural feature $\mathcal{F}$ is the base pair (i,j) the restricted partition function $Q_{\mathcal{F}} = Q^{ij}$ has to be computed. This can be done analogously to the *outside*-MFE algorithm presented for Zuker's suboptimal structures in Section 3.3. Therefore, $Q^{ij} = \bar{Q}^{ij} + \hat{Q}^{ij}$ is divided into two parts. The first part, $\bar{Q}^{ij}$ which comprises all structures where (i,j) is an

3 Despite the fact that both algorithms share the same asymptotic time complexity, actual execution time on a real physical machine may differ a lot. Not only may multiplication take longer on a CPU than an addition, but, depending on the type of CPU, there may also be differences in the number of clock cycles any other operation requires for floating point numbers compared to integer numbers.

exterior base pair, thus not enclosed by any other pair (k,l) with $k < i < j < l$, can be computed straight forward:

$$\bar{Q}^{ij} = Q_{1,i-1} \cdot Q^{B}_{i,j} \cdot Q_{j+1,n}. \tag{17}$$

All remaining cases, where (i,j) is the enclosed part of an interior loop, or the left, the right, or an intermediate part of a multi branch loop, contribute to $\hat{Q}^{ij}$. Taken together, the equilibrium probability p_{ij} of a base pair (i,j) evaluates to

$$\begin{aligned} P_{i,j} = {} & \frac{Q_{1,i-1} \cdot Q^{B}_{i,j} \cdot Q_{j+1,n}}{Q_{i,j}} \\ & + \sum_{\substack{(k,l) \\ k<i<j<l}} P_{k,l} \frac{Q^{B}_{ij}}{Q^{B}_{k,l}} \cdot e^{-\beta J(k,l,i,j)} \\ & + \sum_{\substack{(k,l) \\ k<i<j<l}} P_{k,l} \frac{Q^{B}_{ij}}{Q^{B}_{k,l}} \cdot e^{-\beta(a+b)} \\ & \cdot \Big(e^{-\beta(i-k-1)c} \cdot Q^{M}_{k+1,i-1} \\ & \quad + e^{-\beta(l-j-1)c} \cdot Q^{M}_{j+1,l-1} \\ & \quad + Q^{M}_{k+1,i-1} \cdot Q^{M}_{j+1,l-1} \Big) \end{aligned} \tag{18}$$

With respect to the sequence length n, the sums in the above equation introduce a quadratic time complexity. An algorithm to compute all p_{ij}, again, introduces quadratic evaluation for all base pairs (i,j) as well. Hence, the recursions of Equation (18) suggests an asymptotic time complexity of $\mathcal{O}(n^4)$. However, by introducing some changes of the order in which the sums are evaluated, and the usage of additional dynamic programming tables, an $\mathcal{O}(n^3)$ algorithm can be constructed [155].

STOCHASTIC SAMPLING FROM THE BOLTZMANN ENSEMBLE The partition function matrices as computed in equation (15) also allow for construction of secondary structure representatives. Particularly, it allows one to sample suboptimal structures s according to their equilibrium probability $p(s)$. This can be achieved by application of a backtracking procedure that operates on the partition function matrices Q, Q^{B}, Q^{M}, and Q^{M1}, rather than the minimum free energy matrices [54]. Therefore, the algorithm has to select the base pair pattern, and thus the decomposition tree, by randomly choosing a decomposition step according to its probability mass. For instance, upon backtracking in $Q_{i,j}$, the decomposition scheme distinguishes two possibilities (see Equation (15), first line). Either j is unpaired, or j is paired with some nucleotide $i \leqslant k < j$. The contribution of both cases to $Q_{i,j}$ are (a) $Q^{\text{j-unpaired}} = Q_{i,j-1}$ and (b) $Q^{\text{j-paired}} = \sum_{i \leqslant k < j} Q_{i,k-1} Q^{B}_{k,j}$, respectively. Thus, the probability of j being unpaired

is $p(\text{j-unpaired}) = \frac{Q_{i,j-1}}{Q_{i,j}}$, while the probability of j being involved in a base pair is given by $p(\text{j-paired}) = \frac{Q^{\text{j-paired}}}{Q_{i,j}}$. Given a random variable $0 \leqslant r \leqslant 1$, the decision of which of both backtracking paths to follow can be done easily. If $Q_{i,j-1} \geqslant r \cdot Q_{i,j}$, j is considered to be unpaired and backtracking has to proceed in $Q_{i,j}$. Otherwise, the pairing partner k that forms a base pair with j has to be evaluated. Using the auxiliary function

$$Z(k,i,j) = Q_{i,j-1} + \sum_{i \leqslant u < k} Q_{i,u-1} Q^{B}_{u,j} \tag{19}$$

the first $i \leqslant k < j$ that fulfills the inequality $r \cdot Q_{i,j} < Z(k,i,j)$ is chosen, and backtracking proceeds in $Q_{i,k-1}$ and $Q^{B}_{k,j}$. The application of this principle to all decompositions of the recursion scheme in equation (15) then results in an $\mathcal{O}(n^2)$ algorithm which constructs secondary structures based on their equilibrium probability.

3.5 RELIABILITY MEASURES AND REPRESENTATIVE STRUCTURES

Base pair probabilities and probabilities of larger structure components like helices may be useful to gain insights into the diversity of the ensemble. However, the resulting number of descriptors can rapidly grow very large, and thus may not be suitable for representation. Instead, a single or a small set of structure representatives is chosen. Here, the problem arises, that such representatives may be misleading since they aim to represent a larger structure space. For instance, sometimes the MFE structure may be a wrong choice as a representative, it is after all only the structure with highest equilibrium probability. There may be other structures with similar probabilities but different base pairs. Furthermore, it is not guaranteed that the RNA actually reaches its equilibrium state within its lifetime, i.e. before degradation.

Hence, measures are necessary to judge the reliability of (parts of) a representative structure. Some of these measures can be used to obtain a *global reliability* that usually depicts how diverse the structures space is. In contrast to that, a nucleotide position wise *local reliability* may tell how trustworthy individual pairing patterns of a particular representative structure are.

Finally, instead of assessing the reliability of a given structure, the approach can be turned around to construct representative structures which are most reliable according to certain requirements. Two examples that implement this kind of construction, the *centroid-*, and the *MEA*-structure, will be discussed below.

STRUCTURAL DIVERSITY OF THE ENSEMBLE A classical and widely used global reliability measure that reflects the diversity of structures within the ensemble Ω in one single number is the *mean equilibrium distance*

$$\langle d \rangle = \sum_{s,t \in \Omega} p(s) \cdot p(t) \cdot d(s,t). \tag{20}$$

The choice of the base pair distance $d_B(s,t)$ as distance measure d between two secondary structures s and t allows to rewrite Equation (20) in terms of base pair probabilities p_{ij}. It turns out that $\langle d \rangle$ is the sum of all pairing probabilities weighted against their propensity of being unpaired

$$\langle d \rangle = \sum_{ij} p_{ij} \cdot (1 - p_{ij}). \tag{21}$$

Considering only a single base pair (i,j), the largest diversity in the structure ensemble requires that its pairing status is undecided, i.e. $p_{ij} = 0.5$. Then, the score $p_{ij} \cdot (1 - p_{ij}) = 0.25$ is maximal, and $\langle d \rangle$ becomes larger the more diverse the complete structure ensemble is.

POSITIONAL ENTROPY Originating from information theory, a commonly used local reliability measure is given by following Shannon's entropy formula [199] to re-

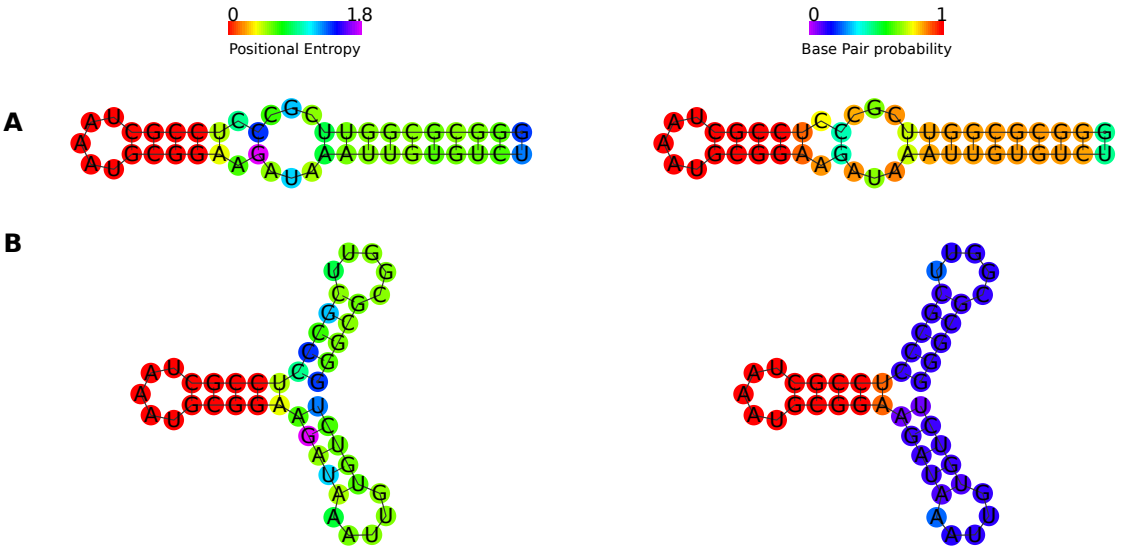


Figure 8: Two methods to assign reliability measures to secondary structure predictions. Exemplary shown for an RNA with sequence GGGCGCGGUUCGCCCUCCGCUAAAUGCGGAAGAUAAAUUGUGUCU and ensemble diversity of 5.95, artificially designed to adopt two distinct low free energy states. On the left side of the figure the positional entropy for each nucleotide is depicted. On the right-hand side base pairs (i, j) are colored after their pairing probability, i.e. p_{ij}, whereas unpaired nucleotides i depict their propensity to stay unpaired, i.e. $q_i = 1 - \sum_j p_{ij}$. **A** shows the MFE structure predicted with `RNAfold` [104, 137] with minimum free energy of -17.7 kcal/mol. **B** represents the designed meta-stable state with free energy of -17.2 kcal/mol.

trieve the so-called positional entropy $S(i)$. This allows for a position-wise assessment of the information whether or not a nucleotide i is mainly paired or unpaired within the ensemble of possible structures:

$$S(i) = -\sum_k p_{ik} \log_2 p_{ik} + (1 - p_{ik}) \log_2 (1 - p_{ik}) \tag{22}$$

Here, $S(i) = 0$ if position i shows no entropy within the ensemble, i.e. it exhibits the same state (paired or unpaired) among the whole ensemble, whereas higher values reflect its tendency to occur in diverse contexts (See Figure 8). Therefore, predictions for positions with low positional entropy can be considered more reliable than those with high entropy.

THE CENTROID STRUCTURE Sometimes, it is hard to obtain a good structure representative, especially in ensembles of high structural diversity. In these cases, it might be useful to look at a very special structure that somehow represents the Boltzmann weighted *mean* of the ensemble. Such a representative is the *centroid structure*. It is

defined as the structure s_c which minimizes its Boltzmann weighted distance d_Ω to all other structures in the ensemble. Hence,

$$d_\Omega(s_c) = \sum_{s \in \Omega} p(s) d(s_c, s). \tag{23}$$

The most simple distance measure d one can use here, is the *base pair distance* $d_{BP}(s_i, s_j)$, i.e. the number of base pairs s_i and s_j have not in common. Interestingly, it turns out that using d_{BP}, this structure can be easily constructed by the set of all base pairs with a probability $p_{ij} > 0.5$.

STRUCTURES THAT MAXIMIZE AN EXPECTED ACCURACY As shown above, the more probable a certain structural feature is, the more reliable will be the prediction of its state.[4] Hence, one can ask to construct a structural representative that consists of the most probable structural features, e.g. base pairs. Such a structure would then yield the highest accuracy, or Maximum Expected Accuracy (MEA). Assuming that the base pair probability p_{ij} serves as a good measure for correctness of the pair (i, j), the following equation has to be maximized

$$EA(s) = \sum_{(i,j) \in s} 2\gamma p_{ij} + \sum_{\substack{i \\ \nexists (i,j) \in s}} q_i \tag{24}$$

with

$$q_i = 1 - \sum_j p_{ij}.$$

Here, γ serves as a factor that allows weighting of paired against unpaired positions.

This, again, leads to an optimization problem that can easily be solved by a Nussinov style DP algorithm (see (4)). Instead of counting the number of base pairs, their probabilities need to be summed up. Similarly, for unpaired positions i their probability to be unpaired q_i must be taken into account. After a successful forward recursion to determine the MEA score, a backtracing procedure can be applied to retrieve a corresponding structure.

4 Here, a feature can be the pairing state of a single nucleotide i. Of course, in this context, the more probable means it is mostly paired or unpaired, i.e. $p_i >> 0.5$ or $q_i >> 0.5$.

3.6 CLASSIFIED DYNAMIC PROGRAMMING APPROACHES

All the DP algorithms discusses so far are variations of a same common scheme to determine certain properties. They generate a search space of candidates, which usually grows exponential with the input size, and evaluate certain features in polynomial time through application of a recursive decomposition of the problem. Of course, such a method can only be applied if Bellman's *Principle of Optimality* holds [15], i.e. "An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision". In other words, an optimal solution for the actual problem can be constructed from optimal solutions of smaller subproblems. For instance, the DP algorithms discussed in the previous sections, count the number of structures, compute the MFE, base pair probabilities, and so on, all from subsolutions of a smaller problem. However, in some analysis it would be better to partition the search space and analyze each of the partitions independently. With such an approach, one could retrieve e.g. the best representative of each partition, or the base pair probabilities of all structures within a certain subclass of structures, or even the probability of the subclasses themselves. Furthermore, one could ask for the number of partitions a particular classification scheme generates. Such approaches are termed *classified dynamic programming* (classified DP), since they subdivide the underlying state space into classes before evaluation. Although some methods that implement classified DP for RNA secondary structure analysis were published within the last decades, a formalism that separates the generator of the search space from the evaluation of each candidate was first introduced in the work on *Algebraic Dynamic Programming* (ADP) of the group around Robert Giegerich [80, 196, 197, 209]. Their formalism allows for an easy way to express complicated combinations of DP approaches, such as classified DP, in terms of grammars. Though, abstractions of functional programming languages and specialized compilers are used to actually implement ADP. While first approaches like ADP as an extension of the functional programming language `Haskell`, or the *GAP-compiler* [196, 197] had a relatively large trade-off between the ease of formalism and performance in terms of runtime, recent tools like *ADP-fusion* [109] produce code that performs close to hand-optimized `C` implementations.

Some remarkable approaches that implement a classified DP scheme are `RNAshapes` [81], `RNAbor` [72], `RNAHeliCes` [111], and `RNA2Dfold` [136], which, except for *RNAshapes*, were originally developed without the explicit notion of classified DP. Below, I introduce a small collection of classified DP algorithms in more detail. The algorithm of `RNA2Dfold` that partitions the secondary structure space into distance classes with respect to two initially chosen reference structures is presented in Part iv, "2D Projections of RNA folding Landscapes", Chapter 9.

CLASSIFICATION ORDER AND COMPLEXITY In general, a classified dynamic programming approach introduces a particular *classification order*, which states how fast the number of classes grows relative to the size of the actual search space. In principle, the previously described algorithms for MFE, or partition function algorithms are classified DP algorithms too. But their order is of type $O(1)$, i.e. their number of classes is bound by a constant, in particular for these examples the number of classes is exactly 1, namely the whole ensemble of secondary structures. Any other classification that distinguishes between a fixed number of features in the search space is of constant order $O(1)$ as well. However, some classifications may result in a growing number of partitions, with respect to the input size. Here, the number of classes may grow *linearly*, in some *polynomial order* k, or even *exponential*, hence their classification orders are $O(n)$, $O(n^k)$, or $O(a^n)$, respectively. Still, the number of classes for any given input can not be larger than the search space, it is usually much smaller. As one would expect, the order of classification has a strong influence on the asymptotic complexity of such algorithms, and therefore, must be carefully investigated.

In detail, any classification $\mathcal{C}$ of order $O(f(n))$ splits up the search space in $f(n)$ partitions. For each of them, some DP analysis $\mathcal{B}$ with asymptotic complexity of polynomial order k, i.e. $\mathcal{B} \in \mathcal{O}(n^k)$, is carried out separately. Classified DP is then the product of the classification algebra $\mathcal{C}$ and evaluation algebra $\mathcal{B}$ denoted by $(\mathcal{C} * \mathcal{B})$ [197]. In the case of RNA secondary structure prediction the total efficiency of the classified analysis[5] is

$$(\mathcal{C} * \mathcal{B}) \in \mathcal{O}(n^2 f(n) \cdot (f(n) \cdot n)^{k-2}). \tag{25}$$

Here, the first term, $n^2 f(n)$, specifies the memorization requirements of $(\mathcal{C} * \mathcal{B})$, while the remaining factor of $(f(n) \cdot n)^{k-2}$ accounts for the work required to obtain a particular subsolution. As a consequence, the efficiency of an algorithm is unchanged[6], if the classification is of order $O(1)$, remains polynomial if $\mathcal{C} \in O(n^k)$, and becomes exponential, if $\mathcal{C} \in O(a^n)$.

DENSITY OF STATES Among the many secondary structures an RNA molecule is capable to adopt, secondary structure prediction algorithms usually compute the one with lowest free energy, i.e. the *ground state*. But for the assessment of how well defined the predicted ground state is, it is crucial to know about alternative structures that also exhibit low free energies. In the previous sections I already presented such measures, e.g. using the partition function to obtain base pair probability derived reliability measures (see Chapter 3.5). Another, though computationally expensive, method was

5 This formula does not necessarily apply to arbitrary classified DP algorithms. The n^2 factor captures the number of possible subsolutions, i.e. each subsequence interval $[i, j]$ of an RNA sequence. Furthermore, it is assumed that the additional work required for any subsolution is determined only by the order of $\mathcal{C}$, hence $f(n)$.

6 Nevertheless, the introduction of a classification scheme will certainly introduce a additional constant factor to the algorithms asymptotic complexity in terms runtime and memory requirements

sketched in Chapter 3.3, the exhaustive enumeration of all suboptimal secondary structures within a certain energy interval δ around the MFE. Once a set of suboptimals is generated, one can derive the diversity and connectedness of structural alternatives. Here, I want to introduce an alternative to these approaches that counts the number of possible structures with free energy ϵ, the so-called *density of states*. Such approaches typically divide the range of free energies into a set of discrete bins, depending on the resolution of the energy evaluation, and the requirements for a particular application. In 1993 Paul G. Higgs explored the thermodynamic differences between structures of real tRNAs and random RNA sequences of the same length [103]. Utilizing a crude algorithm to compute the density of states, he found that not only real tRNAs have more folded configurations than random RNAs, but that the ground state of an actual tRNA has fewer low free energy competitors. However, his method was only applicable to relatively short sequences and required enormous amounts of memory, since it first exhaustively enumerates all possible structures, and only then evaluates their free energies to cluster them into discrete bins according their energy.

In 1996, Cupal et al. presented the first DP algorithm to compute the density of states for an RNA [46]. Using the common decomposition scheme of RNA secondary structures (see Figure Fig. 7), their algorithm computes the density of states $N(\epsilon)$ for all ϵ of fixed resolution $0.01\ \mathrm{kcal/mol}$. Although the notion of classified DP was not common at this time, their algorithm constitutes a nice example. Here, the classification $\mathcal{C}$ is based on free energy, in particular each class consists of all structures with energy ϵ. To make the idea clear, I use the simple Nussinov energy model (see Chapter 3.1) with energy score $S(k,j)$ for each base pair (k,j) for the following equation. Under this model the number of structures $N(\epsilon)$ with energy ϵ can be computed by

$$\begin{aligned} N_{i,j}(\epsilon) = {} & N_{i,j-1}(\epsilon) \\ & + \sum_{\epsilon'} N_{i,k-1}(\epsilon') + N_{k+1,j-1}(\epsilon - \epsilon' - S(k,j)) \end{aligned} \tag{26}$$

Since the minimum free energy of an RNA sequence scales linearly with the sequence length [69], the number of ϵ-classes does as well, i.e. $\mathcal{C} \in O(n)$. According to Equation (25), the algorithm that Cupal et al. proposed has an asymptotic time complexity of $\mathcal{O}(n^5)$ and space requirements of $\mathcal{O}(n^3)$, because the energy evaluation using the nearest neighbor energy model requires time proportional to n^3, i.e. $\mathcal{B} \in O(n^3)$.

RNA SHAPES It has been stated before, that predicting only a single secondary structure representative is usually not sufficient to effectively understand the potential role of an RNA. For some *tRNAs*, for instance, the MFE structure predicted with the nearest neighbor energy model is a rod-like helix interspaced with some small interior loops and bulges, rather than the commonly known cloverleaf structure. Only an inspection of a relatively large set of suboptimal structures reveals, among others, the central multi loop from which four helices emanate. However, even for such small RNAs, one

```
CGUCUUAAACUCAUCACCGUGUGGAGCUGCGACCCUUCCCUAGAUUCGAAGACGAG
((((((...(((..(((...)))))))...(((..((.....))..)))))))))..
```

Shape Level	Description	Result
1	Most accurate - all loops and all unpaired	`[_[_[]]_[_[]_]]_`
2	Nesting pattern for all loop types and unpaired regions in external loop and multiloop	`[[_[]][_[]_]]`
3	Nesting pattern for all loop types but no unpaired regions	`[[[]][[]]]`
4	Helix nesting pattern in external loop and multiloop	`[[][[]]]`
5	Most abstract - helix nesting pattern and no unpaired regions	`[[][]]`

Figure 9: Abstract Shapes of an RNA. On top, an RNA sequence and its corresponding structure in *dot-bracket* notation is shown. Below are the results of the five distinguished abstract shape levels for secondary structures according to Giegerich et al. [81].

is easily overwhelmed with the large variety of possible structures, and it can hardly be decided which of them might be biochemically relevant.

A solution to this dilemma was presented by the *shape abstraction* of RNAs as introduced in 2004 by Giegerich et al. [81]. Their `RNAshapes` algorithm abstracts from individual base pairs and their actual location on the sequence, while retaining the nestedness of helices and hairpin loops. For each of the resulting shapes a representative structure with minimal free energy, a so-called *shrep*, is computed. Offering five levels of abstraction, see Figure 9, this approach allows for a fast visual inspection of structural alternatives. For instance cloverleaf like structures are clustered together into the same shape, no matter where the location of the multi loop is, or how long the branching helical arms are. Although the number of possible shapes of an RNA structure grows exponential with the sequence length, it does so way slower than actual number of secondary structures [81]. Furthermore, the `RNAshapes` approach was extended to compute equilibrium shape probabilities, which allow to asses the proportional 'volume' of structures of a given shape. This can help to identify the most probable candidate shape, represented by its shrep. It may even help to identify RNA switches, i.e. RNAs that can exist in at lest two distinct structures with low free energy, since one would expect that the corresponding shape classes of their meta-stable states subsume a relatively large portion of the probability density compared to other classes [236].

From the computational point of view, the `RNAshapes` algorithm is a classified DP approach, where the number of classes grows exponentially with the input size n, i.e. $\mathcal{C} \in O(a^n)$. Hence, its asymptotic time complexity $(C * \mathcal{B}) \in O(b^n)$ is exponential, too, though with a small base b. The first implementation of `RNAshapes` was done in an ADP framework using the programming language `Haskell`. This rendered the program rather slow and limited its application to RNAs with sequence lengths up to

about 150 nt. However, by using clever heuristics, and omitting shape classes of low probability, the latest `RNAshapes` implementation is speed up substantially. Therefore, their method becomes applicable to RNAs with lengths up to 800 nt [117]

RNABOR For the sake of riboswitch detection, Freyhult et al. developed an *ab-initio* RNA secondary structure space partitioning that clusters structures according to their base pair distance to an initially chosen reference structure [72]. The basic idea behind this approach is, that next to the MFE structure, a biologically relevant meta-stable state should (i) be somewhat distant to the MFE itself in terms of structural similarity, (ii) exhibit a relatively low free energy, and (iii) be separated from the MFE structure by a sufficiently large energy barrier to show a switch like behavior.

Conceptually, the classes introduced by their algorithm, `RNAbor`, are determined *a priori* by choosing a reference structure s_1. Any structure $s \in \Omega$ with $d_{BP}(s, s_1) = \kappa$ is assigned to a corresponding *distance class* α_κ. The algorithm then determines several thermodynamic properties, such as the MFE, a corresponding secondary structure, and the partition function for each of the resulting classes. Since the base pair distance between any two secondary structures grows linearly with the sequence length[7], the number of distance classes grows linearly as well, i.e. $\mathcal{C} \in O(n)$. Therefore, the overall time complexity of distance class partitioning with respect to a single reference structure becomes $(\mathcal{C} * \mathcal{B}) \in O(n^5)$, with memory requirements of $O(n^3)$.

7 The linear growth of the distance classes can be easily seen. According to the definition of secondary structures the maximum number of base pairs for any sequence of length n is bound by $n/2$. Assuming that there exist two structures with maximal number of base pairs s_i and s_j which are maximally apart from each other in terms of base pairs, i.e. $s_i \cap s_j = \emptyset$, their distance is at most $n/2 + n/2 = n$.

3.7 NUCLEIC ACID INTERACTIONS

Many RNA folding problems involve not only the prediction of secondary structures of a single RNA, i.e. *intra*-molecular base pairs, but deal with the interaction of two or more sequences, and thus require the prediction of *inter*-molecular base pairs. Several non-coding RNAs (ncRNAs) are involved in gene silencing and repression in a post-transcriptional manner. In bacteria, the most prominent class of such RNAs are the small RNA (sRNA) which, amongst other things, bind to their target mRNA via an antisense sequence [243, 211]. In plants and animals, such RNAs are known as small interfering RNA (siRNA), and miRNA [154, 63, 220, 161]. After formation of the RNA dimer helices, the resulting complex is recognized by the cells molecular machinery and subsequently degraded [251]. Consequently, they constitute an RNA-mediated down-regulation of gene expression. Furthermore, many other biological processes rely on RNA-RNA interaction. RNA editing, rRNA modification, splicing, and developmental regulation are performed and regulated by small nucleolar RNA (snoRNA)s, snRNAs, and miRNAs [237, 13, 153].

Whereas siRNAs are almost perfectly complementary to their target sequence, this is not necessarily true for miRNA and sRNA. They are able to form more or less complex structures with interwoven inter- and intra-molecular helices [237, 153, 165]. Therefore, to better understand the interaction between two RNAs, in particular the competition between intra- and inter-molecular base pairs, methods for RNA-RNA secondary structure prediction are necessary. An extension to the approaches described below, that takes the hybridization of the two naturally occurring nucleic acids RNA and DNA into account is presented in Part iv, Chapter 8, with my work on *"Prediction of RNA/DNA hybrid structures"*.

INTER-MOLECULAR PAIRS ONLY The most simple way to predict RNA dimer interactions is to neglect intra-molecular base pairs at all. Such approximative methods are described in literature many times and have been implemented in the programs *RNAhybrid*, *RNAduplex*, *GUUGle*, *RNAplex*, *DINAMelt*, and many others [185, 51, 79, 213]. However, there is no biologically motivated reason to assume that the involved RNAs only form inter-molecular base pairs. Though, these methods are generally much faster than any other thermodynamics based approach with asymptotic run times spanning from $\mathcal{O}(n \log n)$ (GUUGle) to $\mathcal{O}(n \cdot m)$ (RNAhybrid, RNAduplex) where n and m are the lengths of both RNA sequences, respectively. In general, these methods are used to scan a (longer) target RNA for potential binding sites with a (shorter) query RNA. Hence, the algorithms allow for RNA-RNA duplex prediction in linear time with respect to the target RNAs length.

SEQUENCE CONCATENATION METHOD Another simple, yet more thorough approach is to simply concatenate both RNAs and predict secondary structures with

the Zuker algorithm (see Section 3.2) for the resulting hybrid sequence [104]. Here, only a handful of small modification are necessary to deal with the loops that contain the artificially introduced *cut point*. In particular, the loop length restriction to at least 3 unpaired nucleotides is invalidated, and the resulting inter-molecular loop motifs are handled as a special kind of exterior loop with an additional destabilizing energy contribution E_{duplex} to account for loss of entropy due to the hybridization of both molecules. As a result, this method provides an $\mathcal{O}((n+m)^3)$ algorithm to predict a certain class of possible RNA-RNA hybrid secondary structures by taking intra- as well as inter-molecular base pairs into account. The program *RNAcofold* implements this method for Zuker's, and McCaskill's algorithm, enabling it to MFE, partition function, base pair probabilities, as well as MEA- and centroid structure [19]. Furthermore, it allows to predict the equilibrium concentrations of duplex structures.

A generalized approach to predict secondary structure hybrids where more than two RNA molecules are involved was already proposed for minimum free energy structures in 1994 [104]. It's partition function variant was published in 2007 by Dirks et al. [56]. Here, the authors presented a method that deals with the combinatorial symmetry issues arising when RNA hybrid structures are predicted. In particular, they suggest corrections to the partition function, that may occur due to over-counting of symmetrical structures.

As noted before, the sequence concatenation method is restricted, as it can not predict RNA-RNA interactions that form pseudoknots in the concatenated sequence. However, such interactions, like the *kissing hairpin* motif, are quite common in nature [31, 73], and may not be excluded for some bioinformatics analysis, see also Fig. 10 of Section 3.8.

HYBRIDIZATION AS A TWO-STEP PROCESS A more sophisticated approach is to consider the binding of one RNA to another a two-step process, (i) removal of intra-molecular base pairs from the the interaction sites, and (ii) hybridization of both molecules. Hence, the total free energy ΔG of such a structure takes the form

$$\Delta G = \Delta G_u + \Delta G_h \tag{27}$$

with the energy contribution necessary to expose the binding site(s) ΔG_u, and the gain in free energy by hybridization at the binding site [167]. For most applications, it is sufficient, to compute ΔG_u only for one of the sequences. Especially in sRNA-mRNA or miRNA-mRNA target prediction, the length of the query sequence, i.e. the sRNA or miRNA, is usually much smaller than that of the target, i.e. the mRNA. In such cases ΔG_u has to be computed only for the target sequence. Two approaches employing this exact scheme have been proposed. However, both of them are limited to only a single interaction site, since they compute ΔG_u just for independent continuous subsequences.

The *RNAup* algorithm [167] computes ΔG_u up to some user specified window size w and the hybridization energy ΔG_h for any possible interaction site according to an *RNAhybrid* like approach. The best RNA-RNA interactions in terms of ΔG are then returned for each position along the target sequence. Although the free energy to expose an unpaired stretch of length w is only computed for the longer RNA sequence, its asymptotic complexity in time and memory are quite high with $\mathcal{O}(n^3 + nw^5)$ and $\mathcal{O}(n^2 + nw^3)$, respectively. A similar method is implemented in the program *IntaRNA* [32]. Utilizing ΔG_u contributions precomputed by *RNAplfold* [18, 21] in $\mathcal{O}(nL^2)$ time, where L is the size of the locally folded subsequence, and enforcing the interaction between both RNAs to start with a user-defined seed makes this approach much faster than *RNAup*. In particular, the authors reduced the memory requirements of their approach to $\mathcal{O}(nm)$, whereas the asymptotic time complexity is $\mathcal{O}(n^2m^2)$[8].

RIP AND ITS LIMITATIONS The generally unrestricted RNA-RNA interaction problem (RIP), i.e. the prediction of RNA-hybrid secondary structures that may form pseudo-knot like structures, was first dealt with independently by Dmitri Pervouchine and Can Alkan et al. in 2004 and 2006, respectively [177, 5]. Although the general, unrestricted RIP is *NP-hard* [5], in both approaches a polynomial time algorithm was developed that enables the computation of a subclass of RNA-RNA hybrid secondary structures. Pervouchine's *IRIS* algorithm utilizes a crude base pair energy model, comparable to Nussinov's maximum matching algorithm. Furthermore, the algorithm restricts the possible base pair interactions to nested intra-molecular pairs, i.e. there are no crossings between any intra-molecular base pairs within each sequence, and nested inter-molecular base pairs. Hence, any two inter-molecular base pairs (i_x, k_y) and (j_x, l_y) imply that if $i < j$, also $k < l$ and vice versa. This enables the IRIS algorithm to compute an optimal RNA-RNA interaction according the simple energy model, with $\mathcal{O}(n^2m^2)$ memory-, and $\mathcal{O}(n^3m^3)$ time requirements. The approach of Alkan et al., on the other hand, discusses the RIP under the more sophisticated nearest neighbor energy model. Along with a prove for the NP-completeness of the general problem, they define all RNA-RNA hybrid structures that comply to the nestedness assumptions as used in the IRIS algorithm as *valid*. Then they develop an algorithm that enables to solve the RIP for valid RNA-RNA hybrids with an asymptotic time complexity of $\mathcal{O}(n^3m^3)$.

Still, both approaches described above only compute a single, yet optimal secondary structure representative according to the energy model used. The computation of the partition function for valid hybrid structures, however, requires an unambiguous decomposition scheme, which was first published in 2009 by Chitsaz et al. [41]. In the same year, Huang et al. independently developed the same decomposition scheme for the partition function [110]. Huang et al. furthermore extended their algorithm

8 Here, the time and memory requirements to pre-compute ΔG_u with *RNAplfold* are assumed to be negligibly small.

with appropriate data structures to compute thermodynamic ensemble features, such as base pair probabilities, and probabilities of hybrid formations.

3.8 BEYOND SECONDARY STRUCTURES

All the algorithms and methods discussed so far make use of the same scheme, the loop decompositions of the nearest neighbor energy model. This model serves very well for most applications with a prediction accuracy of up to 75% [135]. Still, it neglects many properties of a folded RNA that could in principle be considered part of the secondary structure too, possibly increasing its predictive power. Most of these properties have been found experimentally by three-dimensional RNA structure identification using *NMR*, *X-Ray* crystallography, or chemical probing [176, 234, 108, 134]. Among these features is a whole new realm of possible pairwise edge-to-edge interactions between nucleotides, the so-called *non-canonical* base pairs. In fact, apart from the *Watson-Crick edge*, a nucleotide has two additional edges available for base pairing. Although non-canonical base pairs do not occur as frequently as the canonical Watson-Crick base pairs, they still give rise to the distinct tertiary structures of many RNAs [131, 210]. Furthermore, due to the ability to take part in base pairing with more than one edge, an RNA is not limited to pairs of bases, but a nucleotide may exhibit hydrogen bonds with more than one nucleotide, e.g. to form base triples [2]. One of the most remarkable structure motifs consisting of base triples is the *G-Quadruplex*, that forms a stack of adjacent G-quartets, i.e. planar assemblies of four guanosines [119].

Last but not least, the presumed nestedness for RNA secondary structure loops as introduced in Section 2.2, does not always hold. There are many examples where an RNA forms base pairs, that would introduce crossing edges in the resulting secondary structure graph, so-called *pseudo-knots*. Two hairpin loops with (partly) complementary loop sequence, for instance, may form a small, Watson-Crick paired helix, the *kissing hairpin* motif [31]. A comprehensive overview of possible loop-loop interactions can be found in the reviews of Batey et al. [14], and Butcher et al. [33].

Despite the fact that the above properties of RNAs show a high relevance in biological and biochemical terms, their inclusion in prediction algorithms introduces an undesirable complexity, and/or poor predictive performance in many cases. Nevertheless, I want to briefly introduce some approaches that approach the problem of *tertiary structure prediction* within this section. In Part iv Chapter 10 *"2D meets 4G: G-Quadruplexes in RNA Secondary Structure Prediction"*, I present a method to include the self-contained tertiary structure motif of *G-quadruplexes* into RNA secondary structure prediction algorithms.

PSEUDO-KNOTS Since the beginning of efficient computational secondary structure prediction it was known, that the nearest neighbor energy model discards a certain class of looped structures, the *pseudo-knots*[190]. First discovered as long-range interactions in RNA viruses [204, 208], these loop motifs arise when a looped region, typically a hairpin loop, pairs with another unpaired part outside its enclosing helix, see Figure 10 for some examples. This allows the RNA to form more compact structures by

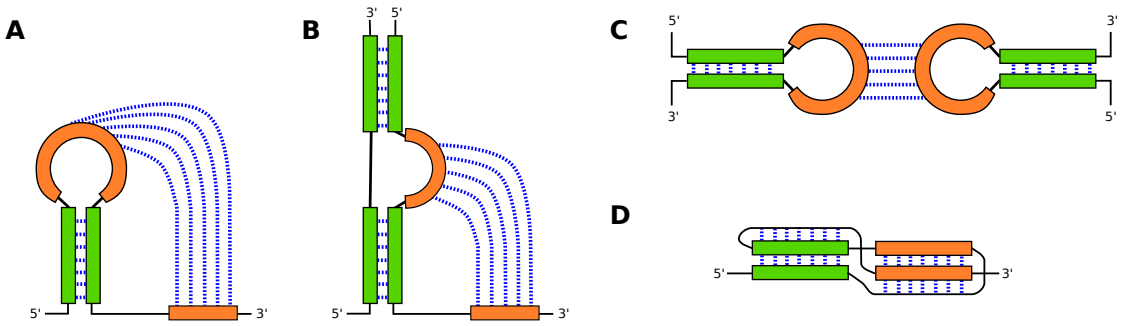


Figure 10: Different types of pseudo-knots with base pairs indicated as blue dashed lines. **A** An *H-type* pseudo knot with initial hairpin and base pairs between its loop region and some part outside its enclosing stem. **B** A bulged region pairs with some part outside the interior loop forming a *B-type* pseudo-knot. Loops where regular interior loop region(s) pair with some other part are generally termed *I-type*. **C** A *kissing hairpin* motif. **D** A densely packed H-type pseudo-knot with triple helices and coaxially stacking stems.

minimizing the amount of unpaired nucleotides, and maximization of the stabilizing effects of base pair stacks. Meanwhile, many more RNA pseudo-knots have been discovered, and their importance in several biochemical contexts has been revealed. They can exhibit enzymatic activity, control gene expression, regulate genome stability, and seem to play an active role in the replication of some RNA Viruses [208, 178, 25, 170].

The topology of pseudo-knots violates the nearest neighbor energy model. In detail, a pseudo-knotted structure exists if any nucleotide p enclosed by a base pair (i, j) with $i < p < j$ forms a base pair (p, q), with $q < i$ or $q > j$. Such 'crossing' base pairs can be arbitrarily interleaved to form more or less complex structures. The enclosing helices of a pseudo-knot may gain an additional stabilizing contribution from stacking on top of each other. Furthermore, the looped regions sometimes form base triples with their enclosing helices. As a consequence, the looped region of pseudo-knots can not easily be decomposed in algorithmical terms. In fact, the problem of predicting arbitrary pseudo-knots has been shown to be *NP-complete* [4, 140]. This is probably the main reason why most secondary structure prediction algorithms simply ignore them, and consider pseudo-knots as part of the tertiary structure instead.

Nonetheless, for some particularly common classes of pseudo-knots polynomial time DP algorithms have been proposed that take their existence into account [192, 4, 140, 55, 184, 186, 206]. These computational apporaches can be devided into *ab-initio* methods, where the pseudo-knot is a part of the underlying decomposition scheme, and *a posteriori* methods, that consider pseudo-knot formation as a hierarchical process where a nested structure forms first and the pseudo-knot is formed only after that, similar to the two-step process of RNA-RNA hybridization discussed in Section 3.7. Still, the parametrization of the variety of loops is challenging. Only a few ex-

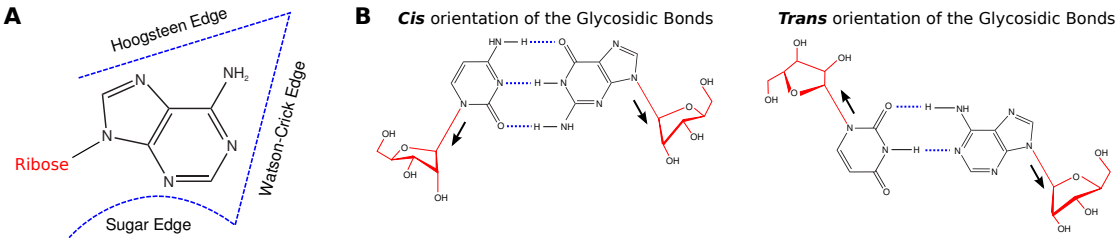


Figure 11: The determinative features of the Leontis-Westhof base pair nomenclature [129]. **A** The three edges of a nucleobase available for pairing, namely the *Watson-Crick* (W) edge, the *Hoogsteen* (H) edge, and the *Sugar* (S) edge. **B** Depending on the relative orientation of the ribose parts of two pairing nucleotides, they pair either in *cis* or in *trans*.

perimental data exist, and pseudo-knots are often further stabilized by cations such as Mg^{2+}, making it difficult to derive a generalized loop model with reasonable thermodynamic parameters [34].

THE BASE PAIR TRILOGY Within the last decades a closer look at experimentally determined three dimensional RNA structures revealed a large variety of possbile base pairs. RNA crystallography showed that not only can RNA form hydrogen bonds with their *Watson-Crick edge*, the one used in Watson-Crick base pairing (see Figure 2), but two additional edges of a nucleotides nitrogenous base are sometimes involved in base pairing, too, namely the *Hoogsteen edge*, and the *Sugar-edge* [128, 247, 129, 130], as depicted in Figure 11. Indeed, the resulting so-called *non-canonical* base pairs are quite frequent and can make up to about 1/3 of an RNAs structure [210].

To bring order into the new realm of base pair possibilities, Leontis and Westhof [129] proposed an RNA base pair nomenclature that categorizes pairs according to both involved edges, and the rotational orientation of the nucleobase relative to each other. This resulted in a distinction between twelve edge-to-edge configurations, see Figure 11. While half of them are sterically compatible with the usual *anti-parallel* orientation of the phosphate backbone, the others require a parallel configuration.

With this nomenclature in hand, tertiary structures of a folded RNA molecule can be decomposed into three-dimensional motifs, whereas their pairwise interactions can easily be added to secondary structure drawings. Unfortunately, despite the effort made in identification and annotation of these types of tertiary motifs [2, 180, 221], their incorporation into existing efficient prediction algorithms appears to be almost impossible for several reasons. First, the additional base pair types introduce a dependence on specific spatial and rotational conformations as well as orientation of the phosphate backbone. But nowadays secondary structure prediction algorithms have no sense of spatial dimensions or rotational degrees of freedom. Furthermore, the base pairs are always assumed to exhibit anti-parallel orientation.

Second, some base pair configurations can only be adopted when certain other conformations are present in the vicinity of the particular pair. This makes them context sensitive, and the nearest neighbor model does not hold anymore. The nearest neighbor model is also easily violated by so-called *base triples*, where a nucleotide forms hydrogen bonds with more than one other nucleotide [2].

Third, since some of the above special base pair types are adopted only when the RNA is interacting with a protein, ligand or other chemical compound [42, 162], it can be hard or even impossible to assess the melting temperature of such a configuration. Thus, obtaining enthalpy- and entropy-values necessary for a thermodynamic parametrization of a putative prediction algorithm is impossible.

Nevertheless, work has been done to circumvent the afforementioned drawbacks. In 2008, Parisien and Major [175] presented the `MC-fold/MC-sym` pineline, an algorithm that is capable of predicting optimal tertiary structures. The authors introduced the term *nucleotide cyclic motif* (NCM) that defines an indivisible set of nucleotides together with their corresponding interactions [127], analogous to the loops of the nearest neighbor energy model. Extraction of these motifs from crystallographic structures and the statistical analysis of their occurence within different overlapping structural contexts finally lead to an effective NCM scoring function. With this in hand, the same algorithm could be used for secondary and tertiary structure prediction. However, the original implementation of `MC-Fold` consists of an exhaustive enumeration of all stem-loop configurations that can be constructed by with the NCMs of a particular input sequence. As a consequence, the runtime of `MC-fold` scales exponentially with the input sequence length, making it applicable only to RNA sequences with no more than 150 nt. This limitation was lifted when zu Siederdissen et al. [263] presented the algorithmic framework to solve the folding problem for the MC-Fold model in polynomial time.

Still, the predictive power of such algorithms can not compete with that of the usual secondary structure prediction algorithms, yet [263]. However, this may be mostly due to the limited data available for tertiary motifs. Currently, the MC-fold model uses statistical measures derived from frequencies of certain motifs in a rather limited set of experimentally determined tertiary structures.

For some *self-contained* tertiary motifs, such as *G-Quadruplexes*, melting curves have been experimentally determined. This not only makes predictions with the MC-fold model more reliable, but also offers promising alternatives. In Chapter 10, for instance, I present a fully parameterized extension of the nearest neighbor energy model to include G-Quadruplexes.

Part II

OF PATHS AND LANDSCAPES

4

TRANSITIONS BETWEEN RNA STRUCTURES

Until now, I only presented static RNA secondary structures, i.e. all methods to predict secondary structures so far yield one or more unrelated representative results. However, an RNA molecules structure always fluctuates between different low energy conformations.s While for the protein coding sequence of some mRNAs its particular secondary structure may not be important, it matters for all ncRNAs, since their structure determines whether or not the molecule is functional. Riboswitches, for instance, usually adopt two distinct structures representing their ON-, and OFF-state. Another, rather extreme, example of RNAs that adopt distinct functional structure states are *viroids*. These tiny plant pathogens, which solely consist of an RNA molecule with a sequence length of about 400 nt, form very stable rod-like secondary structures. It has been shown, that during their life-cycle they switch between at least three meta-stable functional states, to perform their reproduction, pathogeneity, and cell-to-cell trafficking [187, 53, 52].

Furthermore, any RNA in a cell comes to life during transcription of its corresponding DNA template[1]. This polymerization process is a discrete, sequential addition of nucleotides to the 3′ end of the nascent RNA, which in turn immediately starts folding back on itself, known as *co-transcriptional folding* (see Chapter 12.2). Therefore, any RNA has to refold at least during transcription to eventually adopt its native state. Whether on not this refolding can happen within the lifetime of the RNA, i.e. before it is degraded by nucleases, can only be predicted by simulating its folding dynamics, which will be discussed in Chapter 6. Hence, to better understand the function of a potentially switching RNA, or the transition from a (transient or persistent) meta-stable state into the native structure, the process of refolding needs to be investigated.

In general, refolding from one structure into another can be divided into *cis*-, and *trans*-induced switching. *Cis*-induced switches do not require additional external factors for their structural transition. Their conformation change is purely driven by the kinetic- and thermodynamic properties of the RNA molecule itself. Hence, every RNA that relaxes its structure towards thermodynamic equilibrium from a meta-stable transient state, e.g. structure adopted during the transcription process or by degradation, is considered to act in *cis* [22, 89, 232, 78]. Furthermore, RNA thermometers [168] which change their conformation along with slight temperature changes also constitute to this class of switches. On the other hand, *trans*-induced switching

1 Most RNA viruses encode for an RNA-dependent RNA polymerase, that allows for the duplication of their genome without the need of a DNA template. Still, the nascent RNA immediately starts folding during the duplication process.

represents the change in structure upon binding of one or more auxiliary molecules. Such molecules may be other RNAs, proteins, or small ligands, e.g. metal-ions and metabolic reactants and products [42, 59, 253]. Therefore, most of the *trans*-induced switches have the ability to act as sensors of their chemical environment.

The principle of how switching occurs in both classes may be interpreted equally, as their switching rates, i.e. the difficulty of refolding, is determined by the energetic barrier separating both states [29, 164, 66]. However, for *trans*-induced switches the binding of auxiliary molecules to the RNA often physically restrains the formation of certain base pairs. This happens either by coating specific regions of the structure and thus blocking base pairs entirely, or by competition between the binding free energy of the ligand and regular base pairs. Additionally, the external factor can serve as a kind of enzyme, usually lowering the barrier to an energetically preferred state. In the following sections we will restrict ourselves to *cis*- induced switches, since (i) to date thermodynamic data on the influence of bound molecules is quite sparse, and (ii) there exists no generic solution for structure predictions of RNAs upon ligand or protein binding yet. However, in Chapter 13 I will sketch an hierarchical method that enables secondary structure predictions to incorporate some environmental constraints in a generic way.

In this chapter, the theoretical background for transition paths of RNA secondary structure conformation changes is presented. After a formalization of *elementary moves*, a scoring measure for refolding paths, the *energy barrier*, will be introduced. Furthermore, heuristic approaches to determine optimal *direct*- and *indirect* folding paths are discussed.

4.1 MOVE SETS, PATHS AND ENERGY BARRIERS

Formally, the transition process between two RNA structures can be considered a discrete sequence of intermediate structure states, a path, where a particular set of elementary modifications serves as the transforming connection between any two consecutive intermediates. The energy barrier of this path is determined by the highest free energy among the intermediates. However, the vast number of possible patterns of base pair formation and dissociation introduces a combinatorial explosion, rendering exhaustive enumeration strategies, that explore all paths, practically impossible. In fact, it has been shown, that finding an optimal (re-)folding path between two secondary structures is *NP-complete* [145, 146]. Hence, heuristic approaches which compute near-optimal solutions to the energy barrier problem have to suffice for most biological applications.

MOVE SETS AND NEIGHBORHOODS The most simple move set $\mathcal{M}$ consists of exactly two reversible modifications, *formation* and *dissociation* of a single base pair. More sophisticated move sets may also include *shift*-moves [66], resembling helix slipping. Even the formation and dissociation of entire helices can be considered elementary moves [114, 133, 111]. However, the more abstract the elementary structural changes are, the more care must be taken to ensure *ergodicity*, i.e. any structure must be accessible from any other structure through a series of applications of moves. Additionally, abstract moves such as helix-moves by definition allow for simultaneous formation or dissociation of entire sets of base pairs. Thereby, they coarse grain the underlying physical process. Hence, optimal paths derived from abstract move sets may show entirely different properties than those using formation and dissociation the most atomic constituent, a single base pair.

Introducing a move set to the ensemble of RNA structures implicitly defines a neighborhood relation. Thus, any structure s_k obtained by application of a single move $m \in \mathcal{M}$ to a structure s_i is an element of the set of neighbors $\mathcal{N}(s_i)$ of s_i:

$$\mathcal{N}(s_i) = \left\{ s_k \mid s_k = m(s_i), m \in \mathcal{M} \right\} \tag{28}$$

SECONDARY STRUCTURE FOLDING PATHS Since the move set defines the number of possible transitions an RNA undergoes during refolding, the chain of intermediate structures forms a path. More formally, a folding path

$$\vec{\mathcal{P}} = (s_i, s_{i+1}, \ldots, s_{i+n-2}, s_j). \tag{29}$$

of length $n = |\vec{\mathcal{P}}| = j - i + 1$ between two secondary structures s_i and s_j is an ordered sequence of valid secondary structures s_k, with $s_k \in \mathcal{N}(s_{k-1})$. In the following, the notation $\mathcal{P}[k] = s_{i+k}$ will be used to denote the k^{th} intermediate structure within path $\mathcal{P}$.

Of course the above definition leaves the path *unrestricted*. There is no upper limit to the length of the path, i.e. n can be arbitrary large. On the other hand, the lower limit for n is readily inherited from the move set $\mathcal{M}$. Additionally, a path $\vec{\mathcal{P}}$ may contain *cycles*, i.e. it may consist of multiple occurrences of the same intermediate structure $s_k = s_l$. Furthermore, the move-set may contain a *neutral element*, i.e. a move that does not change the base pair pattern. This would result in paths where two consecutive intermediates may be exactly the same, i.e. $\exists k \mid s_k = s_{k+1}$. In the following, I will only consider *cycle-free* folding paths and move-sets *without neutral elements*.

DIRECT AND INDIRECT PATHS As a convention, one usually distinguishes *direct paths* from *indirect paths*. By definition, a direct path $\vec{\mathcal{P}}^D$ is *minimal* with respect to the move set. Therefore there exists no other path $\vec{\mathcal{P}}$ with shorter length, i.e. $\nexists \mathcal{P} \mid |\vec{\mathcal{P}}| < \vec{\mathcal{P}}^D$. A direct path can be constructed from the start-structure s_i by either adding base pairs not present in s_i but in the stop-structure s_j, or by removal of base pairs not present in s_j but in s_i. Hence

$$\begin{aligned} \vec{\mathcal{P}}^D &= (s_i, \ldots, s_k, \ldots, s_j) \text{ with} \\ s_k &= \left\{(p,q) \mid (p,q) \in \left(s_i \cap s_j\right) \cup \left(s_i \triangle s_j\right)\right\} \text{ and} \\ |s_k \cap s_i| &\geqslant |s_{k+1} \cap s_i| \text{ and} \\ |s_k \cap s_j| &\leqslant |s_{k+1} \cap s_j| \end{aligned} \tag{30}$$

where the symmetric difference $s_i \triangle s_j$ is the set of base pairs that are unique to either s_i or s_j (see Chapter 2.2). Since we restricted ourselves to cycle-free paths without neutral moves, all direct paths between s_i and s_j have the same length $n = |\vec{\mathcal{P}}^D| = |s_i \triangle s_j|$. Moreover, any direct path $\vec{\mathcal{P}}^D$ is *minimal* with respect to the underlying move set, i.e. there is no other path $\vec{\mathcal{P}}$ with $|\vec{\mathcal{P}}| < |\vec{\mathcal{P}}^D|$. For indirect paths, there is no such restriction, of course. Their minimal length is limited to that of a direct path with the same start- and stop-structure. However, their maximal length is only bound by the number of structures compatible with the same RNA sequence.

SADDLE POINTS AND ENERGY BARRIERS Any secondary structure s_k of a path $\vec{\mathcal{P}}$ is associated with a free energy $E(s_k)$, according to a specific energy model for RNA structures. Therefore, $\vec{\mathcal{P}}$ represents a path through a one-dimensional space, where each s_k is a position, and $E(s_k)$ is its height. Along a path there may be mountains to climb and valleys to descend to, in terms of free energy. Any structure that exhibits the highest free energy along the path $\vec{\mathcal{P}}$ is a *saddle point* $S(\vec{\mathcal{P}})$s with

$$E(S(\vec{\mathcal{P}})) = \max_{0 \leqslant l \leqslant |\vec{\mathcal{P}}|} E(\mathcal{P}[l]). \tag{31}$$

Hence, the energy difference between the start structure and the saddle point can be considered the amount of *activation energy* to allow the transition from the start- to

the stop-structure. Finding paths where the lowest amount of activation energy is required, i.e. paths with the lowest energy barrier

$$B(\vec{\mathcal{P}}) = E(S(\vec{\mathcal{P}})) - E(\mathcal{P}[0]) \tag{32}$$

is of great interest in RNA structural biology. Such paths are optimal in terms of free energy change, and therefore considered most likely. However, even finding optimal direct refolding paths is a hard to solve problem, in fact it is *NP-hard* [145]. Consequently, approaches are restrained to heuristic methods to avoid exhaustive enumeration of all possible paths.

Below, I will present some of the most widely used heuristics that deal with the problem of finding optimal folding paths. While most of the available methods restrict themselves to direct folding paths, only a handful exist that lift this restriction. Yet allowing to insert and later remove temporary stabilizing base pair not present in either the start- or stop-structure can drastically reduce the energy barrier of a path. The same is true for temporary removal of base pairs which later on have to be inserted again. For this reason, it is crucial to discuss those methods as well.

4.2 PREDICTION OF DIRECT FOLDING PATHS

MORGAN-HIGGS ALGORITHM Among the first to come up with an efficient heuristic to compute near optimal direct folding paths is the work of Morgan and Higgs [164]. Their greedy algorithm creates a direct folding path between two secondary structures s_i and s_j. The order in which base pairs are added and removed is determined by the number of conflicts, whereas fewer conflicts lead to earlier insertions, in particular, the algorithm can be formulated as follows:

1. Set $s_A = s_i$
2. Find $(p, q) \in s_j$ with least incompatible pairs $(k, l) \in s_A$
3. Remove all incompatible pairs from s_A, i.e. $\forall (k, l) \in s_A$ incompatible with (p, q), do $s_A = s_A \ (k, l)$
4. Insert (p, q) into s_A, i.e. $s_A = s_A \cup (p, q)$
5. If additional pairs $(r, s) \in s_j$ can be inserted after removal of (k, l), insert them as well: $\forall (r, s) \in s_j$ compatible with s_A, do $s_A = s_A \cup (r, s)$
6. If $s_A \neq s_j$, goto 2.

In each step of removal or addition of a base pair, the energy of s_A is evaluated and used to update the saddle point energy $E(S) = max_{s_A} E(s_A)$. In cases where step 2. of the algorithm encounters more than one base pair (p, q) with equal number of incompatible pairs, it chooses one randomly. For such cases, the authors suggest to

repeat the algorithm a number of times and follow each new route only as long as the energy of the newly computed saddle point is lower than the one obtained by a previous random choice. While the algorithm is very fast it relies on the simplistic assumption that each base pair can be treated the same. Therefore, it does not account for differences in free energy change induced by addition/removal of different base pairs and thus, tends to overestimate the energy barrier.

THE FINDPATH ALGORITHM In 2001, Flamm et al. presented a new, breadth-first search algorithm that computes near optimal direct folding paths based on the more realistic nearest neighbor model of RNA secondary structures [67]. In each step of the path construction between s_i and s_j, the algorithm generates all neighbors $\mathcal{N}(s_k)$ of the current structure s_k that are closer to the s_j in terms of the move set distance, and determines their free energy. From the resulting set of possible neighboring structures, the one with lowest free energy is chosen as next intermediate and the algorithm proceeds from there until the target structure is reached. In principle, this resembles a nearest neighbor model extension of the Morgan-Higgs algorithm, that follows only energetically favorable paths instead of removing and adding base pairs according to their number of collisions. However, Flamm et al. introduced a parameter m that allows to keep track of the best m folding paths found so far. This transforms the whole algorithm into a breadth-first search strategy controlled by m. Setting $m = 1$ degenerates to a greedy search for the optimal refolding path, which is essentially the Morgan-Higgs algorithm with free energy scores for the choice of the next intermediate structure instead of a base pair collision list. On the other hand, the algorithm becomes exhaustive if $m = n_p$ where n_p is the number of possible paths. This makes it a fast and powerful tool to determine close to optimal folding paths. Hence it yields good approximations of energy barriers along direct paths between any two secondary structures s_i and s_j.

FLOODING ALGORITHM OF THE BARRIERS PROGRAM A non-heuristic approach that relies on an exhaustive enumeration of all secondary structures is implemented in the `barriers` program [68]. Although it is actually designed to compute the barrier tree [207] of the free energy landscape (see also Chapter 5), its algorithm is capable to easily compute the optimal folding path between two local minima as a by-product. Since it constitutes an exhaustive approach it will always compute the exact energy barrier between two secondary structures. However, in general this exhaustiveness is the actual drawback of this method, since the number of structures compatible with an RNA grows exponentially in the sequence length. Thus, the `barriers` program is limited to RNAs with lengths of about 100 nt on current computers, and, due to its

exponentially increasing large input, is not expected to be applicable to much longer RNA sequences in the near future[2].

4.3 TAKING DETOURS

Although the aforementioned algorithms to compute direct folding paths perform quite well, they usually suffer from overestimation of the energy barrier. Since an RNA upon refolding from s_i to s_j does not 'know' about being restricted to direct paths, it is more likely that it follows the energetically most optimal. The intermediate structures the RNA adopts during its refolding process may therefore consist of additional base pairs, not present in either s_i or s_j. Alternatively, base pairs present in both s_i and s_j may be opened during the transition to be later on inserted again. As long as deviating from the direct path contributes favorably to the intermediates stability this may drastically lower the predicted energy barrier. Below, two approaches on how to compute indirect folding paths are presented. A third method can be found in Part iv, *2D Projections of RNA folding Landscapes* (Chapter 9), which is described in more detail in part v, Chapter 11.

MORGAN-HIGGS - REVISITED Within their publication of the direct path heuristic, Morgan and Higgs already suggested an algorithm capable to construct indirect paths. Using the single link cluster (SLC) method they compute optimal indirect folding paths from s_i to s_j that consist of concatenated optimal direct paths between pairs of auxiliary structures s_k and s_l, in which the first direct path starts with s_i, and the last ends with s_j [164]. A problem of this approach is, however, that it is not clear which set of auxiliary structures one should choose. The authors of this method used the Nussinov model to sample *a set of ground-state structures* from the resulting partition function. But this method suffers from the possibility to generate structures too far away from potentially arbitrarily chosen start and stop structures s_i and s_j. Furthermore, as shown in Part v, Chapter 11, Boltzmann sampling from the whole ensemble, regardless of the energy model applied, does usually not give enough insight into its structural diversity. Therefore, other strategies have to be applied. Another possibility to obtain low free energy structure might the use of Zuker-type suboptimals or suboptimal structures within a specific energy band around the MFE, as introduced in 3.3. However, again, if any of the two structures s_i and s_j are, in terms of base pairs, sufficiently far away from the generated set, the resulting paths might depart too much from any meaningful result. In addition, the resulting set of auxiliary structures obtained by such methods can grow very fast, rendering the method impractical.

2 Moore's law predicts a doubling of computational power every two years or so. Increasing the length of an RNA by only a single nucleotide, however, almost doubles the number of possible secondary structures (actually its a factor of 1.8. Hence, about every two years, the RNA used as an input may grow by one single nucleotide.

RNATABUPATH A more recent method that aims to predict optimal indirect folding paths is implemented in the *RNATabuPath* algorithm [58]. The algorithm is a semi-greedy path construction between the start structure s_i and the stop structure s_j. The successive transitions from one structure s_k to its neighbor $s_{k+1} \in \mathcal{N}(s_k)$ are determined by a fitness function $F(s_{k+1}) = E(s_{k+1}) + w \cdot d_{BP}(s_{k+1}, s_j)$. The fitness function F takes into account both, the free energy $E(s_{k+1})$ of the next intermediate structure of the path, as well as its distance to the target structure in terms of base pairs difference $d_{BP}(s_{k+1}, s_j) = |s_{k+1} \triangle s_j|$. This latter term is further controlled by a weighting factor w which essentially determines how far the path may deviate from a direct path between s_i and s_j. Hence, it serves as a *guiding potential* that balances the importance of choosing a low energy neighbor versus reaching the target s_j. Although, low values of w allow for paths that deviate very much from a direct path, and large values of w result in fast convergence toward s_j, there is no guarantee that the algorithm eventually reaches its stop structure s_j. Therefore, the authors employed an oscillation strategy, that adapts the value of w during path construction. In particular, w is increased if the distance to the target has not been improved for a number of iterations, and lowered if the distance decreases again, while it is ensured that w stays within a user-defined range of $w \in [w_{min}, w_{max}]$. Furthermore, the algorithm utilizes a so-called *tabu-list* of previously inserted or removed base pairs, mainly to avoid introducing cycles along the path. As long as a base pair is within this list, it may not be modified, i.e. inserted or removed again. Base pairs will be removed from the tabu-list after a certain number of steps. Although this method seems to produce good results for the examples tested in their publication, the initial choice of the weighting factor range $[w_{min}, w_{max}]$ is crucial for the algorithms performance. Furthermore, the algorithm deterministically creates just a single trajectory, i.e. it explores only a single path. Thus, to obtain a diverse set of near optimal solutions the algorithm has to be applied multiple times with different initial conditions for $[w_{min}, w_{max}]$.

5 THE BIG PICTURE

5.1 SECONDARY STRUCTURE FREE ENERGY LANDSCAPE

The entire ensemble of secondary structures of an RNA forms a high-dimensional free energy landscape. Again, each structure s_i corresponds to a particular location in this space, and the free energy $E(s_i)$ determines the height. However, in contrast to paths, as discussed in the previous chapter, each structure s_i has a multitude of neighbors (see Fig. 12). Strictly speaking, a folding path $\vec{\mathcal{P}}$ is a trajectory through the secondary structure free energy landscape

$$\mathcal{L} = (\Omega, \mathcal{M}, E). \tag{33}$$

Here, Ω is the set of conformations, the move set $\mathcal{M}$ defines the topology, and the energy function $E : s_i \rightarrow \mathbb{R}$ associates each conformation $s_i \in \Omega$ with a fitness value, i.e. its height within the landscape.

Since each conformation may have a large number of neighbors[1], the resulting landscape forms a high-dimensional space. Similar to paths, the resulting landscape usually consists of valleys, mountains, ridges, funnels, and plateaus.

The secondary structure free energy landscape can then be analyzed in various ways, e.g. to extract shortest paths between two secondary structures (see previous Chapter 4), or to simulate RNA folding kinetics (see Chapter 6). However, since it is such a high-dimensional space, this landscape can not be visualized easily. But methods exist, that lump large portions of the landscape into a single point, i.e. a macro-state and therefore allow to visually inspect certain aspects of the landscapes underlying properties. Prominent examples of such visualizations are the *barrier tree* [66] and the 2D projection based on two reference structures, as implemented in the *RNA2Dfold* algorithm [136]. The latter is an *ab-initio* method that projects the high-dimensional space into two dimensions using two representative reference structures. It is presented in detail in Part iv, *2D Projections of RNA folding Landscapes* (Chapter 9) and will be briefly introduced in this chapter, section 5.3.

CONSTRUCTION OF THE LANDSCAPE The landscape itself is usually constructed by *exhaustive enumeration* methods. This means, that for a landscape analysis all secondary structures have to be generated, stored somehow, and assigned a free energy score. But this, in turn, limits the applicability to relatively short RNAs, since even an

1 The unfolded state of an RNA with sequence length n, may have about $n^2/2$ neighbors. Each of them results from the formation of one of the $n^2/2$ possible base pairs.

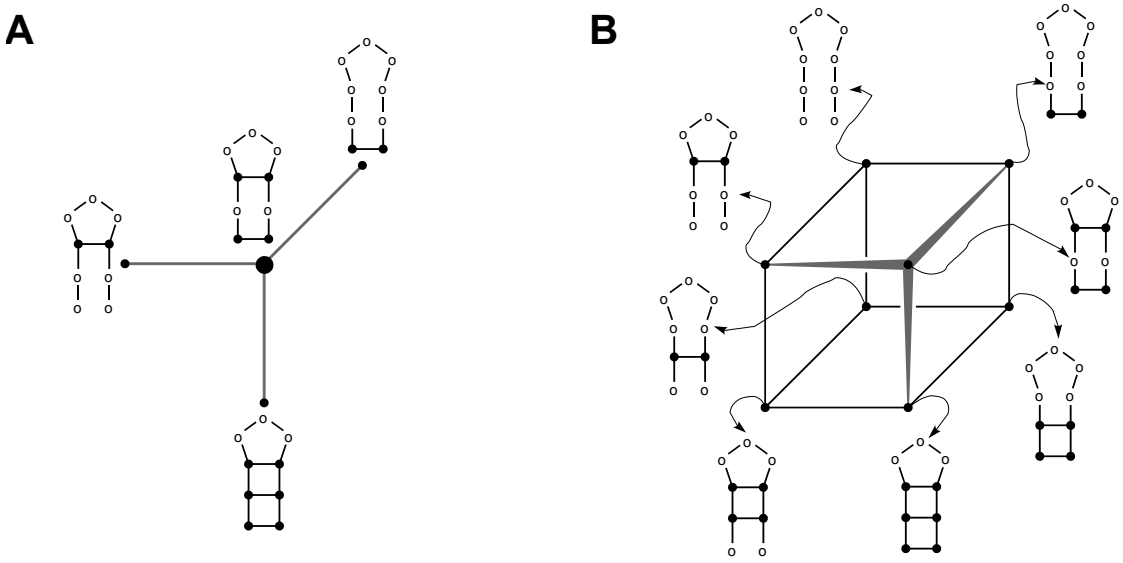


Figure 12: The Neighborhood of a Secondary Structure. **A** Considering a single secondary structure (in the center of this figure), each elementary move, such as the *formation* and *dissociation* of a single base pair, transforms a structure into one of its many neighbors. **B** Any neighbor, again, has neighbors itself. Consequently, the structure ensemble Ω spans a high-dimensional space of structure conformations.

RNA with sequence length 50 may have compatible secondary structures in the order of tens of billions. Nevertheless, most parts of the actual landscape might not contribute to the solution space of a particular problem, for instance finding an optimal refolding path. Such parts may be those, whose structures exhibit free energies above a certain threshold, for instance positive free energy. Furthermore, structures with so-called *lonely base pairs*, i.e. base pairs not surrounded by any other pair may usually be omitted from the landscape construction, since such base pairs are widely considered unstable. Therefore, leaving out certain parts of the landscape makes exhaustive landscape construction feasible for RNAs up to the length of 100 nt. However, omitting structures with higher free energies can result in a landscape representation with unconnected 'islands', canceling out entire sets of solutions to particular landscape analyses. On the other hand, limitation to *canonical secondary structures*, i.e. those without lonely base pairs, introduces the necessity to adopt the move set. Thus, for any restraint of the secondary structure space elaborate measures might have to be taken. Utilization of *Boltzmann sampling* from the entire structure ensemble is usually not advisable and even inapplicable to generate large enough parts of the landscape, even for extremely large sample sizes. Depending on the ruggedness, and depth and distribution of the local minima within the secondary structure space, this method strongly tends to produce only spots close deep to folding funnels, since their surroundings take up large volumes in the probability space of the ensemble (see Chapter 3.4, and

Chapter 12 in Part V of this work). Therefore, a program often used to generate a (reduced) set of secondary structures is *RNAsubopt* of the *ViennaRNA package*, which, amongst others, implements the suboptimal secondary structure prediction algorithm according to Wuchty et al. [255] (see Chapter 3.3).

5.2 BARRIER TREES AND GRADIENT BASINS

A compact and convenient, yet physically meaningful way to reduce the enormous secondary structure free energy landscape is the use of *barrier trees* [252]. This abstraction consists of trees with local minima at their leaves, which form the kernel of so-called *gradient basins*, and minimal saddle points as inner nodes connecting them. Thereby, these gradient basins subsume entire sets of structures from which a *gradient walk* would always end up in the local minimum that constitutes the basin. Here, a gradient walk is the repeated application of a mapping $\gamma : \Omega \to \Omega$, that assigns each structure s_k to its neighbor $s_l \in \mathcal{N}(s_k)$ with lowest free energy $E(s_l) < E(s_k)$. After a finite number t of mappings each structure s_k is eventually assigned to a unique local minimum $z = \gamma^\infty(s_k) = \gamma^t(s_k)$. As a consequence, the local minima represent attractors of the mapping γ, whereas the basins of attraction $\mathcal{B}(z) = \left\{ s_k \in \Omega \mid \gamma^\infty(s_k) = z \right\}$ form a partition of Ω. The flooding algorithm implemented in the *barriers* program [252] processes an energy sorted list of all suboptimal structure states. It identifies all local minima and their connecting minimal saddle points to construct the coarse grained representation of the landscape, the barrier tree. A visual description of the flooding algorithm is given in Figure 13, and an example for the barrier tree representation can be found in Figure 14.

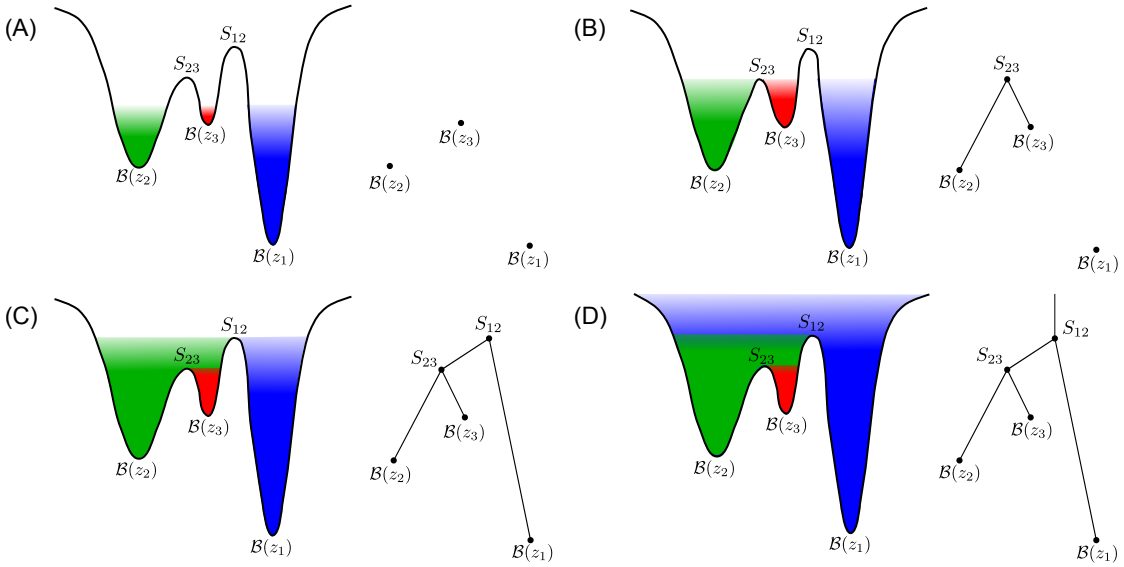


Figure 13: The *flooding algorithm* [68] constructs a *barrier tree* from an energy sorted list of secondary structures. **A** Local minima, i.e. structure states where all neighbors have a higher free energy, form the leaves of the tree. Any other structure is assigned to the local minimum that is reachable through a *gradient walk*, thereby contributing to the *gradient basins* $\mathcal{B}(z_1)$, $\mathcal{B}(z_2)$, and $\mathcal{B}(z_3)$. **B** As the free energy level of the input structures rises, *saddle points* appear that connect two gradient basins (S_{23}). The saddle points form the inner nodes of the barrier tree, connecting two gradient basins with each other. **C** Structures that would be assigned to a gradient basin that is already connected to another one, are assigned to the energetically deeper of both. **D** When the input is completely processed, a compact representation of the high-dimensional free energy landscape is available, the *barrier tree*, whose node lengths represent the refolding energy barriers between the local minima, i.e. the leaves.

5.3 2D PROJECTIONS

In contrast to the aforementioned method of barrier tree construction, the *RNA2Dfold* algorithm [136] allows for an *ab-initio* construction of a coarse grained landscape representation. This eliminates the necessity of an exhaustive enumeration of secondary structures. The method relies on two initially chosen reference structures s_i and s_j which are used to partition the secondary structure free energy landscape into distance classes with respect to the references. In detail, any structure s_k is placed in a particular κ,λ-neighborhood, where $d_{BP}(s_k, s_i) = \kappa$ and $d_{BP}(s_k, s_j) = \lambda$. A classified DP algorithm is utilized to compute specific thermodynamic properties, such as minimum free energy and partition function, for each of the resulting distance classes. Although the algorithm has an asymptotic time complexity of $\mathcal{O}(n^7)$ and memory requirements of $\mathcal{O}(n^4)$, an elaborate implementation in the program *RNA2Dfold* exploits the sparsity of the dynamic programming matrices, making it applicable to RNAs with sequence lengths of up to about 400 nt. The resulting thermodynamic properties can readily be used to depict a projection of the high-dimensional landscape into two-dimensions, where the x-axis represents the base pair distance to reference structure s_i, and the y-axis the distance to s_j (see Figure 14). By carefully choosing the two references, e.g. the MFE structure and a meta-stable state, a representation of lowest free energy for each distance class already provides a lower-bound on the refolding barrier between the two references.

Since this method does not enumerate all structures but uses a classified DP recursion scheme, only thermodynamic properties of the distance classes are available. However, stochastic backtracking can be used to sample structural representatives from the distance classes according to their intrinsic equilibrium probabilities. This allows to create structurally diverse sample sets, which would not be possible via regular stochastic backtracking approaches. Furthermore, based on the distance class partitioning, a method to predict indirect refolding paths between two secondary structures, as well as an approach to simulate RNA folding kinetics are presented in Part v.

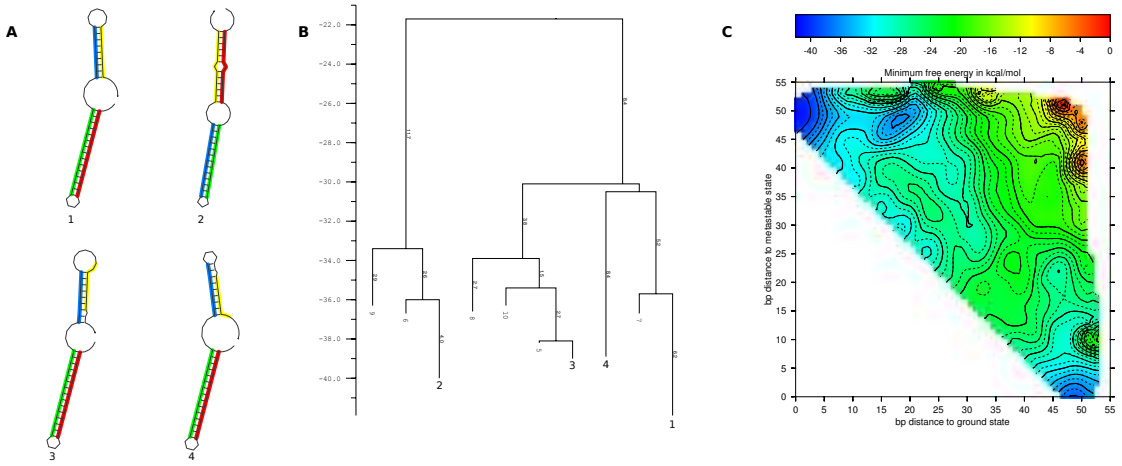


Figure 14: Different representations of an artificially designed riboswitch taken from Xayaphoummine et al. [258]. **A** Secondary structures drawings of the MFE structure (1), the meta stable state that is adopted by co-transcriptional folding (2), and two structures similar to the MFE structure but with slightly shifted 5′ stem (3,4). **B** The corresponding barrier tree with a resolution of 10 macro states. **D** The 2D projection of the MFE representatives obtained from `RNA2Dfold` with the MFE structure (1) and the meta stable state (2) as reference, depicted as large blue areas in the upper left and lower right corner, respectively.

Part III

RNA FOLDING KINETICS

6

RNA FOLDING AS A DYNAMIC PROCESS

In the previous chapter, RNA folding paths and the concept of the secondary structure free energy landscape were introduced. However, both of them only give a rather static view on the transition space of an RNA. To provide information about the involved dynamical behavior when an RNA consecutively adopts distinct secondary structures, each transition has to be linked with a rate or probability. In theory, RNA folding dynamics can be predicted using *molecular dynamics* (MD) simulations. Using the positions, velocities and accelerating forces of the atoms of the RNA, successive snapshots of the changing state space can be obtained through integration of Newton's laws of motion [7, 226]. However, the involved computations of the interacting forces is implementation-wise demanding and the simulation itself computationally expensive. Furthermore, an RNA contains a large number of atoms, so does its surrounding solution Hence, in most practical cases RNA folding dynamics is simulated on a much more coarse grained level, such as the secondary structure. This is where the definition of the energy landscape comes in handy. Each point in the landscape represents a distinct microstate of the RNAs secondary structure system. The states height in the landscape represents their equilibrium probabilities given by each associated free energy (see Equation 16). Transitions within this space of conformations, the canonical ensemble, can now be thought of as a stochastic kinetic process [66]. Therefore, equilibrium statistical mechanics (statistical thermodynamics) can be applied and they can be modeled by a Markov process.

This chapter starts off with some thoughts on the assessment of transition rates between secondary structure states, and discusses two complementary approaches on RNA folding kinetics. The first approach uses the Markov property of the folding process to directly compute the time evolution of the entire Markov process. For this purpose a solution to the *master equation* which describes the Markov process has to be obtained. This immediately allows for a global view on the folding dynamics over time.

The second method, the *Markov Chain Monte Carlo* (MCMC) method, allows for sampling of a single folding trajectory through the vastness of the energy landscape. Therefore, it can be used to follow an RNA molecule along its folding path. This unfortunately may be its weakness as well. Since the folding process is regarded as a stochastic process many folding trajectories are required to obtain a global view of the system, even for short RNA molecules. However, the number of folding paths grows exponentially with sequence length. Therefore, the number of required trajectories

grows very fast as well, rendering the total computational demands for a Monte Carlo simulation very high.

Furthermore, *coarse graining* approaches for secondary structure landscapes will be introduced in this chapter. Such methods represent a useful tool to circumvent the issue of too many states by lumping sets of *micro-states* into more abstract *macro-states*. Folding simulations can then be performed on the (much smaller) system of macro states, thereby enabling application to longer RNA molecules.

6.1 TRANSITION RATES

From now on, we consider RNA folding kinetics as a stochastic process, close to conventional stochastic kinetics of chemical reactions [66, 74]. In particular, it can be modeled as a continuous-time Markov process. The population density[1] $P_i(t)$ of a conformation s_i changes over time t. The rates of influx k_{ji} and outflux k_{ij} from, resp. to another conformation s_j can be used to describe the transition process with the master equation

$$\frac{dP_i(t)}{dt} = \sum_{j \neq i} (P_j(t)k_{ji} - P_i(t)k_{ij}). \tag{34}$$

The elementary moves within our conformation space only allow for transitions between neighboring conformations. Therefore, the transition rates between most conformations are zero $k_{ij} = 0, \text{if } s_j \notin \mathcal{N}(s_i)$ (see Chapter 4.1 for the definition of $\mathcal{N}$), rendering the transition rate matrix $R = \{k_{ij}\}$ sparse. Since in equilibrium $\frac{dP_i(t)}{dt} = 0$, both terms on the right hand side of Equation (34) must be equal. This immediately leads to a relation that is commonly known as *detailed balance*:

$$P_j(t)k_{ji} = P_i(t)k_{ij}. \tag{35}$$

In (16) we learned that given the free energy $E(s_i)$ of a secondary structure state s_i and the partition function Q the states equilibrium probability is given as

$$P_i(t) = \frac{e^{-\beta E(s_i)}}{Q}. \tag{36}$$

Together with the criterion of detailed balance this dictates that the transition rates between two distinct states s_i and s_j have to be consistent with their difference in free energy $\Delta G_{ij} = E(s_j) - E(s_i)$. In particular any set of transition rates satisfying the following relation is acceptable

$$\frac{k_{ji}}{k_{ij}} = \frac{e^{-\beta E(s_i)}}{e^{-\beta E(s_j)}} = e^{\beta(E(s_j)-E(s_i))} = e^{\beta \Delta G_{ij}}. \tag{37}$$

Once all k_{ij} for $i \neq j$ are determined, the self transitions k_{ii} are computed by

$$k_{ii} = 1 - \sum_{j \neq i} k_{ij}. \tag{38}$$

Another crucial requirement for the transition rates is that there exists a series of consecutive transitions that connects any state s_i with any other state s_j. This basic assumption of statistical mechanics, the *ergodic hypothesis*, not only ensures that the

1 This is the probability to observe a conformation s_i at time t: $P_i(t) = \text{Prob}\{\mathfrak{X}(t) = s_i\}$

system contains no purely *absorbing* state. But it ensures, that the time a simulation trajectory resides in a specific region of the state space is proportional to the volume of this region, and therefore equilibrium can be reached eventually.

Still, the above preconditions leave a large number of feasible ways to compute transition rates, of which the two most commonly used definitions will be presented below.

METROPOLIS RULE The earliest expression to compute transition rates in statistical physics was introduced by Metropolis et al. [158] in 1953. Here, the assumption is that a particle within the system immediately changes it state from s_i to a neighbor s_j if the new state is energetically more favorable, i.e. $E(s_j) \leqslant E(s_i)$. Otherwise, the transition is simply determined by the energy difference ΔG of both states. Of course, in the real world it actually requires some time τ_0 to physically change the configuration, but this time is usually set to unity and therefore often not present in the expression of the Metropolis rule: [158]

$$k_{ij} = \begin{cases} \tau_0^{-1} e^{-\beta \Delta G_{ij}} & \text{if } E(s_j) > E(s_i), \\ \tau_0^{-1} & \text{otherwise} \end{cases} . \tag{39}$$

KAWASAKI RULE In 1966, Kyozi Kawasaki presented an additional rule to calculate the spin diffusion constant for time-dependent Ising models [121].

$$k_{ij} = e^{-\frac{1}{2}\beta \Delta G} \tag{40}$$

Since the stochastic characterization of RNA folding kinetics as presented in this work follows the general theories for such Ising models[2] this rule can also be applied. In contrast to the Metropolis rule in (39), here, transitions to energetically more favorable states are not treated all equally. This rule rather introduces symmetry in the forward and reverse transition rates. Applied to RNA folding kinetics, it was found that the folding performance improved substantially without changing the character of transition paths when the Metropolis rule is substituted by the Kawasaki rule [66].

2 Named after the German physicist Ernst Ising, these mathematical models represent the most simple approaches to describe the kinetics of transitions between discrete states of a system. Starting with an initial state (or a vector of population densities) the statistical mechanics behind the Ising model always ensures relaxation towards equilibrium.

6.2 SOLVING THE MASTER EQUATION

RNA folding kinetics can now be regarded as a multi-molecule system where state transitions of the molecules are described by the master equation (34). This allows to follow the systems relaxation towards equilibrium starting from an initial population density of its states over time. For a system with N distinct states, the master equation essentially describes a set of N coupled linear ordinary differential equations with constant coefficients given by the rate matrix $R = \{k_{ij}\}$. So we can rewrite (34) in vector form with $\vec{p}(t) = \left(p_1(t) \dots p_i(t) \dots p_N(t) \right)^T$ to obtain

$$\frac{d}{dt}\vec{p}(t) = R\vec{p}(t), \tag{41}$$

where the formal solution, given an initial population density $\vec{p}(0)$ follows from integration of this expression

$$\vec{p}(t) = \vec{p}(0) \cdot e^{t \cdot R}. \tag{42}$$

Looking at the above solution, the first observation that can be made is that matrix R is in the exponent. But it is not immediately clear what the meaning of a matrix in the exponential function is. However, using the convergent Taylor series of the exponential function

$$e^t = 1 + t + \frac{t^2}{2!} + \frac{t^3}{3!} + \dots \tag{43}$$

the *matrix exponential* can be formally defined as

$$e^{t \cdot R} = I + tR + \frac{t^2 R^2}{2!} + \dots = \sum_{i=0}^{\infty} \frac{R^i}{i!} \tag{44}$$

with identity matrix $I = R^0$ [191, 16]. Still, application of this power series to real data can be enormously expensive. Therefore, other numerical solutions have to be exploited in order to efficiently compute the matrix exponential [159, 160] for a system with N states and its constant $N \times N$ rate matrix R.

MATRIX DECOMPOSITION METHOD One of the most efficient methods for large matrices and repeated evaluation of e^{tR} is based on decompositions of the rate matrix R [160]. In cases where R is symmetric a similarity transformation of the form

$$R = ABA^{-1} \tag{45}$$

in conjunction with the power series definition of (44) leads to

$$\begin{aligned} e^{tR} &= I + ABA^{-1} + \frac{1}{2}ABA^{-1}ABA^{-1} + \frac{1}{6}ABA^{-1}ABA^{-1}ABA^{-1} + \dots \\ &= I + ABA^{-1} + \frac{1}{2}AB^2A^{-1} + \frac{1}{6}AB^3A^{-1} + \dots \\ &= A(I + B + \frac{1}{2}B^2 + \frac{1}{6}B^3 + \dots)A^{-1} \\ &= Ae^{tB}A^{-1}. \end{aligned} \tag{46}$$

The idea behind this transformation is to find a matrix A which makes e^{tB} easy to compute. Although the transition rate matrix R itself is not symmetric, the reversibility of the Markov process, as required by the detailed balance condition, makes it symmetrizable. In particular, a symmetric matrix $S = V\Lambda V^{-1}$ exists, where V is the matrix of eigenvectors of S, and $\Lambda = diag(\lambda_1, \dots, \lambda_N)$ is a matrix with the eigenvalues of S in the diagonal.[3] Now, the naïve approach is to take $A = V$ and $B = \Lambda$ which reduces the matrix exponential to the trivial case of exponentials of N scalars:

$$\begin{aligned} e^{tS} &= Ae^{tB}A^{-1} = Ve^{t\Lambda}V^{-1}, \text{with} \\ e^{t\Lambda} &= diag(e^{\lambda_1 t}, \dots, e^{\lambda_N t}) \end{aligned} \tag{47}$$

The remaining difficulty is to substitute all transformations back to successfully compute $\vec{p}(t)$. However, a major problem with this approach is that any rounding errors that occur during the transformations as well as the actual computations can be magnified enormously. This is especially the case, if $cond(V)$ is large, rendering V close to singular. Thus, V has to be carefully checked to assess whether the matrix decomposition method can be applied to a particular problem or not.

With the above method in hand, Markov processes of some 10.000 states can be handled efficiently on today's computers, e.g. with the *treekin* program [252] which implements a master equation solver according to the matrix decomposition method presented before. But still, in the context of RNA folding kinetics where the number of structure states grows exponentially with sequence length, only very short sequences can be analyzed. In the following, two methods are presented which aim to solve this problem for longer sequences by (i) approximation through sampling, and (ii) reduction of the state space into *macro states*.

3 Note, that S is a symmetrized version of R and therefore, the eigenvalues and eigenvectors of S and R are the same.

6.3 MARKOV CHAIN MONTE CARLO METHOD

In almost all cases where stochastic simulation of the kinetics of a system is attempted, the system is very large in terms of distinct states. This is especially true for a system of RNA secondary structures which grows exponentially in the RNAs sequence length. Apart from the vast number of transition rates to compute, a direct solution of the master equation (34) for such large systems can not be obtained within reasonable time. Furthermore, depending on the method used to compute the matrix exponential, large transition rate matrices may introduce numerical instability. This might be the cause why physicists deal with this problem for several decades by approximating the global behavior of the system through averages over samples of individual states. This sampling process is called *Monte Carlo Scheme*. It produces a time-ordered path of states by performing a non-deterministic walk through the underlying landscape. Each consecutive state is randomly selected from the set of neighbors of a previous state according to individual acceptance probabilities. This scheme was introduced by Metropolis et al. in 1953 [158] and later extended by W.K. Hastings in 1970 to the general case [99]. Therefore, this method is also known as *Metropolis-Hastings algorithm*[4] and can be described by the following simple rules:

1. Select an initial state s_i
2. Choose a neighboring state s_j (according to the probability distribution of all possible neighbors)
3. compute the transition rate k_{ij} to this neighbor (e.g. with Metropolis rule)
4. generate a random number r with $0 < r < 1$
5. if $r < k_{ij}$, set $s_i = s_j$
6. proceed with 2.

The simulation ends after a certain simulation time t has passed, where the simulation time correlates with the number of transitions that take place. Note, that a self transition, i.e. several rounds where the system does not leave s_i due to successive rejections of several neighbors, is also considered a valid transition that increases simulation time. This probability is given as

$$p(s_i \to s_i) = 1 - \sum_{j \neq i} p(s_i \to s_j) \tag{48}$$

4 In this algorithm, the probabilities to select a specific neighbor along the path only depends on the current state. Accordingly, the resulting set of paths are specific realizations of the underlying Markov chain, hence the name *Markov Chain Monte Carlo* (MCMC) method.

where $p(s_i \to s_j)$ is the transition rate from s_i to s_j. Finally, the state s_i obtained at the end of the simulation is used to update the statistical average global system properties. After multiple applications of the MCMC algorithm, the global population density is simply derived from the arithmetic average of the obtained samples.

It can be easily seen that due to a continuous rejection of moves the MCMC algorithm may be stuck in its main loop without updating its current state s_i for many iterations. This can severely slow down an actual implementation of the algorithm. More importantly, the Metropolis-Hastings algorithm as described above can only be applied to systems where the number of neighbors stays the same for all states, since the transition rate $p(s_i \to s_j)$ is the product of the *a priori* probability to attempt the move $\mathcal{A}(s_i \to s_j) = \frac{1}{N}$ and the probability to accept the move $\mathcal{P}(s_i \to s_j)$, where N is the number of neighbors. However, N is most likely not a constant for different secondary structures in the RNA secondary structure landscape, therefore detailed balance is not preserved [65]. However, the *Gillespie* algorithm [84] can be used to circumvent this problem, and, as a by-product, to speed up the simulation. It transforms the simulation into a *rejectionless* Monte Carlo algorithm by scaling the probabilities accordingly

$$\widehat{p}(s_i \to s_j) = \frac{1}{\sum_j p(s_i \to s_j)} \cdot p(s_i \to s_j). \tag{49}$$

In contrast to a direct solution of the master equation, the Monte Carlo approach is very economical in terms of memory consumption. There is no need to compute and store a transition matrix. The acceptance probabilities for a transition to any neighboring state may be computed on-the-fly.

First computational methods aiming to predict reliable secondary structures and/or (re-)folding paths by taking RNA folding kinetics into account date back almost as long as thermodynamic secondary structure predictions. In 1984, Nussinov and Pieczenik [172] already proposed a crude algorithm for secondary structure prediction that takes the hierarchical nature of helix nucleation into account. By using a Monte Carlo simulation that at any step inserts helices according to their equilibrium probability, this simplistic method was refined and even allowed to incorporate pseudo-knot interactions and folding during transcription [148, 1, 93]. However, these early approaches considered formation of helices only, ignoring their dissociation capabilities. Incorporation of the reversibility of helix formation was first achieved in 1995 by the genetic algorithm of Gultyaev et al. [91, 92].

In 2000, two efficient algorithms implementing Monte Carlo simulations were made available by Flamm et al. [66] and Isambert et al. [114]. Flamm et al. considered a structure s_i to have a set of neighboring states $s_j \in \mathcal{N}(s_i)$ and transition rates k_{ij} between them. The neighborhood relation was defined by a set of three moves: (1) opening a base pair, (2) closing a base pair, and (3) shifting a base pair through a helix stem. Transition rates k_{ij} were then determined by the free energy change ΔG_{ij}

associated with the particular move that transforms s_i into s_j, e.g. using the Metropolis rule. Isambert et al., on the other hand, coarse grained the structure space such that any state consists of a set of compatible helices. Whereas local zipping/unzipping within a helix is considered to happen extremely fast with rate $k_0 \simeq 10^8 s^{-1}$, transitions between the possible conformations of helices are assigned Arrhenius-like rates $k = k_0 \cdot \exp(-\Delta G/kT)$ with the thermal energy kT to capture the crossing of the energy barrier ΔG separating them. In addition to the free energy of the stacking interactions they include the conformational entropy of the entire structure into the energy evaluation, hence determination of ΔG, which in turn allowed them to include pseudo-knotted interactions as well.

6.4 COARSE GRAINING OF THE LANDSCAPE

For biologically interesting lengths of RNAs the analysis of their folding kinetics via a direct solution of the master equation for the complete secondary structure space is infeasible. Therefore, approaches that analyze kinetic effects in multi-molecule systems, i.e. ensembles of RNAs, were usually realized by simulating a larger set of single trajectories. However, this was about to change in 2004 when Wolfinger et al. presented their work on "Efficient computation of RNA folding dynamics" [252]. Based on an exhaustive enumeration of secondary structures [255] they partitioned the energy landscape into (i) local minima, (ii) their basins of attraction, so-called gradient basins, and (iii) the saddle points separating them. Using a flooding algorithm their program *barriers* computes the above partitioning that can conveniently be visualized by its corresponding barrier tree [66, 68]. While generating the coarse grained representation, *barriers* already computes the exact transition rates between the resulting macro states, i.e. the gradient basins. This enables to assess the folding behavior of RNAs between the gradient basins instead of every possible secondary structure. The relatively low number of macro states, again, opened computation of the kinetics by numerical integration of the master equation.

MICRO AND MACRO STATES In principle, any partitioning of the energy landscape is an abstraction of the underlying state space Ω, i.e. the set of *micro states* $s_i \in \Omega$, into a smaller set of *macro states* $\mathcal{A} = \{\alpha_1, \alpha_2, \ldots\}$, with $\Omega = \cup_j \alpha_j$. The thermodynamic properties of any such macro state α_j are then determined by the 'atomic' states $s_i \in \alpha_j$ it consists of. Hence, to each α_j the partition function

$$Q_{\alpha_j} = \sum_{s_i \in \alpha_j} e^{-\beta E(s_i)} \tag{50}$$

and its corresponding free energy

$$G(\alpha_j) = -\beta^{-1} \ln Q_{\alpha_j} \tag{51}$$

can be assigned. In thermal equilibrium, the probability $p(\alpha_j)$ of a macro state α_j then depends on the fraction of states it subsumes, and evaluates to

$$p(\alpha_j) = \frac{\sum_{s_i \in \alpha_j} e^{-\beta E(s_i)}}{\sum_{s_i \in \Omega} e^{-\beta E(s_i)}} = \frac{Q_{\alpha_j}}{Q}. \tag{52}$$

However, for a reasonable partitioning that can be used for RNA folding kinetics simulation some specific assumptions have to hold. First, any macro state is assumed to be in thermal equilibrium itself. That is, the probability to observe a specific instance $s_i \in \alpha$ within a macro state α is only governed by its intrinsic equilibrium probability and independent of any time passed

$$\text{Prob}[s_i|\alpha] = \frac{e^{-\beta E(s_i)}}{Q_\alpha}. \tag{53}$$

Second, it must be possible to compute effective transition rates between any two macro states.

COMPUTING TRANSITION RATES BETWEEN MACRO STATES In cases where the partitioning of the energy landscape produces macro states that introduce a kind of neighborhood relation, the most simple model for transition rates between two neighboring macro states α_i and α_j may be the application of transition rate rules for micro states, e.g. the Metropolis rule. Accordingly, the rate from α_i to α_j only depends on their difference in free energy $\Delta G_{\alpha_i,\alpha_j} = G(\alpha_j) - G(\alpha_i)$:

$$k_{\alpha_i,\alpha_j} = \begin{cases} e^{-\beta \Delta G_{\alpha_i,\alpha_j}} & \text{if } G(\alpha_j) > G(\alpha_i), \\ 1 & otherwise \end{cases} \tag{54}$$

But this model certainly performs poorly, compared to folding kinetics on the full landscape, and can only serve as a crude approximation, since it omits too many microscopic properties captured within the macro states.

Another simple, yet straightforward approximation for the transition rate between two macro states α_i and α_j directly connected by a saddle point is the Arrhenius law

$$k = k_0 e^{-\beta E_a} \tag{55}$$

with activation energy E_a, and a scaling factor k_0. Here, E_a can be approximated by the saddle point energy $E[\alpha_i, \alpha_j]$. However, in most cases two macro states may be connected not only by a single saddle point but through a multitude of possible paths. Therefore, this approximation inevitably neglects entropic terms arising from the degeneracy of paths.

A more thorough approximation can be made by taking into account the microscopic changes [252] during a transition. The macro state kinetics is determined by the corresponding master equation

$$\frac{dP_{\alpha_i}}{dt} = \sum_{\alpha_j \in \mathcal{A}} k_{\alpha_j,\alpha_i} P_{\alpha_j}(t) \tag{56}$$

with $P_{\alpha_i}(t) = \sum_{s_x \in \alpha_i} P_x(t)$. Furthermore, the transition rate from α_i to α_j should reflect transitions between their micro states

$$k_{\alpha_i,\alpha_j} = \sum_{s_x \in \alpha_i} \sum_{s_y \in \alpha_j} k_{xy} \mathrm{Prob}[s_x|\alpha_i] \text{ for } \alpha_i \neq \alpha_j \tag{57}$$

where $\mathrm{Prob}[s_x|\alpha_i]$ is the probability to observe s_x given the system is within macro state α_i, and k_{xy} the *micro-rate* from s_x to s_y. Substitution of $\mathrm{Prob}[s_x|\alpha_i]$, see Equation (53), directly leads to

$$k_{\alpha_i,\alpha_j} = \frac{1}{Q_{\alpha_i}} \sum_{s_x \in \alpha_i} \sum_{s_y \in \alpha_j} k_{xy} e^{-\beta E(s_x)}. \tag{58}$$

Q_{α_i} can be easily computed during the construction of α_i, rendering the computation of the micro-rates k_{xy} the only remaining problem. If s_x and s_y are direct neighbors, i.e. $s_y \in \mathcal{N}(s_x)$, one of the beforementioned transition rules, e.g. Metropolis rule, can be readily applied. For all other cases, more sophisticated methods need to be developed. The *barriers* algorithm presented in the publication of Wolfinger et al. elegantly solves this issue by computing the micro-rates on-the-fly during the construction of the barrier tree [252].

LIMITATIONS AND ALTERNATIVES Albeit the advantage of making RNA folding kinetics prediction feasible the drawbacks of the aforementioned methods are twofold. First, the Monte Carlo approaches follow the trajectory of only a single molecule through the vast space of possible conformations. Thus the simulation has to be repeated hundreds to thousands times to obtain an overall picture of possible structural transformations and its transient states. Depending on the length of the RNA and the ruggedness of the underlying energy landscape it is even possible to miss certain folding pathways despite repeating the simulation considerably often. Second, the 'global' approach, that treats the RNA folding kinetics problem as a Markov process, is feasible only for small state spaces and, additionally, relies on a pre-computed secondary structure free energy landscape. Even though partitioning methods exist that lump states of the landscape into so-called macro states, they still rely on exhaustive enumeration of all secondary structures. And this, in turn, is only possible for small RNAs with sequence lengths of at most up to some 100 nt, since the number of secondary structures grows exponentially with the sequence length.

As an alternative, simulations can be performed on a representative subset of the energy landscape. Tang et al. [217, 218] already proposed a method for ab-initio coarse graining of the energy landscape. The resulting graph-like representation consists of nodes representing sampled states of the free energy landscape while the weighted edges connecting them define the neighborhood relation. Edge weight calculations are then based on barrier height estimation, typically using the Morgan-Higgs heuristics. Although these methods are usually much faster and can deal with RNAs longer then 100 nt they still suffer from the problem of generating the right structure representatives, a problem that gets increasingly difficult for longer RNAs.

An novel method which ab-initio partitions the secondary structure landscape to solve the master equation of a macro state system will be presented in Part v, *Unpublished work*, Chapter 12.1. Utilization of a classified DP approach and stochastic backtracking within the resulting partitions circumvents the need for exhaustive enumeration of secondary structures. While the number of partitions grows only quadratic with the RNAs sequence length, the approach also allows to limit the number of macro states even further. By merging those states that are far away from regions of 'interest' in terms of base pair distance, the number of states can be kept relatively small, making this method applicable to considerably longer RNA sequences. Additionally,

the presented method allows the analysis of the folding behavior upon changing landscapes in terms of successive nucleotide addition, as required for co-transcriptional folding.

Part IV

PUBLISHED WORK

7

VIENNARNA PACKAGE 2.0

Ronny Lorenz, Stephan H. Bernhart, Jing Qin, Christian Höner zu Siederdissen, Andrea Tanzer, Fabian Amman, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker.
ViennaRNA Package 2.0.
in *Algorithms for Molecular Biology* 2011, **6**:26.
doi: 10.1186/1748-7188-6-26

RL designed the study, designed and re-implemented the algorithms, and wrote substantial parts of the paper.

Lorenz *et al. Algorithms for Molecular Biology* 2011, **6**:26
http://www.almob.org/content/6/1/26

AMB ALGORITHMS FOR MOLECULAR BIOLOGY

SOFTWARE ARTICLE Open Access

ViennaRNA Package 2.0

Ronny Lorenz[1*], Stephan H Bernhart[1], Christian Höner zu Siederdissen[1], Hakim Tafer[2], Christoph Flamm[1], Peter F Stadler[2,1,3,4,5,6] and Ivo L Hofacker[1,3,7*]

Abstract

Background: Secondary structure forms an important intermediate level of description of nucleic acids that encapsulates the dominating part of the folding energy, is often well conserved in evolution, and is routinely used as a basis to explain experimental findings. Based on carefully measured thermodynamic parameters, exact dynamic programming algorithms can be used to compute ground states, base pairing probabilities, as well as thermodynamic properties.

Results: The `ViennaRNA` Package has been a widely used compilation of RNA secondary structure related computer programs for nearly two decades. Major changes in the structure of the standard energy model, the *Turner 2004* parameters, the pervasive use of multi-core CPUs, and an increasing number of algorithmic variants prompted a major technical overhaul of both the underlying `RNAlib` and the interactive user programs. New features include an expanded repertoire of tools to assess RNA-RNA interactions and restricted ensembles of structures, additional output information such as *centroid* structures and *maximum expected accuracy* structures derived from base pairing probabilities, or *z-scores* for locally stable secondary structures, and support for input in `fasta` format. Updates were implemented without compromising the computational efficiency of the core algorithms and ensuring compatibility with earlier versions.

Conclusions: The `ViennaRNA Package 2.0`, supporting concurrent computations `via OpenMP`, can be downloaded from http://www.tbi.univie.ac.at/RNA.

Background

A typical single stranded-nucleic acid molecule has the propensity to form double helical structures causing the molecule to fold back onto itself. Simple rules of complementary base pairing govern this process, which results in a regular pattern of Watson-Crick and GU pairings (helices) and intervening stretches of less regularly ordered nucleotides (loops), collectively known as the molecule's *secondary structure*. Secondary structure elements may be placed in close spatial proximity allowing additional non-covalent interactions. These are not as frequent and often are energetically less favorable compared to canonical base pairs, thus rendering the 3-dimensional tertiary structure of an RNA to be dominated by the underlying scaffold of the secondary structure. The canonical base pairing governs not only the thermodynamics but also the folding kinetics, which can be approximated as a hierarchical process in which secondary structure is formed before tertiary structure [1].

The dominance of base pairing and the confinement to a single interaction partner makes it possible to model RNA (and DNA) secondary structures at a purely combinatorial level, completely ignoring both atom-scale details and spatial embeddings. Formally, an RNA secondary structure is a (labeled) graph whose nodes represent nucleotides. The edge set contains edges between consecutive nodes $(i, i + 1)$ representing the phosphate backbone as well as edges between base pairs. For the latter, the following conditions must hold:

1. base pair edges are formed only between nucleotides that form Watson-Crick or GU base pairs;
2. no two base pair edges emanate from the same vertex, i.e., a secondary structure is a matching;
3. base pair edges span at least three unpaired bases;
4. if the vertices are placed in 5' to 3' order on the circumference of a circle and edges are drawn as straight lines, no two edges cross.

* Correspondence: ronny@tbi.univie.ac.at; ivo@tbi.univie.ac.at
[1]Institute for Theoretical Chemistry and Structural Biology, University of Vienna, Währingerstraße 17/3, A-1090 Vienna, Austria
Full list of author information is available at the end of the article

The last condition ensures that the graph is outerplanar and therefore excludes so-called pseudo-knots. Matching problems usually have cost functions determined by edge-weights. The earliest predictions of RNA secondary structures in the early 1970s indeed used such simple energy models [2]. Detailed melting experiments, however, soon showed that a different, more complex type of energy function is necessary to properly model the thermodynamics of nucleic acid structures. Instead of individual base pairs, the energy contributions are dominated by base-pair stacking and the destabilizing entropic effects of unpaired "loops". Sequence-dependent energy parameters for these building blocks contribute to a very good approximation additively to the folding energy [3]. Over the last two decades, this additive standard energy model has been repeatedly refined and updated, see e.g. [4-9].

The RNA folding problem is solvable by means of dynamic programming. The simplest version, known as *maximum circular matching problem*, accounts for base pairing energies only [10,11]. In the early 1980s Nussinov and Jacobson [12] and Michael Zuker with collaborators [13,14] demonstrated that the loop-based energy model is also amenable to the same algorithmic ideas. Their work made computational RNA structure prediction accurate and efficient enough for practical use, resulting in the first versions of `mfold`. A decade later, John McCaskill realized that the dynamic programming recursions can be adapted to compute the partition function of an equilibrium ensemble of RNA molecules [15], paving the way for efficient computational access to accurate thermodynamic modelling without exceeding an asymptotic time complexity of $\mathcal{O}(n^3)$.

The secondary structure model of RNA perfectly fits together with modern genomics and transcriptomics since it works at the same level of abstraction, treating nucleotides as basic entities. With the increasing availability of RNA sequence data, and the realization that many of the functional RNAs have evolutionary well-conserved secondary structures, many research groups developed a plethora of specialized tools for various aspects of RNA bioinformatics. As an alternative to the direct measurement of thermodynamic parameters, for instance, machine learning approaches employing stochastic context free grammars (SCFG) were introduced e.g. in the `infernal` suite [16,17]. The algorithmic work horses of the SCFG approach, the Cocke-Younger-Kasami (CYK), the inside and the outside algorithms, are also dynamic programming schemes. They are, in fact, very close cousins of the minimum free energy and partition function folding algorithms. The `contrafold` tools in fact recently bridged the apparent gap between the thermodynamic and the machine learning approach to RNA bioinformatics proposing to learn a parameter set for a SCFG that structurally matches the standard energy model [18].

Several other tools implement dynamic programming based RNA secondary structures prediction: `UNAfold` [19] is the successor of the original `mfold` program and adds suppport for predicting RNA-RNA hybridization. `RNAstructure` [20] started as a reimplementation of `mfold` with a graphical user interface in Windows, but is now available for other platforms and has added several additional algorithms such as partition function folding and suboptimal structures. The `NUPACK` suite [21] focuses on folding of several interacting RNA strands and design problems. The group around Kiyoshi Asai developed several tools focusing the usage of centroid and maximum expected accuracy (MEA) estimators, see e.g. [22]. Ye Ding's `Sfold` program [23] was the first to introduce stochastic structure sampling. The group around Robert Giegerich provides several RNA related tools, notably the `RNAshapes` [24] program.

The `Vienna RNA Package` [25] has its roots in a series of large-scale simulation studies aiming at an understanding of adaptive evolution on rugged fitness landscapes [26-28] and the statistical properties of the sequence-structure relationships of RNA [29-31] rather than the detailed analysis of individual RNA molecules of biological interest. The primary design goals for its implementation in the early 1990s, therefore, were twofold. First and foremost, the basic folding algorithms were to be implemented so as to be as efficient as possible in their usage of both CPU and memory resources. The core algorithms are accessible as a `C` library, which later on was also equipped with `Perl` bindings to facilitate interoperability with this commonly used scripting language. Secondly, the interactive programs were to be used mostly in (shell-script) pipelines, hence they use a simple command-line interface and, where possible, they read from and write to a stream. This feature made it easy to construct a suite of web services [32] providing easy access to most functionalities of the `Vienna RNA Package`. With the rising tide of first genomics and then transcriptomics data, the need for both efficient implementation and easy incorporation into pipelines remained, even though the focus gradually shifted from large-scale simulation to large-scale data analysis. Little has changed in the core folding algorithms in the 17 years since the first publication [25] of the package. On the other hand, a variety of variants have been included such as consensus structure prediction from alignments or scanning versions capable of dealing with local structures in genome-scale data sets. The systematic overhaul of the `Vienna RNA Package` documented here was

Lorenz *et al. Algorithms for Molecular Biology* 2011, **6**:26
http://www.almob.org/content/6/1/26

largely triggered by the publication of improved parametrizations of the energy model, which affected nearly every component in the library, and by the progress in computer technology, which led to the widespread deployment of shared-memory multi-core processors. In order to exploit these hardware features a restructuring of the RNA library to make it thread-safe and hence fit for use in concurrent computations was required. Beyond these technical improvements, the `Vienna RNA Package 2.0` features a number of additions to its algorithmic repertoire, an improved API to `RNAlib`, and an expanded toolkit of auxiliary programs.

Interactive tools

Since its first release, the `ViennaRNA Package` included interactive command-line tools which enable users to access the high performance implementations of the algorithms via a command-line interface. To ensure scalability of the use-cases all programs were developed with the objective of handling input- and output-streams, facilitating their integration into *UNIX pipes*. Thus pre- and post-processing of the input/output data can proceed without the need of intermediate input- or output-files. Most programs of the `ViennaRNA Package` furthermore are able to operate in *batch mode*, handling large sets of input data with a single call. By default, the programs of the `ViennaRNA Package` generate an output that is meant to be easily parsable while keeping it human-readable.

The core of the package provides several variants of the RNA folding recursion: energy minimization, partition function and base pairing probabilities, backtracing of suboptimal structures, alignment-based as well as scanning versions. The decision whether a certain functionality is implemented as a separate stand alone program or as an optional command-line switch is based on the compatibility of I/O formats and internal data structures. Table 1 presents the implemented model variants as well as the data formats for each program, whereas Figure 1 illustrates example program calls together with their corresponding output. In the following paragraphs, we provide a comprehensive summary of programs included in the `ViennaRNA Package`.

Folding

The main secondary structure prediction tool is `RNAfold`, which computes the minimum free energy (MFE) and backtraces an optimal secondary structure. Using the `-p` option, `RNAfold` also uses McCaskill's algorithm [15] to compute the partition function, the matrix of base pairing probabilities, and the centroid structure. The `RNAfold` output is a string representation of the structure and the folding energy written to the standard output stream. With the `-p` option, it also creates a PostScript file containing the base pairing probability matrix. Circular RNA sequences are rare in nature and appear infrequently in practical applications. With the `–circ` option this case is handled as a post-processing for the forward recursion and a preprocessing of the backward recursions without compromising the performance of the folding algorithms for linear RNAs [33]. Constraints can be supplied to the folding algorithms enforcing that individual positions are paired, unpaired, or paired with specific partners.

The program `RNAsubopt` can be used to generate suboptimal structures. Using command-line options, it can switch between three different ways of generating them: by default, it generates the complete set of suboptimal structures within a certain energy band, the size of which can be chosen using the `-e` option [34]. With the `-p` option it uses stochastic backtracking [35] from the partition function to generate a Boltzmann-weighted random sample of structures, effectively providing the functionality of `sfold` [23]. Finally, the `-z` option generates suboptimal secondary structures according to Zuker's algorithm [36]. The resulting set consists, for each basepair (i, j) that can be formed by the input structure, of the energetically most favorable structure that contains the (i, j)-pair. This option implements a feature that has been used frequently in applications of the `mfold` package.

`RNALfold` [37] is a "scanning" version of the folding programs that can be used to calculate local stable substructures of very long RNA molecules. Local in this context means that the sequence interval spanned by a base pair is limited by a user-defined upper bound (set by the `-L` option). Scanning versions of RNA folding programs conceptutally perform computations for all sequence-windows of a fixed size. Algorithmically, they are faster than the naïve approach by re-using partial results for overlapping windows. `RNALfold` does not come with a partition function version because the global partition function with restricted base pair span is of limited interest in practical applications. Instead, a separate program, `RNAplfold` [38], computes the base pairing probability averaged over all sequence windows that contain the putative pair. This tool can also be used to compute the local accessibilities, i.e., the probabilities that sequence intervals are single-stranded in thermodynamic equilibrium (option `-a`).

`RNA2Dfold` [39] implements energy minimization, partition function computations, and stochastic backtracing for the two dimensional projection of the secondary structure space that is defined by the base pair distances from the two prescribed reference structures. The restricted ensembles of secondary structures are useful in particular for tracing refolding pathways and to compute lower bounds of energy barriers between

Table 1 Main features of the interactive programs provided by the ViennaRNA Package 2.0

Program	Energy model variants										Data formats		
	intramolecular bp	*intermolecular bp*	*structure constraint*	*canonical structures*	*circular sequence*	*dangling end model (s)*	*centroid structure*	*MEA structure*	*suboptimal structures*	*base pair probabilities/ partition function*	*input format (s)*	*text output file(s)*	*PostScript plot(s)*
single sequence analysis (global variant)													
RNAfold	+	-	+	+	+	0,1,2,3	+	+	-	+	F,V	-	+
RNAsubopt	+	+	+	+	+	0,1,2,3	NA	NA	B,E,Z	-	F,V	-	-
RNAcofold	+	+	+	+	NA	0,1,2,3	-	-	-	+	F,V	+	+
RNAup	+	+	+	+	-	0,2	-	-	-	+	V	+	-
RNAduplex	-	+	-	+	-	0,1,2,3	-	-	E	-	V	+	-
RNA2Dfold*	+	-	-	-	+	0,2	-	+	B	-	V	-	-
RNAPKplex*	+	-	-	+	-	0,1,2,3	-	-	E	-	F,V	-	+
RNAplex*	-	+	+	-	-	2	NA	NA	E	-	V,W	+	+
RNAsnoop*	+	+	+	+	-	2	NA	NA	E	-	V,W	+	+
single sequence analysis (local variant)													
RNALfold	+	-	-	+	-	0,1,2,3	-	-	-	-	F,V	-	-
RNAplfold	+	-	-	+	-	0,2	-	-	-	-	F,V	+	+
comparative analysis (global variant)													
RNAalifold	+	-	+	+	+	0,2	+	+	B	+	C,S	-	+
RNAaliduplex	-	+	-	+	-	0,1,2,3	-	-	E	-	C,S	+	+
comparative analysis (local variant)													
RNALalifold*	+	-	-	+	-	0,1,2,3	-	-	-	+	C,S	+	+
Misc. analysis/Utilities													
RNAeval	+	+	NA	NA	+	0,1,2,3	NA	NA	NA	NA	F,V	-	-
RNAplot	NA	NA	NA	NA	+	NA	NA	NA	NA	NA	F,V	-	+
RNAheat	+	-	-	+	-	0,2	-	-	-	-	F,V	-	-
RNAinverse	+	-	NA	NA	-	0,1,2,3	NA	NA	NA	NA	V	-	-
RNApaln	+	-	-	+	-	0,1,2,3	NA	NA	NA	+	V	+	+
RNApdist	+	-	-	-	-	0,1,2,3	NA	NA	NA	+	V	+	+
RNAdistance	+	-	NA	NA	NA	NA	NA	NA	NA	NA	V	+	+

The characters + and - show presence and absence of a certain feature, while NA indicates that the feature is not applicable in a given context. Abbreviations of input file formats are (C)lustal-format, (F)asta-format, (S)tockholm-format, and (V)iennaRNA-format. Support for prediction of suboptimal structures may be implemented as (B)oltzmann weighted sampling, exhaustive (E)numeration of all structures in a given energy band, and (Z)uker-style suboptimal structures. Programs marked by an asterisk (*) were not included in a previous release of the ViennaRNA Package.

Lorenz *et al. Algorithms for Molecular Biology* 2011, **6**:26
http://www.almob.org/content/6/1/26

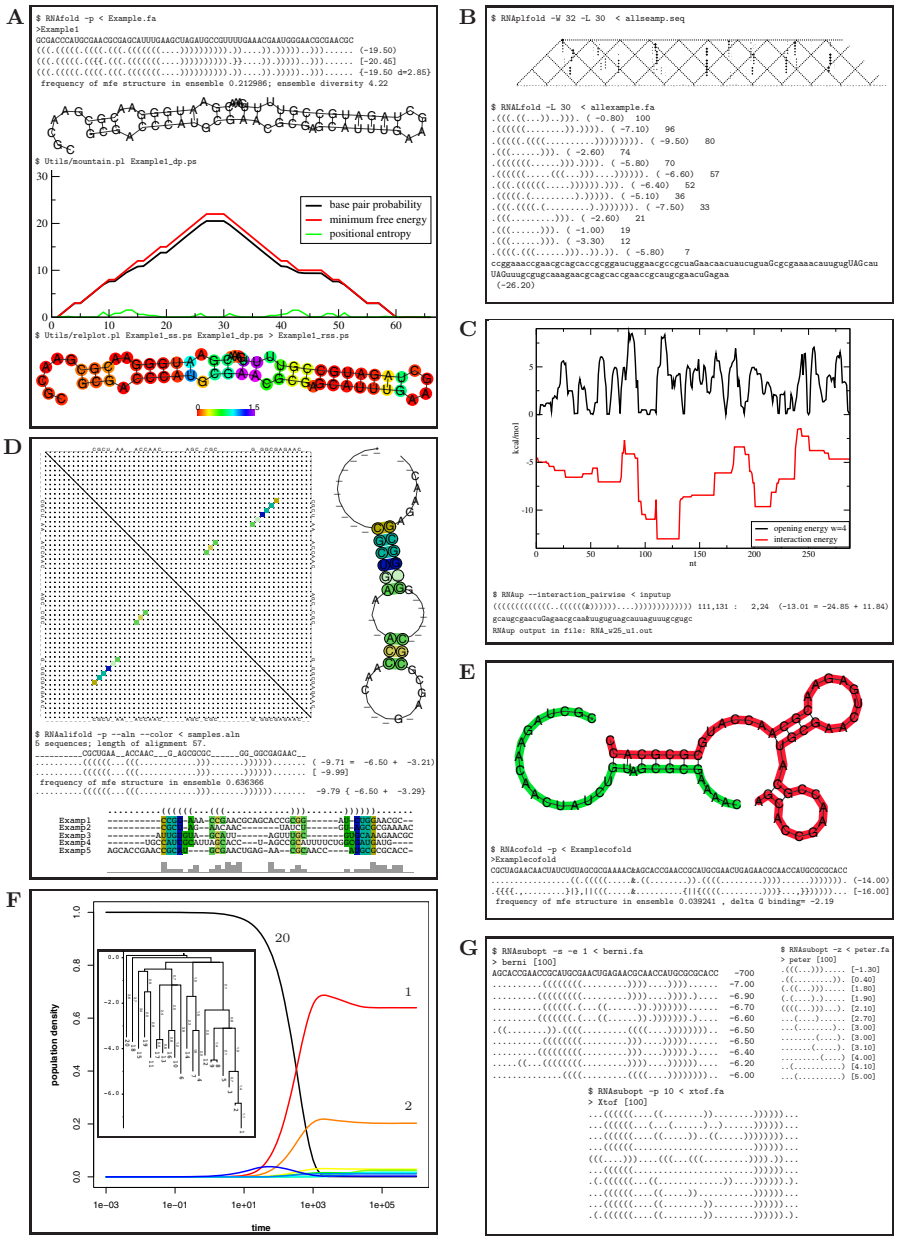


Figure 1 Example calls of programs included in the `ViennaRNA Package` and their corresponding output. (**A**) Single sequence analysis using `RNAfold`. (**B**) Locally optimal secondary structures and base pair probabilities using `RNAplfold` and `RNALfold`. (**C**) Interaction thermodynamics of two RNA sequences computed by `RNAup`. (**D**) Consensus structures and base pair probabilities for RNA sequence alignments obtained from `RNAalifold`. (E) Secondary structure of an RNA dimer calculated by `RNAcofold`. (F) Folding kinetics using `RNAsubopt` in conjunction with the external programs `barriers` and `treekin`. (**G**) Suboptimal secondary structures generated by `RNAsubopt`. For a detailed description see the appendix.

alternative conformations of an RNA molecule. Although `RNA2Dfold` is based upon the usual dynamic programming recursion of energy-directed folding, the asymptotic time complexity is multiplied by a factor of $k^2 \cdot l^2$, where k and l are maximum base pair distances to the first and the second reference structure, resp. Hence, the overall time complexity for a sequence of length n is $\mathcal{O}(n^7)$. The memory requirements of $\mathcal{O}(n^4)$ are also higher than for the regular secondary structure prediction scheme. However, since the implementation

uses a sparse matrix approach, the prefactor of time and memory complexity is very small, making the program applicable for RNA sequence lengths of up to about 400 - 600 nt.

RNA-RNA interactions

Several programs focus on various aspects of the hybridization structure of two RNA molecules, using different levels of detail. The programs `RNAcofold` [40] and `RNAup` [41] are two complementary programs with the highest level of detail available within the `ViennaRNA Package`. `RNAup` first computes local opening energies for both molecules and then computes interaction energies, looking for the best interaction site of two molecules. `RNAcofold`, on the other hand, concatenates two molecules and computes a common secondary structure using modified energies for the loop that contains the cut. `RNAcofold` thus can generate arbitrary many binding sites, but does not allow pseudoknotted configurations, while `RNAup` covers only a single interaction site, which however may form a complex pseudoknotted configuration. The partition function version of `RNAcofold` can be used to investigate the concentration dependency of dimerization, similar to [42]. On the other hand, `RNAup` is mostly geared towards investigations of the binding of regulatory RNA molecules with their target RNAs.

`RNAPKplex` is at present the only component of the `Vienna RNA Package` that explicitly predicts pseudoknotted RNA structures [43]. As an "intramolecular variant" of `RNAup` it computes accessibilities and then identifies regions that can form stable base pairs.

Although optimized for speed, the full-fledged folding algorithms are not fast enough for genome-wide applications. `RNAduplex`, similar to Rehmsmeier's `RNAhybrid` [44], ignores intramolecular structures and all multi-branch loops in its search for thermodynamically favorable interaction regions. `RNAplex` [45] achieves a massive gain in speed by simplifying the energy model for interior loops to an affine gap cost model, effectively reducing the folding problem to a variant of local sequence alignment. The accuracy of this approach can be further improved by reading in accessibilities (as computed by `RNAplfold`) and incorporating them into the scoring model [46].

The specialized programs `RNAsnoop` [47] for the prediction of target sites of H/ACA snoRNAs, and `RNALfoldz` [48] for the evaluation of predicted local secondary structures, use SVMs to further classify the output of the RNA folding routines.

Consensus structures and alignments

A central issue for the comparative analysis of RNA sequences is the computation of a consensus structure. Starting from a sequence alignment, this can be achieved using the same algorithmic framework as folding a single sequence. More precisely, energy contributions can be added up in a columnwise manner to yield an effective energy model for the alignment as a whole [49]. The `Vienna RNA Package` provides alignment-based variants for several of the algorithms discussed above: `RNAalifold` [50] computes global consensus structures both in MFE and partition function mode, a scanning version of long sequence alignments is `RNALalifold`. `RNAaliduplex` is designed to facility the search for conserved RNA-RNA interaction sites in large alignment data sets. The `alidot` program [51,52], finally, extracts local conserved structures given a sequence alignment and secondary structure predictions for each of the aligned sequences. By default, consensus structure prediction is dominated by the thermodynamic parameters and sequence covariation. Thus, phylogenetic support for conservation of secondary structure is included only as a small bonus energy term. A much more sophisticated substitution model for paired regions based on the RIBOSUM scoring scheme [53] can be invoked with the `-R` option.

The `Vienna RNA Package` does not contain its own optimized implementation for the *simultaneous* folding and alignment of two RNA sequences, i.e., of the Sankoff algorithm [54]. We refer to the well-established software tools `FoldAlign` [55], or `DynAlign` [56] for this task. A simplified version of the Sankoff algorithm underlies `pmcomp` [57,58], a facility to align pre-computed base-pairing probability matrices, although this tool is now included mostly for backward compatibility. An improved and much more efficient implementation is provided by the `locarna` package [59] developed in cooperation with Rolf Backofen and Sebastian Will and distributed separately.

With `RNApaln` and `RNApdist` the package also provides tools to align and compare base pair probability patterns using modified string alignment algorithms. Tree editing distances and corresponding pairwise alignments can be computed with `RNAdist`.

Miscellaneous tools

Concerning sequence design, we ship the program `RNAinverse` [25]. It generates a sequence that folds into the input structure by mutating a start sequence. More efficient versions of inverse folding algorithms have become available over the last decade, see e.g. `INFO-RNA` [60], `RNA Designer` [61] and the recent `NUPACK` design algorithms [62]. Nevertheless, `RNAinverse` remains useful for some applications as it is designed for search for solutions as close as possible to the starting sequence. `RNAswitch` [63] takes a pair of secondary structures as input and finds a sequence that

Lorenz *et al. Algorithms for Molecular Biology* 2011, **6**:26
http://www.almob.org/content/6/1/26

has both input structures as near ground states. The possibility to design bistable RNAs may be useful e.g. for synthetic biology.

A closer look at the dynamics of RNA folding a available through `kinfold` [64], a rejectionless Monte Carlo simulation algorithm generating trajectories of subsequent secondary structures. Kinetic information can also be obtained from the exhaustive enumeration of suboptimal structures using `RNAsubopt` in conjunction with the `barriers` package [64,65]. The latter is not restricted to RNA landscapes and hence distributed separately from the `Vienna RNA Package`.

Auxiliary Programs

In addition to the prediction and analysis tools, the `ViennaRNA Package` provides utility programs and scripts that mainly assist in processing input- and output data. `RNAeval` computes the energy of a given structure formed by a given sequence and can in particular be used to re-compute energies for a given pair of sequence and structure with different energy models. The `Perl` script `refold.pl` generates single structure predictions using a previously computed consensus structure as constraint.

`RNAplot` can be used to generate a graphical representation of the an input sequence/structure pair [66]. Several `Perl` scripts can be used to further manipulate PostScript output produced by the various components of the `Vienna RNA Package`. Conventional structure drawings can be rotated with `rotate_ss.pl`. The `relplot.pl` script includes reliability annotation into secondary structure plots, `colorrna.pl` uses the conservation of alignments for coloring consensus structure plots, while `coloraln.pl` does the same with an alignment. Mountain plots can be produced with `mountain.pl` and `cmount.pl` from single and consensus structures, respectively.

Many tools in RNA bioinformatics use `mfold`'s "connectivity" (.ct) file format. The dot-bracket representation used consistently by the `Vienna RNA Package` can converted into this format using `b2ct` and `ct2b.pl`, resp.

The ViennaRNA Webserver

The `ViennaRNA Webserver` [32] facilitates an easy to use form based web browser interface to most of the programs included in the ViennaRNA Package and additional tools. It combines the call of the appropriate command-line tools with post-processing steps to obtain a visualization of the output. The webtools echo the command-lines used to call components of the `Vienna RNA Package`; this feature can be used to get more familiar with the individual tools. The webserver also provides an interface to the `barriers` and `treekin` program allowing the analysis of folding landscapes and structural refolding kinetics. The backbone of the `ViennaRNA Webserver` has been upgraded so that all calculations with the webserver profit from the increased performance of the new `ViennaRNA Package`.

Modifying the energy parameters of the model

The energy model implemented in `ViennaRNA Package 2.0` follows the structure of the *Turner 2004* energy parameters as described in [9] with a few very minor deviations. Compared to previous parametrizations, the *Turner 2004* model introduced additional look-up tables for certain free energies and for loop entropies in response to more precise measurements of certain loop types. For the sake of computational efficiency a few peculiar rules were deliberately ignored, however. Details on these discrepancies, which do not affect the overall accuracy of predictions (see below), are provided in the appendix.

All programs of the `ViennaRNA Package` can read in energy parameters from a human-readable text file allowing the user to replace the default *Turner 2004* parameter set. This can either be a user-supplied parameter file or one of several parameter compilations that are shipped with the package. Of particular interest are parameters for DNA folding. Here we provide a parameter set compiled by Douglas Turner and David Mathews [67] from published data, incorporating in particular earlier work by the group of John SantaLucia [68]. While the Turner parameters are based almost exclusively on thermodynamic measurements, there has been increasing interest in optimizing parameters such as to maximize prediction accuracy, see e.g. [69]. As an example for such trained parameters we provide the *Andronescu* parameter set from ref. [70].

To maintain backward compatibility we also ship *Turner '99* energy parameter files containing the basic contributions used in previous versions of the `ViennaRNA Package`. These parameter files, however, will not always produce results identical to earlier versions of the package. Affecting mainly the computation of consensus structures, these differences are mainly owed to a different handling of non-standard base pairs (i.e., base pairs other than Watson-Crick and GU). The current implementation assumes that the energy contribution of a loop with non-standard base pairs or non-standard nucleotides equals the least stabilizing contribution from the same loop type with canonical nucleotides and pairs only. Small differences may also appear in partition function computations as a consequence of round-off errors.

Since the structure of the energy model has changed in `ViennaRNA Package` 2.0, energy parameter files for

versions 1.8.5. and earlier will not work with the new version of the package. Such old-style user-supplied parameter files can be converted to the new file format using the RNAparconv utility.

Additional output options

More information gathered through the course of the folding algorithms can be included in the output. RNAfold and RNAalifold, for instance, optionally provide further information about the reliability of folding results. When evaluating ensemble properties with the partition function, most programs now also compute the *centroid* structure [71], i.e., the structure with the smallest average base pair distance to all other structures in the ensemble. When base pair probabilities are computed, the maximum expected accuracy (MEA) structure [18,72] is also available. The RNALfold/RNALfoldz program now features an add-on to calculate the *z-score* for the predicted local secondary structures [48]. This makes results comparable between sequences with different nucleotide compositions and facilitates the choice of a reasonable cutoff thresholds to decrease the number of structure hits.

Program options and documentation

Each of the command-line tools provides the option -h or –help to print a brief overview of its general behavior as well as a list of all available parameter options including their description. To obtain more detailed information or even exemplary use-case scenarios for a certain program of the ViennaRNA Package, a *UNIX manpage* is provided for each of them.

An important change in the new release is the compliance to the GNU standard regarding the format of command-line options. Short options consist of a single character preceded by a minus sign, e.g. -p, while long options are strings of two or more characters preceded by two minus signs, e.g. –noLP. This change will break backward compatibility wherever command-line tools from older versions of the package were used. This can be easily fixed by inserting the second dash in long options.

Input file formats

A plethora of different file formats have been introduced by the many tools and databases relevant to RNA bioinformatics. The ViennaRNA Package has also contributed to this unpleasant diversity with its own native formats. Originally designed for simulation pipelines in which no meta-data is attached to sequence or structure data, it expects input items (sequences and/or structures) as single strings uninterrupted by white spaces or line breaks. FASTA-like headers can optionally be used to specify an identifier for the data item(s). Secondary structures are also specified as strings, using the three characters (,), and. to denote nucleotides that are paired with a partner upstream or downstream, or that are unpaired, resp. In addition to uniquely determining a pseudoknot-free secondary structure, this notation has the advantage of providing a compact annotation of the sequence or alignment to which the structure refers. The dot-parentheses-format is meanwhile used also in many unrelated tools e.g. [18,21,61,73-79]. Similar annotation strings are used to specify constraints as input to folding algorithms.

The requirement to write input items on a single line usually requires data format conversions for the interactions with most other bioinformatics tools. These usually read and write *FASTA* format [80], which allows white spaces and line breaks arbitrarily interspersed within a sequence. An improved handling of data input now provides full FASTA support for all tools that require only sequences or sequence alignments as input. This should considerably facilitate the use of the ViennaRNA Package. More complex input structures are still required for the tools that compute RNA-RNA interactions, in particular RNAup and RNAcofold.

Programs that process alignment data used clustal format [81] in previous versions of the package. Due to the wide-spread use of the STOCKHOLM format in RNA bioinformatics, e.g. in the Rfam - RNA family database [82]), support for *.stk files has been added.

There are currently no plans to include support for input formats that use heavy markup such as Genbank [83] files or XML-based formats such as BioXSD [84] or RNAML [85].

RNAlib -

API to fast and reliable algorithms

The algorithms implemented in the ViennaRNA Package are not only accessible by means of the interactive programs outlined in the previous section but also directly in the form of a C/C++ library. This makes them readily available for third-party programs and, with the help of included Perl-interface, to elaborate scripting pipelines.

OpenMP thread-safe C/C++ API

Multi-core CPUs have become standard components in off-the-shelf PC hardware. In order to allow the ViennaRNA Package to make use of this increase of computational power, several changes had to be introduced into the API functions of the RNAlib. Although it is possible to parallelize the core folding algorithms [86,87] this requires substantial overheads so that the gain is small unless massively parallel architectures are used. On the other hand, computationally demanding applications of RNA folding typically require the

Lorenz *et al. Algorithms for Molecular Biology* 2011, **6**:26
http://www.almob.org/content/6/1/26

processing of large numbers of input sequences, a task that trivially can be parallelized. The only requirement for enabling concurrent computation on shared memory multi-core systems using OpenMP [88] is that the core algorithms are independent of shared global variables and thus thread-safe. In particular the variables referring to the energy parameters are now deprecated and replaced by additional functions or parameters which have to be passed to functions. A few remaining global variables, which are inaccessible through RNAlib, were made *thread-private* using OpenMP, allowing simultaneous function calls to operate on private copies of these variables. Using the OpenMP framework, third party applications are therefore now able to call RNAlib interfaces, such as MFE or partition function algorithms, in parallel. Limitations concerning the use of different energy models used in concurrent computations are described in detail in the API reference manual. For backward compatibility, the old functions of the previous API remain included in RNAlib but are marked as deprecated. Thus, programs which were developed for binding against the previous versions of RNAlib up to 1.8.5 are still working without limitations when linked against the new library.

The reference manual

Documentation is an important issue for the usability of the RNAlib API. In previous versions of the ViennaRNA Package, this was addressed by maintaining, in addition to the source code, a *texinfo*-based reference manual containing introductions into the particular problem sets and describing the related library functions. In order to keep this documentation up to date and to decrease the developers' effort in maintaining the manual, we opted to use *in-source* documentation that (a) helps developers who interact with the source code directly and (b) enable to use the doxygen documentation program to generate a comprehensive and always up-to-date reference manual automatically. An HTML and a PDF version are included in the package.

PERL bindings

Scripting language bindings to the C functions in the RNAlib are made using the SWIG interface compiler. With the ViennaRNA Package, we include bindings for the most important library functions made accessible for the script language Perl. This allows a very easy access to e.g. the folding functions and thus a rapid design of functional pipelines or small programs that exploit the potential of the ViennaRNA Package. Using the SWIG environment bindings for other (scripting) languages including Python and JAVA can be implemented quite easily.

Performance

We assess the performance of the ViennaRNA Package 2.0 both in terms of computational efficiency and in terms of prediction accuracy. We emphasize that it is not the purpose of this contribution to compare thermodynamics-based prediction algorithms against other approaches to RNA structure prediction. For such a benchmark we refer to the literature, e.g. [18,89,90].

In order to investigate the impact of the energy parameters, and in particular of our small changes to the *Turner 2004* model, we use a test set comprising all 1817 non-multimer sequence/structure pairs in the RNAstrand database [73] without pseudoknots in the reference structure. For each sequence, the MFE secondary structure was calculated with RNAfold 2.0, RNAfold 1.8.5, UNAFold 3.8 [19], and RNAstructure 5.2 [20]. All use a nearest neighbor energy model and a variant of Zuker's dynamic programming algorithm. As expected, the new version of RNAfold performs better than the old one. Somewhat surprisingly, however, RNAfold 2.0 also performs slightly better than RNAstructure 5.2 and UNAFold 3.8, despite the fact that we neglected a few peculiarities of the most recent energy model, see Figure 2, Additional File 1 and the implementation details in the appendix. The average performance indicators are compiled in Table 2. We emphasize, however, that the performance of the algorithms differs widely across RNA families and no single implementation provides consistently superior results. Detailed data can be found in Additional File 2.

Despite the increase in the number of parameters from *Turner '99* to *Turner 2004* we observe virtually no difference in the runtime and memory consumption between RNAfold 1.8.5 and RNAfold 2.0. Similar comparisons can be made for other components of the ViennaRNA Package. The computational speed of RNAfold compares quite favorably to that of the competing implementations, Figure 2B, although all the implementations of thermodynamic folding algorithms use essentially the same energy model and algorithmic framework, and hence have the same asymptotic runtime and memory consumption.

Discussion

The ViennaRNA Package has been a useful tool for the RNA bioinformatics community for almost two decades. Quite a few widely-used software tools and data analysis pipelines have been built upon this foundation, either incorporating calls to the interactive programs or directly interfacing to RNAlib. Numeric characteristics of secondary structures, such as Gibbs free energy ΔG, Minimum free energy (MFE), ensemble diversity or probabilities of MFE structures in the ensemble, have been widely used as features for machine learning classification, e.g. in

Lorenz *et al. Algorithms for Molecular Biology* 2011, **6**:26
http://www.almob.org/content/6/1/26

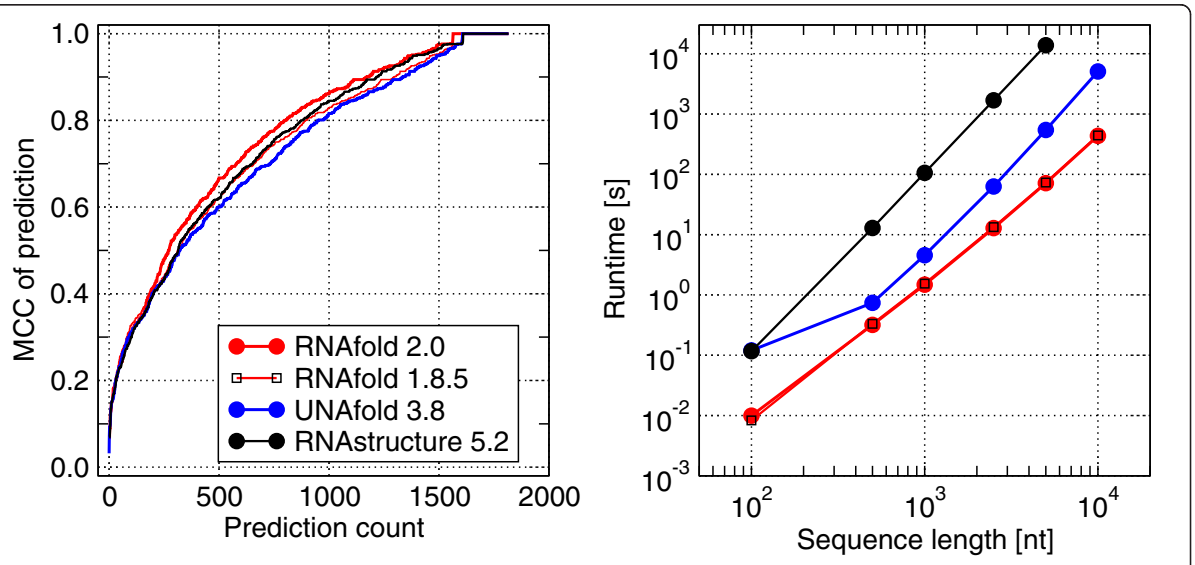


Figure 2 Performance comparison of RNAfold 2.0 to other secondary structure prediction software. (A) Accuracy of thermodynamic folding programs in terms of cumulative distribution of the Matthews correlation coefficient (MCC). `RNAfold 2.0` outperforms the other secondary structure prediction programs on the `RNAstrand` dataset: more of its predictions fall into the region of higher performance values. Both versions of `RNAfold` were run with `-d2` option. For `UNAFold` and `RNAStructure` default options were used. Performance distributions of Sensitivity, Positive predictive value (PPV) and F-measure are shown in Additional File 1. The averaged overall accuracies can be taken from table 2. (**B**) Comparison of runtimes for MFE structure predictions. Measurement was performed on an Intel® Core™ 2 6600 CPU running at 2.4 GHz. Shown are averaged running times for random sequences of lengths 100 nt (100 samples), 500 nt (100 samples), 1000 nt (100 samples), 2500 nt (20 samples), 5000 nt (16 samples) and 10000 nt (16 samples). While the compared programs `RNAfold 2.0`, `RNAfold 1.8.5` and `UNAfold 3.8` were capable of predicting an MFE structure for all tested samples in a relatively small time frame, `RNAstructure 5.2` was omitted from predictions for the 10000 nt sample set due to its time requirements.

microRNA precursor and target detection [91-94]. The non-coding RNA gene finder `RNAz` [95,96], the snoRNA detector `snoReport` [97], and `RNAstrand` [98], a tool that predicts the reading direction of structured RNAs from a multiple sequence alignment, combine thermodynamic properties computed with `RNAlib` functions and a machine learning component. `RNAsoup` [99] takes advantage of the programs `RNAfold`, `RNAalifold` and some other tools provided by the `ViennaRNA Package` for a structural clustering of ncRNAs. The siRNA design program `RNAxs` [100] employs the site accessibility predictions offered by `RNAplfold`, as does IntaRNA [60], a program to predict RNA interaction sites. Several secondary structure prediction tools, such as `CentroidFold` [22], `McCaskill-MEA` [101], or `RNAsalsa` [102], use base pair probabilities predicted by `RNAfold -p` as input, while the `LocARNA` package [59] uses them for structural alignment. The motif-based comparison and alignment tool `ExpaRNA` [103] and the tree alignment program `RNAforester` [75] also rely on the algorithms provided by `RNAlib`. Since its initial publication [25], no comprehensive description [104] of the `ViennaRNA Package` has appeared. Release 2.0 now implements the latest energy model, provides many new and improved functionalities, and - as we hope - is even easier and more efficient to use due to a thread-safe architecture, an improved API, a more consistent set of options, and a much more detailed documentation. Care has been taken to ensure backward compatibility so that `ViennaRNA Package 2.0` can be readily substituted for earlier versions.

Table 2 Averaged performance measures for thermodynamic folding algorithms

	Sensitivity	Specificity	MCC	F-measure
RNAfold 2.0	0.739	0.792	0.763	0.761
RNAfold 1.8.5	0.711	0.773	0.740	0.737
UNAFold	0.692	0.766	0.727	0.724
RNAStructure	0.715	0.781	0.745	0.742

Availability and Requirements

The source code of the `ViennaRNA Package` as well as the current reference manual can be downloaded from http://www.tbi.univie.ac.at/RNA.

Appendix

Energy model implementation details

The most important technical innovation is the use of the **2004 - improved nearest neighbor model** by Mathews et al. [9] as the default parameter set in all

Lorenz *et al. Algorithms for Molecular Biology* 2011, **6**:26
http://www.almob.org/content/6/1/26

free energy calculations. This entails not only an update of all free energy evaluating sections in each affected program, but also major changes in the structure of the parameter sets. In particular, several additional energy parameters for the different loop types (hairpin loops, interior loops and multi-branch loops) were introduced.

In order to keep the number of energy parameters and thus the complexity of the energy model small, we refrained from implementing exceptional contributions for some highly specialized configurations. In particular the following special cases are not incorporated in our folding recursions:

1. *All-C* loop penalty, i.e., a penalizing contribution for loops consisting of unpaired cytosine only;
2. Additional stabilizing *GU-closure term* that is applied only in the context of hairpin loops, enclosed by a *GU* (not *UG*) base pair which is preceded by two *G*s;
3. A special intramolecular helix formation of the four consecutive base pairs *GC, GU, UG* and *CG*, which has a single tabulated contribution of -4.12 kcal/mol.
4. Consideration of an auxilary contributing factor that reflects the *number of states of a bulge loop*, i.e. the number of all possible bulges with identical sequence.
5. *Average asymmetry* correcting penalty *in multi-branch loops* which constitutes the mean difference in unpaired nucleotides on both sides of the branching stems;
6. Extra penalty for *three-way branching loops with less then two unpaired nucleotides*;

Adapting the dynamic programming recursions to also take into account these loop configurations resulted in an increase of time and memory requirements without a compensating benefit in terms of prediction accuracy. The data-set we used for measuring the prediction performance also did not reveal any significant unfavorable effect of our simplification of the model. However, free energy evaluation of a given sequence/structure pair, as done by `RNAeval`, may introduce these extra cases in the near future as an additional parameter, such as logarithmic multi-branch loop evaluation.

All our folding algorithms assume `-d2` as the default dangling-end model, allowing a single nucleotide to contribute with all its possible favorable interactions. The dangling-end/helix-stacking model suggested by the Turner'04 parameters is realized with the `-d3` option. An additional model allowing a single nucleotide to be involved in at most one favorable interaction but ignoring helix-stacking can be chosen with `-d1`, while `-d0` deactivates dangling-end and helix-stacking contributions altogether.

Performance

The base pair positions along the RNA sequence were taken as predicted properties for all of the performance measurements. Thus, all base pairs in the reference structure contribute to the number of *true positives* (TP). The number of *false positives* (FP) is obtained by counting all base pairs that are in the predicted but not in the reference secondary structure. Along with that, all base pairs present in the reference but not in the prediction result are regarded as *false negatives* (FN). These numbers are then used to compute the *sensitivity*, also known as *true positive rate* (TPR), and *precision*, also known as *positive predictive value* (PPV) [105].

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

To combine these performance measures into one single value, we used the Matthews Correlation Coefficient (MCC) [106] and the F_1-score (F-measure), i.e. the harmonic mean of *precision* and *true positive rate.*

$$\mathrm{MCC} = \frac{\mathrm{TP} \cdot \mathrm{TN} - \mathrm{FP} \cdot \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

$$\mathrm{F}_1 = 2 \cdot \frac{\mathrm{PPV} \cdot \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}}$$

Since the total number of possible base pairings is bound by $\frac{1}{2} \cdot n \cdot (n-1)$, with sequence length n, we estimated the number of *true negative* (TN) which is required for calculating the MCC by its upper bound of $\mathrm{TN} = \frac{1}{2} \cdot n \cdot (n-1) - \mathrm{TP}$.

Detailed description of Figure 1

Example calls of programs included in the `ViennaRNA Package` and their corresponding output. (**A**) `RNAfold` output on a small example sequence. Top: On-screen output - mfe, ensemble representation, and centroid structure as dot-parenthesis (Vienna) representations. Numbers in brackets denote the energies, and the centroid's mean distance to the ensemble. Below: postscript output as generated by the above programm call. The mountain plot and the generating program call are in the center of the sub figure. The bottom shows positional entropy derived reliabilty information color coded into the secondary structure drawing.

(B) Example output of programs for local folding. Top: Dot plot as generated by `RNAplfold`. The plot is a cut out along the diagonal of a quadratic dot plot (see e.g. part **(D)** of this figure). At the bottom, an example output of `RNALfold` is shown. Local optimal

substructures are shown in dot-parenthesis notation together with their energy and the index of their first base.

(C) Example output of `RNAup`. At the bottom the best interacting site between the two input molecules is shown. The xmgrace generated picture above shows the energy necessary to open a window of 4 consecutive bases and the interaction energy that can be achieved when the probe molecule is bound to the target molecule in black and red, respectively.

(D) `RNAalifold` output. At the top and bottom pictures generated by `RNAalifold` are shown. The conservation of the base pairs is encoded in a color scheme. Red means only one of the 6 possible base pairs is present, ochre means two, green 3 and so on. Paler colors indicate that some of the sequences cannot form a base pair at the respective position in the alignment. The top right corner shows a dot plot. Every dot symbolizes a base pair, the size of the dots at the upper right triangle is proportional to the respective base pair probabilities, while on the lower left triangle the mfe structure is depicted. On the top right the conservation annotated consensus structure drawing can be seen, while on the bottom the annotated alignment is shown. The center of the subfigure shows the on-screen output of `RNAalifold`. As in the ordinary fold case, the minimum free energy structure, a representation of the ensemble structure and the centroid structure are shown. The energies are split into a thermodynamic part (first) and the conservation part, which are summed to give the total predicted score.

(E) `RNAcofold` output. At the top the secondary structure drawing of the minimum free energy folding of the two molecules is shown. The molecules are color coded to make it easier to tell them apart. The "&" character in the on-screen output below is the separator between the two sequences. In addition to the mfe and the ensemble representation with their energies, the binding energy is shown.

(F) Output for kinetics (using `RNAsubopt` output fed into the external programs `barriers` and `treekin`). The diagram shows the change in population from the start, where state 20 is populated, towards the equilibrium state 1. The inner picture shows the barrier tree upon which the relative concentrations of the diagram are based. The 20 lowest suboptimal structures and the paths connecting them are depicted, together with the barrier heights.

(G) Output of the three versions of `RNAsubopt`. Left: Output of the Wuchty algorithm, all structures within a certain energy band are shown. Right: Zuker algorithm, showing the best structures for every possible base pair. Bottom: Stochastic backtracking, random structures drawn according to their probability in the ensemble.

Additional material

Additional file 1: Performance comparison (Sensitivity, PPV, F-measure).

Additional file 2: Detailed performance comparison.

Acknowledgements
This work is dedicated to Peter Schuster on the occasion of his 70th birthday.
We thank all major contributers to earlier versions of the `ViennaRNA Package` whose implementation of algorithms and programs helped to create this versatile and comprehensive software collection (in alphabetical order):
Wolfgang Beyer, Sebastian Bonhöffer, Martin Fekete, Walter Fontana, Ulrike Mückstein, Wolfgang Schnabel, Manfred Tacker, and Stefan Wuchty.
We are grateful to all beta testers who were unrestrained in reporting bugs on preliminary versions of `ViennaRNA Package 2.0`. This work was funded, in part, by the austrian GEN-AU projects "regulatory non coding RNA", "Bioinformatics Integration Network III", the Austrian FWF project "SFB F43 RNA regulation of the transcriptome", the European Union FP-7 project QUANTOMICS and Deutsche Forschungsgemeinschaft (grant No. STA 850/7-1 under the auspices of SPP-1258 "Small Regulatory RNAs in Prokaryotes").

Author details
[1]Institute for Theoretical Chemistry and Structural Biology, University of Vienna, Währingerstraße 17/3, A-1090 Vienna, Austria. [2]Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany. [3]Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark. [4]Max Planck Institute for Mathematics in the Sciences, Inselstraße 22 D-04103 Leipzig, Germany. [5]Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, D-04103 Leipzig, Germany. [6]Santa Fe Institute, 1399 Hyde Park Rd, Santa Fe, NM 87501, USA. [7]Research Group Bioinformatics and Computational Biology, Faculty of Computer Science, University of Vienna, Währingerstraße 17/3, A-1090 Vienna, Austria.

Authors' contributions
Work on the `Vienna RNA Package` is coordinated by ILH. The design and structure of version 2.0 resulted from discussion of RL with ILH, PFS, CF, SHB and CHzS. Implementation and performance analysis was performed by RL with contributions of HT (`RNAplex and RNAsnoop`). CHzS provided the converted new energy parameter files. Detailed documentation for the RNAlib was done by RL and SHB based on pre-existing sources. The manuscript was written by RL with major contribution by SHB, PFS, and ILH. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Received: 22 August 2011 Accepted: 24 November 2011
Published: 24 November 2011

References

1. Thirumalai D, Lee N, Woodson SA, Klimov DK: **Early Events in RNA Folding.** *Annu Rev Phys Chem* 2001, **52**:751-762.
2. Tinoco I Jr, Uhlenbeck OC, Levine MD: **Estimation of Secondary Structure in Ribonucleic Acids.** *Nature* 1971, **230**:362-367.
3. Tinoco I Jr, Borer PN, Dengler B, Levine ND, Uhlenbeck OC, Crothers DM, Gralla J: **Improved estimation of secondary structure in ribonucleic acids.** *Nature* 1973, **246**:40-41.
4. Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, Neilson T, Turner DH: **Improved free-energy parameters for predictions of RNA duplex stability.** *Proc Natl Acad Sci, USA* 1986, **83**:9373-9377.
5. Jaeger JA, Turner DH, Zuker M: **Improved predictions of secondary structures for RNA.** *Proc Natl Acad Sci USA* 1989, **86**:7706-7710.
6. He L, Kierzek R, SantaLucia J, Walter AE, Turner DH: **Nearest-Neighbor Parameters for GU Mismatches.** *Biochemistry* 1991, 30.

7. Xia T, SanatLucia J Jr, Burkard ME, Kierzek R, Schroeder SJ, Jiao X, Cox C, Turner DH: **Parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick pairs.** *Biochemistry* 1998, **37**:14719-14735.
8. Mathews DH, Sabina J, Zuker M, Turner DH: **Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure.** *J Mol Biol* 1999, **288(5)**:911-940.
9. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH: **Incorporating chemical modification constraints into a dynamic programming algorithm for prediction ofRNA secondary structure.** *Proc Natl Acad Sci USA* 2004, **101**:7287-7292.
10. Nussinov R, Piecznik G, Griggs JR, Kleitman DJ: **Algorithms for Loop Matching.** *SIAM J Appl Math* 1978, **35**:68-82.
11. Waterman MS, Smith TF: **RNA secondary structure: A complete mathematical analysis.** *Mathematical Biosciences* 1978, **42**:257-266.
12. Nussinov R, Jacobson AB: **Fast algorithm for predicting the secondary structure of single stranded RNA.** *Proc Natl Acad Sci USA* 1980, **77**:6309-6313.
13. Zuker M, Stiegler P: **Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information.** *Nucleic Acids Res* 1981, **9**:133-148.
14. Zuker M, Sankoff D: **RNA secondary structures and their prediction.** *Bull Math Biol* 1984, **46**:591-621.
15. McCaskill JS: **The Equilibrium Partition Function and Base Pair Binding Probabilities for RNA Secondary Structure.** *Biopolymers* 1990, **29**:1105-1119.
16. Eddy SR: **A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure.** *BMC Bioinformatics* 2002, **3**:18.
17. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009, **25**:1335-1337.
18. Do CB, Woods DA, Batzoglou S: **CONTRAfold: RNA secondary structure prediction without physics-based models.** *Bioinformatics* 2006, **22**:90-98.
19. Markham NR, Zuker M: **UNAFold: software for nucleic acid folding and hybridization.** *Methods Mol Biol* 2008, **453**:3-31.
20. Reuter JS, Mathews DH: **RNAstructure: software for RNA secondary structure prediction and analysis.** *BMC Bioinformatics* 2010, **11**:129-129.
21. Zadeh JN, Steenberg CD, Bois JS, Wolfe BR, Pierce MB, Khan AR, Dirks RM, Pierce NA: **NUPACK: Analysis and design of nucleic acid systems.** *J Comput Chem* 2011, **32**:170-3.
22. Hamada M, Kiryu H, Sato K, Mituyama T, Asai K: **Prediction of RNA secondary structure using generalized centroid estimators.** *Bioinformatics* 2009, **25(4)**:465-473.
23. Ding Y, Lawrence CE: **A statistical sampling algorithm for RNA secondary structure prediction.** *Nucleic Acids Res* 2003, **31**:7280-7301.
24. Reeder J, Giegerich R: **RNA secondary structure analysis using the RNAshapes package.** *Curr Protoc Bioinformatics* 2009, Chapter 12:Unit12.8..
25. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P: **Fast Folding and Comparison of RNA Secondary Structures (The Vienna RNA Package).** *Monatsh Chem* 1994, **125(2)**:167-188.
26. Fontana W, Schuster P: **A computer model of evolutionary optimization.** *Biophys Chem* 1987, **26**:123-147.
27. Fontana W, Schnabl W, Schuster P: **Physical aspects of evolutionary optimization and adaption.** *Phys Rev A* 1989, **40**:3301-3321.
28. Fontana W, Griesmacher T, Schnabl W, Stadler PF, Schuster P: **Statistics of Landscapes Based on Free Energies, Replication and Degradation Rate Constants of RNA Secondary Structures.** *Monatsh Chem* 1991, **122**:795-819.
29. Bonhoeffer S, McCaskill JS, Stadler PF, Schuster P: **RNA Multi-Structure Landscapes. A Study Based on Temperature Dependent Partition Functions.** *Eur Biophys J* 1993, **22**:13-24.
30. Fontana W, Konings DAM, Stadler PF, Schuster P: **Statistics of RNA Secondary Structures.** *Biopolymers* 1993, **33**:1389-1404.
31. Schuster P, Fontana W, Stadler PF, Hofacker IL: **From Sequences to Shapes and Back: A case study in RNA secondary structures.** *Proc Roy Soc Lond B* 1994, **255**:279-284.
32. Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL: **The Vienna RNA websuite.** *Nucleic Acids Res* 2008, **36(Web Server)**:70-74.
33. Hofacker IL, Stadler PF: **Memory Efficient Folding Algorithms for Circular RNA Secondary Structures.** *Bioinformatics* 2006, **22**:1172-1176.
34. Wuchty S, Fontana W, Hofacker IL, Schuster P: **Complete Suboptimal Folding of RNA and the Stability of Secondary Structures.** *Biopolymers* 1999, **49**:145-165.
35. Tacker M, Stadler PF, Bornberg-Bauer EG, Hofacker IL, Schuster P: **Algorithm Independent Properties of RNA Structure Prediction.** *Eur Biophy J* 1996, **25**:115-130.
36. Zuker M: **On finding all suboptimal foldings of an RNA molecule.** *Science* 1989, **244**:48-52.
37. Hofacker IL, Priwitzer B, Stadler PF: **Prediction of Locally Stable RNA Secondary Structures for Genome-Wide Surveys.** *Bioinformatics* 2004, **20**:186-190.
38. Bernhart S, Hofacker IL, Stadler PF: **Local RNA Base Pairing Probabilities in Large Sequences.** *Bioinformatics* 2006, **22**:614-615.
39. Lorenz R, Flamm C, Hofacker IL: **2D Projections of RNA folding Landscapes.** In *German Conference on Bioinformatics 2009, Volume 157 of Lecture Notes in Informatics.* Edited by: Grosse I, Neumann S, Posch S, Schreiber F, Stadler PF. Bonn: Gesellschaft f. Informatik; 2009:11-20.
40. Bernhart SH, Tafer H, Mückstein U, Flamm C, Stadler PF, Hofacker IL: **Partition Function and Base Pairing Probabilities of RNA Heterodimers.** *Algorithms Mol Biol* 2006, **1**:3.
41. Mückstein U, Tafer H, Bernhart SH, Hernandez-Rosales M, Vogel J, Stadler PF, Hofacker IL: **Translational Control by RNA-RNA Interaction: Improved Computation of RNA-RNA Binding Thermodynamics.** In *Bioinformatics Research and Development, Volume 13 of Communications in Computer and Information Science.* Edited by: Elloumi M, Küng J, Linial M, Murphy R, Schneider K, Toma C. Springer; 2008:114-127.
42. Dimitrov RA, Zuker M: **Prediction of Hybridization and Melting for Double-Stranded Nucleic Acids.** *Biophys J* 2004, **87**:215-226.
43. Beyer W: **RNA Secondary Structure Prediction including Pseudoknots.** *Master's thesis* University Vienna; 2010 [http://www.tbi.univie.ac.at/newpapers/Master_theses2010.html].
44. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R: **Fast and effective prediction of microRNA/target duplexes.** *RNA* 2004, **10**:1507-1517.
45. Tafer H, Hofacker IL: **RNAplex: a fast tool for RNA-RNA interaction search.** *Bioinformatics* 2008, **24(22)**:2657-2663.
46. Tafer H, Ammann F, Eggenhoffer F, Stadler PF, Hofacker IL: **Fast Accessibility-Based Prediction of RNA-RNA Interactions.** *Bioinformatics* 2011, **27**:1934-1940.
47. Tafer H, Kehr S, Hertel J, Hofacker IL, Stadler PF: **RNAsnoop: efficient target prediction for H/ACA snoRNAs.** *Bioinformatics* 2010, **26(5)**:610-616.
48. Gruber AR, Bernhart SH, Zhou Y, Hofacker IL: **RNALfoldz: Efficient prediction of thermodynamically stable, local secondary structures.** In *German Conference on Bioinformatics 2010, Volume 173 of Lecture Notes in Informatics.* Edited by: Schomburg D, Grote A. Bonn: Gesellschaft f. Informatik; 2010:12-21.
49. Hofacker IL, Fekete M, Stadler PF: **Secondary Structure Prediction for Aligned RNA Sequences.** *J Mol Biol* 2002, **319**:1059-1066.
50. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF: **RNAalifold: improved consensus structure prediction for RNA alignments.** *BMC Bioinformatics* 2008, **9**:474.
51. Hofacker IL, Fekete M, Flamm C, Huynen MA, Rauscher S, Stolorz PE, Stadler PF: **Automatic Detection of Conserved RNA Structure Elements in Complete RNA Virus Genomes.** *Nucl Acids Res* 1998, **26**:3825-3836.
52. Hofacker IL, Stadler PF: **Automatic Detection of Conserved Base Pairing Patterns in RNA Virus Genomes.** *Comp & Chem* 1999, **23**:401-414.
53. Klein R, Eddy S: **RSEARCH: Finding homologs of single structured RNA sequences.** *BMC Bioinformatics* 2003, **4**:44.
54. Sankoff D: **Simultaneous solution of the RNA folding, alignment, and proto-sequence problems.** *SIAM J Appl Math* 1985, **45**:810-825.
55. Havgaard JH, Lyngs RB, Gorodkin J: **The foldalign web server for pairwise structural RNA alignment and mutual motif search.** *Nucleic Acids Research* 2005, **33(suppl 2)**:W650-W653.
56. Mathews DH, Turner DH: **Dynalign: an algorithm for finding the secondary structure common to two RNA sequences.** *Journal of Molecular Biology* 2002, **317(2)**:191-203.
57. Hofacker IL, Bernhart SHF, Stadler PF: **Alignment of RNA Base Pairing Probability Matrices.** *Bioinformatics* 2004, **20**:2222-2227.
58. Hofacker IL, Stadler PF: **The Partition Function Variant of Sankoff's Algorithm.** In *Computational Science - ICCS 2004, Volume 3039 of Lecture Notes in Computer Science* Edited by: Bubak M, van Albada G, Sloot PA, Dongarra J 2004, 728-735, [Kraków, June 6-9, 2004]..

59. Will S, Missal K, Hofacker IL, Stadler PF, Backofen R: **Inferring Non-Coding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering.** *PLoS Comp Biol* 2007, **3**:e65.
60. Busch A, Backofen R: **INFO-RNA-a fast approach to inverse RNA folding.** *Bioinformatics* 2006, **22(15)**:1823-1831.
61. Andronescu M, Aguirre-Hernández R, Condon A, Hoos HH: **RNAsoft: a suite of RNA secondary structure prediction and design software tools.** *Nucleic Acids Research* 2003, **31(13)**:3416-3422.
62. Zadeh JN, Wolfe BR, Pierce NA: **Nucleic acid sequence design via efficient ensemble defect optimization.** *J Comput Chem* 2011, **32**:439-452.
63. Flamm C, Hofacker IL, Maurer-Stroh S, Stadler PF, Zehl M: **Design of Multi-Stable RNA Molecules.** *RNA* 2001, **7**:254-265.
64. Flamm C, Fontana W, Hofacker IL, Schuster P: **RNA Folding at Elementary Step Resolution.** *RNA* 2000, **6**:325-338.
65. Flamm C, Hofacker IL, Stadler PF, Wolfinger MT: **Barrier Trees of Degenerate Landscapes.** *Z Phys Chem* 2002, **216**:155-173.
66. Bruccoleri RE, Heinrich G: **An improved algorithm for nucleic acid secondary structure display.** *Computer applications in the biosciences: CABIOS* 1988, **4**:167-173.
67. Turner DH, Mathews DH: **NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure.** *Nucleic Acids Res* 2010, **38(Database)**:280-282.
68. SantaLucia J, Hicks D: **The thermodynamics of DNA structural motifs.** *Annu Rev Biophys Biomol Struct* 2004, **33**:415-440.
69. Andronescu MS, Pop C, Condon AE: **Improved free energy parameters for RNA pseudoknotted secondary structure prediction.** *RNA* 2010, **16**:26-42.
70. Andronescu M, Condon A, Hoos HH, Mathews DH, Murphy KP: **Efficient parameter estimation for RNA secondary structure prediction.** *Bioinformatics* 2007, **23**:i19-i28.
71. Ding Y, Chan CY, Lawrence CE: **RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble.** *RNA* 2005, **11(8)**:1157-1166.
72. Knudsen B, Hein J: **Pfold: RNA secondary structure prediction using stochastic context-free grammars.** *Nucleic Acids Research* 2003, **31(13)**:3423.
73. Andronescu M, Bereg V, Hoos HH, Condon A: **RNA STRAND: the RNA secondary structure and statistical analysis database.** *BMC Bioinformatics* 2008, **9**:340-340.
74. Sczyrba A, Krüger J, Mersch H, Kurtz S, Giegerich R: **RNA-related tools on the Bielefeld Bioinformatics Server.** *Nucleic Acids Research* 2003, **31(13)**:3767-3770.
75. Höchsmann M, Töller T, Giegerich R, Kurtz S: **Local similarity in RNA secondary structures.** *Proc IEEE Comput Soc Bioinform Conf* 2003, **2**:159-168.
76. Aksay C, Salari R, Karakoc E, Alkan C, Sahinalp S: **taveRNA: a web suite for RNA algorithms and applications.** *Nucleic acids research* 2007, **35(suppl 2)**: W325.
77. Parisien M, Major F: **The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data.** *Nature* 2008, **452**:51-55.
78. Darty K, Denise A, Ponty Y: **VARNA: Interactive drawing and editing of the RNA secondary structure.** *Bioinformatics* 2009, **25(15)**:1974-1975.
79. Höner zu Siederdissen C, Bernhart SH, Stadler PF, Hofacker IL: **A Folding Algorithm for Extended RNA Secondary Structures.** *Bioinformatics* 2011, **27**:129-136.
80. Pearson W, Lipman D: **Improved tools for biological sequence comparison.** *Proceedings of the National Academy of Sciences of the United States of America* 1988, **85(8)**:2444.
81. Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, McWilliam H, Valentin F, Wallace I, Wilm A, Lopez R, *et al*: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23(21)**:2947.
82. Gardner P, Daub J, Tate J, Nawrocki E, Kolbe D, Lindgreen S, Wilkinson A, Finn R, Griffiths-Jones S, Eddy S, *et al*: **Rfam: updates to the RNA families database.** *Nucleic acids research* 2009, **37(suppl 1)**:D136.
83. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Research* 2009, **37(suppl 1)**:D26-D31.
84. Kalaš M, Puntervoll P, Joseph A, Bartaševičiūte E, Töpfer A, Venkataraman P, Pettifer S, Bryne J, Ison J, Blanchet C, *et al*: **BioXSD: the common data-exchange format for everyday bioinformatics web services.** *Bioinformatics* 2010, **26(18)**:i540.
85. Waugh A, Gendron P, Altman R, Brown JW, Case D, Gautheret D, Harvey SC, Leontis N, Westbrook J, Westhof E, Zuker M, Major F: **RNAML: a standard syntax for exchanging RNA information.** *RNA* 2002, **8(6)**:707-717.
86. Hofacker IL, Huynen MA, Stadler PF, Stolorz PE: **Knowledge Discovery in RNA Sequence Families of HIV Using Scalable Computers.** In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR.* Edited by: Simoudis E, Han J, Fayyad U, Menlo Park. CA: AAAI Press; 1996:20-25.
87. Fekete M, Hofacker IL, Stadler PF: **Prediction of RNA base pairing probabilities using massively parallel computers.** *J Comp Biol* 2000, **7**:171-182.
88. Dagum L, Menon R: **OpenMP: an industry standard API for shared-memory programming.** *IEEE Computational Science and Engineering* 1998, **5**:46-55.
89. Gardner P, Giegerich R: **A comprehensive comparison of comparative RNA structure prediction approaches.** *BMC bioinformatics* 2004, **5**:140.
90. Zakov S, Goldberg Y, Elhadad M, Ziv-Ukelson M: **Rich parameterization improves RNA structure prediction.** *Research in Computational Molecular Biology* Springer; 2011, 546-562.
91. Enright AJ, John B, Gaul U, Tuschl T, Sander C, Marks DS: **MicroRNA targets in Drosophila.** *Genome Biol* 2003, 5.
92. Thadani R, Tammi MT: **MicroTar: predicting microRNA targets from RNA duplexes.** *BMC Bioinformatics* 2006, **7**:Suppl 5.
93. Rusinov V, Baev V, Minkov IN, Tabler M: **MicroInspector: a web tool for detection of miRNA binding sites in an RNA sequence.** *Nucleic Acids Res* 2005, **33(Web Server)**:696-700.
94. Hertel J, Stadler PF: **Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data.** *Bioinformatics* 2006, **22(14)**:197-202.
95. Washietl S, Hofacker IL, Stadler PF: **Fast and reliable prediction of noncoding RNAs.** *Proc Natl Acad Sci USA* 2005, **102(7)**:2454-2459.
96. Gruber AR, Findeiss S, Washietl S, Hofacker IL, Stadler PF: **RNAZ 2.0: IMPROVED NONCODING RNA DETECTION.** *Pac Symp Biocomput* 2010, **15**:69-79.
97. Hertel J, Hofacker IL, Stadler PF: **SnoReport: computational identification of snoRNAs with unknown targets.** *Bioinformatics* 2008, **24(2)**:158-164.
98. Reiche K, Stadler PF: **RNAstrand: reading direction of structured RNAs in multiple sequence alignments.** *Algorithms Mol Biol* 2007, **2**:6-6.
99. Kaczkowski B, Torarinsson E, Reiche K, Havgaard JH, Stadler PF, Gorodkin J: **Structural profiles of human miRNA families from pairwise clustering.** *Bioinformatics* 2009, **25(3)**:291-294.
100. Tafer H, Ameres SL, Obernosterer G, Gebeshuber CA, Schroeder R, Martinez J, Hofacker IL: **The impact of target site accessibility on the design of effective siRNAs.** *Nat Biotechnol* 2008, **26(5)**:578-583.
101. Kiryu H, Kin T, Asai K: **Robust prediction of consensus secondary structures using averaged base pairing probability matrices.** *Bioinformatics* 2007, **23**:434-441.
102. Stocsits RR, Letsch H, Hertel J, Misof B, Stadler PF: **Accurate and efficient reconstruction of deep phylogenies from structured RNAs.** *Nucleic Acids Research* 2009, **37(18)**:6184-6193.
103. Heyne S, Will S, Beckstette M, Backofen R: **Lightweight comparison of RNAs based on exact sequence-structure matches.** *Bioinformatics* 2009, **25(16)**:2095-2102.
104. Upper D: **The unsuccessful self-treatment of a case of "writer's block".** *Journal of Applied Behavior Analysis* 1974, **7(3)**:497.
105. Dowell R, Eddy S: **Evaluation of several lightweight stochastic context-free grammars for RNA secondary structure prediction.** *BMC bioinformatics* 2004, **5**:71.
106. Matthews B: **Comparison of the predicted and observed secondary structure of T4 phage lysozyme.** *Biochimica et Biophysica Acta (BBA)-Protein Structure* 1975, **405(2)**:442-451.

doi:10.1186/1748-7188-6-26
Cite this article as: Lorenz *et al.*: ViennaRNA Package 2.0. *Algorithms for Molecular Biology* 2011 **6**:26.

8

PREDICTION OF RNA/DNA HYBRID STRUCTURES

Ronny Lorenz, Ivo L. Hofacker, and Stephan H. Bernhart.
Folding RNA/DNA hybrid duplexes.
in *Bioinformatics* 2012 **Vol. 28** no. 19, pages 2530-2531.
doi: 10.1093/bioinformatics/bts466

SB designed the study. RL designed and implemented the algorithm. All authors have shared writing the paper.

BIOINFORMATICS APPLICATIONS NOTE Vol. 28 no. 19 2012, pages 2530–2531
doi:10.1093/bioinformatics/bts466

Structural bioinformatics

Advance Access publication July 24, 2012

Folding RNA/DNA hybrid duplexes

Ronny Lorenz[1,*], Ivo L. Hofacker[1,2] and Stephan H. Bernhart[2,*]

[1]Institute for Theoretical Chemistry, University of Vienna, [2]Research group BCB, Faculty of Computer Science, University of Vienna, Währinger Straße 17, 1090 Vienna, Austria and [3]Department of Bioinformatics, University of Leipzig, Haertelstraße 16-18, 04109 Leipzig, Germany

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: While there are numerous programs that can predict RNA or DNA secondary structures, a program that predicts RNA/DNA hetero-dimers is still missing. The lack of easy to use tools for predicting their structure may be in part responsible for the small number of reports of biologically relevant RNA/DNA hetero-dimers.

Results: We present here an extension to the widely used `ViennaRNA Package` (Lorenz *et al.*, 2011) for the prediction of the structure of RNA/DNA hetero-dimers.

Availability: http://www.tbi.univie.ac.at/~ronny/RNA/vrna2.html

Contact: ronny@tbi.univie.ac.at, berni@bioinf.uni-leipzig.de

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on April 10, 2012; revised and accepted on July 19, 2012

1 INTRODUCTION

Nucleic acids have many important functions in biological systems, such as information carriers, catalysts and regulators. Both variants, DNA and RNA, can form complex structures by base pair interactions. The pattern of Watson–Crick or wobble base pairs a molecule forms is called the secondary structure. A main difference between RNA and DNA is due to their usage in biological systems: while DNA builds a dimer with its reverse complement almost all the time, RNA, being mostly single stranded, is more prone to fold back onto itself. However, if DNA is single stranded (e.g. during replication, transcription, repair or recombination), it will also form intramolecular base pairs. There are several examples where RNA and DNA interact via base pairing, generating a structure called R-loop. They protect CpG (CG dinucleotide regions in the genome) islands from being methylated (Ginno *et al.*, 2012). The stability of DNA/mRNA hybrids is reported to have an influence on the copy number of repeats (McIvor *et al.*, 2010) and can impair transcription elongation, promoting transcription-associated recombination (Huertas and Aguilera, 2003). Moreover, during the intensive search for new functional transcripts, rendered possible by next-generation sequencing technology, at least 1000 novel lincRNAs were identified that may act as regulators of transcription, and one possibility to do this is via base pairing between RNA and DNA (see Guttman and Rinn, 2012 for a recent review). As a final biological example, DNA/RNA hybrids play a role within the CRISPR pathway in bacteria and archaea (Howard *et al.*, 2011). On a technical side, RNA/DNA dimers are often used in micro-array or PCR experiments. The ability to computationally investigate the properties of the structures of interacting RNA and DNA molecules would help in the analysis of these dimers. While a number of programs exist that can predict RNA or DNA secondary structures (For a review, see e.g. Reeder *et al.*, 2006) or even their dimers, there is, to our knowledge, no program that also includes the possibility to predict the full secondary structure of RNA/DNA hetero-dimers. This may be due to the lack of compiled energy parameters, as there are only stacking energies available today. We introduce here a possibility to predict the secondary structure of such hetero-dimers within the widely used `ViennaRNA Package`.

*To whom correspondence should be addressed.

2 APPROACH

2.1 Adaptation of the `ViennaRNA` package

For the prediction of RNA/DNA hetero-dimers, three distinct energy parameter sets (RNA, DNA and mixed) are necessary. In order to keep the changes within the inner recursions of the algorithm minimal, we use 8 instead of 4 bases and 24 instead of 6 types of Watson–Crick or wobble base pairs for the energy computations. The look-up tables for the energy parameters grow accordingly. However, this does not significantly contribute to memory consumption. Finally, we demand a fixed order for the input sequences: RNA first, DNA last.

2.2 Energy parameters

Since we are only considering RNA/DNA hetero-dimers, not co-polymers, we only need a subset of the energy parameters necessary for the full computations in the RNA or DNA case: we do not encounter stacking of a RNA on a DNA base, for example. In contrast, stacking parameters of RNA/DNA base pairs, multi-loop penalties and the interior loop parameters are needed. For stacked pairs, i.e. perfect helices, the parameters have been determined experimentally (Martin and Tinoco, 1980; Sugimoto *et al.*, 1995). Stacking parameters are the most important part of the energy model, as they account for the majority of RNA/DNA binding free energy. Unfortunately, for other mixed RNA–DNA loops no experimental data are available.

As no high-quality dataset of RNA/DNA structures exists, parameters cannot be trained, and the performance of different parameter sets cannot be evaluated. The data we have are not sufficient to chose a function for the derivation of the energy

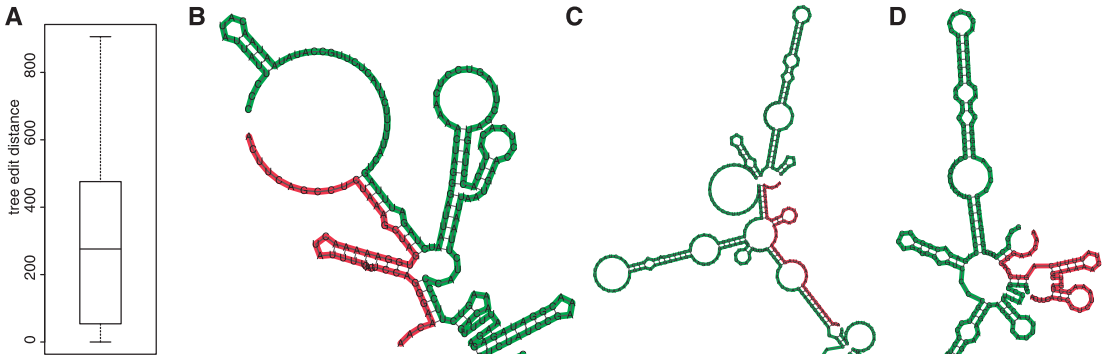


Fig. 1. Using hybrid instead of RNA parameters leads to very different structure prediction as summarized in a boxplot of the tree edit distances (**A**) for all interactions between ANRIL and a putative target region at human chr 1, 36 269 747 to 36 272 043 (hg19). (**B–D**) Predicted secondary structures for one example binding site (light: ANRIL), (**B**) DNA parameters, (**C**) RNA parameters and (**D**) mixed parameters

parameters (Supplementary Fig. S1), so we simply use the average of RNA and DNA $E_{R/D} = (E_R + E_D)/2$ for all missing values. We are aware that this formula gives an at best crude approximation of the actual folding energies, but note that new parameters can easily be incorporated. We hope that the availability of prediction tools for RNA/DNA hetero-dimers will encourage experimentalists to provide the missing energy parameters.

2.3 Inclusion into the `ViennaRNA package`

Our approach to just adapt the energy computations enabled us to easily include the RNA/DNA hetero-dimer support into all parts of the `ViennaRNA Package` that compute dimer structures, namely, `RNAcofold` (Bernhart *et al.*, 2006), `RNAup` (Mückstein *et al.*, 2008) and `RNAduplex` (Tafer and Hofacker, 2008). RNA/DNA hetero-dimers are supported for both the minimum free energy as well as the partition function-based calculations. Furthermore, all programs can read in user-supplied energy parameters. In particular, this allows to make immediate use of any newly determined parameters for mixed RNA–DNA loops, or enables scientists to easily evaluate different energy models as soon as sufficient RNA/DNA structure data are available.

3 RESULTS

In the absence of programs that can deal with RNA/DNA dimers, researchers have resorted to simply treating both parts of the dimer as RNA. To test the effect of applying the right (DNA or RNA) parameters to the single molecules and using explicit RNA/DNA interactions, we chose ANRIL, a long non-coding RNA that is suspected to have a DNA target sequence, and one of its putative targets. In order to simulate the size of the transcription bubble, we chose a sliding window approach with DNA pieces of length 50, step size 25 for our computations. Even though ANRIL (1500 nt) is much longer than the pieces of DNA bound to it, we sometimes observe a dramatic effect on the structure of the hybrid molecule. The tree edit distance of the hybrid structures predicted with RNA parameters and with our hybrid approach rises up to 900, and the structure of the binding site can vary greatly (Fig 1).

Another example for the usefulness of the mixed parameters is the strand dependency of R-loop formation (Reddy *et al.*, 2011). A $r(GAA)_n$ trinucleotid repeat can form R-loops while its reverse complement $r(UUC)_n$ cannot. The non-symmetric mixed parameters can explain that: addition of another trinucleotid will give –4.1 kcal/mol for $r(GAA)$, but only –2.6 kcal/mol for $r(UUC)$. In contrast to RNA or DNA parameters, the mixed parameters can also explain the changes in R-loop formation of other trinucleotides investigated (see Supplementary Fig. S2).

Funding: This work was funded by the Austrian 'SFB 43 RNA regulation of the transcriptome'. This work was funded by the Austrian GEN-AU project 'Bioinformatics Integration Network III'.

Conflict on Interest: none declared.

REFERENCES

Bernhart,S.H. *et al.* (2006) Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol. Biol.*, **1**, 3.

Ginno,P.A. *et al.* (2012) R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol Cell.*, **45**, 814–825.

Guttman,M. and Rinn,J.L. (2012) Modular regulatory principles of large non-coding RNAs. *Nature*, **482**, 339–346.

Howard,J.A. *et al.* (2011) Helicase dissociation and annealing of RNA–DNA hybrids by *Escherichia coli* cas3 protein. *Biochem. J.*, **439**, 85–95.

Huertas,P. and Aguilera,A. (2003) Cotranscriptionally formed DNA:RNA hybrids mediate transcription elongation impairment and transcription-associated recombination. *Mol. Cell*, **12**, 711–721.

Lorenz,R. *et al.* (2011) ViennaRNA package 2.0. *Algorithms Mol. Biol.*, **6**, 26–26.

Martin,F.H. and Tinoco,I. (1980) DNA–RNA hybrid duplexes containing oligo (da:ru) sequences are exceptionally unstable and may facilitate termination of transcription. *Nucleic Acids Res.*, **8**, 2295–2299.

McIvor,E.I. *et al.* (2010) New insights into repeat instability: role of RNA–DNA hybrids. *RNA Biol.*, **7**, 551–558.

Mückstein,U. *et al.* (2008) Translational control by RNA–RNA interaction: improved computation of RNA–RNA binding thermodynamics. In Elloumi,M., Küng,J., Linial,M., Murphy,R., Schneider,K. and Toma,C. (eds) *Bioinformatics Research and Development*. Vol. 13, Communications in Computer and Information Science, Springer, pp. 114–127.

Reddy,K. *et al.* (2011) Determinants of r-loop formation at convergent bidirectionally transcribed trinucleotide repeats. *Nucleic Acids Res.*, **39**, 1749–1762.

Reeder,J. *et al.* (2006) Beyond Mfold: recent advances in RNA bioinformatics. *J. Biotechnol.*, **124**, 41–55.

Sugimoto,N. *et al.* (1995) Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry*, **34**, 11211–11216.

Tafer,H. and Hofacker,I.L. (2008) RNAplex: a fast tool for RNA–RNA interaction search. *Bioinformatics*, **24**, 2657–2663.

9

2D PROJECTIONS OF RNA FOLDING LANDSCAPES

Ronny Lorenz, Christoph Flamm, Ivo Hofacker.
2D Projections of RNA folding Landscapes.
in *In proceedings, Lecture Notes in Informatics, German Conference on Bioinformatics 2009* GCB'09. pages 11–20, GI, Bonn, Germany, 2009

RL (in parts) designed the study, designed and implemented the algorithm, and wrote substantial parts of the paper.

2D Projections of RNA folding Landscapes

Ronny Lorenz, Christoph Flamm, Ivo L. Hofacker
{ronny, xtof, ivo}@tbi.univie.ac.at

Institute for Theoretical Chemistry
University of Vienna, Währingerstraße 17, 1090 Wien, Austria

Abstract: The analysis of RNA folding landscapes yields insights into the kinetic folding behavior not available from classical structure prediction methods. This is especially important for multi-stable RNAs whose function is related to structural changes, as in the case of riboswitches. However, exact methods such as *barrier tree* analysis scale exponentially with sequence length. Here we present an algorithm that computes a projection of the energy landscape into two dimensions, namely the distances to two reference structures. This yields an abstraction of the high-dimensional energy landscape that can be conveniently visualized, and can serve as the basis for estimating energy barriers and refolding pathways. With an asymptotic time complexity of $\mathcal{O}(n^7)$ the algorithm is computationally demanding. However, by exploiting the sparsity of the dynamic programming matrices and parallelization for multi-core processors, our implementation is practical for sequences of up to 400 nt, which includes most RNAs of biological interest.

1 Introduction

Structure formation of RNA molecules is crucial for the function of non-coding RNAs (ncRNAs) as well as for coding mRNAs with regulatory elements like riboswitches and attenuators. Some RNAs possess distinct meta-stable structures with different biological activity. A prime example are riboswitches that regulate gene expression depending on the presence or absence of a small ligand molecule. The pathogenity of viral agents like viroids is achieved by distinct meta stable structures of their 'genome'. While efficient RNA folding algorithms such as `mfold` [Zuk89] or the Vienna RNA package [HFS$^+$94] can be used to compute most equilibrium properties of an RNA molecule, they provide little information on folding dynamics. In this case one has to resort to either stochastic simulation of the folding process [FHMS$^+$01, IS00, GFW$^+$08] or analysis of the energy landscape based on enumeration or sampling. In particular the `barriers` program [FHSW02] is used to find all local minima in the energy landscape and their connecting transition states and energy barriers. The algorithm is based on a complete enumeration of all low-energy conformations [WFHS99] in the landscape and therefore scales exponentially with sequence length. In contrast, the `paRNAss` tool [GHR99] relies on sampling structures and clustering in order to detect multi-stable RNAs but gives no information on energy barriers.

The dynamic programming (DP) approach to RNA folding can also be extended to obtain

more information on the energy landscape: Cupal et al. [CHS96] proposed an algorithm that computes the density of states, i.e. the number of structures that fall into a particular energy bin, by extending the usual DP table into a third dimension corresponding to the energy bins. The `RNAbor` algorithm [FMC07], uses the base pair distance to a reference structure as the additional dimension and computes the optimal secondary structure as well as partition function for each distance class (δ-neighborhood).

Both approaches can be viewed as a one-dimensional projection of the high dimensional energy landscape, which however results into a drastic loss of information. Here, we describe a related method that employs the distances to two reference structures in order to compute a 2D projection which retains enough information to predict qualitative folding behavior and is easy to visualize. In particular, we define a κ, λ - *neighborhood* to be all secondary structures s with $d_{\text{BP}}(s_1, s) = \kappa$ and $d_{\text{BP}}(s_2, s) = \lambda$, where $d_{\text{BP}}(s_a, s_b)$ is the base pair distance of s_a and s_b, and proceed to compute minimum free energy (MFE) structure, as well as partition function and Boltzmann weighted structure samples for each κ, λ - *neighborhood* .

2 Methods

2.1 Minimum free energy algorithm

In the following we will write (i, j) to denote a base pair between the ith and jth nucleotide. A secondary structure s is regarded as a set of base pairs, the *base pair distance* between two structures is defined as $d_{\text{BP}}(s_1, s_2) = |s_1 \cup s_2| - |s_1 \cap s_2|$ and equals the number of base pairs present in either but not both structures. We will write $s[i, j] = \{(p, q) \in s : i \leq p < q \leq j\}$, to identify the substructure on the sequence interval $[i, j]$. $E(s)$ denotes the free energy of structure s.

For reference, we reproduce below the classic recurrences for MFE folding, which we will extend to the κ, λ - *neighborhood* in the following section. Note that the recursions employ an unambiguous decomposition of secondary structures as implemented in the Vienna RNA package [HFS+94].

$$
\begin{aligned}
F_{i,j} &= \min \left\{ F_{i,j-1}, \min_{i<k\leq j} F_{ik} + C_{k+1,j} \right\} \\
C_{i,j} &= \min \left\{ \mathcal{H}(i,j), \min_{i<k<l<j} C_{kl} + \mathcal{I}(i,j;k,l), \min_{i<u<j} M_{i+1,u} + \hat{M}_{u+1,j-1} + a \right\} \\
M_{i,j} &= \min \left\{ \min_{i<u<j} (u-i-1)c + C_{u+1,j} + b, \min_{i<u<j} M_{i,u} + C_{u+1,j} + b, M_{i,j-1} + c \right\} \\
\hat{M}_{i,j} &= \min \left\{ \hat{M}_{i,j-1} + c, C_{ij} + b \right\}
\end{aligned}
\tag{1}
$$

The upper triangular matrices $F_{i,j}$, $C_{i,j}$, $M_{i,j}$ and $\hat{M}_{i,j}$ contain the optimal folding energy on the sequence interval $[i, j]$, optimal energy given that (i, j) form a pair, given that i and j

reside in a multi-loop, and for multi-loop components with exactly one stem in the interval $[i, j]$, respectively. $\mathcal{H}(i, j)$ denotes the energy of a haipin-loop closed by (i, j), $\mathcal{I}(i, j, p, q)$ the energy of an interior-loop closed by (i, j) and (p, q). The parameters a, b, and c contain the penalties for closing a multi-loop, for adding a multi-loop component, and enlarging a multi-loop by one unpaired base. We use the energy parameters as tabulated by the Turner group [MSZT99]. After filling the matrices the MFE structure is found by backtracking in the usual manner.

2.2 Minimum free energy κ, λ-neighbors

For a given RNA sequence $\mathcal{S}$ and two fixed reference structures s_1 and s_2, the MFE version of the κ, λ - *neighborhood* algorithm computes energetically optimal structures $s_{\text{opt}}^{\kappa,\lambda} \in S^{\kappa,\lambda}$ where $S^{\kappa,\lambda} = \{s \mid d_{\text{BP}}(s_1, s) = \kappa \wedge d_{\text{BP}}(s, s_2) = \lambda\}$ is the κ, λ - *neighborhood* of reference structure s_1 and s_2. We extend the recursions (1) such that for each entry of the energy matrices F,C,M and $\hat{M}$ the optimal energy contribution of substructures $s[i, j]$ with $d_{\text{BP}}(s_1[i, j], s[i, j]) = \kappa$ and $d_{\text{BP}}(s_2[i, j], s[i, j]) = \lambda$ are computed. This leads to two additional dimensions in the energy matrices denoted by $F^{\kappa,\lambda}$,$C^{\kappa,\lambda}$,$M^{\kappa,\lambda}$ and $\hat{M}^{\kappa,\lambda}$. Since closing base pairs may lead to an increase of the base pair distance to both reference structures s_1 and s_2, additional decomposition constraints have to be introduced in the recurrences.

A hairpin loop closed by (i, j), for example, contributes to $C_{i,j}^{\kappa,\lambda}$ only if the substructure $s[i, j]$ consisting of the single pair $\{(i, j)\}$ only has distances κ and λ to the two substructures $s_1[i, j]$ and $s_2[i, j]$, respectively. Thus, we introduce the shorthand

$$\mathfrak{H}(i, j, \kappa, \lambda) = \begin{cases} \mathcal{H}(i, j) & \text{if } d_{\text{BP}}(s_1[i, j], \{(i, j)\}) = \kappa,\ d_{\text{BP}}(s_2[i, j], \{(i, j)\}) = \lambda \\ \infty & \text{else} \end{cases} \tag{2}$$

For non-hairpin loops, we introduce five terms $\delta_1^x - \delta_5^x$, where the superscript x is either 1 or 2, denoting the reference structure.

$$\delta_1^x(i, j) = d_{\text{BP}}(s_x[i, j], s_x[i, j-1]) \tag{3}$$
$$\delta_2^x(i, j, u) = d_{\text{BP}}(s_x[i, j], s_x[i, u-1] \cup s_x[u, j]) \tag{4}$$
$$\delta_3^x(i, j, p, q) = d_{\text{BP}}(s_x[i, j], \{(i, j)\} \cup s_x[p, q]) \tag{5}$$
$$\delta_4^x(i, j, u) = d_{\text{BP}}(s_x[i, j], \{(i, j)\} \cup s_x[i+1, u] \cup s_x[u+1, j-1]) \tag{6}$$
$$\delta_5^x(i, j, u) = d_{\text{BP}}(s_x[i, j], s_x[u, j]) \tag{7}$$

Each of the δ in eqs. (3-7) covers a distinct case in the energy minimization recursions, and denotes the minimal distance to the reference structure incurred when decomposing a substructure into two parts (since base pairs in the reference structure crossing the decomposition splitting positions j, u, p and q must be opened). For example $\delta_1^x(i, j)$ equals 1, if j is paired in the structure interval $s_x[i, j]$ of the reference structure s_x and 0 otherwise.

Decompositions into more than one substructure lead to additional combinatorial possibilities. They are taken into account by minimizing over $(\omega, \hat{\omega})$ pairs, where the sum $(\omega + \hat{\omega})$

reflects the residual of the base pair distance between the substructures and the references. Thus, the recursions to compute $E(s_{\text{opt}}^{\kappa,\lambda}) = F_{1,n}^{\kappa,\lambda}$ are:

$$
\begin{aligned}
F_{i,j}^{\kappa,\lambda} &= \min \left\{ \begin{array}{l} F_{i,j-1}^{\kappa-\delta_1^1(i,j),\lambda-\delta_1^2(i,j)}, \\ \min\limits_{i\le u<j} \min\limits_{\substack{\omega_1+\hat{\omega}_1=\kappa-\delta_2^1(i,j,u) \\ \omega_2+\hat{\omega}_2=\lambda-\delta_2^2(i,j,u)}} F_{i,u-1}^{\omega_1,\omega_2} + C_{u,j}^{\hat{\omega}_1,\hat{\omega}_2} \end{array} \right. \\
C_{i,j}^{\kappa,\lambda} &= \min \left\{ \begin{array}{l} \mathfrak{H}(i,j,\kappa,\lambda), \\ \min\limits_{i<p<q<j} \left\{ C_{p,q}^{\kappa-\delta_3^1(i,j,p,q),\lambda-\delta_3^2(i,j,p,q)} + \mathcal{I}(i,j,p,q) \right\}, \\ \min\limits_{i<u<j} \min\limits_{\substack{\omega_1+\hat{\omega}_1=\kappa-\delta_4^1(i,j,u) \\ \omega_2+\hat{\omega}_2=\lambda-\delta_4^2(i,j,u)}} \left\{ M_{i+1,u}^{\omega_1,\omega_2} + \hat{M}_{u+1,j-1}^{\hat{\omega}_1,\hat{\omega}_2} + a \right\} \end{array} \right. \\
M_{i,j}^{\kappa,\lambda} &= \min \left\{ \begin{array}{l} M_{i,j}^{\kappa-\delta_1^1(i,j-1),\lambda-\delta_1^2(i,j)} + c \\ \min\limits_{i\le u<j} \left\{ (u-i)\cdot c + C_{u,j}^{\kappa-\delta_5^1(i,j,u),\lambda-\delta_5^2(i,j,u)} + b \right\}, \\ \min\limits_{i\le u<j} \min\limits_{\substack{\omega_1+\hat{\omega}_1=\kappa-\delta_2^1(i,j,u) \\ \omega_2+\hat{\omega}_2=\lambda-\delta_2^2(i,j,u)}} \left\{ M_{i,u-1}^{\omega_1,\omega_2} + C_{u,j}^{\hat{\omega}_1,\hat{\omega}_2} + b \right\}, \end{array} \right. \\
\hat{M}_{i,j}^{\kappa,\lambda} &= \min \left\{ \begin{array}{l} C_{i,j}^{\kappa,\lambda} + b \\ \hat{M}_{i,j-1}^{\kappa-\delta_1^1(i,j),\lambda-\delta_1^2(i,j)} + c, \end{array} \right.
\end{aligned}
\tag{8}
$$

2.3 Time and memory complexity

Regarding the time complexity of the algorithm, a contribution of $\mathcal{O}(n^3)$, where n denotes RNA sequence length is implicit due to the underlying MFE folding algorithm. The additional degrees of freedom of the multi-loop decompositions in $C_{i,j}^{\kappa,\lambda}$ and $M_{i,j}^{\kappa,\lambda}$ increase the complexity by a factor of $\kappa \cdot \lambda$. The extension of the dynamic programming matrices by two further dimensions κ and λ additionally requires quadratically more effort. If the maximum distance values of κ and λ is limited to $\kappa \le d_1$ and $\lambda \le d_2$, the time complexity becomes $\mathcal{O}(n^3 \cdot d_1^2 \cdot d_2^2)$. Since the maximum number of base pairs on a sequence of length n is $\sim \frac{n}{2}$, the maximum achievable base pair distance between any two structures is bounded by n. Thus, the total asymptotic time complexity of the κ, λ - *neighborhood* algorithm results in $\mathcal{O}(n^7)$ for any distance boundaries d_1 and d_2.

A similar argument holds for the memory complexity which is $\mathcal{O}(n^2 \cdot d_1 \cdot d_2) = \mathcal{O}(n^4)$. Thus, the memory increase compared to regular MFE folding is $d_1 \cdot d_2 \le n^2$.

2.4 Partition function of the κ, λ - *neighborhood*

A modification of algorithm (8) to compute the partition function

$$Q^{\kappa,\lambda} = \sum_{s_x \in S^{\kappa,\lambda}} e^{-E(s_x)/kT} \tag{9}$$

for each κ, λ - *neighborhood* according to the algorithm of McCaskill et al. [McC90] is straight forward. The energy contributions are Boltzmann weighted and all sums/minimizations are replaced by products/sums. This can be done, as the recursions (8) perform unique decompositions and therefore already constitute a partitioning.

Since clustering of the complete secondary structure space into κ, λ - *neighborhoods* is a partitioning too, $\sum_{\kappa,\lambda} Q^{\kappa,\lambda} = Q$, where $Q = \sum_s e^{-E(s)/kT}$ is the partition function of the complete ensemble of all secondary structures.

The Boltzmann probabilities of a κ, λ - *neighborhood* in the complete ensemble and for a structure $s_x \in S^{\kappa,\lambda}$ inside a κ, λ - *neighborhood* become

$$P(S^{\kappa,\lambda}) = \frac{Q^{\kappa,\lambda}}{Q} \quad \text{and} \quad P(s_x \in S^{\kappa.\lambda}) = \frac{e^{-E(s_x)/kT}}{Q^{\kappa,\lambda}} \tag{10}$$

Stochastic backtracking yields a Boltzmann weighted sample of representative structures.

2.5 Sparse matrix approach and Parallelization

Some properties of the κ, λ - *neighborhood* can be used to improve the runtime and reduce memory requirements. Due to the definition of the κ, λ - *neighborhood* of two structures s_1 and s_2, there exist combinations of κ, λ distance pairs which do not contribute to any solution. For example, there is no κ, λ-neighbor with $\kappa + \lambda < d_{\text{BP}}(s_1, s_2)$. An increase or decrease of the base pair distance of a structure s to one of the reference structures implicitly changes the base pair distance to the other reference. In particular, if $d_{\text{BP}}(s_1, s_2) =$ even (resp. odd), then $\kappa + \lambda =$ even (resp. odd). This checkerboard-like pattern of the κ, λ - *neighborhood* roughly halves the number of entries in the extended dimensions κ and λ actually needed for the calculations. Furthermore, the maximum distance d_{max} to any reference structure in any substructure $s[i, j]$ of length $m = j - i + 1$ is constrained to $d_{\text{max}} < m$. These observations introduce sparsity in the dynamic programming matrices. Hence, two-dimensional matrices $F, C, M, \hat{M}$ with lists of triples, containing energy E, distance κ and distance λ at each matrix entry can be used. By iterating over the list instead of all κ, λ combinations, impossible structure formations are avoided.

Further runtime improvements can be obtained through parallelization by noting that all entries of the matrices F, C, M and $\hat{M}$ with $j - i =$ const. can be computed concurrently if the matrices are filled in diagonal order, see [FHS00].

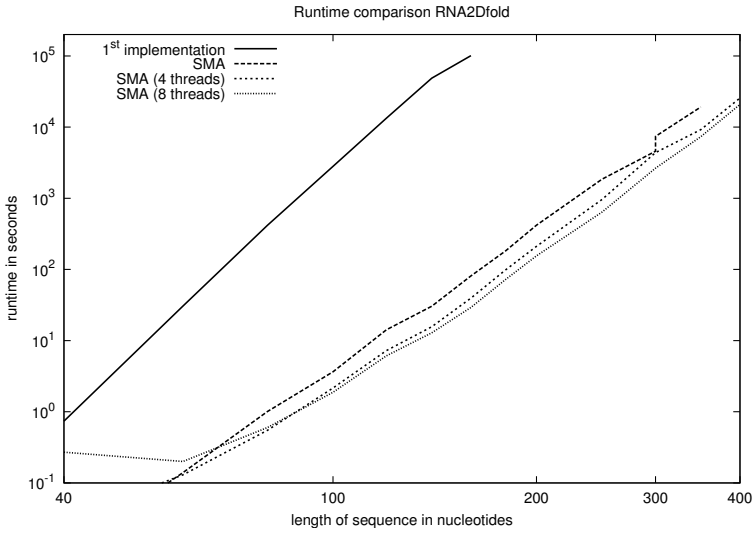


Figure 1: Runtimes of MFE calculation for the complete κ, λ - *neighborhood* . Timings are given for naïve approach (1^{st} implementation) and the sparse matrix approach (SMA) using 1, 4 and 8 threads on a dual quad-core Intel® Xeon® E5450 @3.00GHz with 32GB RAM. Runtimes are means of 15 random sequences. Reference structures used were the MFE structure and the open chain. With 8 processor cores, a sequence of 400 nt can be processed in about $5.8h$.

3 Results

3.1 Implementation

The partition function as well as the MFE version of the κ, λ - *neighborhood* algorithm was implemented in `ISO C` and will be available as a stand-alone program `RNA2Dfold` in one of the next releases of the Vienna RNA Package. A release candidate is available from `http://www.tbi.univie.ac.at/~ronny/RNA/`. The implementation provides most of the command line options of `RNAfold` such as different dangling end models and temperature. Given an RNA sequence $\mathcal{S}$ and two reference structures s_1 and s_2, `RNA2Dfold` computes for each κ, λ - *neighborhood* the MFE structure $s_{\text{opt}}^{\kappa,\lambda}$ and its free energy, the probabilities $P(S^{\kappa,\lambda})$, $P(s_{opt}^{\kappa,\lambda} \in S^{\kappa.\lambda})$, the probability of $s_{opt}^{\kappa,\lambda}$ in the complete ensemble and the Gibbs free energy $\Delta G^{\kappa,\lambda}$. The maximum values d_1 and d_2 with $\kappa \leq d_1$ and $\lambda \leq d_2$ can be specified by the user.

For parallelization we used *OpenMP* which allows efficient use of modern multi-core systems while requiring only small changes to the serial version of the source code. The performance gain from exploiting sparsity as well as parallelization is demonstrated in Fig. 1. The resulting speedups for 4 and 8 cores were 2.0 and 2.9, respectively. On modern multi-core systems `RNA2Dfold` can easily compute the MFE structures and partition functions for all κ, λ - *neighborhoods* for RNA sequences up to about 400 nt. This length range covers functional RNAs such as riboswitches and viroids.

3.2 2D Projection of the energy landscape

The probability densities and partition functions calculated by `RNA2Dfold` can be used for several secondary structure space analysis. One of the possible applications of the κ, λ - *neighborhood* algorithm is the prediction of metastable structure states and the detection of bi-stable RNA switches. Typically, the MFE structure is used as the first reference structure s_1. A meta-stable state, suitable as second reference structure s_2, can be obtained e.g. from a first run of `RNA2Dfold` using the open chain as second reference, and selecting s_2 from the $s_{\text{opt}}^{\kappa,\lambda}$. We note that this provides an alternative to the `paRNAss` approach for detecting RNA switches that avoids sampling errors. Computing the κ, λ - *neighborhood* of s_1 and s_2 and plotting the MFE values, probability densities and/or the Gibbs free energy of the partitioned landscape as a two dimensional height map reveals a qualitative picture of the roughness of the landscape. In the examples of Fig. 2 both RNAs can be clearly recognized as bi-stable switches. Molecules with more than 2 long-lived meta-stable states should exhibit additional minima in the interior of the height map. Furthermore, the height map yields a lower bound on the energy barrier between s_1 and s_2, and indicates the difficulty of refolding from s_1 to s_2 and vice versa.

3.3 A heuristic for finding non direct refolding paths

Most existing approaches utilize heuristics that consider only direct (minimal length) refolding paths between two states [MH98, FHMS$^+$01]. Since direct paths allow no detours, potentially stabilizing base pairs which are not in either of the two ground states cannot be formed. This can lead to intermediate structures with energetically unfavorable loop motifs and thus unnecessarily high energy barriers. In contrast, a refolding path with guaranteed minimal barrier can be obtained from the `barriers` program [FHSW02]. Since the approach is based of exhaustive enumeration of the energy landscape, it is limited to short sequences, typically less than a 100 nt.

The κ, λ - *neighborhood* can be used as base for various heuristics estimating the refolding path and energy barrier. Note however that taking the representatives $s_{\text{opt}}^{\kappa,\lambda}$ from a series of adjacent κ, λ-neighbors does usually not yield a continuous path of adjacent structures. Nevertheless, the height map already provides a lower bound for the height of the transistion state and therefore for the energy barrier too. Direct path heuristics perform poorly when the two structures are far apart. Therefore, a natural extension is to construct an intermediate structure s_m, termed *mesh-point*, thus splitting the path construction problem between the two reference structures s_1 to s_2 into two path constructions from s_1 to s_m and from s_m to s_2. The problem is of course to find suitable mesh points, and the κ, λ - *neighborhoods* turn out to be an excellent starting point for this. The `Pathfinder` algorithm given below (3.1) connects such mesh points using the direct path heuristic from [FHMS$^+$01]. The method produces indirect paths, since the mesh points need not lie on a shortest path between s_1 and s_2.

After computing the κ, λ - *neighborhood* of the start (s_1) and target (s_2) structure, we test

Algorithm 3.1 Pseudo-code of the Pathfinder(s_a, s_b, iter) algorithm where s_a is the start structure, s_b is the stop structure, iter is the maximal number of iterations

```
MeshpointHeap ← ∅ /* initialize min-order mesh-point heap */
bestpath ← DirectPath(s_a, s_b) /* get best refolding path so far */
while Meshpoints available ∧ |MeshpointHeap| < m do
    s ← Meshpoint structure /* sample a mesh point structure */
    path ← DirectPath(s_a, s) + DirectPath(s, s_b)
    if Barrier(path) < Barrier(bestpath) then
        insert(MeshpointHeap, (s, path, Barrier(path)))
    end if
end while
if iter > 0 then
    for 0 . . . m do
        (s, path) ← pop(MeshpointHeap)
        path ← Pathfinder(s_a, s, iter − 1) + Pathfinder(s, s_b, iter − 1)
        if Barrier(path) < Barrier(bestpath) then
            bestpath ← path
        end if
    end for
else
    (s, path) ← pop(MeshpointHeap)
    bestpath ← path
end if
return bestpath
```

as mesh-points all MFE structures $s_{\text{opt}}^{\kappa,\lambda}$ where $\kappa+\lambda \leq \gamma$ with a constant γ. This constraint limits the maximal deviation from a direct path and allows an adjustable exploration of the underlying energy landscape. Clearly, it is possible to recursively subdivide the problem further if required (see Pseudocode). With this simple approach the `Pathfinder` algorithm is able to find refolding paths with energy barriers very close or identical to those of an exhaustive search using the `barriers` program [FHSW02]. Results for an artificially designed RNA switch of 45 nt length revealed a barrier height of 10.7 kcal/mol (see Fig. 2A) which is the same as found with a barrier tree analysis. In contrast to that, a direct path generated according to [FHMS$^+$01] predicts an energy barrier of 13.33 kcal/mol. As mentioned before, the heightmap already provides a lower bound for the energy barrier. Here, direct paths are bounded by at least 13.3 kcal/mol, while indirect paths are bounded by 10.0 kcal/mol. Refolding between the aptamer- and non-aptamer fold of the *add*-riboswitch [RLGM07] (Fig. 2B) also shows the same energy barrier of 6.77 kcal/mol for both, Pathfinder and barrier tree analysis, while a direct path exhibits 7.28 kcal/mol.

4 Conclusion

We introduced a method for a unique partitioning of the RNA secondary structure space, in which structures are lumped together according to their base pair distances to two reference structures. In effect, this provides a 2D projection of the high-dimensional folding space. To overcome the high time complexity of $\mathcal{O}(n^7)$ our implementation exploits the sparseness of the dynamic programming matrices as well as *OpenMP* parallelization. The

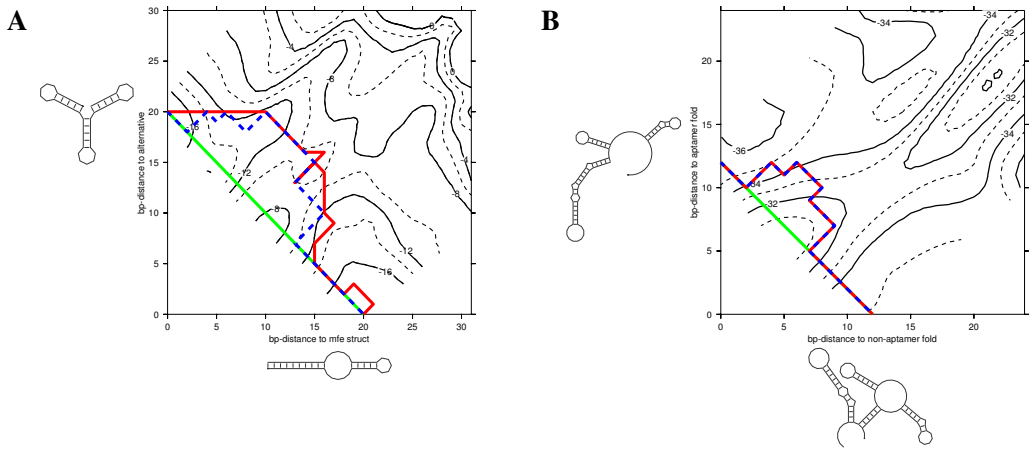


Figure 2: Gibbs free energy height map of all κ, λ - *neighborhoods* and projection of the refolding paths generated by `barriers` (red line) and the `Pathfinder` (blue dashed line) without recursive refinement. Mesh-points are taken from the κ, λ - *neighborhood* of both reference structures. **A**: MFE- and alternative structure of an artificial RNA switch with sequence `GGGCGCGGUUCGCCCUCCGCUAAAUGCGGAAGAUAAAUUGUGUCU` and meta stable structure conformations `(((((.....)))))((((((.....))))))((((((.....))))))` *(MFE structure)* and `((((((((((.....((((((.....))))))..... ))))))))))` *(alternative structure)*. **B**: Aptamer- and non-aptamer fold of an *add*-riboswitch [RLGM07]. In contrast to direct paths (straight diagonal green line), the `Pathfinder` solution is as good as the (optimal) solution generated by `barriers` in both cases. In **B**, identical refolding paths are obtained for `Pathfinder` and `barriers` analysis.

resulting program is fast enough to treat RNA molecules up to 400 nt which covers most biologically interesting cases such as riboswitches and viroids.

The κ, λ - *neighborhoods* provide both a qualitative picture of the energy landscape, as well as a convenient starting point for more detailed exploration. As an example we show that it can be used to suggest excellent intermediate nodes for the construction of refolding paths, resulting in a fast heuristic that often gives optimal results. Such heuristics are needed e.g. for kinetic folding strategies like `Kinwalker` [GFW$^+$08]. Furthermore, the height maps could provide the starting point for methods that recognize RNA switches or for coarse grained folding simulations.

Acknowledgments

This work has been funded, in part, by the Austrian GEN-AU projects “bioinformatics integration network III” and “non coding RNA”.

References

[CHS96] J. Cupal, I.L. Hofacker, and P.F. Stadler. Dynamic programming algorithm for the density of states of RNA secondary structures. *Computer Science and Biology*, 96(96):184–186, 1996.

[FHMS+01] C. Flamm, I.L. Hofacker, S. Maurer-Stroh, F. Stadler, and M. Zehl. Design of multistable RNA molecules. *RNA*, 7:254–265, 2001.

[FHS00] Martin Fekete, Ivo L. Hofacker, and Peter F. Stadler. Prediction of RNA base pairing probabilities using massively parallel computers. *J. Comp. Biol.*, 7:171–182, 2000.

[FHSW02] Christoph Flamm, Ivo L. Hofacker, Peter F. Stadler, and Michael T. Wolfinger. Barrier Trees of Degenerate Landscapes. *Z. Phys. Chem.*, 216:155–173, 2002.

[FMC07] E. Freyhult, V. Moulton, and P. Clote. Boltzmann probability of RNA structural neighbors and riboswitch detection. *Bioinformatics*, 23(16):2054, 2007.

[GFW+08] M. Geis, C. Flamm, M.T. Wolfinger, A. Tanzer, I.L. Hofacker, M. Middendorf, C. Mandl, P.F. Stadler, and C. Thurner. Folding kinetics of large RNAs. *Journal of Molecular Biology*, 379(1):160–173, 2008.

[GHR99] R. Giegerich, D. Haase, and M. Rehmsmeier. Prediction and visualization of structural switches in RNA. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 126, 1999.

[HFS+94] I.L. Hofacker, W. Fontana, P.F. Stadler, S. Bonhoeffer, M Tacker, and P. Schuster. Fast Folding and Comparison of RNA Secondary Structures. *Monatshefte f. Chemie*, 125:167–188, 1994.

[IS00] H Isambert and E D Siggia. Modeling RNA folding paths with pseudoknots: application to hepatitis delta virus ribozyme. *Proc Natl Acad Sci U S A*, 97(12):6515–20, Jun 2000.

[McC90] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990.

[MH98] S.R. Morgan and P.G. Higgs. Barrier heights between ground states in a model of RNA secondary structure. *Journal of Physics A-Mathematical and General*, 31(14):3153–3170, 1998.

[MSZT99] D.H. Mathews, J. Sabina, M. Zuker, and D.H. Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *Journal of Molecular Biology*, 288(5):911–940, 1999.

[RLGM07] R. Rieder, K. Lang, D. Graber, and R. Micura. Ligand-induced folding of the adenosine deaminase A-riboswitch and implications on riboswitch translational control. *Chembiochem*, 8(8), 2007.

[WFHS99] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–165, February 1999.

[Zuk89] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244(4900):48–52, April 1989.

10

2D MEETS 4G: G-QUADRUPLEXES IN RNA SECONDARY STRUCTURE PREDICTION

Ronny Lorenz, Stephan H. Bernhart, Jing Qin, Christian Höner zu Siederdissen, Andrea Tanzer, Fabian Amman, Ivo L. Hofacker, and Peter F. Stadler.
2D meets 4G: G-Quadruplexes in RNA Secondary Structure Prediction.
in *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2013 **Volume 10**, Issue: 4, pages 832-844.
doi: 10.1109/TCBB.2013.7

All authors contributed to the design of the study, RL designed and implemented the algorithms, and wrote substantial parts of the paper.

832 IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 10, NO. 4, JULY/AUGUST 2013

2D Meets 4G: G-Quadruplexes in RNA Secondary Structure Prediction

Ronny Lorenz, Stephan H. Bernhart, Jing Qin, Christian Höner zu Siederdissen, Andrea Tanzer, Fabian Amman, Ivo L. Hofacker, and Peter F. Stadler

Abstract—G-quadruplexes are abundant locally stable structural elements in nucleic acids. The combinatorial theory of RNA structures and the dynamic programming algorithms for RNA secondary structure prediction are extended here to incorporate G-quadruplexes using a simple but plausible energy model. With preliminary energy parameters, we find that the overwhelming majority of putative quadruplex-forming sequences in the human genome are likely to fold into canonical secondary structures instead. Stable G-quadruplexes are strongly enriched, however, in the 5′UTR of protein coding mRNAs.

Index Terms—Dynamic programming, RNA folding, `ViennaRNA Package`

1 Introduction

Guanosine-rich nucleic acid sequences readily fold into four-stranded structures known as G-quadruplexes. DNA quadruplexes are, for instance, an important component of human telomeres [1], they appear to be strongly overrepresented in the promoter regions of diverse organisms, and they can associate with a variety of small molecule ligands, see [2], [3] for recent reviews. SNPs in G-quadruplexes, finally, have been implicated as a source of variation of gene expression levels [4]. RNA quadruplexes have also been implicated in regulatory functions. Conserved G-quadruplex structures within the 5′UTR of the human TRF2 mRNA [5] and eukaryotic MT3 matrix metalloproteinases, for example, repress translation [6]. Another well-studied example is the interaction of the RGG box-domain fragile X mental retardation protein (FMRP) to a G-quartet-forming region in the human semaphorin 3F (S3F) mRNA [7], [8]. A recent review of G-quadruplex-based translation regulation is [9]. A functional RNA G-quadruplex in the 3′UTR was recently described as a translational repressor of the proto-oncogene PIM1 [10]. A mechanistic study of this effect, which seems to be widely used in the cell [11], [12] can be found, e.g., in [13]. Most recently, G-quadruplexes were also reported in several long non-coding RNAs [14]. G-quadruplexes are potentially of functional importance in the 100 to 9,000 nt G-rich telomeric repeat-containing RNAs (TERRAs) [15].

Quadruplex structures consist of stacked associations of G-quartets, i.e., planar assemblies of four Hoogsteen-bound guanines. As in the case of base pairing, the stability of quadruplexes is derived from π-orbital interactions among stacked quartets. The centrally located cations that are coordinated by the quartets also have a major influence on the stability of quadruplex structures.

DNA quadruplexes are structurally heterogeneous: Depending on the glycosidic bond angles, there are 16 possible structures and further combinatorial complexity is introduced by the relative orientations of the backbone along the four edges of the stack [17]. RNA quadruplexes, in contrast, appear to be structurally monomorphic forming parallel-stranded conformations (Fig. 1, left) independently of surrounding conditions, i.e., different cations and RNA concentration [18]. Here, we restrict ourselves to the simpler case of RNA quadruplexes.

From a bioinformatics perspective, G-quadruplex structures have been investigated mostly as genomic sequence motifs. The `G4P Calculator` searches for four adjacent runs of at least three Gs. With its help, a correlation of putative quadruplex forming sequences and certain functional classes of genes was detected [19]. Similarly,

- *R. Lorenz, C. Höner zu Siederdissen, and F. Amman are with the Department of Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria. E-mail: {ronny, choener, fabian}@tbi.univie.ac.at.*
- *S.H. Bernhart is with the Bioinformatics Group, Department of Computer Science and the Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany. E-mail: berni@bioinf.uni-leipzig.de.*
- *J. Qin is with the Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany, and the Interdisciplinary Center for Bioinformatics, University of Leipzig. E-mail: qin@bioinf.uni-leipzig.de.*
- *A. Tanzer is with the Department of Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria, and the Center for Genomic Regulation (CRG), Dr. Aiguader, 88, 08003 Barcelona, Spain. E-mail: at@tbi.univie.ac.at.*
- *I.L. Hofacker is with the Department of Theoretical Chemistry, University of Vienna, Währingerstrasse 17, A-1090 Vienna, Austria, and the Center for RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, Frederiksberg C, Denmark. E-mail: ivo@tbi.univie.ac.at.*
- *P.F. Stadler is with the Bioinformatics Group, Department of Computer Science and the Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstrasse 16-18, D-04107 Leipzig, Germany; the Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany, the Center for RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, Frederiksberg C, Denmark; the Fraunhofer Institute for Cell Therapy and Immunology, Perlickstrasse 1, D-04103 Leipzig, Germany; and the Department of Theoretical Chemistry of the University of Vienna, and the Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501. E-mail: studla@bioinf.uni-leipzig.de.*

Manuscript received 7 Sept. 2012; revised 7 Dec. 2012; accepted 15 Jan. 2013; published online 22 Jan. 2013.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBBSI-2012-09-0243.
Digital Object Identifier no. 10.1109/TCBB.2013.7.

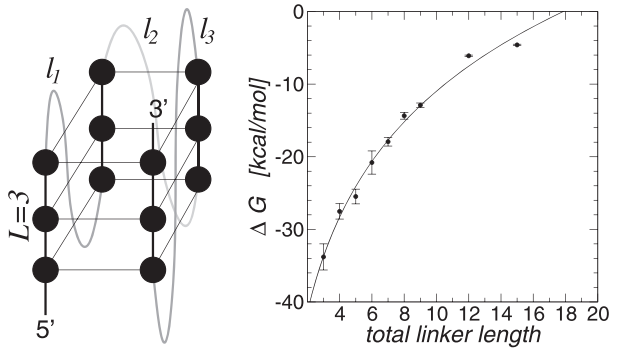


Fig. 1. RNA quadruplexes form parallel arrangements with $L = 2...5$ layers. Folding energies for $L = 3$ depend mostly on the total length ℓ of the linker sequences: The data from [16] fit well to an energy model of the form $\Delta G = a + b \ln \ell$ (solid line). Spearman's rank correlation for the fit is $\rho \approx 0.972$, mean square error 4.2 kcal/mol.

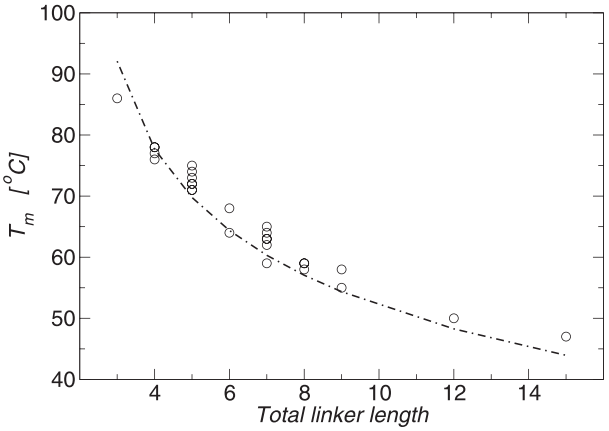


Fig. 2. Melting temperatures as a function of the total linker length for three-layer G-quadruplexes. Experimentally determined melting temperatures from [16] (circles) are compared with the predictions made with the G-quadruplex aware version of `RNAfold` (dash-dotted line).

`quadparser` [20] recognizes the pattern (1) below. It was used, e.g., in [21] to demonstrate the enrichment of quadruplexes in transcriptional regulatory regions. A substantial conservation of such sequence patterns in mammalian promoter regions is reported in [22]. The web service `QGRS Mapper` uses a similar pattern and implements a heuristic scoring system [23], see also [24] for a review. A Bayesian prediction framework based on Gaussian process regression was recently introduced to predict melting temperatures of quadruplex sequences [25].

The formation of RNA quadruplexes necessarily competes with the formation of canonical secondary structures. Hence, they cannot be fully understood in isolation. In this contribution, which is a revised and expanded version of [26], we describe how G-quadruplex structures can be incorporated into RNA secondary structure prediction algorithms and survey the genomic distribution of stable G-quadruplex structures.

2 Energy Model

Thermodynamic parameters for RNA quadruplexes can be derived from measurements of UV absorption as a function of temperature [27], analogous to melting curves of secondary structures. While the stability of DNA G-quadruplexes strongly depends on the arrangement of loops [28], [29], this does not appear to be the case for RNA. RNA not only forms mostly parallel-stranded stacks for G-quartets but their stability also exhibits a rather simple dependence of the loop length [16]. In further contrast to DNA [30], they appear to be less dependent on the nucleotide sequence itself.

A G-quadruplex with $2 \le L \le 5$ stacked G-quartets and three linkers of length $l_1, l_2, l_3 \ge 1$ has the form

$$G_L N_{l_1} G_L N_{l_2} G_L N_{l_3} G_L. \tag{1}$$

It is commonly assumed that $1 \le l_i \le 7$ [25], although in vitro data for DNA suggest that longer linkers are possible [31]. For $L = 2$, the existence of quadruplexes with $1 \le \ell_i \le 2$ was reported [32]. For $L = 3$, detailed thermodynamic data are available only for the 27 cases with $1 \le l_1, l_2, l_3 \le 3$ and for some longer symmetric linkers $l_1 = l_2 = l_3$ [16], see Fig. 1b.

To our knowledge, no comprehensive data are available for $L \ge 4$. It appears reasonable to assume that the stacking energies are additive. The energetic effect of the linkers appears to be well described in terms of the total linker length ℓ [16]. As shown in Fig. 1b, the free energy depends approximately logarithmically on ℓ. In this contribution, we are mostly concerned with the algorithmic issues of including G-quadruplexes into thermodynamic folding programs. In particular, we ignore here the strong dependence of quadruplex stability on the potassium concentration, see, e.g., [33]. We thus resort to the simplified temperature-dependent energy function

$$\begin{aligned} E[L, \ell, T] &= a(T)(L-1) + b(T) \ln(\ell - 2) \\ a(T) &= H_a + T S_a \\ b(T) &= H_b + T S_b, \end{aligned} \tag{2}$$

if the pattern (1) is matched and $E = \infty$ otherwise. The parameters are fitted to match the free energies in Fig. 1, i.e., $a(37\,^\circ\mathrm{C}) = -18$ kcal/mol, and $b(37\,^\circ\mathrm{C}) = 12$ kcal/mol. The enthalpy $H_a = -119.34$ kcal/mol is estimated from the melting experiments reported in [16]. Furthermore, we assume that the contribution of the linkers is purely entropic, i.e., $H_b = 0$ kcal/mol. To test the consistency of the parameters, we compare in Fig. 2 the measured melting temperatures of 27 sequences reported in [16] with melting temperatures predicted by `RNAfold`, i.e., the temperature at which the probability of the G-quadruplex equals 0.5. We observe an excellent agreement and conclude that the simple energy model here is adequate. Not surprisingly, reducing the estimates for the quadruplex stacking energy a and the entropic penalty b for the linker sequences favors the prediction of quadruplexes over alternative secondary structures. A quantitative sensitivity analysis can be found in the electronic supplement.

G-quadruplex structures can be located within loops of more complex secondary structures. Fig. 3, for instance, shows the $L = 2$, $l_1 = l_2 = l_3 = 2$ quadruplex in a hairpin of the semaphorin 3F RNA [7]. It seems natural to treat G-quadruplexes inside multiloops similar to their branching helices: Each unpaired base incurs a penalty a and each

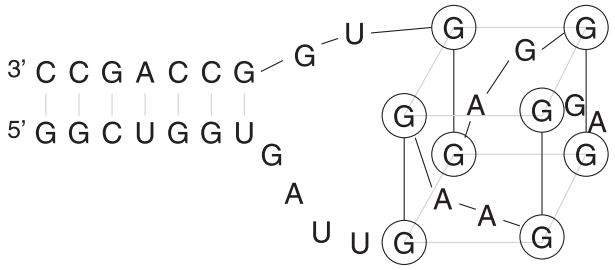


Fig. 3. Structure of the G-quadruplex in a hairpin of human semaphorin 3F RNA that binds the RGG box domain of fragile X mental retardation protein (FMRP). Redrawn based on [7].

G-quadruplex within a loop is associated with an additional "loop strain" b. For the interior-loop case in Fig. 3, only stabilizing mismatch contributions of the enclosing pair and a penalty for the stretches of unpaired bases are used. Sterical considerations for this case suggest that a G-quadruplex is flanked by a stretch of at least three unpaired nucleotides or has at least one unpaired nucleotide on either side.

3 COMBINATORICS

RNA secondary structures consist of mutually noncrossing base pairs and unpaired positions. Thus, they can be represented as strings composed of matching parentheses (base pairs) and dots. This "dot-parenthesis" notation is used by the `ViennaRNA Package` [34]. G-quadruplexes constitute an extra type of structural element. The semaphorin hairpin, Fig. 3, can therefore be written as

```
GGCUGGUGAUUGGAAGGGAGGGAGGUGGCCAGCC
((((((((....++..++..++..++..))))))))
```
(3)

using the symbol $+$ to mark the bases involved in G-quartets. This string representation uniquely identifies all G-quartets since the first run of $+$ symbols determines L for the 5′-most quadruplex, thus determining the next three G-stacks that are separated by at least one "`.`" and must have the same length. It follows immediately that the number of secondary structures with G-quadruplexes is still smaller than 4^n, an observation that is important, e.g., for the evolvability of RNAs [35].

Combinatorial model: G-structure. A more detailed estimate can be obtained from a detailed combinatorial model of secondary structures with G-quadruplexes. This model also forms the basis for folding algorithms outlined in the following section.

For simplicity, any quadruplex with $L \geq 2$ stacked G-quartets and three linkers of length l_1, l_2, and $l_3 \geq 1$ are allowed in any context with the following exception: If (i, j) is a base pair that encloses a sequence of quadruplexes, then at least one of three conditions is satisfied: 1) $i+1$ and $j-1$ are both unpaired; 2) $i+1$, $i+2$, and $i+3$ are unpaired or 3) $j-3$, $j-2$, and $j-1$ are unpaired. A secondary structure of length n is modeled as a noncrossing partial matching (matching with isolated vertices) on n vertices such that each base pair is of length at least 3. A G-structure is a secondary structure with quadruplexes as defined above.

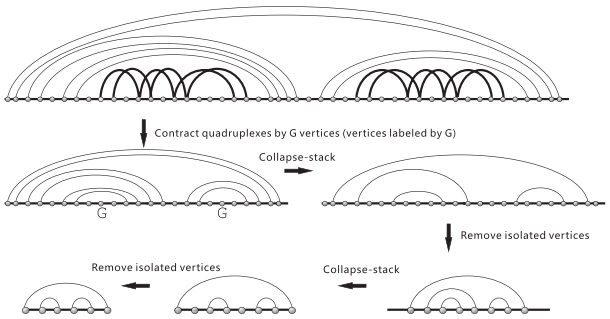


Fig. 4. Extraction of a shape from a G-structure. The shape of a given G-structure is obtained by (P1) contracting each G-quadruplex into a single vertex labeled "G," and (P2) iteratively collapsing each stack to a single arc and removing any isolated vertices from the resulting structure. Procedure (P2) is applied when neither isolated vertices nor stacks with more than one arc are left.

We say that a G-structure is τ-canonical if all stacks consist of at least τ arcs.

Generating functions. The enumeration of G-structures is based on the notion of shapes, that is, noncrossing matchings in which each stack consists of exactly one arc. The shape of an arbitrary G-structure s is obtained by 1) contracting each G-quadruplex to a single vertex labeled "G," and 2) iteratively collapsing each stack to a single arc and then removing any unpaired bases from the resulting structure as illustrated in Fig. 4. It corresponds to the coarsest shape abstraction in the sense of [36].

The central objects of interest are the noncrossing shapes over $2n$ vertices with t "1-arcs" that link adjacent points, i.e., t hairpin loops. Their number $\mathbf{s}_{n,t}$ can be expressed in terms of the number $\mathbf{m}_t$ of noncrossing matchings with t arcs, where $\mathbf{m}_{2t}$ is the tth Catalan number. The well-known generating function for these matchings is $\mathbf{M}(u) = (1 - \sqrt{1-4u})/(2u)$, yielding

$$\mathbf{S}(u,e) = \sum_{n,t} \mathbf{s}_{n,t} u^n \, e^t = \frac{1+u}{1+2u-ue} \mathbf{M}\left(\frac{u(1+u)}{(1+2u-ue)^2}\right) \tag{4}$$

for the shapes. For each shape, we can count the number of all possible G-structures with this shape by "inflating" the shape in a controlled manner Fig. 5.

To obtain an arbitrary τ-canonical G-structure with a given shape, we first insert at most one red isolated vertex into each interval with the exception of the 1-arcs. The allowed insertion points are marked in the first diagram in Fig. 5. Next, one green vertex is inserted within each of the 1-arc. The next step is to expand each arc into a stack by adding $t \geq 0$ parallel arcs. If $t \geq 1$, a blue vertex is inserted immediately before, immediately after, or on both sides of each inserted arc to separate the original arc and the newly inserted ones from each other. Finally, each arc is inflated to a stack of at least τ arcs. Each of the insertion steps is associated with determining for a known number of possible positions whether or not a colored vertex is inserted. Thus, each step corresponds to simple substitution in the generating function.

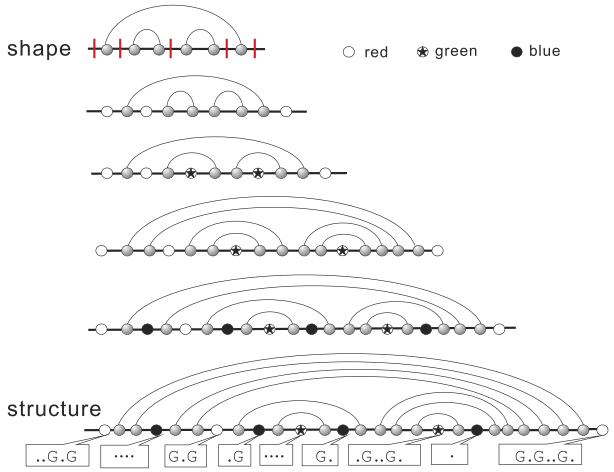


Fig. 5. From a shape to a τ-canonical G-structure. Each G-quadruplex is shown as a vertex labeled by G. Three different symbols represent the red, green, and blue vertices, respectively. In the online version, the figure will appear in color.

The colored vertices correspond to substructures that entirely consist of G-quadruplexes and unpaired bases. The colors encode additional restrictions. The basic building block in each colored vertices is a $\mathcal{P}_0$-structure that entirely consist of G-quadruplexes and unpaired bases such that 1) its leftmost and rightmost bases are contained in some quadruplex (could be the same) and 2) any two consecutive quadruplexes, if exist, are separated by at least one unpaired base. Red vertices are replaced by either 1) a sequence of at least one unpaired base, 2) a $\mathcal{P}_0$ structure with two possibly empty of unpaired bases attached at both ends. Green vertices become either 1) a sequence of at least three unpaired bases or 2) a $\mathcal{P}_0$ with nonempty sequences of unpaired bases attached to both ends, or 3) a $\mathcal{P}_0$ structure a sequence of at least three unpaired bases attached at exactly one end. Blue vertices, finally, are replaced by either 1) a nonempty sequence of unpaired bases, or 2) a $\mathcal{P}_0$ with possibly empty sequences of unpaired bases attached at both ends. It suffices, therefore, to derive the generating function $\mathbf{P}_0$ for $\mathcal{P}_0$. It is straightforward to express $\mathbf{P}_0$ in terms of the generating function $\mathbf{Q}$ for the individual G-quadruplexes.

Making these ideas precise, we obtain

Theorem 1. *The generating function of τ-canonical G-structures is*

$$\begin{aligned}\mathbf{G}^{\tau}(x) &= \sum_n \mathbf{g}_n^{\tau} x^n \\ &= \mathbf{P}_3(x)\mathbf{S}\left(\frac{x^{2\tau}\cdot \mathbf{P}_3^2(x)}{(1-x^2)-x^{2\tau}(2\mathbf{P}_1(x)+\mathbf{P}_1^2(x))}, \frac{\mathbf{P}_2(x)}{\mathbf{P}_3(x)}\right)\end{aligned} \tag{5}$$

with the auxiliary functions

$$\mathbf{Q}(x) = \frac{x^{11}}{(1-x^4)(1-x)^3} \qquad \mathbf{P}_0(x) = \frac{(1-x)\,\mathbf{Q}(x)}{1-x-x\mathbf{Q}(x)}$$

$$\mathbf{P}_1(x) = \frac{x}{1-x} + \frac{\mathbf{P}_0(x)}{(1-x)^2} \qquad \mathbf{P}_3(x) = \frac{1}{1-x} + \frac{\mathbf{P}_0(x)}{(1-x)^2}$$

$$\mathbf{P}_2(x) = \frac{x^3}{1-x} + \frac{x^2+2x^3-2x^4}{(1-x)^2}\mathbf{P}_0(x).$$

Asymptotic results for the sequence $\mathbf{g}_n^{\tau}$ can be obtained from an analysis of the singularities of $\mathbf{g}(x)$, see [37] and the online material[1] for details. For $\tau = 1, 2$, we obtain

$$\mathbf{g}_n^{\tau} \sim k_{\tau}\, n^{-3/2}\, (\rho_{\tau}^{-1})^n. \tag{6}$$

We used `Maple`, version 11, to obtain the numerical values $\rho_1^{-1} \approx 2.2903$ and $\rho_2^{-1} \approx 1.8643$. The same functional form is obtained for structures without G-quadruplex by simply setting $\mathbf{Q}(x) = 0$. This recovers earlier results $\hat{\rho}_1^{-1} \approx 2.2887$ and $\hat{\rho}_2^{-1} \approx 1.8489$ [38]. The shape spaces of RNAs with and without G-quadruplexes, thus, are quite similar. In particular, we can conclude that the inclusion of quadruplexes does not affect qualitative properties of the sequence-structure map such as the existence of extensive neutral networks, or the shape-space covering property [35].

4 RNA Folding Algorithms

Energy minimization. Dynamic programming algorithms for secondary structure prediction are based on a simple recursive decomposition: Any feasible structure on the interval $[i, j]$ has the first base either unpaired or paired with a position k satisfying $i < k \leq j$. The condition that base pairs do not cross implies that the intervals $[i+1, k-1]$ and $[k+1, j]$ form self-contained structures whose energies can be evaluated independent of each other. In conjunction with the standard energy model [39], which distinguishes hairpin loops, interior loops (including stacked base pairs), and multiloops, this leads to the recursions diagrammatically represented in Fig. 6 (ignoring the cases involving black blocks). This algorithmic approach was pioneered, e.g., in [40], [41] and is also used in the `ViennaRNA Package` [34].

G-quadruplexes form closed-structural elements on well-defined sequence intervals. Thus, they can be treated just like substructures enclosed by a base pair so that the additional ingredients in the folding algorithms are the energies G_{ij} (free energy of the most stable quadruplex so that the pattern (1) exactly matches the interval $[i, j]$) and the partition functions Z_{ij}^G (defined as the sum of the Boltzmann factors of all distinct quadruplexes on the interval $[i, j]$). As a consequence of (1), we have $G_{ij} < \infty$ and $Z_{ij}^G > 0$ only if $|j - i| < 4L_{\max} + \ell_{\max}$. All possible quadruplexes on the interval $[i, j]$ can be determined and evaluated in $\mathcal{O}(L_{\max}^2 \ell_{\max}^2)$ time so that these arrays can be precomputed in $\mathcal{O}(n(L_{\max} + \ell_{\max})L_{\max}^2 \ell_{\max}^2)$, i.e., in linear time.

The standard recursions for RNA secondary structure prediction can now be extended by extra terms for quadruplexes, see Fig. 6. The simplest strategy would be to add G-quadruplexes as an additional type of base-pair enclosed structures. This would amount to using standard interior loop parameters also for cases such as Fig. 3. Hence, we use the somewhat more elaborate grammar of Fig. 6, which introduces the quadruplexes in the form of additional cases into the multiloop decomposition. An advantage of this method is that one can use different parameter

1. For a complete formal proof, see http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/12-006/.

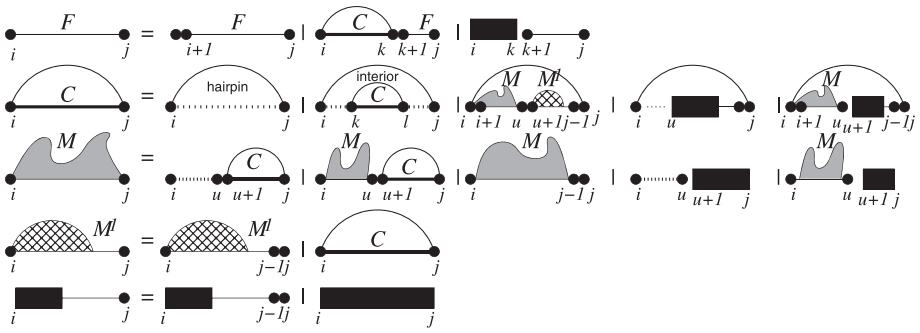


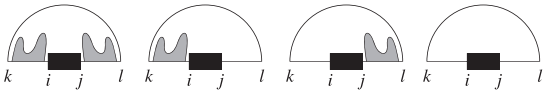
Fig. 6. Extension of recursions of the `ViennaRNA Package` to accommodate G-quadruplexes. This grammar treats G-quadruplexes with multiloop like energies also in an interior-loop-like context.

values to penalize the inclusion of quadruplexes and helical components into a multiloop. Clearly, the grammar is still unambiguous, i.e., every structure has an unique parse. Thus, it can be used directly to compute partition functions.

Base pairing probabilities. A straightforward generalization of McCaskill's algorithm can be used to compute the probabilities P_{ij} of all possible base pairs (i, j). The probability P^G_{ij} of finding a G-quadruplex delimited by positions i and j can then be written as

$$P^G_{ij} = \frac{Z_{1,i-1} Z^G_{ij} Z_{j+1,n}}{Z} + \sum_{\substack{k<i-1 \\ l>j+1}} P_{kl} \mathbb{P}\big\{\text{quadruplex}[i,j] \big| (k,l)\big\}. \quad (7)$$

The conditional probabilities $\mathbb{P}\{\cdots\}$ in turn are composed of the four individual cases depending on the placement of the components of the generalized multiloop enclosed by (k, l) relative to the interval $[i, j]$.



This decomposition translates to the recursion

$$\begin{aligned} &\mathbb{P}\{\text{quadruplex}[i,j] \big| (k,l)\} \\ &\quad = \frac{Z^M_{k+1,i-1} Z^G_{ij} Z^M_{j+1,l-1}}{Z^B_{kl}} + \frac{Z^M_{k+1,i-1} Z^G_{ij} \hat{b}^{l-j-1}}{Z^B_{kl}} \\ &\quad + \frac{\hat{b}^{i-k-1} Z^G_{ij} Z^M_{j+1,l-1}}{Z^B_{kl}} + \frac{\hat{b}^{i-k-1} Z^G_{ij} \hat{b}^{l-j-1}}{Z^B_{kl}}, \end{aligned} \quad (8)$$

where $\hat{b} = \exp(-b/RT)$. From the P^G_{ij}, it is straightforward to compute the probability of a particular quadruplex as

$$p([i, L, l_1, l_2, j]) = \frac{\exp(-E[L, \ell])}{Z^G_{ij}} P^G_{ij}, \quad (9)$$

where $l_3 = j - i + 1 - 4L - l_1 - l_2$. Summing up the probabilities of all quadruplexes that contain a particular contact $i' : j'$ of two guanosines in a layer finally yields the probability of the G:G contact $i' : j'$.

Fig. 7 shows an example of the graphical output of `RNAfold`. In the minimum energy case, we use a very simple modification of the standard layout [42] treating each quadruplex like a local hairpin structure, explicitly indicating the G-G pairs. Quadruplexes are shown in addition to the individual G-G pairs as shaded triangles in the base pair probability dot plots. From the base pairing probabilities, we also compute MEA [43] and centroid structures.

By definition, the centroid structure X minimizes the expected base pair distance to the other structures within the Boltzmann-weighted ensemble. In the absence of G-quadruplexes, X consists of all base pairs (i, j) with $p_{ij} > 0.5$. A certain ambiguity arises depending on whether X is interpreted as a list of base pairs that may contain incomplete quadruplexes, or whether quadruplexes are treated as units. Here, we insert a quadruplex if $P^G_{ij} > 0.5$, and represent it by the most stable quadruplex with endpoints i and j. The same representation is used for MEA structures where we extend the maximized expected accuracy to $\text{EA} = \sum_{(i,j)\in S} 2\gamma(P_{i,j} + P^G_{ij}) + \sum_i P^u_i$ with $P^u_i = 1 - \sum_j P_{ij} - \sum_{k\le i\le l} P^G_{kl}$, accordingly.

Scanning variants. For large RNA molecules with a length of more than, say, 500 nt it is often more useful to consider locally stable secondary structure elements than to compute globally optimal structures. The most straightforward approach to this end is to simply restrict the base pair span, i.e., to omit all base pairs with a span $j - i$ larger than some threshold S. This generates a global structure with local base pairs only. In the `ViennaRNA Package`, variations of this approach are used. In `RNALfold` [45], structures of local minimum free energy are computed. In essence, the energy of these structures can not be lowered by adding another base $5'$ or $3'$. The partition function

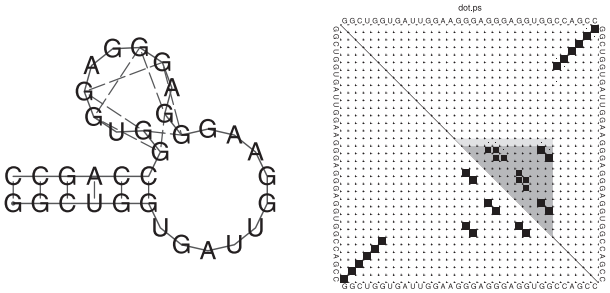


Fig. 7. Representation of minimum free energy structure (l.h.s.) and base pairing probability matrix (r.h.s.) of the semaphorin hairpin (see Fig. 3), respectively. The probabilities of the two possible G-quadruplex conformations are indicated by triangles with low probability in light and high probability in dark gray.

prediction in RNAplfold [46] conceptually uses a sliding window approach with step size 1 and averages the base pair probabilities over all windows in which a base pair can appear. Since G-quadruplexes are inherently local, it is straightforward to include them in the same way as in the global folding algorithms.

RNA dimers. The intermolecular interactions involved in the hybridization of two RNA molecules are governed by the same rules that determine the secondary structure of a single RNAs. Several distinct approaches are used depending on the types of RNA-RNA interactions that are to be included. The most straightforward solution is to concatenate the two RNA sequences (with lengths n_1 and n_2) and to use the same dynamic programming recursions as for a single sequence. Only the energy contributions are modified for all loops that involve the cut between the two molecules [47]. The result is a common secondary structure that may contain arbitrary many intermolecular interactions and can be computed in $\mathcal{O}((n_1+n_2)^3)$ time. Intramolecular G-quadruplexes are straightforward to include. However, separate energy contributions have to be assigned to intermolecular G-quadruplexes; these may depend on the linker that contains the cut.

RNA-RNA hybrids with a single interaction region can be handled by decomposing the interaction energy into three components: The energies used to make the interaction region unpaired in each of the two partners and the energy gained from the hybridization itself. The unpairing energies can be computed efficiently by a modified version of McCaskill's algorithm implemented in RNAup [48] and intaRNA [49]. The computation of the hybridization energies is analogous to RNAhybrid [50] for regular base pairs. It can be easily extended to intermolecular G-quadruplexes by direct enumeration of quadruplexes that can be formed by combining sequence windows from both partners. Such "2-component" quadruplexes are of practical interest because it has recently been shown that the binding of G-rich small RNAs to G-rich regions in reporter mRNAs can lead to the formation of an intermolecular RNA G-quadruplex that inhibits translation in living cells [51].

A much more general, but also computationally more expensive approach to treat RNA-RNA interactions is the RNA-RNA interaction prediction (RIP) model [53], [54], which allows arbitrary secondary structures of both interaction partners and any crossing-free set of intermolecular base pairs that avoids the so-called tangle configurations. The most convenient starting point to include quadruplexes into this type of algorithms is the unambiguous grammars for the RIP problem that are required to compute the partition function [55], [56], [57]. First, we observe that quadruplexes that are entirely formed by one of the two interacting sequences (i.e., intramolecular structures) are subsumed by the terms for noninteracting secondary structures, which we have seen how to compute in Fig. 6. Intermolecular quadruplexes are again local in each of both sequences and conceptually analogous to intermolecular base pairing interactions. Thus, they can be included as an additional alternative in all productions that expose an intermolecular base pair (symbol K in [56, Fig. 7]). In the modified grammar of [57, Fig. 4], it is most natural to include the intermolecular quadruplexes into the decomposition of the intermolecular hybrids (symbol K in [57], symbol E in Fig. 9), where quadruplexes again appear as additional alternatives to the single intermolecular base pairs. As in the case of single-sequence folding, the possible intermolecular quadruplexes can be precomputed so that the computational complexity of these algorithms stays $\mathcal{O}(n^4)$ in memory and $\mathcal{O}(n^6)$ in time. A simple version of RIP grammar with pseudoknots is sketched in the appendix. We note, however, that the inclusion of intermolecular quadruplexes in this way leads to several new types of loops partially delimited by quadruplexes for which parameters would need to be estimated. We have not attempted to provide an implementation of a RIP algorithm with quadruplexes.

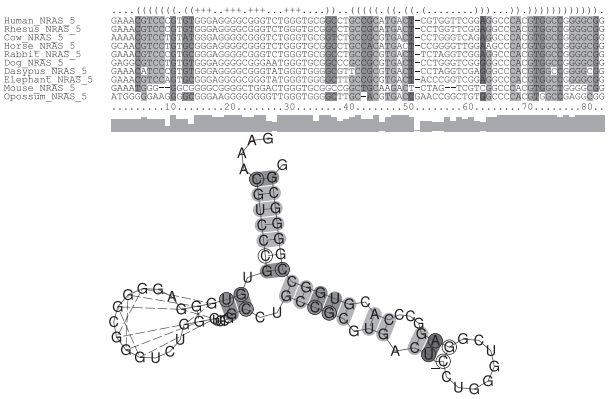


Fig. 8. Consensus structure of the $5'$-most part of the $5'$UTR of the NRAS mRNA, exhibiting a conserved G-quadruplex with $L=3$ that modulates translation of the *NRAS* proto-oncogene [44]. Shades of gray indicate the number, from light gray (1) to dark gray (3), of different base pair types in the alignment. Substitutions in stem regions are indicated by circles in the secondary structure drawing.

Consensus structures. Can be readily obtained for a given multiple sequence alignment. The idea is to apply the dynamic programming recursions to alignment columns. The energy contributions are determined as the average of the corresponding contributions to the individual sequences [58]. In addition, small contributions are added to favor pairs of columns with consistent (e.g., $\mathrm{GC} \rightarrow \mathrm{GU}$) and compensatory mutations ($\mathrm{AU} \rightarrow \mathrm{GC}$) because these provide direct evidence for selection acting to preserve base pairing. Similarly, penalties are added if one or a few sequences cannot form a base pair. We refer to [59] for details of the scoring model implemented in RNAalifold. Here, we extend it by a simple system of penalties for mutations that disrupt quadruplexes. Non-G nucleotides incur an energy E' in the outer layers of the quadruplex and $2E'$ in the inner layers as they affect one or two stacking interactions, respectively. An example of a consensus structure prediction is shown in Fig. 8.

Implementation details. The implementation of G-quadruplex folding in RNAfold, RNAalifold, RNAcofold, and RNALfold essentially follows the extended grammar shown in Fig. 6, distinguishing the energy contribution of G-quadruplexes and their adjacent unpaired nucleotides in the external loop from those enclosed by base pairs. Here, we restricted the algorithms to consider G-quadruplexes

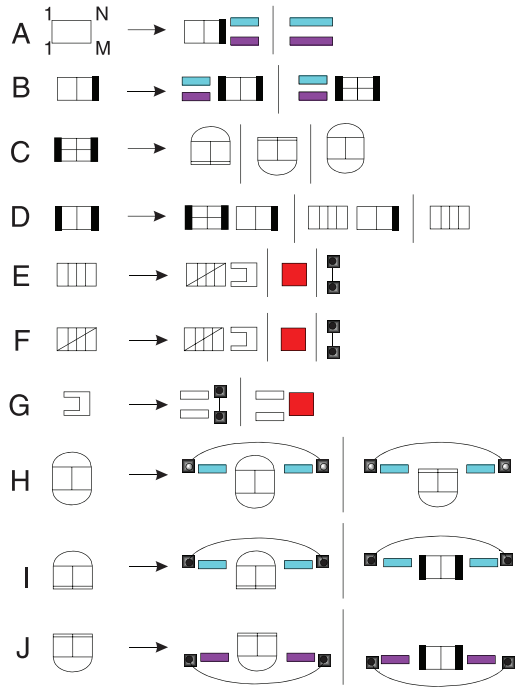


Fig. 9. RIP grammar with quadruplexes. Large rectangular symbols refer to different types of interaction structures. Outer arcs indicate enclosure by a base pair in one or both interacting subsequences. Flat rectangles are secondary structures with intramolecular quadruplexes in one of the two interacting molecules. The shaded versions (light for the first, dark for the second interaction partner) are nonempty, while the unshaded ones in production G may also be empty. Shaded squares denote intermolecular quadruplexes. After removing terminal secondary structures (productions A and B), the remainder is a tight structure of type C and D. The "doubletight" structure (C) is included by an intermolecular base pair. Otherwise, the absence of tangles allows to split of either a "doubletight" structure (C) or a maximal intermolecular stem-loop region without any intramolecular base pairs. This maximal structure E is further deconstructed analogous to a simple, unbranched stem-loop structure following the lines of `RNAhybrid` [50] or `RNAplex` [52]. The intermolecular quadruplexes are included at the stage like a single intermolecular base pair.

with a maximum layer size of $L = 7$ and a maximum total linker length of $\ell = 15$ only. The energies of all possible G-quadruplexes are precomputed, storing the energy of the most stable quadruplex for each pair of endpoints in the triangular matrix G. As this matrix will be very sparse for most inputs, a sparse matrix optimization is possible, but not yet implemented. In the backtracing part, we re-enumerate quadruplexes with given endpoints whenever necessary. Base pairing probabilities are computed as outlined above. Since there cannot be a conflict with canonical base pairs, we store P^G_{ij} as part of the base pairing probability matrix. The probabilities of individual G-G contacts are computed by enumeration as a postprocessing step. We also adapted the `RNAeval` and `RNAplot` programs so that sequence/structure pairs can be parsed and re-evaluated according to the extended grammar.

Runtime performance. The runtime of `RNAfold` with the extended grammar of Fig. 6 was compared to the implementation of the standard model. For both, energy minimization and partition function, virtually no difference was observed. For short sequences of about 200 nt, the additional preprocessing steps incur a minor but negligible runtime overhead.

Accuracy. Unfortunately, no large-scale set of benchmark data is available for RNA quadruplexes. Thus, to estimate our false-positive prediction rate, we use the `Rfam` database, version 11.0 [60]. This large set of structured RNAs can be assumed to be mostly free of G quartets, although quadruplex structures are not annotated at all in this resource so that a few unknown and/or unannotated may be hidden. For the 45,511 sequences in the seed alignments `RNAfold` predicted 640 quadruplexes, giving an upper bound on the false-positive rate of 1.4 percent. Using `RNAalifold`, we predict a consensus G-quadruplex for 17 of 2,185 Rfam seed alignments, which corresponds to a false-positive rate of 0.7 percent. We will discuss some of these "false positives" together with a set of RNA quadruplexes described in the literature at the end of the next section. Although the number of the known positive examples is too small to derive a meaningful quantitative estimate for the sensitivity, we will see that the thermodynamic predictions are in excellent agreement with the available experimental data.

5 Genome Screening

Occurrence and stability of G-quadruplexes in genomes. Sequence motifs of the form (1) that can in principle form quadruplex structures are very abundant in most genomes, see, e.g., [19], [20], [21]. Distinct G-quadruplexes can have substantial overlaps. Each quadruplex with $L > 2$ layers, for instance, contains 2^4 distinct quadruplexes with $L - 1$ layers. Additional variants can be formed when a linker sequence begins or ends with one or more Gs. In a genomic context, we are interested primarily in the number of loci that can harbor a quadruplex rather than their local diversity. We therefore count only quadruplex structures that cannot be embedded into a larger quadruplex. Equivalently, we count multiple overlapping quadruplexes of the same or smaller linker length as one quadruplex, represented by the one with most layers and the largest extent. Note that this still allows partial overlaps among quadruplexes.

The possible G-quadruplex loci were surveyed using the counting rules detailed above for several large genomic data sets: all bacterial ($n = 1{,}867$) and archaea ($n = 129$) genomes from the `NCBI` RefSeq database, all genomes provided by `wormbase` [61] ($n = 19$) and `flybase` [62] ($n = 12$) and all primate genomes hosted by `Ensembl` [63] ($n = 10$). To make the results comparable to each other, we normalized them to obtain the number of putative sites per Megabase (Fig. 10A). The majority (>90 percent) of putative sites for any of the screened genomes were two-layer G-quadruplexes (data not shown).

Since the pattern scan method does not score the G-quadruplexes according to their stability and sequence context, it results in a large number of potentially unstable G-quadruplexes with very long linker lengths and a small number of layers. Biologically, these are probably not relevant. We, therefore, investigated for which of these loci a quadruplex is more stable than alternative secondary structures. To this end, we used all putative sites of the

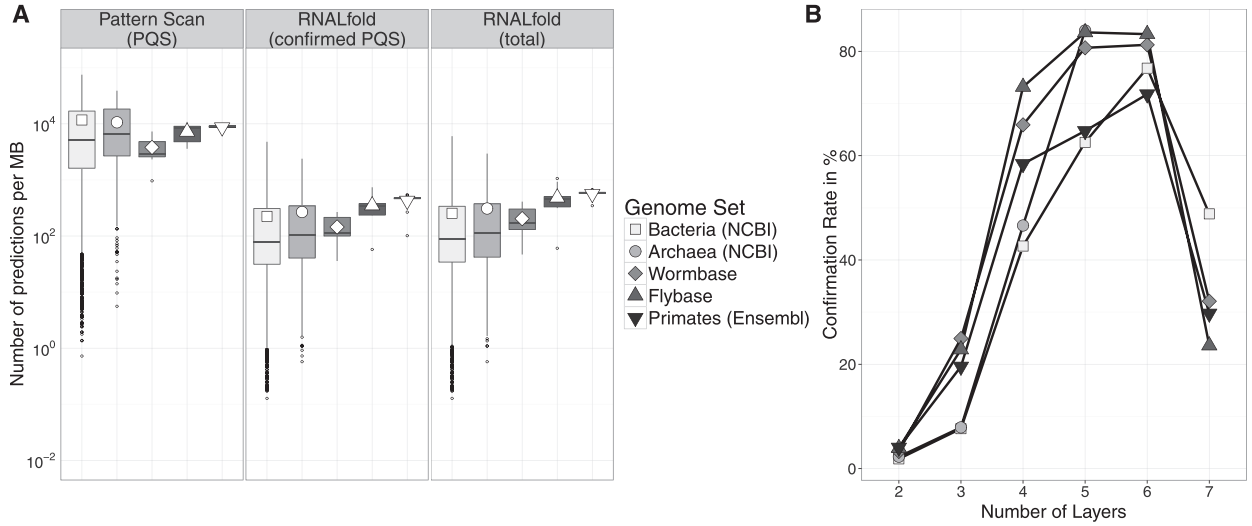


Fig. 10. Abundance and stability of putative G-quadruplexes. (A) Boxplot showing the number of putative G-quadruplex clusters/islands (PQS) identified by the pattern scan approach, number of clusters confirmed by `RNALfold` evaluation (confirmed PQS) and the total number of unique thermodynamically stable G-quadruplexes obtained by the `RNALfold` screening. Mean values are depicted by symbols. (B) Confirmation rates of the putative sites, i.e., ratio of confirmed and total number of putative sites. Each PQS that contains at least one thermodynamically stable G-quadruplex according to the `RNALfold` results is considered to be confirmed. Data are split to visualize the different rates for each layer size.

pattern scan as input for an `RNALfold` screen. We extended each putative quadruplex locus by 250 nt of flanking sequence on both the $5'$ and the $3'$ side. A pattern scan locus was considered "confirmed" if at least one stable G-quadruplex was predicted by `RNALfold`. These predictions were used to create annotation files for potentially stable G-quadruplexes and to count the number of pattern scan sites that are supported by the thermodynamic stability prediction. The average confirmation rates of 1.96 percent (bacteria), 2.38 percent (archaea), 3.92 percent (wormbase), 4.79 percent (flybase), and 4.91 percent (primates) show that the majority of the putative quadruplex loci are more likely to form canonical secondary structures than G-quadruplexes. Not surprisingly, the confirmation rate is very small, in particular, for G-quadruplexes with only $L = 2$ layers. In contrast, the rates rapidly increase for > 2 layers (Fig. 10B) due to their increased stability gained through additive stacking interactions in our energy model. These contribute little to the overall density of quadruplexes because they are much rarer than the $L = 2$ candidates.

Enrichment analysis. Our genome-wide surveys compute potential RNA G-quadruplexes throughout the entire genomic DNA sequence. To determine whether there is evidence for a widespread function of G-quadruplexes in transcripts, we investigated in detail the distribution of putative and confirmed quadruplexes relative to the annotation of bacterial, archaeal, and primate genomes.

More precisely, we considered the enrichment of G-quadruplexes in subsets of the genome relative to the genomic average. To this end, we partitioned the genome into "genic" regions (annotated genes) and the "intergenic" regions between them. For primates, the genic partition is further split into "exonic" and "intronic" regions. The protein coding "exonic" partition is again divided into "coding," "5UTR," "3UTR," and the remaining "exonic" regions are labeled as "nc-exonic." Since G-quadruplexes are directional, we considered the annotation also in a strand-specific manner. The total length of all items in a genomic partition is therefore twice the genome size.

Due to extensive overlap of transcripts and thus different transcript elements, e.g., exon-exon, intron-exon, or intron-intron overlap, we project annotations onto the genome in a hierarchical fashion: "coding," "5UTR," "3UTR," "nc-exonic," "intronic." If, for instance, five transcripts overlap on the same strand and the first nucleotide they have in common resides in a $5'$ UTR of one transcript, a coding region of the second one and in introns of the remaining three, then this base is labeled "coding" according to our hierarchy rule. For labeling all nucleotides following this procedure, we used the tool `multiIntersectBed` of the `BedTools` suite [64]. We then intersected the G-quadruplex predictions from both the pattern scan and the subsequent `RNALfold` screen with the genomic partitions using `intersectBed` [64] and calculated nucleotide-based coverages for each partition. Enrichment or depletion of a partition was calculated by normalizing the coverage to global genome coverage.

Fig. 12 summarizes the results. In bacteria, G-quadruplexes are enriched in the genic regions and depleted in the intergenic regions. This pattern was observed for the `RNALfold` results as well as the pattern scan results. For the latter, the effect seems to be slightly less pronounced.

In primates, we observe the same trend. Intergenic regions, which are the biggest fraction of the genome, are slightly depleted in G-quadruplexes, whereas genic regions, the complement to intergenic, are slightly enriched.

Within the genic partition, exonic regions are enriched further, whereas introns are only slightly above genomic background. This can be explained by the higher GC content of protein coding exons which lead to an increased probability for forming G-quadruplexes in the exons. The exonic partition is enriched by a factor 2.04 (median) in the

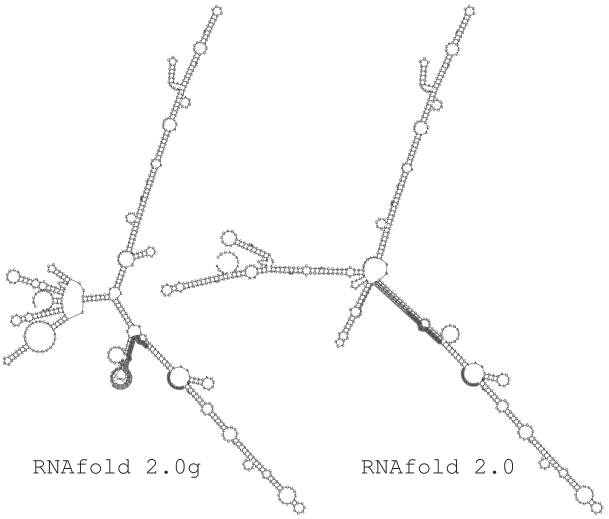


Fig. 11. Secondary structure predictions for RNA component of human telomerase (acc no. *AF221907*) with (l.h.s., computed with `RNAfold 2.0g`) and without (r.h.s., computed with `RNAfold 2.0`) quadruplexes. The template (medium gray), the $5'$ – part of P1 helix (dark gray), and the G-quadruplex forming sequence (light gray) are marked.

pattern scan, but only 1.49 fold (median) in the `RNALfold` screen. This is indicative of a selection pressure against stable G-quadruplex structures within the coding regions, presumably because these could interfere with the ribosome during translation. The discrepancy between the abundance of quadruplex patterns and stable quadruplexes is slightly more pronounced in coding exons compared to noncoding exons. The median enrichment factor drops from 2.18 to 1.33 (difference of 0.85) for coding exons and from 1.71 to 1.00 (difference of 0.71) in noncoding exons.

As expected, the enrichment of quadruplex patterns is even higher in 5′UTR (median: 3.37-fold) and increases further among the thermodynamically stable G-quadruplexes identified by `RNALfold` (median: 5.12-fold). This finding is in agreement with numerous previously reported examples, e.g., the G-quadruplexes in the 5′UTR of TRF2 [5], the eukaryotic MT3 matrix metalloproteinases [6], and the human telomerase RNA hTERC [65]. The strong enrichment of stable G-quadruplexes in 5′UTRs strongly suggests that they are an abundant functional feature of mRNAs.

The occurrence and abundance of both G-quadruplexes and stable RNA secondary structures strongly depend on the GC content of the sequences analyzed. To assess this effect, we calculate the GC enrichment for each partition in each genome as the ratio of local GC content of the respective partition to the genome wide GC content and compared the results to pattern scan and `RNALfold` screen. The enrichment of possible G-quadruplexes increases systematically with the GC content across all partitions (Fig. 13). This effect is much smaller for the `RNALfold` predictions. It is well known that GC content is indicative of a variety of biological functional genome elements, such as DNA-protein interaction, promoter function and epigenetic controlled gene regulation [66], which fulfill their function on the DNA level rather then on the RNA level. The observation that the density of thermodynamically stable G-quadruplexes is less dependent of GC content, i.e., that they are found abundantly over a wide range of GC contents, lends further support to the statement that G-quadruplexes frequently exert their function at the RNA level.

Specific examples. Several experimentally known RNA G-quadruplexes are predicted by the current version, including the semphorin hairpin of Fig. 7 and the quadruplex in human telomerase RNA [65], Fig. 11. A G-quadruplex in the 5′- terminal region of human telomerase RNA hTERC appears to obstruct the formation of the P1 helix, which is essential as boundary of the template. Hence, the quadruplex impairs the telomerase activity. Our computational results clearly reflect both the formation of the G-quadruplex in the 5′- part and its conflict with the P1 helix (see Fig. 11). In addition, the probability that the template nucleotides are unpaired is reduced when the G-quadruplex is included, suggesting that the effect of quadruplex formation at least in part is explained by inaccessibility of the template sequence.

We furthermore reinvestigated 17 mRNA sequences for which the effect of putative quadruplexes on protein expression has been determined by means of a luciferase luminescence assay or circular dichroism measurements, see Table 1 for the references. We predicted the MFE

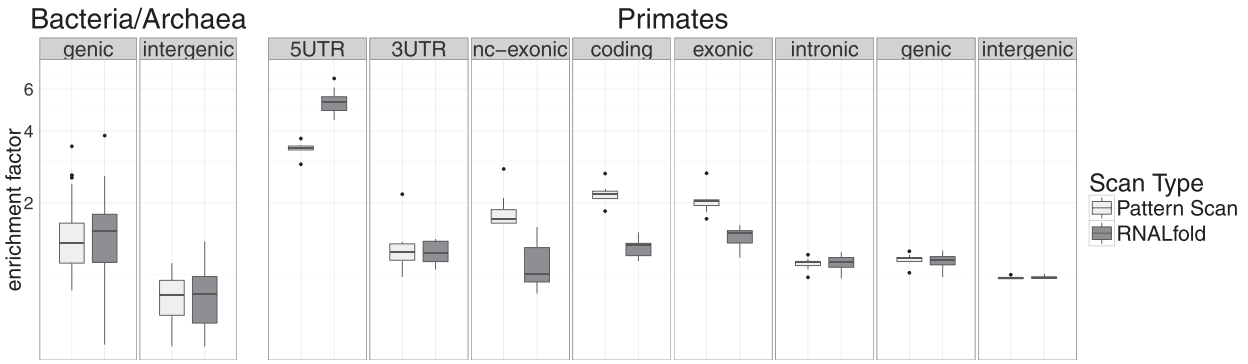


Fig. 12. Boxplot comparing the enrichment of predicted G-quadruplexes for sequence pattern search and `RNALfold`, and between different genomic partitions. Genic regions are enriched in G-quadruplexes compared to the genomic background, for the most bacteria and primates.

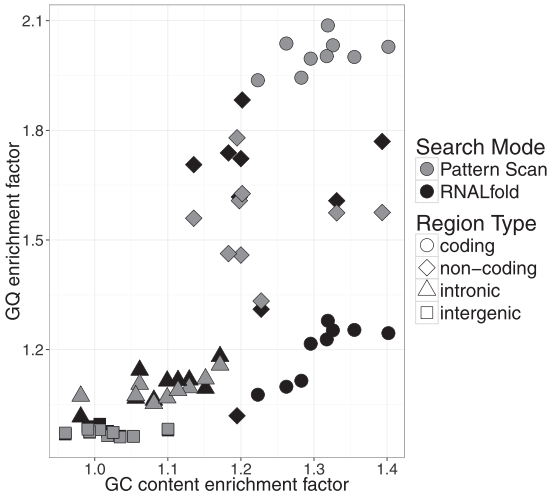


Fig. 13. Scatterplot of G-quadruplex versus GC-content enrichment. Enrichment was calculated as G-quadruplex density and GC-content for each genomic partition in relation to the complete genomic sequence background. The pattern scan results are more dependent on GC content compared to the RNALfold results.

structure with RNAfold of both the full-length mRNA and the putative G-quadruplex site PQS listed in the literature only. In addition, we used RNALfold to determine whether the quadruplex appears in a locally stable secondary structure and hence can be identified, e.g., in a genome-wide screen. With the exception of NCAM2 and THRA, the predictions for the PQS only and for locally stable secondary structures completely agree with the experiments. The misprediction of a two-layered instead of a three-layered G-quadruplex can be explained 1) by the very large total linker length of 16 in the case of NCAM2 that makes this structure undetectable for RNAfold and RNALfold and 2) large total linker lengths in combination with a concurrent canonical structure within the PQS of THRA. Even the MFE predictions for the full length mRNA transcripts agree well with the experimental data. Since mRNAs are packed with protein, in vivo long-range structures can be expected to be suppressed in mRNAs in general; hence, we already expect a priori that the local structures predicted by RNALfold gives better results than the global structures computed by RNAfold. Our model also clearly predicts that canonical base pairs dominate in the three negative examples (TNFSF12, MAP3K11, DOC2B). A simple scan for G-rich sequence motifs thus will overpredict RNA quadruplex structures.

Regarding the Rfam prediction set used above to assess accuracy, at least some of the putative false positive predictions might be real, however. The best example is the family RF00523 of prion protein (PrP) mRNA pseudoknots, which was predicted to play a role in PrP translation [72], [73]. Quadruplex forming nucleic acids are known to bind to PrP [74], suggesting that there could be a direct feedback between PrP and its mRNA via the predicted quadruplex.

Other true positives include the quadruplexes in some of the mammalian telomerase RNAs as discussed above, and the IRES (internal ribosomal entry site) in the VEGF mRNA IRES [71]. The latter suggests that the quadruplexes predicted in the n-myc IRES (RF00226) and in several viral IRES elements might also be true positives. We suspect that the quadruplex structure is not predicted from the Rfam alignment in these examples because quadruplexes are not part of Rfam models and hence the manually curated alignments have been built to favor alternative, conventional secondary structures. The XIST intron (RF02266) is likely a true positive. It features a conserved quadruplex that does not interfere with the annotated structure. Another likely case is the antisense transcript FMR1-AS1 (RF02118), as it is suspected that FMR1 itself is regulated by a G-quadruplex [9]. Additionally, possible quadruplex structures in rRNA genes have been described for yeast

TABLE 1
Comparison of Experimental and Computational Results for mRNA Sequences with PQS in Their 5′UTR

Gene		length		Layers	Experiment			RNAfold		RNALfold
name	ID	mRNA	PQS				Ref.	mRNA	PQS	mRNA
TRF2	NM_005652.3	2979	18	3	+++	*in vivo*	[5]	3	3	3
MT3-MMP	NM_005941.4	6347	20	3	+++	*in vivo*	[6]	3	3	3
EBAG9	NM_198120.1	1494	45	3	+	*in vivo*	[12]	3	3	3
FZD2	NM_001466.3	3834	60	3	+	*in vivo*	[12]	2	3	2
BARHL1	NM_020064.3	1907	46	3	+++	*in vivo*	[12]	–	3	3
NCAM2	NM_004540.3	5006	54	3	+++	*in vivo*	[12]	–	2	2
THRA	NM_199334.3	2400	49	3	+	*in vivo*	[12]	–	2	2
AASDHPPT	NM_015423.2	2880	38	3	+	*in vivo*	[12]	3	3	3
TNFSF12	NM_003809.2	1407	54	3	–	*in vivo*	[12]	–	–	–
MAP3K11	NM_002419.3	3574	55	3	–	*in vivo*	[12]	–	–	–
DOC2B	NM_003585.3	2022	56	3	–	*in vivo*	[12]	–	–	–
NRAS	NM_002524.4	4454	18	3	+++	*in vivo*	[44]	3	3	3
ZIC-1	NM_003412.3	5231	27	3	+++	*in vivo*	[67]	3	3	3
BCL-2	ENST00000333681	3209	25	3	+++	*in vivo*	[68]	–	3	3
CHST2	NM_004267	4137	25	3	+++	*in vivo*	[69]	3	3	3
ADAM10	NM_001110	3927	30	3	+++	*in vitro*	[70]	3	3	3
VEGFA	NM_001171623	3677	18	2	+	*in vivo*	[71]	2	2	2

The column "Experiment" refers to the presence of a G-quadruplex in vivo. The symbols $+++$, $+$, and $-$ indicate the presence of a G-quadruplex, the formation of a G-quadruplex under sufficiently large potassium concentration, and the absence of a quadruplex, respectively. The prediction results of RNAfold and RNALfold display the layer size of a predicted G-quadruplex within the PQS or—if none was predicted.

[75] and a significant enrichment has been reported for long noncoding RNAs [14].

6 Discussion

We have shown in this contribution that structural elements such as G-quadruplexes that correspond to uninterrupted sequence intervals can be included in a rather straightforward way into the standard dynamic programming recursions—provided a corresponding extension of the energy model can be devised. We describe here an extension of the most important and widely used RNA secondary structure prediction. This includes in particular energy minimization and the computation of the partition function. Extensions to local folding algorithms and to the computation of consensus structures from sequence alignments are also straightforward applications of the modified grammar (Fig. 6). It is less obvious how to handle quadruplexes in RNA-RNA interactions since our recursions consider local G-quadruplexes only. A simple adaptation of the `RNAup` approach, however, also covers such situations of trans-G-quadruplexes.

A survey of G-quadruplexes in genomic sequence data shows that thermodynamically G-quadruplexes are substantially enriched in exonic regions and in particular in 5′UTRs. At the same time, the stability of G-quadruplexes is reduced in protein-coding sequences. Taken together, these observations strongly suggest that stable G-quadruplexes in 5′UTRs are abundant features of mRNAs. Most likely, they function as regulators of translation.

The G-quadruplex-aware programs are currently available as a separate branch (version number with the suffix "g") of the `ViennaRNA Package` using a very simple energy function for the quadruplexes that reproduces the few available experimental data at least semiquantitatively. Following further optimization of the code, the algorithmic extensions will be integrated in the main version of the package in the near future. The main problem for practical applications of quadruplex-aware RNA folding tools is our limited knowledge of the energy function in particular for $L \neq 3$ and for asymmetric linkers. Even with the crude energy function employed here, it becomes clear that the overwhelming majority of putative genomic quadruplex sequences will fold into a canonical secondary structure rather than G-quadruplex structures.

Online Supplement

Supplemental data and source code are available from http://www.bioinf.uni-leipzig.de/publications/supplements/12-025 and www.tbi.univie.ac.at/RNA/.

Acknowledgments

This work was supported in part by the German Research Foundation (STA 850/7-2, under the auspices of SPP-1258 "Sensory and Regulatory RNAs in Prokaryotes"), the Austrian GEN-AU projects "regulatory non coding RNA," "Bioinformatics Integration Network III," and the Austrian FWF project "SFB F43 RNA regulation of the transcriptome."

References

[1] K. Paeschke, T. Simonsson, J. Postberg, D. Rhodes, and H.J. Lipps, "Telomere End-Binding Proteins Control the Formation of G-Quadruplex DNA Structures *in Vivo*," *Nature Structural Molecular Biology,* vol. 12, pp. 847-854, 2005.

[2] J.E. Johnson, J.S. Smith, M.L. Kozak, and F.B. Johnson, "*In Vivo Veritas*: Using Yeast to Probe the Biological Functions of G-Quadruplexes," *Biochimie,* vol. 90, pp. 1250-1263, 2008.

[3] H.M. Wong, L. Payet, and J.L. Huppert, "Function and Targeting of G-Quadruplexes," *Current Opinion Molecular Therapeutics,* vol. 11, pp. 146-155, 2009.

[4] A. Baral, P. Kumar, R. Halder, P. Mani, V.K. Yadav, A. Singh, S.K. Das, and S. Chowdhury, "Quadruplex-Single Nucleotide Polymorphisms (Quad-SNP) Influence Gene Expression Difference among Individuals," *Nucleic Acids Research,* vol. 40, pp. 3800-3811, 2012.

[5] D. Gomez, A. Guédin, J.L. Mergny, B. Salles, J.F. Riou, M.P. Teulade-Fichou, and P. Calsou, "A G-Quadruplex Structure within the 5′-UTR of TRF2 mRNA Represses Translation in Human Cells," *Nucleic Acids Research,* vol. 38, pp. 7187-7198, 2010.

[6] M.J. Morris and S. Basu, "An Unusually Stable G-Quadruplex within the 5′-UTR of the MT3 Matrix Metalloproteinase mRNA Represses Translation in Eukaryotic Cells," *Biochemistry,* vol. 48, pp. 5313-5319, 2009.

[7] L. Menon and M.R. Mihailescu, "Interactions of the G Quartet Forming Semaphorin 3F RNA with the RGG Box Domain of the Fragile X Protein Family," *Nucleic Acids Research,* vol. 35, pp. 5379-5392, 2007.

[8] M. Bensaid, M. Melko, E.G. Bechara, L. Davidovic, A. Berretta, M.V. Catania, J. Gecz, E. Lalli, and B. Bardoni, "FRAXE-Associated Mental Retardation Protein (FMR2) Is an RNA-Binding Protein with High Affinity for G-Quartet RNA Forming Structure," *Nucleic Acids Research,* vol. 37, pp. 1269-1279, 2009.

[9] A. Bugaut and S. Balasubramanian, "5′-UTR RNA G-Quadruplexes: Translation Regulation and Targeting," *Nucleic Acids Research,* vol. 40, pp. 4727-4741, 2012.

[10] A. Arora and B. Suess, "An RNA G-Quadruplex in the 3′ UTR of the Proto-Oncogene PIM1 Represses Translation," *RNA Biology,* vol. 8, pp. 802-805, 2011.

[11] J.L. Huppert, A. Bugaut, S. Kumari, and S. Balasubramanian, "G-Quadruplexes: The Beginning and End of UTRs," *Nucleic Acids Research,* vol. 36, pp. 6260-6268, 2008.

[12] J.D. Beaudoin and J.P. Perreault, "5′-UTR G-Quadruplex Structures Acting as Translational Repressors," *Nucleic Acids Research,* vol. 38, pp. 7022-7036, 2010.

[13] M. Wieland and J.S. Hartig, "RNA Quadruplex-Based Modulation of Gene Expression," *Chemistry Biology,* vol. 14, pp. 757-763, 2007.

[14] G.G. Jayaraj, S. Pandey, V. Scaria, and S. Maiti, "Potential G-Quadruplexes in the Human Long Non-Coding Transcriptome," *RNA Biology,* vol. 9, pp. 81-86, 2012.

[15] B. Luke and J. Lingner, "TERRA: Telomeric Repeat-Containing RNA," *EMBO J.,* vol. 28, pp. 2503-2510, 2009.

[16] A.Y. Zhang, A. Bugaut, and S. Balasubramanian, "A Sequence-Independent Analysis of the Loop Length Dependence of Intramolecular RNA G-Quadruplex Stability and Topology," *Biochemistry,* vol. 50, pp. 7251-7258, 2011.

[17] M. Webba da Silva, "Geometric Formalism for DNA Quadruplex Folding," *Chemistry,* vol. 13, pp. 9738-9745, 2007.

[18] D.H. Zhang and G.Y. Zhi, "Structure Monomorphism of RNA G-Quadruplex That Is Independent of Surrounding Condition," *J. Biotechnology,* vol. 150, pp. 6-10, 2010.

[19] J. Eddy and N. Maizels, "Gene Function Correlates with Potential for G4 DNA Formation in the Human Genome," *Nucleic Acids Research,* vol. 34, pp. 3887-3896, 2006.

[20] J.L. Huppert and S. Balasubramanian, "Prevalence of Quadruplexes in the Human Genome," *Nucleic Acids Research,* vol. 33, pp. 2908-2916, 2005.

[21] Y. Zhao, Z. Du, and N. Li, "Extensive Selection for the Enrichment of G4 DNA Motifs in Transcriptional Regulatory Regions of Warm Blooded Animals," *FEBS Letters,* vol. 581, pp. 1951-1956, 2007.

[22] A. Verma, K. Halder, R. Halder, V.K. Yadav, P. Rawal, R.K. Thakur, F. Mohd, A. Sharma, and S. Chowdhury, "G-Quadruplex DNA Motifs as Conserved Cis-Regulatory Elements," *J. Medicinal Chemistry,* vol. 51, pp. 5641-5649, 2008.

[23] O. Kikin, L. D'Antonio, and P.S. Bagga, "QGRS Mapper: A Web-Based Server for Predicting G-Quadruplexes in Nucleotide Sequences," *Nucleic Acids Research,* vol. 34, pp. W676-W682, 2006.

[24] A.K. Todd, "Bioinformatics Approaches to Quadruplex Sequence Location," *Methods,* vol. 43, pp. 246-251, 2007.

[25] O. Stegle, L. Payet, J.-L. Mergny, D.J.C. MacKay, and J.L. Huppert, "Predicting and Understanding the Stability of G-Quadruplexes," *Bioinformatics,* vol. 25, pp. i374-i382, 2009.

[26] R. Lorenz, S.H. Bernhart, F. Externbrink, J. Qin, C. Höner zu Siederdissen, F. Amman, I.L. Hofacker, and P.F. Stadler, "RNA Folding Algorithms with G-Quadruplexes," *Proc. Brazilian Symp. Bioinformatics (BSB '12),* M.C.P. de Souto and M.G. Kann, eds., pp. 49-60, 2012.

[27] J.L. Mergny and L. Lacroix, "UV Melting of G-Quadruplexes," *Current Protocols in Nucleic Acid Chemistry,* Unit 17.1, Wiley, 2009.

[28] A. Bugaut and S. Balasubramanian, "A Sequence-Independent Study of the Influence of Short Loop Lengths on the Stability and Topology of Intramolecular DNA G-Quadruplexes," *Biochemistry,* vol. 47, pp. 689-697, 2008.

[29] D.H. Zhang, T. Fujimoto, S. Saxena, H.Q. Yu, D. Miyoshi, and N. Sugimoto, "Monomorphic RNA G-Quadruplex and Polymorphic DNA G-Quadruplex Structures Responding to Cellular Environmental Factors," *Biochemistry,* vol. 49, pp. 4554-4563, 2010.

[30] A. Guédin, A. De Cian, J. Gros, L. Lacroix, and J.L. Mergny, "Sequence Effects in Single-Base Loops for Quadruplexes," *Biochimie,* vol. 90, pp. 686-696, 2008.

[31] A. Guédin, J. Gros, A. Patrizia, and J.-L. Mergny, "How Long Is Too Long? Effects of Loop Size on G-Quadruplex Stability," *Nucleic Acids Research,* vol. 38, pp. 7858-7868, 2010.

[32] C.T. Lauhon and J.W. Szostak, "RNA Aptamers that Bind Flavin and Nicotinamide Redox Cofactors," *J. Am. Chemical Soc.,* vol. 117, pp. 1246-1257, 1995.

[33] A. Joachimi, A. Benz, and J.S. Hartig, "A Comparison of DNA and RNA Quadruplex Structures and Stabilities," *Bioorganic Medicinal Chemistry,* vol. 17, pp. 6811-6815, 2009.

[34] R. Lorenz, S.H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler, and I.L. Hofacker, "ViennaRNA Package 2.0," *Algorithms Molecular Biology,* vol. 6, article 26, 2011.

[35] P. Schuster, W. Fontana, P.F. Stadler, and I.L. Hofacker, "From Sequences to Shapes and Back: A Case Study in RNA Secondary Structures," *Proc. Royal Soc. London B,* vol. 255, pp. 279-284, 1994.

[36] R. Giegerich, B. Voss, and M. Rehmsmeier, "Abstract Shapes of RNA," *Nucleic Acids Research,* vol. 32, pp. 4843-4851, 2004.

[37] P. Flajolet and R. Sedgewick, *Analytic Combinatorics.* Cambridge Univ. Press, 2009.

[38] I.L. Hofacker, P. Schuster, and P.F. Stadler, "Combinatorics of RNA Secondary Structures," *Discrete Applied Math.,* vol. 88, pp. 207-237, 1998.

[39] D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker, and D.H. Turner, "Incorporating Chemical Modification Constraints into a Dynamic Programming Algorithm for Prediction of RNA Secondary Structure," *Proc. Nat'l Academy Sciences USA,* vol. 101, pp. 7287-7292, 2004.

[40] M. Zuker and P. Stiegler, "Optimal Computer Folding of Large RNA Sequences Using Thermodynamics and Auxiliary Information," *Nucleic Acids Research,* vol. 9, pp. 133-148, 1981.

[41] J.S. McCaskill, "The Equilibrium Partition Function and Base Pair Binding Probabilities for RNA Secondary Structure," *Biopolymers,* vol. 29, pp. 1105-1119, 1990.

[42] R.E. Bruccoleri and G. Heinrich, "An Improved Algorithm for Nucleic Acid Secondary Structure Display," *Computer Application Biosciences,* vol. 4, pp. 167-173, 1988.

[43] C.B. Do, D.A. Woods, and S. Batzoglou, "CONTRAfold: RNA Secondary Structure Prediction without Physics-Based Models," *Bioinformatics,* vol. 22, no. 14, pp. e90-e98, 2006.

[44] S. Kumari, A. Bugaut, J.L. Huppert, and S. Balasubramanian, "An RNA G-Quadruplex in the 5'UTR of the NRAS Proto-Oncogene Modulates Translation," *Nat'l Chemical Biology,* vol. 3, pp. 218-221, 2007.

[45] I.L. Hofacker, B. Priwitzer, and P.F. Stadler, "Prediction of Locally Stable RNA Secondary Structures for Genome-Wide Surveys," *Bioinformatics,* vol. 20, pp. 191-198, 2004.

[46] S. Bernhart, I.L. Hofacker, and P.F. Stadler, "Local RNA Base Pairing Probabilities in Large Sequences," *Bioinformatics,* vol. 22, pp. 614-615, 2006.

[47] S.H. Bernhart, H. Tafer, U. Mückstein, C. Flamm, P.F. Stadler, and I.L. Hofacker, "Partition Function and Base Pairing Probabilities of RNA Heterodimers," *Algorithms Molecular Biology,* vol. 1, article 3, 2006.

[48] U. Mückstein, H. Tafer, J. Hackermüller, S.B. Bernhard, P.F. Stadler, and I.L. Hofacker, "Thermodynamics of RNA-RNA Binding," *Bioinformatics,* vol. 22, pp. 1177-1182, 2006.

[49] A. Busch, A. Richter, and R. Backofen, "IntaRNA: Efficient Prediction of Bacterial sRNA Targets Incorporating Target Site Accessibility and Seed Regions," *Bioinformatics,* vol. 24, pp. 2849-2856, 2008.

[50] M. Rehmsmeier, P. Steffen, M. Höchsmann, and R. Giegerich, "Fast and Effective Prediction of MicroRNA/Target Duplexes," *RNA,* vol. 10, pp. 1507-1517, 2004.

[51] K. Ito, S. Go, M. Komiyama, and Y. Xu, "Inhibition of Translation by Small RNA-Stabilized mRNA Structures in Human Cells," *J. Am. Chemical Soc.,* vol. 133, pp. 19153-19159, 2011.

[52] H. Tafer, F. Ammann, F. Eggenhoffer, P.F. Stadler, and I.L. Hofacker, "Fast Accessibility-Based Prediction of RNA-RNA Interactions," *Bioinformatics,* vol. 27, pp. 1934-1940, 2011.

[53] D. Pervouchine, "IRIS: Intermolecular RNA Interaction Search," *Proc. Int'l Conf. Genome Informatics,* vol. 15, pp. 92-101, 2004.

[54] C. Alkan, E. Karakoc, J. Nadeau, S. Sahinalp, and K. Zhang, "RNA-RNA Interaction Prediction and Antisense RNA Target Search," *J. Computational Biology,* vol. 13, pp. 267-282, 2006.

[55] H. Chitsaz, R. Salari, S. Sahinalp, and R. Backofen, "A Partition Function Algorithm for Interacting Nucleic Acid Strands," *Bioinformatics,* vol. 25, pp. i365-i373, 2009.

[56] F.W.D. Huang, J. Qin, C.M. Reidys, and P.F. Stadler, "Partition Function and Base Pairing Probabilities for RNA-RNA Interaction Prediction," *Bioinformatics,* vol. 25, pp. 2646-2654, 2009.

[57] F.W.D. Huang, J. Qin, C.M. Reidys, and P.F. Stadler, "Target Prediction and a Statistical Sampling Algorithm for RNA-RNA Interaction," *Bioinformatics,* vol. 26, pp. 175-181, 2010.

[58] I.L. Hofacker, M. Fekete, and P.F. Stadler, "Secondary Structure Prediction for Aligned RNA Sequences," *J. Molecular Biology,* vol. 319, pp. 1059-1066, 2002.

[59] S.H. Bernhart, I.L. Hofacker, S. Will, A.R. Gruber, and P.F. Stadler, "RNAalifold: Improved Consensus Structure Prediction for RNA Alignments," *BMC Bioinformatics,* vol. 9, article 474, 2008.

[60] S.W. Burge, J. Daub, R. Eberhardt, J. Tate, L. Barquist, E.P. Nawrocki, S.R. Eddy, P.P. Gardner, and A. Bateman, "Rfam 11.0: 10 Years of RNA Families," *Nucleic Acids Research,* vol. 41, pp. D226-D232, 2012.

[61] T.W. Harris, I. Antoshechkin, T. Bieri, D. Blasiar, J. Chan, W.J. Chen, N. De La Cruz, P. Davis, M. Duesbury, R. Fang, J. Fernandes, M. Han, R. Kishore, R. Lee, H.M. Müller, C. Nakamura, P. Ozersky, A. Petcherski, A. Rangarajan, A. Rogers, G. Schindelman, E.M. Schwarz, M.A. Tuli, K. Van Auken, D. Wang, X. Wang, G. Williams, K. Yook, R. Durbin, L.D. Stein, J. Spieth, and P.W. Sternberg, "WormBase: A Comprehensive Resource for Nematode Research," *Nucleic Acids Research,* vol. 38, pp. D463-D467, 2010.

[62] P. McQuilton, S.E.S. Pierre, J. Thurmond and the FlyBase Consortium, "FlyBase 101—The Basics of Navigating FlyBase," *Nucleic Acids Research,* vol. 40, pp. D706-D714, 2012.

[63] P. Flicek, M. Amode, D. Barrell, K. Beal, S. Brent, D. Carvalho-Silva, P. Clapham, G. Coates, S. Fairley, S. Fitzgerald, L. Gil, L. Gordon, M. Hendrix, T. Hourlier, N. Johnson, A.K. Kähäri, D. Keefe, S. Keenan, R. Kinsella, M. Komorowska, G. Koscielny, E. Kulesha, P. Larsson, I. Longden, W. McLaren, M. Muffato, B. Overduin, M. Pignatelli, B. Pritchard, H.S. Riat, G.R. Ritchie, M. Ruffier, M. Schuster, D. Sobral, Y.A. Tang, K. Taylor, S. Trevanion, J. Vandrovcova, S. White, M. Wilson, S.P. Wilder, B.L. Aken, E. Birney, F. Cunningham, I. Dunham, R. Durbin, X.M. Fernández-Suarez, J. Harrow, J. Herrero, T.J. Hubbard, A. Parker, G. Proctor, G. Spudich, J. Vogel, A. Yates, A. Zadissa, and S.M. Searle, "Ensembl 2012," *Nucleic Acids Research,* vol. 40, pp. D84-D90, 2012.

[64] A.R. Quinlan and I.M. Hall, "Bedtools: A Flexible Suite of Utilities for Comparing Genomic Features," *Bioinformatics,* vol. 26, pp. 841-842, 2010.

[65] J. Gros, A. Guédin, J.-L. Mergny, and L. Lacroix, "G-Quadruplex Formation Interferes with P1 Helix Formation in the RNA Component of Telomerase hTERC," *ChemBioChem,* vol. 9, pp. 2075-2079, 2008.

[66] S.C.J. Parker, E.H. Margulies, and T.D. Tullius, "The Relationship between Fine Scale DNA Structure, GC Content, and Functional Elements in 1% of the Human Genome," *Genome Informatics,* vol. 20, pp. 199-211, 2008.

[67] A. Arora, M. Dutkiewicz, V. Scaria, M. Hariharan, S. Maiti, and J. Kurreck, "Inhibition of Translation in Living Eukaryotic Cells by an RNA G-Quadruplex Motif," *RNA,* vol. 14, pp. 1290-1296, 2008.

[68] R. Shahid, A. Bugaut, and S. Balasubramanian, "The BCL-2 5′ Untranslated Region Contains an RNA G-Quadruplex-Forming Motif That Modulates Protein Expression," *Biochemistry*, vol. 49, pp. 8300-8306, 2010.

[69] K. Halder, M. Wieland, and J.S. Hartig, "Predictable Suppression of Gene Expression by 5′-UTR-Based RNA Quadruplexes," *Nucleic Acids Research*, vol. 37, no. 20, pp. 6811-6817, 2009.

[70] S. Lammich, F. Kamp, J. Wagner, B. Nuscher, S. Zilow, A.-K. Ludwig, M. Willem, and C. Haass, "Translational Repression of the Disintegrin and Metalloprotease ADAM10 by a Stable G-Quadruplex Secondary Structure in Its 5′-Untranslated Region," *J. Biological Chemistry*, vol. 286, pp. 45063-45072, 2011.

[71] M.J. Morris, Y. Negishi, C. Pazsint, J.D. Schonhoft, and S. Basu, "An RNA G-Quadruplex Is Essential for Cap-Independent Translation Initiation in Human VEGF IRES," *J. Am. Chemical Soc.*, vol. 132, pp. 17831-17839, 2010.

[72] P.R. Wills, "Potential Pseudoknots in the PrP-Encoding mRNA," *J. Theoretical Biology*, vol. 159, pp. 523-527, 1992.

[73] I. Barrette, G. Poisson, P. Gendron, and F. Major, "Pseudoknots in Prion Protein mRNAs Confirmed by Comparative Sequence Analysis and Pattern Searching," *Nucleic Acids Research*, vol. 29, pp. 753-758, 2001.

[74] P. Cavaliere, B. Pagano, V. Granata, S. Prigent, H. Rezaei, C. Giancola, and A. Zagari, "Cross-Talk between Prion Protein and Quadruplex-Forming Nucleic Acids: A Dynamic Complex Formation," *Nucleic Acids Research*, vol. 41, pp. 327-339, 2013.

[75] J.A. Capra, K. Paeschke, M. Singh, and V.A. Zakian, "G-Quadruplex DNA Sequences Are Evolutionarily Conserved and Associated with Distinct Genomic Features in *Saccharomyces cerevisiae*," *PLoS Computational Biology*, vol. 6, article e1000861, 2010.

Ronny Lorenz completed his diploma thesis in computer sciences with a special focus on bioinformatics from the Department of Computer Science, University of Leipzig. He is currently working toward the PhD degree in the area of molecular biology at the Institute for Theoretical Chemistry, Vienna, Austria. His research interests focus on algorithmics in RNA secondary structure prediction, and in-silico folding kinetics of growing RNA transcripts and metastable RNAs.

Stephan H. Bernhart studied chemistry from the University Vienna. He received the PhD degree with Peter Schuster on Algorithms for RNA secondary structure prediction in 2007. He is currently a postdoc in the Department of Computer Science and the Junior Research Group Transcriptome Bioinformatics at the University of Leipzig.

Jing Qin received the PhD degree from CFC, Nankai University, China, in the area of applied discrete mathematics. She is currently a postdoctoral fellow in MPI for Mathematics in the Sciences, Germany. Her main research interests include analytic combinatorics, graph algorithms, stochastic analysis, and their applications in mathematical biology.

Christian Höner zu Siederdissen received the diploma in applied computer science in the natural sciences from Bielefeld University, Germany. He is currently working toward the PhD degree in molecular biology at the Institute for Theoretical Chemistry, Vienna, Austria. His research interests include optimization problems in biology, Bayesian statistics, and functional programming.

Andrea Tanzer studied biology at the University of Vienna and received the PhD degree in computer science from the University of Leipzig in 2009. From 2010-2012, she was postdoc in Roderic Guigó's lab at the CRG in Barcelona working in the ENCODE Project. Since September 2012, she has been a research assistant at the Institute for Theoretical Chemistry, University of Vienna.

Fabian Amman studied biology at the University Vienna. He is currently working toward the PhD degree at the Institute for Theoretical Chemistry, University of Vienna. His research interests include gene regulation by bacterial small RNA and next-generation sequencing.

Ivo L. Hofacker is a professor in the Departments of Chemistry and of Computer Science, University of Vienna, where he heads the Institute for Theoretical Chemistry and the Research Group Bioinformatics and Computational Biology, respectively. His research interests focus on the prediction of RNA structures, noncoding RNA annotation, and the analysis of folding landscapes.

Peter F. Stadler received the PhD degree in chemistry from University of Vienna in 1990 and then was an assistant and associate professor for theoretical chemistry at the same school. In 2002, he moved to Leipzig as a full professor for bioinformatics. Since 1994, he is an external professor at the Santa Fe Institute. He has been an external scientific member of the Max Planck Society since 2009 and a corresponding member abroad of the Austrian Academy of Sciences since 2010.

Part V

UNPUBLISHED WORK

11

FINDING OPTIMAL REFOLDING PATHS

As pointed out in Chapter 4, the prediction of optimal refolding paths between two secondary structures is an important challenge. Intermediate structures of a path may point into directions that lead to a better understanding of the functional role of an RNAs structure. More importantly, the saddle point, i.e. the structure with highest free energy along the path, yields the energy barrier that separates both structures. This in turn determines the folding rate required for RNA folding kinetics, cf. Chapter 6.1. Since the prediction of optimal refolding paths is an *NP-complete* problem [145], heuristic approaches are used instead of exhaustive ones. Yet, there is only a handful of such algorithms, and most often they only consider direct paths. Therefore, they tend to overestimate actual refolding energy barriers, and consequently predictions based thereof are most likely at odds with each other.

In this chapter, I discuss a generic iterative algorithm for near optimal indirect refolding path prediction. Based on supporting points in the secondary structure free energy landscape, the algorithm aims to minimize the refolding barrier by constructing indirect paths between two secondary structures out of direct paths between those structures and the supporting point. By keeping the generation of supporting points flexible, this algorithm enables a variety of applications where a compromise between optimality of the prediction and computation speed is required. Although many of the following ideas were already mentioned briefly in Chapter 9, I want to emphasize here the potential of distance class partitioning of RNA secondary structure free energy landscapes, especially for refolding path prediction. Therefore, the first section starts with a recapitulation of the idea behind distance class partitioning, followed by a description of the possibilities it opens. Finally, the indirect refolding path construction of Chapter 9 is presented from a more generalized point of view in the second part of this chapter.

11.1 DISTANCE CLASS PARTITIONING REVISITED

The set of secondary structures which is compatible with an RNA sequence, the structure ensemble Ω, spans a hyperdimensional space, the secondary structure free energy landscape. Here, each secondary structure $s \in \Omega$ represents a point in the landscape, its free energy $E(s)$ determines its height, and finally two structures s_i and s_j are considered neighbors if they differ in exactly one base pair, i.e. they have base pair distance $d_{BP}(s_i, s_j) = 1$. Ordinary MFE structure prediction, as introduced in Chapter 3.2 determines the global minimum of this landscape, i.e. the deepest point. However,

alternative low free energy structures with distinct structural features are generally of interest as well, as they provide understanding of the potential role of an RNA. For this purpose, the generation of suboptimal secondary structures within a certain energy band around the MFE proves useful. Still, once structural alternatives are known, it needs to be investigated whether or not they are easily adopted, and how likely a switch from one into another is. For most RNAs with sequence length of biological importance, it is infeasible to enumerate relevant suboptimal structures, since $|\Omega|$ grows exponentially with sequence length n. Even partition function computation intertwines all the information of alternative structures with the complete ensemble, such that refolding barriers can not be derived from it.

This is where classified DP comes into play, in particular distance class partitioning. This method projects the high-dimensional energy landscape into lower dimensions using a partitioning of structures with respect to their base pair distance to a set of initially chosen reference structures. The number of references therefore determines the dimensionality of the projection of the energy landscape. Within each partition thermodynamic properties are determined, such as MFE representative, partition function, and so on. An algorithm that implements the idea of distance class partitioning with one reference structure is `RNAbor` [72], whereas `RNA2Dfold`, as presented in Chapter 9, uses two references to obtain a 2D projection. Nevertheless, the following ideas can be easily applied to projections of higher dimensionality, i.e. distance class partitionings with more than 2 reference structures, as well. In fact, increasing the number of references has a large impact on a number of applications, since it allows to triangulate certain areas of interest within the energy landscape. Though, increasing the number of reference structures just by one results in an increase of the algorithms asymptotic time complexity in the order of n^2, which is why I restrict the description below to 2D projections.

DETECTING ALTERNATIVE LOW FREE ENERGY STRUCTURES As mentioned before, low free energy states within the ensemble of all secondary structures compatible with a particular sequence can be obtained from suboptimal structure prediction. However, the resulting set of meta-stable structures alone is not sufficient to decide whether or not the RNA is likely to ever adopt them. Distance class partitioning, on the other hand, also enables the determination of low free energy states. Using a single reference structure, for instance the MFE structure, a classified MFE prediction produces a set of m optimal structure representatives which differ in exactly $1, 2, \ldots, m$ base pairs, i.e. there is no other structure with the same base pair distance to the reference and lower free energy. This immediately yields a set of candidate structures that can be filtered according to their free energy and distance to the MFE structure. Here, one would assume that a meta-stable state of biological importance differs from the MFE in some more or less complex structure motifs, i.e. its distance to the MFE needs to be sufficiently large. Furthermore, a candidate should exhibit a low free energy, and

needs to be sufficiently separated from its neighborhood. Hence, it should be surrounded by representative structures with higher free energy. Of course, each MFE representative of a particular distance class can not provide a complete picture of all the alternatives the distance class subsumes. It just provides the most probable structure within the distance class. However, the addition of further reference structures, such as candidates obtained from a projection with lower dimensionality, leads to a more fine grained characterization of their structural environment.

An example for such a procedure is provided with an artificial RNA sequence known to form two stable secondary structures [258]. This particular 73 nt long switching RNA molecule was designed such that it adopts a meta-stable secondary structure (Figure 15B) right after transcription, and only refolds into its ground state (Figure 15A) after a long period of time. As depicted in Figure 15, two alternative meta-stable states can already be identified by visual inspection of the results from a distance class partitioning with respect to the MFE structure. The most distant of them happens to be the meta-stable structure the RNA was designed to form. Using this structure as a second reference for a 2D projection then clearly reveals its separation from from the remaining low free energy conformations Still, this projection is hiding other alternatives, since each spot in the 2D landscape constitutes just one of many possible structures in the corresponding distance class. Therefore, either additional reference structures are necessary to obtain a more fine grained projection of the energy landscape, or the set of potentially interesting candidates can be iteratively extended with results from 2D projections using the MFE structure and previous candidates. Here, the latter approach is considered the more useful, since additional reference structures render the distance class partitioning method almost infeasible.

DISTANCE CLASSES AND ENERGY BARRIERS In the previous paragraph, I already made use of a measure of separation for distance classes. In particular, meta-stable structures were identified as such if their neighborhood consists of high free energy structures. But no rationale for this assumption has been provided yet. Since a meta-stable structure can be considered of biological importance if it is not too easily transformed into the MFE structure (or another meta-stable state), refolding energy barriers are of utmost importance. Energy barriers again can be extracted from (near) optimal refolding paths, but their prediction is *NP-complete* even when restricting the problem to direct paths. Consequently, to support the assumption that meta-stable candidate structures derived from the method discussed above are indeed meta-stable, additional computational effort seems to be necessary. However, a closer look at the properties of distance classes reveals that they already provide a lower bound on the refolding energy barrier. This can be easily realized by the fact that the intermediate structures of any refolding path between two structure representatives in the projection must either be a structure representatives themselves, or they have a higher free energy compared to the representative of the distance class they belong to. Therefore,

replacing the intermediate structures by their corresponding distance class MFE representative yields a lower bound on the actual refolding barrier. Any distance class partitioning can be considered a graph, where MFE representatives represent the nodes, and edges are drawn between neighboring distance classes. A shortest path algorithm can then be used to extract the lower bound of the energy barrier between any pair of representatives, see Figure 16 for an example.

11.2 THE PATHFINDER ALGORITHM

Although heuristic methods that predict optimal direct refolding paths are relatively fast, they perform poorly when the two structures are far apart. This is mainly due to the fact that the direct, i.e. shortest, paths are not necessarily the most optimal. In fact, in almost all cases where an RNA refolds from one structure into another, base pairs have to be opened before others can effectively form. Therefore, the RNA structure first has to loose stability in terms of free energy, before free energy can be gained again from newly formed structural elements. The intermediate loss determines the energy barrier that separates both structures. However, the RNA tends to compensate for the loss of free energy by forming intermediate base pairs that only exist during the transition process. Refolding paths that consider these additional base pairs are termed *indirect* paths, and lead to much more realistic estimates of the actual refolding energy barrier, cf. Chapter 4.3.

In the following, I describe a heuristic method that allows to predict near optimal (indirect) refolding paths between two secondary structures s_1 and s_2 in an iterative, flexible manner, the `Pathfinder` algorithm. In particular, I want to stress here the flexibility it provides, since the core algorithm itself was already presented, though in combination with distance class partitioning, as part of Chapter 9. Thus I will first briefly introduce the general idea of the method, before proceeding to its variations that constitute its flexibility.

THE PATHFINDER ALGORITHM The algorithm is initialized with a guess for an optimal folding path between s_1 and s_2, e.g. obtained from a direct path heuristic. Using this guess as the best refolding path known so far, the algorithm extracts the saddle point s_b, i.e. the structure with highest free energy along the path. Then it proceeds by generating a set of stabilizing points $s \in S$ in order to split the path construction problem into two direct path constructions, one from s_1 to s, and the other from s to s_2. Of course, the free energy of the stabilizing points s has to be lower than that of s_b of the best refolding path known so far, i.e. $E(s) < E(s_b)$. The method(s) how to generate S will be discussed further below. If the concatenation of both direct paths yields a saddle point with lower free energy, hence lower refolding energy barrier, it is accepted as a new optimal path, and s becomes a member of the set of optimal stabilizing points $\widehat{S}$. The number of memorized optimal stabilizing points can furthermore be limited

by a constant $|\widehat{S}| < m$ if $|S|$ becomes large. The algorithm proceeds for the remaining stabilizing points in S until all of them are tested, and returns the stabilizing point that results in best refolding path among all constructions. If required, an iterative refinement can be applied that uses each optimal stabilizing point $\hat{s} \in \widehat{S}$ to refine the direct paths connecting it to s_1 and s_2. Therefore, the `Pathfinder` algorithm is applied again on both subpaths from s_1 to $\hat{s}$, and $\hat{s}$ to s_2, respectively. The pseudocode for the Pathfinder heuristic that utilizes a *min-order heap* data structure for $\widehat{S}$ is given in Algorithm 1.

Algorithm 1 Pathfinder(s_1, s_2, i, m):
Pseudo-code of the `Pathfinder` algorithm, where s_1 is the start structure, s_2 is the stop structure of the predicted path, i the maximum number of iterations, and m the number of best solutions kept in each iteration.

```
bestpath ← DirectPath(s1, s2) /* get initial guess of an optimal refolding path */
S ← GetStabilizingPoints(bestPath) /* initialize the set of stabilizing points */
Ŝ ← ∅ /* initialize min-order stabilizing point heap */
while (|S| > 0) ∧ (|Ŝ| < m) do
   s ← pop(S) /* draw a stabilizing point and remove it from S */
   path ← DirectPath(s1, s) + DirectPath(s, s2)
   if Barrier(path) < Barrier(bestpath) then
      insert(Ŝ, (s, path, Barrier(path)))
   end if
end while
if i > 0 then
   for 0 ... m do
      (s, path) ← pop(Ŝ)
      path ← Pathfinder(s1, s, i − 1, m) + Pathfinder(s, s2, i − 1, m)
      if Barrier(path) < Barrier(bestpath) then
         bestpath ← path
      end if
   end for
else
   (s, path) ← pop(Ŝ)
   bestpath ← path
end if
return bestpath
```

So far, it has not been mentioned how stabilizing points in S are generated in the first place (see above). This is because it constitutes the flexible part of the algorithm, i.e. stabilizing point generation is independent of the path construction algorithm[1]. In Chapter 9, I used the MFE representatives of the distance class partitioning derived from taking s_1 and s_2 as reference structures. However, any other method that is able to generate alternative local minima of the secondary structure free energy landscape

1 In principle, using a fixed set of near ground state structures for S, and omitting the iterative refinement part, the `Pathfinder` algorithm is similar to the indirect path construction algorithm of Morgan and Higgs, presented in Chapter 4.3

may suffice. Boltzmann sampling from (possibly a different set of) distance classes, for instance can be used instead of using MFE representatives. Below, I introduce some additional variations for the generation of S, that constitute a computationally less expensive alternative to the distance class approach. All the presented methods have in common, that they use the saddle point s_b of the currently known best refolding path as a source for energy directed routes $\vec{P}(s_b)$ through the energy landscape. The destinations of these routes then provide structures suitable for stabilizing points of the `Pathfinder` algorithm.

GRADIENT WALKS The most straight-forward approach of an energy directed route through the secondary structure free energy landscape is to perform a *gradient walk* $\vec{P}^G(s_1)$. Starting at structure s_1, subsequent structures s_{i+1} of the route are selected from the set of neighbors $\mathcal{N}(s_i)$ of their predecessor s_i, if they constitute the neighbor with lowest free energy, i.e.

$$\begin{aligned} \vec{P}^G(s_1) &= (s_1, \ldots, s_i, s_{i+1}, \ldots, s_n) \text{ where } s_{i+1} \in \mathcal{N}(s_i) \\ &\quad \text{and} \\ E(s_{i+1}) < E(s_i) \wedge E(s_{i+1}) &= \min_{s \in \mathcal{N}(s_i)} E(s) \end{aligned} \tag{59}$$

The last structure s_n obtained by a gradient walk $\vec{P}^G(s_1)$ is finally used as supporting point in the Pathfinder algorithm. If the choice of s_{i+1} is ambiguous, one of the lowest free energy neighbors of s_i is selected randomly, for instance by taking the first in lexicographical order of the corresponding *dot-bracket* string [68]. Alternatively, the gradient walk can be branched at this point to consequently yield a set of solutions instead of a single one.

ENERGY DIRECTED RANDOM WALKS Since the previously mentioned method is de facto deterministic, and in most cases leads to only a single solution, an energy directed random walk poses a more sophisticated approach. Therefore, a Monte Carlo scheme can be applied, similar to the *Gillespie* algorithm, as discussed in Chapter 6.3. As a result, the obtained route $\vec{P}^M(s_1)$ consists of intermediate structures, where an intermediate structure s_{i+1} is selected randomly from the neighborhood of s_i according to its proportional weight among all alternative neighbors $s \in \mathcal{N}(s_i)$. Here, the weight

is determined by the free energies $E(s)$ of the structures s. Hence, the probability $p(s_i \to s)$ to accept a particular $s \in \mathcal{N}(s_i)$ as successor of s_i is

$$p(s_i \to s) = \frac{1}{\sum_{s_j \in \mathcal{N}(s_i)} \tilde{p}(s_i \to s_j)} \cdot \tilde{p}(s_i \to s) \tag{60}$$

with

$$\tilde{p}(s_i \to s) = \frac{1}{|\mathcal{N}(s_i)|} \begin{cases} e^{-\beta \cdot (E(s) - E(s_i))} & \text{if } E(s_i) < E(s) \\ 1 & \text{else.} \end{cases}$$

In order to avoid cycles in the resulting route $\vec{P}^{M}(s_1)$, previously visited structures can be penalized with a pseudo energy ε, e.g. $\varepsilon = 2 \cdot kT$, to weaken their acceptance probability. The energy directed random walk ends at structure s_n, as soon as a certain stop criterion is fulfilled, for instance the length of the route reached a size limit $|\vec{P}^{M}(s_1)| = l$. Alternatively, the random walk can be finalized if s_n represents a local minimum, i.e. $\forall s \in \mathcal{N}(s_n) : E(s) > E(s_n)$, and the probability to reject any of its neighbors

$$p^{R}(s_n) = 1 - \sum_{s \in \mathcal{N}(s_n)} \tilde{p}(s_n \to s) \tag{61}$$

surmounts a certain threshold $p^{R}(s_n) > \mu$. Accordingly, the local minimum s_n is considered deep enough with respect to its neighbors. Again, s_n is used as an adequate stabilizing point for the `Pathfinder` algorithm, and repeated realizations of the random walk may be used to produce a whole set of solutions.

SIMULATED ANNEALING Optionally, the previously described energy directed random walk can be transformed into a *simulated annealing* approach [123, 37]. Therefore, the acceptance probabilities $p(s_i \to s)$ are initially computed using a high thermodynamic temperature $T = T_{max}$, and with each Monte Carlo step the system is successively cooled down with rate $0 < r_T < 1$. Once the thermodynamic temperature falls below a minimum value $T < T_{min}$ the Monte Carlo simulation is stopped, and the last structure visited is used as stabilizing point. Of course, this stop criterion overrules those of the regular random walk without simulated annealing described above.

The acceptance probability of a neighboring structure is essentially determined by the factor $exp(-\Delta G/RT)$, where ΔG is the change in free energy between the current structure and the neighbor, R is the Gas constant and T the thermodynamic temperature of the state system. Thus, the probabilities for energetically good and poor neighbors are aligned with higher T, and, as a consequence, the simulated annealing approach allows for faster exploration of the state space compared to a regular random walk.

RESULTS

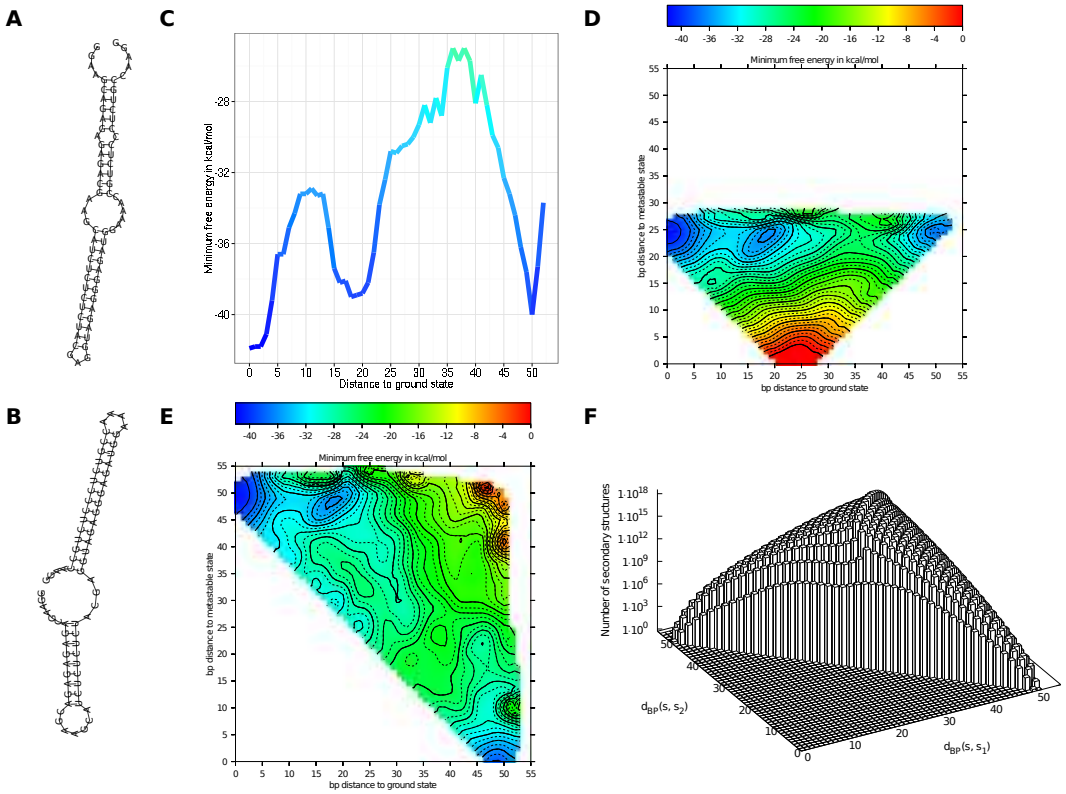


Figure 15: Distance class partitioning of secondary structure free energy landscapes, exemplarily depicted for an RNA sequence, taken from Xayaphoummine et al. [258], that was deliberately designed to adopt more than on stable low free energy structure. **A** and **B**, show both secondary structure states of interest, the MFE structure, and the meta-stable state, respectively. In **C**, a plot of the MFE representatives for a distance class partitioning with one reference structure is shown. Here, the structure with lowest free energy among the entire ensemble Ω, the *ground state*, is used as a reference. Apart from the ground state, two additional low free energy states become visible. **D** Shows the MFE representatives of a 2D projection of the landscape into distance classes with respect to two initially chosen references. For this purpose, the ground state and the completely unfolded chain are taken as reference structures. As in the case of a single reference, the auxiliary low free energy structures can be identified, though with a better resolution. **E** Substitution of the unfolded chain with the most distant meta stable state obtained from the analysis in D, reveals the separating structure of the landscape between both reference structures. **F** The number of secondary structure for each distance class derived from taking the ground state and the meta stable state as reference structures.

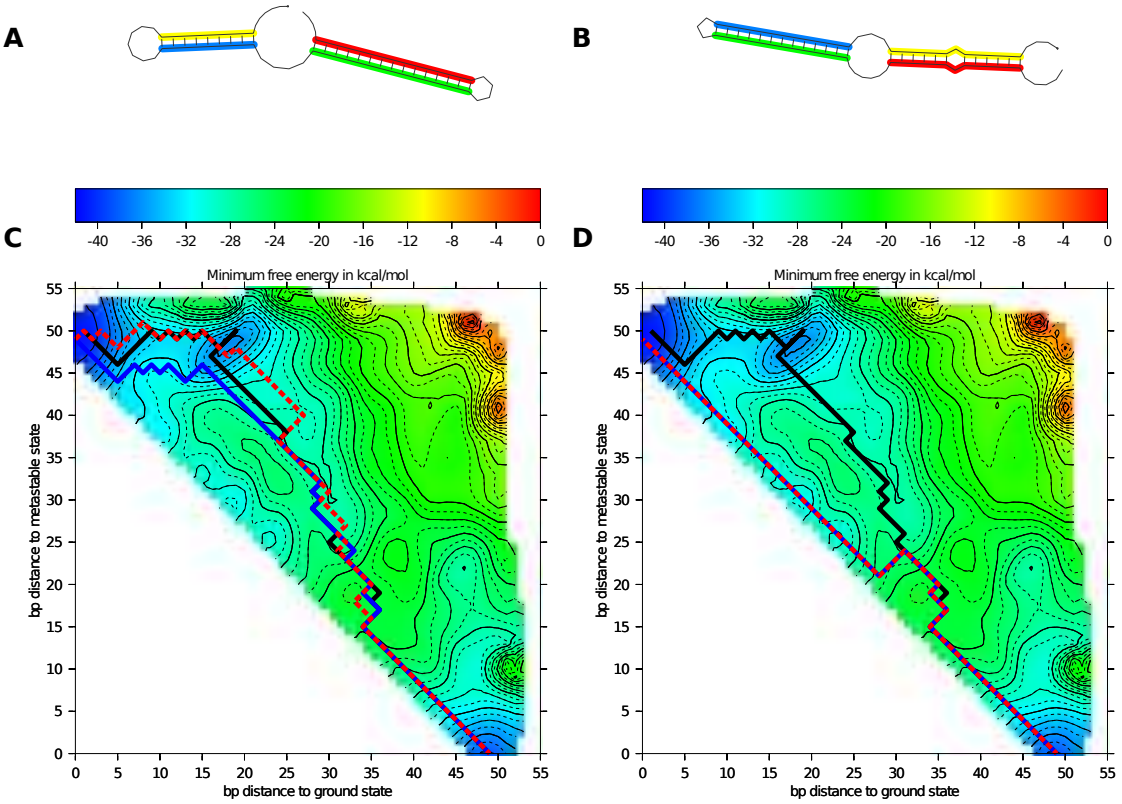


Figure 16: Barrier estimation and prediction of near optimal refolding paths. Using the same example as in Figure 15, two modes of near optimal refolding path predictions using distance class partitioning, and the more general variant of the `Pathfinder` algorithm are compared against an exhaustive method that finds the global optimum utilizing the `barriers` program [68], shown as a black line. **A** Distance class partitioning. The red dashed line is reflects the estimate by application of a shortest-path algorithm on the graph of MFE representatives. The blue line shows a near optimal path obtained from the `Pathfinder` algorithm using distance class partitioning without iterative refinement. Although the paths take distinct routes through the actual landscape, all of them correctly predict the limiting saddle point that separates both states with an energy barrier of 19.70 kcal/mol. **B** The `Pathfinder` algorithm using a energy directed random walks. Here, many iterations ($i = 50$) are necessary for the approach to find the correct saddle point of the transition. This behavior was observed by using the regular Gillespie scheme and keeping the best $s = 50$ solutions in each iteration (red curve), as well as for the simulated annealing approach, with the same number of s and an initial temperature of $T = 353.15K$, a stop temperature of 283.15, and a cooling rate of 0.99 per simulation step. Within this setting, both methods find the same near optimal refolding path.

12 FOLDING KINETICS ON COARSE GRAINED LANDSCAPES

As pointed out in Chapter 6.4, prediction of folding kinetics by a direct solution of the master equation is feasible only for systems with a low number of states. The RNA secondary structure free energy landscape is not such a system. Coarse graining approaches can be used to lump sets of related secondary structures into so called macro-states. Using sophisticated transition rates between the macro-states, the resulting partition of the huge structural ensemble becomes manageable again. Though, details may be lost due to the abstraction. Still, Wolfinger et al. [252] successfully applied a coarse graining in terms of gradient basins to enable the prediction of RNA folding kinetics by solving the corresponding master equation.

In this chapter, I present an alternative to the coarse grained RNA folding kinetics approaches discussed so far, the *2Dkin* algorithm. Utilizing the distance class partitioning of the RNA energy landscape as presented in Chapter 9, the thermodynamic properties of the distance classes will be used to effectively predict folding dynamics in the coarse grained partitioning. Thus, in contrast to many other methods, an exhaustive enumeration of structure states to generate the landscape before lumping states together is not required. The coarse graining is rather constructed *ab-initio*, where Boltzmann weights of the resulting macro states are computed via DP. Transition rates between any two neighboring macro states are then estimated from Boltzmann weighted sampling. This renders this approach applicable to RNAs with much larger sequence length, compared to existing methods. Furthermore, the distance class partitioning also allows for a straightforward application to chain growth scenarios, such as *co-transcriptional folding*, which will be discussed in Section 12.2.

12.1 KINETICS ON 2D PROJECTIONS OF THE ENERGY LANDSCAPE

Before introducing the ideas behind the `2Dkin` approach, let me recapitulate the idea of distance class partitioning into $\kappa\lambda$-neighborhoods (again). Given an RNA sequence S and two compatible reference secondary structures s_1, s_2, all structures s_i sharing the same pair distances $\kappa = d_{BP}(s_1, s_i)$ and $\lambda = d_{BP}(s_2, s_i)$ are grouped together into class $\alpha_{\kappa\lambda}$. The *RNA2Dfold* algorithm as presented in Chapter 9 can be used to effectively compute thermodynamic properties of such a partitioning. In particular, its classified DP approach allows for computation of the volume of each class α, i.e. their partition functions Q_α and corresponding free energy $G_\alpha = -1/\beta \ln Q_\alpha$. Furthermore, structural representatives can be obtained from each class α via stochastic sampling, see 3.4.

TRANSITION RATES BETWEEN DISTANCE CLASSES To obtain transition rates between any two macro states, i.e. distance classes, is the most difficult part of any coarse graining approach. As discussed in Chapters 6.1 and 6.4, the rate is essentially determined by the energy barrier along an optimal refolding path between two structures. Therefore, the simplest ansatz for the distance class partitioning is to consider the MFE representatives of each class only, and compute pairwise refolding paths between them. The saddle point, i.e. the intermediate structure with highest free energy, along the resulting paths then determines the separating energy barrier, and transition rates can be derived, e.g. from Arrhenius law. However, such an approach unavoidably neglects all other micro states within the distance class, and the multitude of transitions. Of course, the other extreme would be to compute transition rates from exhaustive enumeration of micro states. However, this would eliminate the benefit of the ab-initio partitioning and turn the approach presented below quite inefficient. Instead, I suggest to estimate transition rates from micro rates between stochastically sampled structural representatives of the distance classes and their corresponding neighbors. Therefore, I adapt the gradient basin rates computation of Wolfinger et al. [252] presented in Chapter 6.4. Given a large enough sample size, the resulting estimates are assumed to converge to the actual rates between the partitions.

In detail, for each distance class $\alpha_{\kappa\lambda}$, a set of representative structures $S_{\alpha_{\kappa\lambda}}$ is sampled according to their equilibrium probabilities. For the micro states $s_i \in \alpha_{\kappa\lambda}$ in the sample set, transition rates k_{ij} to their individual neighbors $s_j \in \mathcal{N}(s_i)$ are computed. Using a move set that consists of formation and dissociation of a single base pair, we again use the Metropolis rule for micro rates computation (see also Chapter 6.1)

$$k_{ij} = \begin{cases} e^{-\beta(E(s_j)-E(s_i))} & \text{if } E(s_i) < E(s_j) \\ 1 & \text{otherwise.} \end{cases} \tag{62}$$

Because each neighbor $s_j \in \mathcal{N}(s_i)$ differs from $s_i \in \alpha_{\kappa\lambda}$ in exactly one base pair, it is immediately clear which distance class it belongs to. For instance, if s_j originated from s_i by the insertion of a base pair (p, q) there are only two scenarios regarding its distance κ to reference structure s_1 compared to that of s_i. If $(p, q) \in s_1$ s_j is 'one step closer' to s_1, hence compared to s_i its κ is smaller by 1. Otherwise, the insertion of base pair $(p, q) \notin s_1$ increases κ by 1. The same is true for removal of a base pair, and the distance λ to the second reference s_2. Therefore, each structure $s_i \in \alpha_{\kappa\lambda}$ has at most 4 neighboring distance classes $\alpha_{\kappa\pm1\lambda\pm1}$ where its structural neighbors $s_j \in \mathcal{N}(s_i)$ belong to. For the sake of readability of the equations, I reduce the $\kappa\lambda$ subscript of distance classes $\alpha_{\kappa\lambda}$ to a single value i, i.e. α_i, below. Furthermore, the notion $\mathfrak{N}(\alpha_i)$ will be used to represent the set of macro states α_j in the direct neighborhood of macro state α, i.e.

$$\mathfrak{N}(\alpha_{\kappa\lambda}) = \left\{\alpha_{\kappa'\lambda'} \mid \kappa' = \kappa \pm 1 \wedge \lambda' = \lambda \pm 1\right\} \tag{63}$$

Using the ideas presented in Chapter 6.4, the actual *macro* rate from distance class α_1 to one of its neighbors $\alpha_2 \in \mathfrak{N}(\alpha_1)$ can be written as

$$k_{\alpha_1,\alpha_2} = \sum_{s_i \in \alpha_1} \text{Prob}[s_i|\alpha_1] \cdot \sum_{\substack{s_j \in \alpha_2 \\ s_j \in \mathcal{N}(s_i)}} k_{ij}. \tag{64}$$

Here, $\text{Prob}[s_i|\alpha_1]$ is the probability to observe micro state s_i given the system is in macro state α_1. Unfortunately, this conditional probability can not be easily computed for arbitrary conditions. However, assuming equilibrium within each macro state, this probability is known to be $\exp(-\beta E(s_i))/Q_{\alpha_1}$, with the partition function of the macro state $Q_{\alpha_1} = \sum_{s \in \alpha_1} \exp(-\beta E(s))$, cf. Chapter 3.4. Of course, this assumption may be problematic, especially for the distance class partitioning. There are no direct neighbors s_j of a micro state $s_i \in \alpha$ that reside in α themselves. Still, the structural relationship between micro states taken from distance classes with low-numbered $\kappa\lambda$ pairs can be considered close, thus equilibrium may be assumed within clusters of neighboring distance classes.

Nevertheless, Equation (64) assumes an exhaustive summation over all members of α_1. Thus, some modifications are necessary to consider the samples $s_i \in S_{\alpha_1}$ drawn from stochastic backtracking within α_1 instead. The first key observation is that each structure $s_i \in S_{\alpha_1}$ was already sampled according to its equilibrium probability within macro state α_1. Therefore, a substitution of α_1 with its corresponding sample set S_{α_1} leads to

$$\widehat{k}_{\alpha_1,\alpha_2} \approx \frac{1}{|S_{\alpha_1}|} \sum_{s_i \in S_{\alpha_1}} \sum_{\substack{s_j \in \alpha_2 \\ s_j \in \mathcal{N}(s_i)}} k_{ij} \tag{65}$$

Of course, the quality of the resulting macro rate $\widehat{k}_{\alpha_1,\alpha_2}$ now heavily depends on the quality and size of the sample set S_α. But, regardless of the sample size, Equation (65) still lacks an important feature for successfully modeling the macro state kinetics of the system, namely *reversibility*. On average, the rate $\widehat{k}_{\alpha_1,\alpha_2}$ results in a good approximation, but sampling errors can easily break detailed balance. This becomes clear if the macro rate from α_1 to its neighbor $\alpha_2 \in \mathfrak{N}(\alpha_1)$ is formulated in terms of S_{α_2}

$$\widetilde{k}_{\alpha_1,\alpha_2} \approx \frac{1}{|S_{\alpha_2}|} \sum_{s_j \in S_{\alpha_2}} \sum_{\substack{s_i \in \alpha_1 \\ s_i \in \mathcal{N}(s_j)}} \frac{\pi_{\alpha_2}}{\pi_{\alpha_1}} k_{ji} \tag{66}$$

where π_{α_x} is the equilibrium probability of the macro state α_x and k_{ji} the micro rate from s_j to s_i. Since both sample sets S_{α_1} and S_{α_2} are independent from each other, it is not guaranteed that for any $s_i \in S_{\alpha_1}$, its neighbor $s_j \in \alpha_2$ is within S_{α_2}, too, and vice versa.

A simple, yet effective, solution is to take the weighted average of both rate approximations to derive the transition rates from the union of both sample sets $S_{\alpha_1} \cup S_{\alpha_2}$.

Consequently, any pair of structures $s_i \in S_{\alpha_1}$ and $s_j \in \alpha_2$ is used to compute the back and forth transition rate, and detailed balance is restored. The total transition rate k_{α_1,α_2} between any two macro states α_1 and α_2 of the (κ,λ)-partitioning can finally be written as

$$\begin{aligned} k_{\alpha_1,\alpha_2} &\approx \frac{|S_{\alpha_1}| \cdot \widehat{k}_{\alpha_1,\alpha_2} + |S_{\alpha_2}| \cdot \widetilde{k}_{\alpha_1,\alpha_2}}{|S_{\alpha_1}| + |S_{\alpha_2}|} \\ &\approx \frac{1}{|S_{\alpha_1}| + |S_{\alpha_2}|} \left\{ \sum_{s_i \in S_{\alpha_1}} \sum_{\substack{s_j \in \alpha_2 \\ s_j \in \mathcal{N}(s_i)}} k_{ij} + \sum_{s_j \in S_{\alpha_2}} \sum_{\substack{s_i \in \alpha_1 \\ s_i \in \mathcal{N}(s_j)}} \frac{\pi_{\alpha_2}}{\pi_{\alpha_1}} k_{ji} \right\} \end{aligned} \tag{67}$$

with

$$k_{\alpha_1,\alpha_1} = - \sum_{\alpha_i \neq \alpha_1} k_{\alpha_1,\alpha_i}.$$

COMPARISON TO GRADIENT BASIN KINETICS Comparison of folding kinetics predictions between two different coarse graining approaches can be rather difficult. While the qualitative picture of the corresponding simulations definitely should agree with each other, the actual population densities of the macro states, i.e. the quantitative picture, may be very distinct. This is due to the different partitionings of the landscape. Some macro states that subsume a large volume of the secondary structure free energy landscape in one approach may correspond to such with small volumes in the other, and vice versa. To test the predictive power and potential of the `2Dkin` approach, I exemplarily use an artificially designed RNA with distinct low free energy states from Xayaphoummine et al. [258]. Starting with an initial population density of 100% for 3 local minima of the state space, a folding kinetics simulation is performed with the well established *Gradient basin* approach of Wolfinger et al. [252], and `2Dkin`. Results are depicted in Figure 17. As can be seen by comparing the barrier tree with the 2D projection, one deep local minimum is missing in the distance class approach. This is due to the fact that the ensemble consists of two structures similar to the MFE structure but with a 3 nt shift in the 5′ stem. Coincidentally, both are members of the same distance class. Thus, only the one with lower free energy is visible in the 2D projection. However, as can be seen in the folding simulations, both methods compare well. The same is true for other examples that are small enough to be feasible with the Gradient basin method (data not shown).

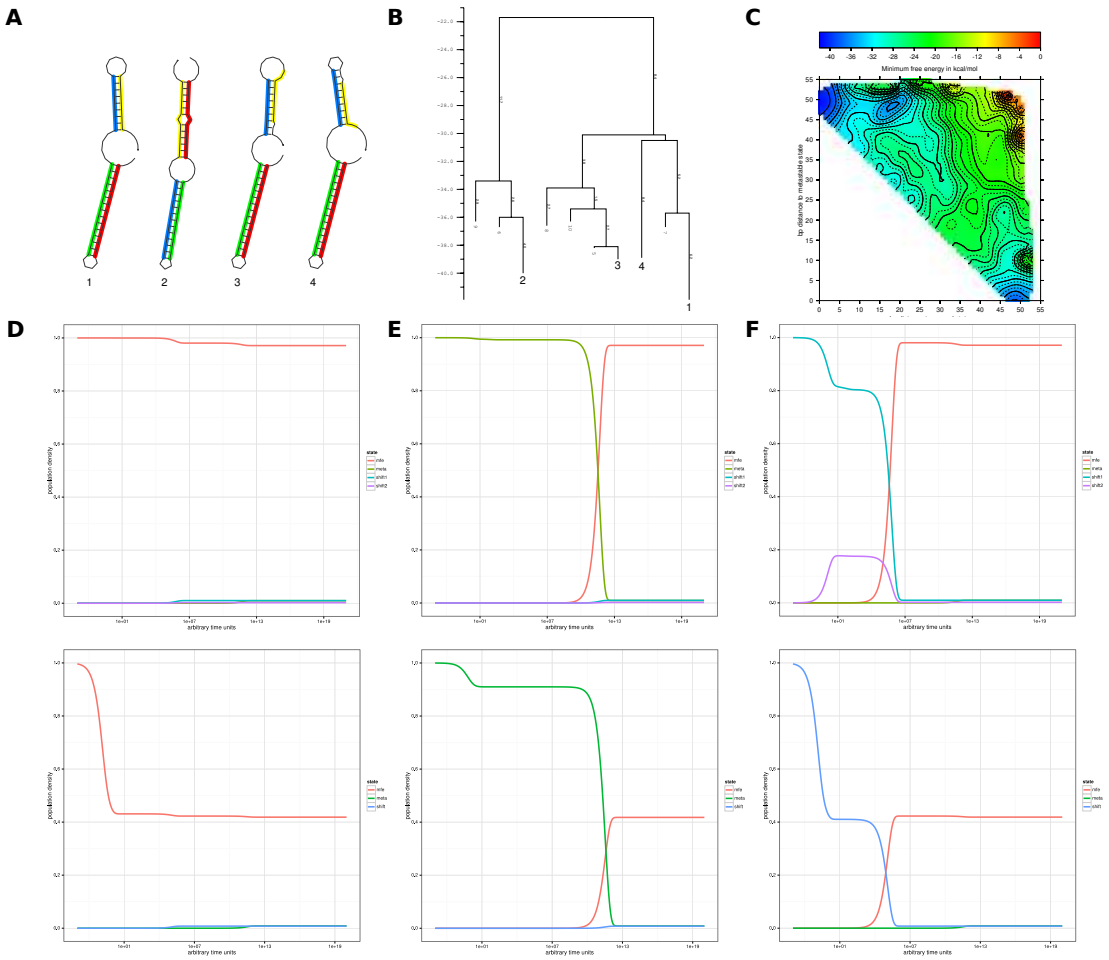


Figure 17: Folding dynamics prediction using distance class partitioning of the `2Dkin` algorithm, compared to the well established *Gradient basin* approach of Wolfinger et al. [252]. The RNA sequence is taken from Xayaphoummine et al. [258]. **A** Secondary structure plots of (1) the MFE structure, (2) the meta stable state that consitutes a cotranscriptionally active kinetic trap, and (3,4) both shifted versions of the MFE structure (see text). **B** The *barrier tree* representation of the corresponding landscape. **C** The distance class representation of the landscape. **D-E** Simulation results for the Gradient basin approach versus the `2Dkin` method starting with an initial population density of 100% for the MFE structure (**D**), the metastable state (**E**), and finally the shifted MFE structure (**F**), respectively. Red, green, and blue curve depict the population density of distance classes containing the MFE structure, the meta stable state, and the shifted MFE structure, respectively. Although quantitatively different, both methods compare very well in terms of quality.

12.2 FOLDING KINETICS ON VARYING LANDSCAPES

As shown in the previous section, the distance class partitioning can be used to efficiently predict coarse grained RNA folding kinetics. Nonetheless, the approaches presented so far only consider static secondary structure energy landscapes, i.e. the landscape does not change. Though, in their native environment an RNA certainly undergoes specific changes that reshapes its corresponding secondary structure energy landscape. The most prominent example of such a change that even accounts for all RNAs produced in any living cell is its fabrication, better known as *Transcription*. Here, the RNA immediately starts folding back on itself as soon as it leaves the transcription bubble. Each nucleotide that is added to the nascent RNA results in a growth of the set of compatible secondary structures Ω. Hence, the induced landscape grows with any nucleotide added as well. The same is true for processes where the RNA is degraded. Any nucleotide removed from the polymers ends removes certain states in the landscape. Another example is the change of the environmental temperature. Since the free energy $\Delta G = \Delta H - T\Delta S$ of a loop in a secondary structure consists of an enthalpic part H and a temperature dependent entropic part S, the free energy $E(s)$ of any entire secondary structure s changes with temperature as well, lifting or deepening its corresponding spot within the landscape. Further examples of environmental changes that induce a change in the underlying energy landscape of an RNA include the binding of small molecules which stabilize a particular structure or structural motif, or the force of unpairedness of certain stretches of the RNA, e.g. caused by the elongation of the ribosome during translation.

Computational tools for the above mentioned scenarios are very limited. This may be due to the fact that even for the analysis and prediction of RNA folding kinetics on static landscapes, only a handful of computational methods exist (reviewed in [65, 125]). Nevertheless, to date, a few implementations of algorithms to predict the *co-transcriptional folding* of an RNA are available [257, 77, 182]. Additionally, some of the Monte-Carlo scheme based RNA folding kinetics approaches explicitly included chain growth in their algorithms by restricting the set of possible transitions according to their availability during RNA polymerization (see Chapter 6.3). However, these methods are limited since they either provide only a single co-transcriptional folding trajectory, or require a large amount of repeated executions of the simulation to achieve an over-all picture of the co-transcriptional folding process.

A generalized method that allows the prediction of RNA folding kinetics on varying energy landscapes was presented by Hofacker et al. in 2010 [106]. For each discrete environmental setting, their *BarMap* approach generates a corresponding energy landscape, or a coarse grained partition thereof. Transition rate matrices are then computed for each of these settings to allow for a direct solution of the master equation (see Chapter 6.2). Furthermore, a sophisticated mapping function $\mathbf{M}$ is used to map states between consecutive energy landscapes. The actual folding simulation for the

dynamic energy landscape finally is constructed of short simulations on the individual landscapes, where the initial conditions of a simulation are derived from the final conditions of its predecessor utilizing **M**. Still, the BarMap approach relies on exhaustive enumeration of the energy landscapes, and is therefore applicable only to RNAs of short sequence length.

In the following, I present two new methods that approach the problem of *co-transcriptional folding* using distance class partitioning. While the first provides a static view on the growing transcript in terms of MFE predictions, the second approach is based on the ideas in the previous section 12 and allows for the prediction of RNA folding kinetics of the growing transcript.

CO-TRANSCRIPTIONAL FOLDING To predict the secondary structure an RNA adopts after being released from the transcription bubble is a difficult task. It requires to consider the interplay between transcript extension by the RNA polymerase and the refolding kinetics of the nascent transcript towards its (new) equilibrium state. If a small energy barrier separates a structure from the ground state of the partial transcript, refolding is fast and the transcript will permanently stay in equilibrium. In this case, a regular MFE structure prediction would be sufficient. For high barriers, on the other hand, refolding may take much longer than the elongation of the polymerase, and the RNA gets trapped in a non-equilibrium state. As a consequence, the transcript can reside in a non-equilibrium state for timespans that easily exceed that of a typical RNA molecule in the cell, which is in the range of 1 minute to 1 hour. Therefore, MFE structure prediction leads to an erroneous result and a potential function of the RNA can be misinterpreted.

Since RNA secondary structure folding kinetics requires much more computational effort compared to a DP secondary structure prediction, several methods exist that try to circumvent the kinetics part with clever heuristics. The *CoFold* algorithm, for instance, uses Zuker's algorithm for MFE structure prediction, but penalizes long range base pairs with a pseudo energy term that grows logarithmically with the distance of the nucleotides involved [182]. This idea is based on the fact that hairpin helices form first, due to their close proximity of base pairs. Many of these short-range base pairs may then be kinetically trapped in local structures, and are therefore not available for long-range interactions in a more complex structure. However, to date a wide variety of functionally important long-range interactions are known, hence such a simple modification of the energy model appears to be at odds with that [8]. A more elaborate approach was presented with the *Kinwalker* algorithm in 2008 by Geis et al. [77]. Utilizing the DP matrices of Zuker's MFE algorithm, their method constructs a folding trajectory of intermediate structures. Therefore, optimal helices obtained from partial backtracking in the MFE matrices are inserted in structure intermediates if (i) they decrease the structures free energy, and (ii) the involved refolding energy barrier allows them to form before more favorable helix alternatives appear due to transcript

elongation. Furthermore, the algorithm extends helical regions whenever possible. Energy barriers are estimated from direct refolding paths between an intermediate structure and the next structure candidate. This renders *Kinwalker* a fast tool applicable to RNAs with sequences length of up to about 1000 nt. However, this heuristic has its drawbacks, too. When the algorithm constructs more complex intermediate structures, backtracking in the MFE matrices is likely to yield structural elements which induce high refolding barriers, although suboptimal alternatives with a much lower barrier might exist. But these alternatives can not be easily generated from the MFE matrices. Additionally, the usage of direct refolding paths for energy barrier estimation tends to overestimate the actual refolding barrier (see Chapter 4.1). Hence, the derived refolding time gets overestimated as well, and substructures are not incorporated into the intermediate structure although they could be.

Here I suggest an alternative to the above approaches, based on classified DP MFE predictions. Simple MFE predictions for each step of the consecutively growing transcript per se do not reveal much insight, except for a consistent decrease in free energy (see Figure 18A) and possibly a very diverse set of intermediate structures. A classified DP approach with distance class partitioning, on the other hand, can help to follow low free energy structure representatives in the energy landscape while the transcript grows. For the above purposes, the (classified) DP matrices have to be filled only once, using the full length sequence. Predictions for the shorter intermediate subsequences can then be readily obtained from optimal subsolutions. As shown in Figure 18B, using a single reference structure already reveals alternative low free energy states. Though, it is almost impossible to infer likely refolding events between the developing distance classes. However, the addition of another reference structure already yields a pretty good picture of which distance classes might be populated the most throughout transcription.

A simple inspection of the MFEs for each distance class in the 2D projection may be sufficient, see Figure 18C, since the distance class partitioning already yields a lower bound on the refolding energy barrier between any two partitions, and the assignment of distance classes between two consecutive transcription steps is straight-forward. With full details, the mapping $\mathbf{M}$ of partitions $\alpha_{\kappa,\lambda}$ obtained for a transcript of length i onto partitions for the transcript of length $i+1$ is

$$\begin{aligned} \mathbf{M}_{i\to i+1}(\alpha_{\kappa,\lambda}) &= \alpha_{\kappa+\delta_1,\lambda+\delta_2} \\ &\text{with} \\ \delta_x &= s_x[1:i+1] - s_x[1:i] \end{aligned} \tag{68}$$

where $s_x[1:i]$ is the set of base pairs (p,q) in reference structure s_x with $1 \leqslant p < q \leqslant i$. A stack of resulting consecutive MFE predictions of the distance classes can also conveniently be visualized in form of a video.

CO-TRANSCRIPTIONAL FOLDING KINETICS As shown in the previously described method, the mapping between distance class partitions of a growing transcript can be obtained simply from counting base pairs in parts of the reference structure. This observation can be employed to extend the RNA folding kinetics approach presented in Section 12 in order to take a growing RNA chain into account. Therefore, the classified DP matrices only need to be filled once for the full length transcript. This yields all partition functions of the partial transcripts as well, hence all requirements for Boltzmann sampling of substructures are already set. The only computational overhead to extend the RNA folding kinetics approach towards co-transcriptional folding is the actual sampling procedure to obtain transition rate matrices of the intermediate distance class partitionings. Folding kinetics upon chain growth is then predicted according to the BarMap approach [106]. Starting with a short transcript, the system is initialized with 100% population density for the unfolded state, and the simulation is performed for the equivalent time the RNA polymerase needs to add a further nucleotide. The population densities of the distance classes derived from a preceding simulation are then mapped to the next partitioning to provide a reasonable start condition. Once the full length transcript is reached, no further mapping is required and the system can be simulated until it eventually reaches equilibrium.

Exemplarily, I show different folding dynamics predictions for an artificially designed short RNA sequence of 73 nt length [258] in Figure 19. This RNA was specifically created and experimentally validated to encode a particular cotranscriptional folding path that leads into a *kinetic trap*, i.e. a meta stable non-equilibrium structure that is structurally very different to the MFE structure. The cotranscriptional folding simulation using `2Dkin` was performed with constant transcription speed, and $1,000$ samples were drawn from each $\kappa\lambda$-neighborhood. Mapping of the arbitrary time units (τ) that are introduced by solving the *master equation* to physical time scales is difficult, and needs to be calibrated against experimental data. Therefore, the simulation was repeated with 3 different transcription speeds of 10^0 nt/τ, 10^{-1} nt/τ, and 10^{-2} nt/τ, respectively. For comparison, the simulation was also repeated for the full length transcript, with the complete initial population $\vec{p}(0)$ being concentrated in the distance class that subsumes the *unfolded state*. An analysis of the folding dynamics starting with the MFE structure, and the meta stable structure that constitutes the kinetic trap for cotranscriptional folding, has been shown above in Figure 17, accompanied by the corresponding 2D projection of the secondary structure free energy landscape.

The cotranscriptional version of *2Dkin* correctly predicts the kinetic trap for relatively slow transcription speed up to 10^{-2} nt/τ. Here, the meta stable state appears as a long-lived non-equilibrium state. The strong dependence on the transcription speed was expected, since the formation of the central hairpin helix of the structure that constitutes the kinetic trap involves exchange of base pairs with other, initially forming helices. The time required to exchange the base pairs has therefore be taken into account. On the other hand, superslow transcription speeds would of course not reveal

the kinetic trap, as the RNA would have enough time to adopt its current equilibrium state within each transcription step (data not shown).

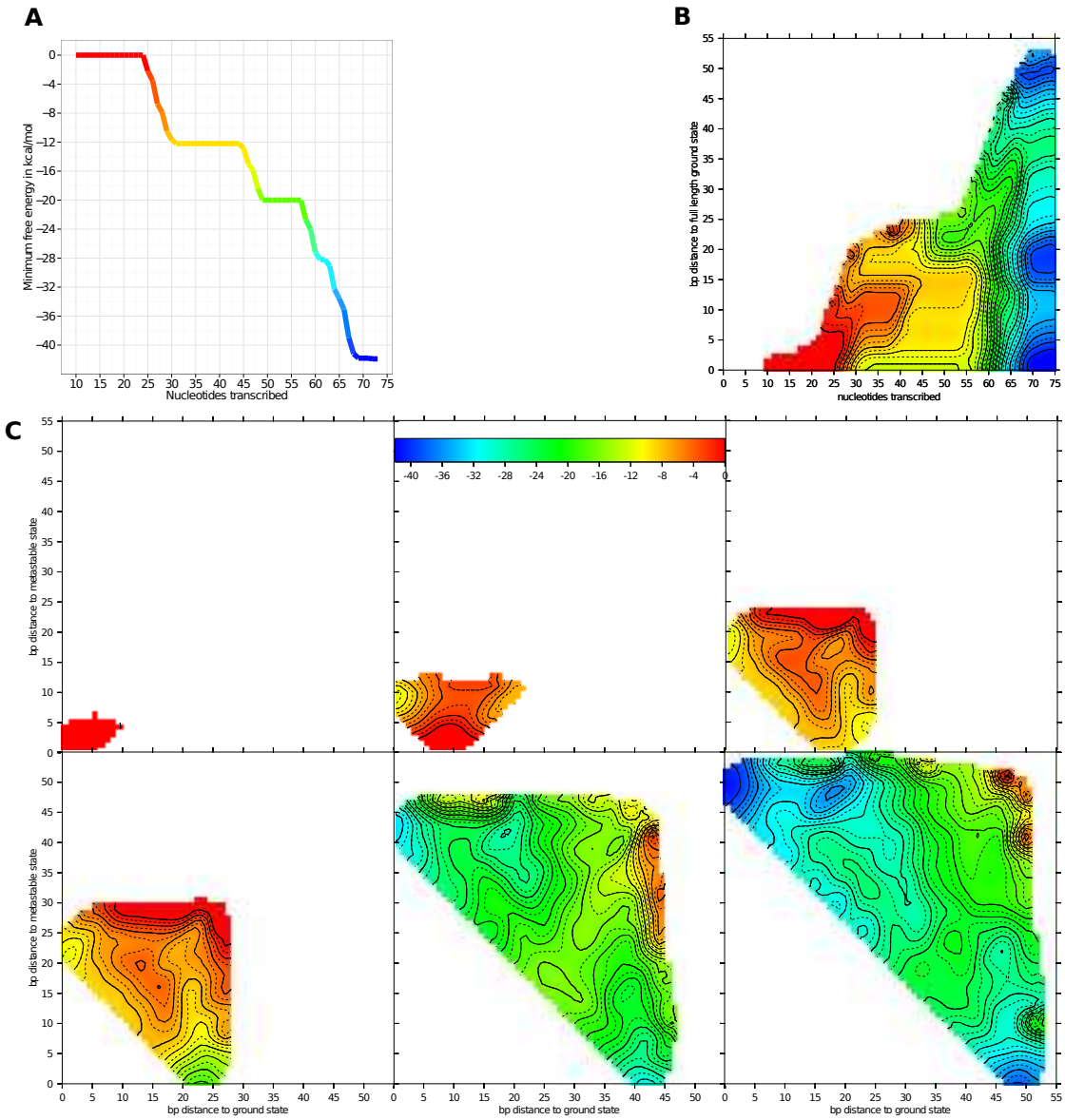


Figure 18: Cotranscriptional folding and the potential of distance class partitioning to detect the kinetically trapped non-equilibrium state of the artificially designed RNA from Xayaphoummine et al. [258]. While regular MFE prediction for each stage of the transcription process (**A**) only shows the obvious drop in MFE, distance class partitioning with respect to the full length MFE structure (**B**) already reveals at least two further low free energy states and some energy barriers separating them. As shown with the sequence of six transcription stages in steps of 10 nt starting with a partial transcript of length 25 nt, the cotranscriptional folding path can be deduced from visual inspection. While initially (35 nt) the most probable structure is very close to the full length MFE structure, an alternative state closer to the kinetic trap emerges (45 nt). At this time the separating energy barrier between both local minima is relatively small with just about 4.9 kcal/mol. Thus, rapid exchange between both local minima can be assumed. After further 10 nt of transcription, the kinetic trap is energetically much more favorable than any other alternative. The longer the transcript becomes, the higher the energy barrier that separates the kinetic trap from the rest of the structure space (65 nt). Finally, the full length transcript gives rise to the actual equilibrium structure that is, however, energetically well separated from the kinetic trap.

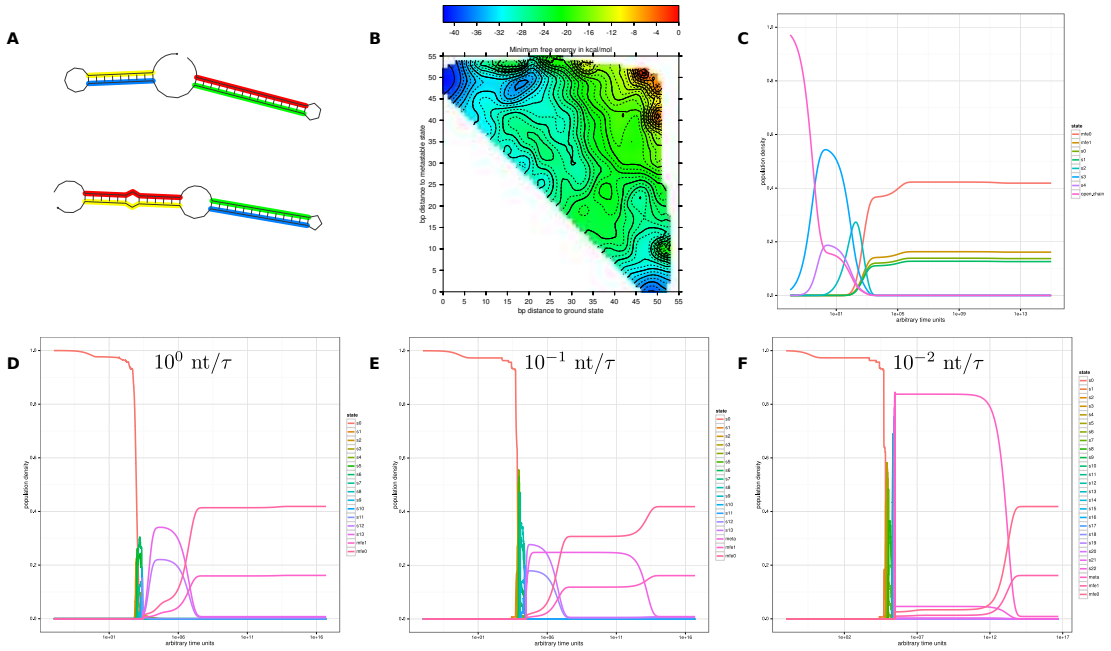


Figure 19: Cotranscriptional folding kinetics using `2Dkin`, exemplarily shown for an artificially designed RNA with encoded kinetic trap taken from Xayaphoummine et al. [258]. **A** The secondary structures of interest: MFE structure (top), and meta stable state that is adopted right after transcription (bottom). **B** `2Dkin` simulation of the full length transcript. Starting with 100% population density for the distance class that subsumes the *unfolded state*.
The remaining plots show the cotranscriptional folding kinetics as predicted with `2Dkin` for transcription speeds of 10^0 nt/τ (**D**), 10^{-1} nt/τ (**E**), and 10^{-2} nt/τ (**F**), respectively. The meta stable state that consitutes the kinetic trap becomes clearly visible as a very long-lived non-equilibrium state in the last two simulations (see also text).

Part VI

DISCUSSION

13 CONCLUSIONS AND OUTLOOK

The long term goal of computational RNA structure prediction is to deduce three dimensional structures and tertiary interactions, to realistically test biological and biochemical hypotheses. Current methods are either very demanding in terms of computational complexity, or do not provide sufficient parametrization yet [175, 203, 94, 263]. Since RNA folding is generally considered a hierarchical process, where canonical base pairs form first and tertiary interactions appear later, most of tertiary structure prediction approaches still depend on secondary structure predictions. Since it is unlikely that algorithms appear that deduce the 3D conformation of RNAs directly from the primary structure, i.e. the RNA sequence, it remains crucial to provide adequate and reliable tools for the prediction of secondary structures.

In this thesis I presented a broad overview of stat-of-the-art DP algorithms for prediction of both RNA secondary structure and folding dynamics. The related work done by me and my colleagues was extensively described in Parts iv and v. This includes the reimplementation of major parts of the `ViennaRNA Package` [137], which is a fast and comprehensive toolbox for almost all secondary structure prediction related problems.

Furthermore, the development of the classified DP algorithm `RNA2Dfold` [136], opens up a variety of perspectives on future research, that would otherwise be infeasible. Its application to predict near optimal *refolding paths*, and therefore near optimal estimations *energy barriers* is described in much detail in Chapter 11. Additionally, the potential to use *ab initio* distance class partitioning to predict coarse grained RNA folding dynamics was discussed in the previous Chapter 12. This allows *in silico* investigation of conformational changes of RNAs with biologically meaningful sequence lengths of up to 400 nt. Here, not only full length transcripts are considered, but the growing chain of the nascent RNA during transcription is taken into account as well.

The work on RNA/DNA hybrid structure prediction, as presented in Chapter 8, opens up new vistas for the understanding of the interplay between RNA and DNA. Previous approaches were limited to predict (hybrid) structures for only a single species of nucleic acid. With the extension of the nucleotide and base pair sets of state-of-the-art secondary structure prediction DP algorithms, this new method allows one to predict co-transcriptionally occurring RNA/DNA hybrids known as *R-loops*. First thought of as by-products of the transcription process, these hybrid structures appear to be involved in DNA damage and genome instability [3]. The incorporation of all types of intramolecular and intermolecular base pairs between RNA and DNA can be used to develop elaborate simulation tools modeling the transcription process

itself. For that purpose, the transcription speed can be controlled through intrinsic properties of RNA/DNA intermolecular base pairs, taking the whole interplay between intramolecular and intermolecular pairs into account, whereas today's methods simply assume a constant speed of RNA chain elongation [66, 114, 257, 77].

With the advent of tools, such as the *MC-fold/MC-sym* pipeline [175] and *RNAwolf* [263], secondary structure prediction became more and more intertwined with tertiary structure prediction. Thus, any improvement made in secondary structure prediction leads to a beneficial contribution to predict tertiary interactions. On the other hand, some ideas taken from 3D approaches can be readily incorporated into existing secondary structure prediction methods, as presented for the non-canonical *G-Quadruplex* motif in Chapter 10.

Below, some of the most urgent remaining problems concerning structure prediction and RNA folding dynamics will be discussed.

13.1 THE FUTURE OF SECONDARY STRUCTURE PREDICTION

TERTIARY STRUCTURE MOTIFS As mentioned in Chapter 3.8, RNA secondary structure prediction is by definition very much limited in terms of predictive power. This is mainly due to the fact that it neglects loops that arise through *non-canonical* base pairing, although they are present in almost every native tertiary RNA structure and can make up about 1/3 of all base pairs [129]. However, the extension of existing secondary structure prediction algorithms to non-canonical base pairs is not straightforward. Some of them require distinct spatial orientations of the nucleobases, that are only possible if the nucleotides in their vicinity are arranged properly. But spatial properties are omitted from secondary structure prediction algorithms entirely. Recently developed algorithms that aim to predict non-canonical base pairs therefore abstract from small loops up to a size of 4 that may contain base triples and are enclosed by base pairs of any kind to entire modules of non-canonical motifs (NCM) [175, 263, 221]. These modules are then combined into more complex structures, analogous to loops in secondary structure prediction algorithms. Therefore, such algorithms can be considered 2.5 dimensional structure prediction algorithms, as the results are somewhere between secondary (2D) and tertiary (3D) structures. The bonus of such an approach is that it can be easily extended to variations of the secondary structure MFE algorithm, such as RNA-RNA interaction prediction, suboptimal structure prediction, and consensus structure prediction.

Unfortunately, the parametrization of a scoring function for the individual motifs is very difficult. Only a very limited set of actual free energies of tertiary structure motifs is available [205, 261, 100, 231]. Therefore, these algorithms rely on statistical properties drawn from the frequency to observe a motif within a given context, e.g. in available 3D structures taken from various databases such as the PDB [17]. Although for some examples this so-called *extended secondary structure model* increases prediction

accuracies, it still performs unsatisfactory for the general case [175, 263]. Nevertheless, by taking structure probing data, such as *SHAPE reactivity* [157], or constraints of already formed canonical pairs as additional input, the prediction accuracies can be increased substantially. For the latter non-canonical motifs have to be predicted in unpaired regions of the structure constraint only, while probing data allows to add favorable(unfavorable) scores to motif configurations that agree(disagree) with the particular probing procedure. This method has already been successfully applied to regular secondary structure prediction algorithms [151, 49, 239].

While presumably some time needs to pass until a larger set of training data for a proper parametrization of the extended secondary structure model is available, conservative methods that include only a limited number of non-canonical tertiary structure motifs into regular secondary structure prediction methods may be a good choice. This method already lead to various specifications of secondary structure loops, such as hairpins of size 3, 4, and 6, and small interior loops. The energy parameters utilized in today's secondary structure prediction algorithms explicitly tabulate the sequence variations of these small loops to assign energy contributions that do not fit the simple loop model [228, 39].

The inclusion of motifs as completely separate self-contained loop was already presented for the *G-Quadruplex* motif in Chapter 10. Although this motif is much larger than the NCMs of the extended secondary structure model, thermodynamic data from melting experiments is available that allows to model its free energy contribution in terms of layer size and linker length [135]. The same approach might be possible for other self-contained, i.e. local, tertiary motifs with a relatively low variation in the nucleotide sequence, such as *C-loops*, *T-loops*, and *kink-turns* [205, 131, 262, 248, 180, 221]. However, to date only a handful of thermodynamic parameters for tertiary structure motifs are available [156, 205, 261, 100, 231], which hopefully will be extended in the near future.

BASE PAIRS WITH EXTENDED NUCLEOTIDE ALPHABET An additional difficulty for RNA structure prediction poses the existence of *non-standard* nucleotides. It is well known that structural RNAs may contain modified nucleotides, created by post-transcriptional processing such as *RNA editing* and *pseudouridylation*, that usually come with a distinct pairing behavior [61, 118, 143, 76]. These modifications extend the RNA alphabet of the standard nucleotides $\Sigma = \{A,G,C,U\}$, and give rise to a large variety of new (non-)canonical base pairs. However, most implementations of secondary structure prediction algorithms simply ignore modified nucleotides and treat them as special nucleotides that stay unpaired in all decompositions. This, of course impairs the outcome of the prediction, since these nucleotides may very well form base pairs with other nucleotides. On the other hand, if the input data lacks an indication of a nucleotide modification the outcome of a secondary structure prediction may also

be heavily distorted, for instance a modified U often found in tRNAs is *pseudouridine*, which is known to form base pairs with any other nucleotide A, G, C, and U [122].

Since thermodynamic parameters for some of them emerged within the past years [250, 254, 112, 122], it might be worthwhile to include these additional base pairs in recent secondary structure prediction algorithms to augment the prediction accuracy for RNA sequences with known nucleotide modifications. In Chapter 8 I already presented an extension for RNA structure prediction algorithms alphabet Σ. Here, the additional nucleotides originate from the DNA alphabet, and the loop type dependent energy contributions of all possible base pairs, RNA-RNA, RNA-DNA, and DNA-DNA are distinguished and tabulated. The same straightforward modification to existing implementations can be used to incorporate base pairs involving modified RNA nucleotides, provided energy parametrization is possible.

THE ENVIRONMENT MATTERS A large simplification made by the nearest neighbor model is the independence of environmental ion concentrations. However, each phosphate group of the backbone of an RNA is negatively charged, rendering it a *poly-anion*. Therefore, positively charged *cations*, such as Na^{+} and Mg^{2+}, can easily bind to the RNA and stabilize its structure.The importance of this stabilizing effect has especially been shown for pseudo-knots, and G-quadruplexes. Still, cations are capable to stabilize regular RNA helices as well [59, 253, 216, 142].

Several models have been proposed [59, 60, 215, 216, 142] that explicitly take the concentrations of certain cations into account. But they are usually used to reassess the energy contribution of a particular secondary structure, instead to predict structures. Although, cation dependence of the stability for self-contained structural motifs such as G-Quadruplexes is local and can therefore directly be incorporated into prediction algorithms, the models propose that the influence on stacks of canonical base pairs depends on the size of the entire helix they are part of. This information, however, gets lost in the recursive decomposition scheme of the nearest neighbor model. Thus, it remains a challenge to develop a model that allows for an adequate estimation of ion concentration dependence while being compatible with the decomposition scheme of secondary structure DP algorithms.

Moreover, in the cellular context, RNAs are not only accompanied by ions in solution. There is a huge amount of other small and large molecules that can bind to specific RNA sequence and/or structure motifs, effectively distorting the underlying secondary structure energy landscape [82, 198, 181]. Bacterial riboswitches, for instance, control gene expression transcriptionally and posttranscriptonally by 'sensing' small ligand molecules within the cell [198, 183]. These ligands usually bind to specific structural 'pockets' that mostly consist of well-defined sequence motifs, and thereby stabilize loop motifs such as hairpins, interior loops, or multi loops [102]. Despite the availability of thermodynamic parameters that describe the ligand contributions to the overall structural stability, these external factors are virtually always omitted from

structure prediction algorithms. Therefore, accurate prediction of meta-stable states for riboswitches is impaired, as the stability of the corresponding structure motifs may be drastically underestimated [86, 82].

HARD AND SOFT CONSTRAINTS In principle, the nearest neighbor energy model can be adapted to situations where ligands with well-defined stabilizing contributions bind to specific sequence-structure motifs. As long as binding is local, the involved loops can be assigned an auxiliary (pseudo)energy that reflects the stabilizing effect. This method can be considered as *soft constraint* for the prediction algorithm. An additional application of soft constraints is their potential to allow *guided secondary structure prediction* using experimental structure probing data. This is especially important, since in the last years several high-throughput methods were developed that indicate the presence of looped or paired regions of RNA structures *in vitro*, and *in vivo* [246, 166, 230, 245] by structure probing. Although these methods do not identify actual pairing partners of a single nucleotide, sophisticated algorithms can be used to deduce secondary structures that align best with the provided probing data [151, 49, 239].

On the other hand, some proteins or enzymes that bind to parts of the RNA, e.g to recruit more complex cellular machinery, tend to mask nucleotides, rendering them inaccessible for the formation of alternative structure configurations. The nucleotides involved may then be forced to stay unpaired, or remain in a particular paired conformation [95, 149, 85]. For RNA structure prediction, these situations can be handled computationally by restricting the possible decomposition paths using so-called *hard constraints*. Consequently, this method removes a set of certain secondary structures from the ensemble, and therefore from the solution space. Of course, special care has to be taken in terms of ergodicity when using hard constraints for RNA refolding analyses.

While hard and soft constraints can be utilized for a large variety of purposes, most actual implementations of secondary structure DP algorithms offer only a limited constraint feature set. Thus, to enable prediction of a more complex interaction of RNAs with its environment, a generalized structure constraint approach needs to be developed.

13.2 RNA FOLDING DYNAMICS

The importance of RNA folding kinetics has been repeatedly discussed throughout this entire thesis. Not only is efficient folding dynamics prediction required to understand the mode of action of an RNA that performs its function through structure transition, like metabolite sensing *riboswitches* or *RNA thermometers* [178, 168], but the emergence of an RNA in the cell itself is heavily linked to folding kinetics. The nascent RNA that leaves the transcription bubble already folds back on itself and requires con-

secutive refolding events towards its native ground state. This process is well known as *Cotranscriptional folding*, and can lead to long-lived non-equilibrium states that may surmount the average life span of an RNA in the cell [125].

Unfortunately, the prediction of the dynamic behavior of RNA folding is a computationally demanding task (discussed in much detail in Chapter 6). While Markov Chain - Monte Carlo (MCMC) methods follow only a single folding trajectory out of exponentially many, holistic approaches that consider the entire ensemble of structures to solve the *master equation* of the underlying Markov process usually depend on exhaustive enumeration of the ensemble. Consequently, existing methods are still mostly applicable to RNAs with relatively short sequence lengths of at most 100 nt.

AB-INITIO COARSE GRAINING Coarse graining of the structure space provides a possible lift to this limitation. However, to successfully apply a coarse graining technique to longer sequences, the properties of the resulting macro-states and the transition rates between them must be computed efficiently. Since exhaustive enumeration of structures is out of question for longer RNA sequences, other means need to be utilized. For instance, folding dynamics may be predicted by restricting the set of possible conformations to just a few representative states and structures closeby. Tang et al. [217, 218] already proposed such a method where macro states and transition rates are approximated through Boltzmann sampling from the ensemble. The most difficult part of such methods is to generate a structure set that adequately represents the actual, vast ensemble of structures. But this is not guaranteed by regular Boltzmann sampling approaches, since very large sample sizes may be required to obtain structures that deviate from the ground state in terms of free energy. This dilemma is depicted exemplarily for two RNA sequence in Figure 20.

As an alternative, classified DP approaches as introduced in Chapter 3.6, can be used to create structurally diverse samples by *ab-initio* partitioning the structure space and sampling from within the generated partitions. In Chapter 12, I already presented a technique that uses distance class partitioning with respect to two reference structures to effectively predict their population densities over time. The proposed idea, however, is not limited to this particular classification. In fact, it can be easily adopted to other classifications, such as distance classes with larger numbers of reference structures, or abstract RNA shapes, as presented in Section 3.6.

DISTORTION OF THE BOLTZMANN ENSEMBLE The drawback of classified DP approaches is, however, the increasing runtime requirements for more sophisticated partitionings, cf. Chapter 3.6. Therefore, it remains to be investigated, if computationally cheaper methods, such as *soft constraints* used to distort the partition function Q, are able to yield diverse enough samples that can be further taken as input for *a posteriori* classifications. In fact, the application of sophisticated pseudo-energy terms that favor

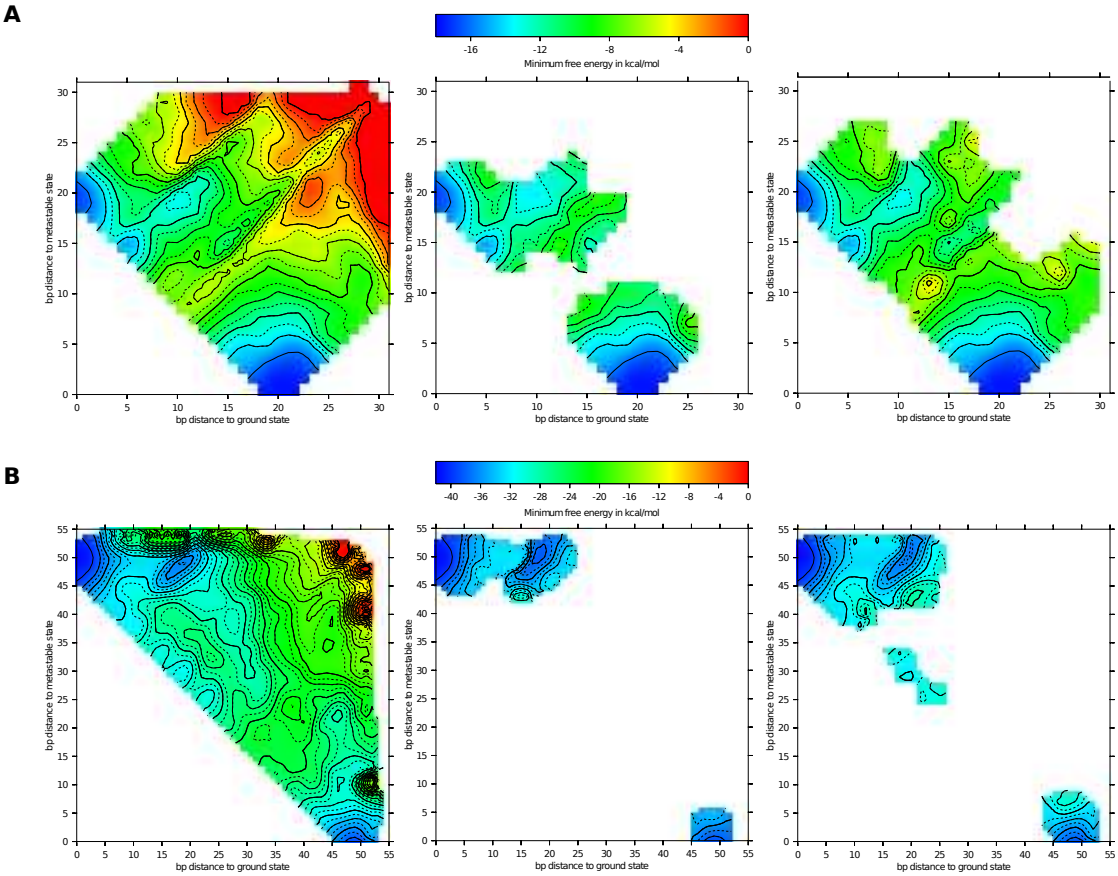


Figure 20: The bottleneck of Boltzmann sampling from the structure ensemble to construct the energy landscape is the sample size N. While for short RNA sequences sample sizes of some million structures may be sufficient to approximate large portions of the actual structure space, they need to be increased to billions of structures for sequences already close to 100 nt. Exemplarily shown are MFE representatives obtained from 2D projections of the structure ensemble for two RNA switches. For both the two corresponding low free energy conformations are used as reference structures (see also Chapter 9). Depicted from left to right are (i) the full structure ensemble, (ii) the ensemble obtained by Boltzmann sampling of $N = 1,000,000$, and (iii) $N = 100,000,000$ structures. **A** Small artificially designed RNA switch with sequence `GGGCGCGGUUCGCCCUCCGCUAAAUGCGGAAGAUAAAUUGUGUCU` and low free energy conformations of -17.7 kcal/mol (MFE), and -17.2 kcal/mol (Meta). Sample sizes of $N = 1,000,000$ and $N = 100,000,000$ generated 1,491, and 7,877 distinct structures, respectively. **B** 73 nt long artificially designed RNA [258] with two low free energy conformations of -40.9 kcal/mol (MFE), and -40.0 kcal/mol (Meta), respectively. The unique structures drawn from Boltzmann sampling amount to 1,777 ($N = 1,000,000$), and 11,702 ($N = 100,000,000$).

structures within certain parts of the secondary structure free energy landscape may be used for that purpose.

An example for such an approach that borrows the ideas of 2D projections of the free energy landscape, as introduced in Chapter 9, is presented below [1]. Instead of using a classified DP algorithm that *ab-initio* partitions the structure space into distance classes according to the reference structures s_1 and s_2, the distorted Boltzmann ensemble

$$Q^D = \sum_s e^{-\beta(E(s)+x(s)+y(s))} \tag{69}$$

is used to draw representative structure samples. Here, the two soft constraining pseudo-energy terms $x(s) = x' \cdot d_{BP}(s, s_1)$, and $y(s) = y' \cdot d_{BP}(s, s_2)$ with constant parameters x' and y' can be used to favor ($x', y' < 0$) or disfavor ($x', y' > 0$) structures according to their distance to the reference structures s_1 and s_2. Consequently, on average Boltzmann sampling generates structures in the vicinity, or far away from the reference structures, respectively. Using this approach it is even possible to align the equilibrium probabilities of s_1 and s_2 with that of the MFE structure s_{MFE}. Therefore, the distance dependent pseudo-energy factors x' and y' have to be chosen such that their equilibrium probabilities $p(s) = \exp(-\beta E(s) + x(s) + y(s))/Q$ are all equal, i.e.

$$\begin{aligned}
e^{-\beta(E(s_1)+x'\cdot 0+y'\cdot d_{BP}(s_1,s_2))} &= e^{-\beta(E(s_2)+x'\cdot d_{BP}(s_1,s_2)+y'\cdot 0)} \\
e^{-\beta(E(s_1)+x'\cdot 0+y'\cdot d_{BP}(s_1,s_2))} &= e^{-\beta(E(s_{MFE})+x'\cdot d_{BP}(s_1,s_{MFE})+y'\cdot d_{BP}(s_2,s_{MFE}))} \\
e^{-\beta(E(s_2)+x'\cdot d_{BP}(s_1,s_2)+y'\cdot 0)} &= e^{-\beta(E(s_{MFE})+x'\cdot d_{BP}(s_1,s_{MFE})+y'\cdot d_{BP}(s_2,s_{MFE}))}.
\end{aligned} \tag{70}$$

Consecutive sampling from distorted Boltzmann ensembles with a good choice of different pairs of x' and y' may then yield a diverse set of secondary structures suitable for landscape approximation and *a posteriori* macro state partitioning.

How well this method scales for short and long RNA sequences, and if it provides the means for adequate approximation of macro states, such as distance classes, to predict coarse-grained folding dynamics will be part of future research. Nonetheless, such an approach highlights the potential of constraints in secondary structure prediction (as discussed in the previous section). Especially, since it is not restricted to two reference structures s_1 and s_2, but allows for an arbitrary number of reference structures, which enables to *triangulate* small subspaces in the high-dimensional secondary structure landscape.

CO-TRANSCRIPTIONAL FOLDING Every RNA in a cellular environment of a living organism is produced by a process called *transcription*, where an enzyme, the *RNA polymerase*, covalently connects single ribonucleotides according to the complementary sequence of the corresponding DNA template. As soon as the nascent RNA leaves the *transcription bubble*, it folds back on itself to adopt an energetically favorable structure.

1 These ideas were developed together with Yann Ponty and Ivo L. Hofacker

However, with every transcription step the sequence is extended by an additional nucleotide, and the optimal structure for the partial transcript constantly changes. Therefore, the RNA undergoes several refolding processes from transient structures to its new native state. Depending on the stability of a transient structure, the transcription speed, and the energy barrier that has to be surpassed for the structure transition, the RNA may become stuck in a so-called *kinetic trap*, a transient non-equilibrium state that may be very long-lived.

As a consequence, pure thermodynamically driven methods, such as MFE prediction, that assume the RNA in equilibrium can lead to very different structures than observed in living organisms. Thus, it is of much importance to provide methods that enable the prediction of secondary structure by taking the transcription process into account. Some approaches for this purpose that use heuristics instead of simulating the entire folding process have already been presented in Chapter 12.2. However, their number is very limited and the heuristic assumptions are sometimes questionable. Therefore, new methods have to be developed that allow for the assessment of native, possibly non-equilibrium structures by explicitly accounting for transitions between transient states, especially for longer RNA transcripts.

The `Kinwalker` approach [77] seems to provide a good foundation for further extensions, since it seems to be possible to address the drawbacks of this method with recent advances in RNA secondary structure bioinformatics. On the one hand, new efficient energy barrier estimation techniques emerged [58, 136], that do not restrain themselves to direct transition paths. The application of such new methods very likely improves the accuracy of predicting whether or not a particular transient structure undergoes larger refolding when further nucleotides are attached to its sequence. On the other hand, the way how `Kinwalker` deduces transient structure candidates definitely needs to be improved. Since the algorithm constructs candidates by merging a current structure with optimal substructures drawn from the MFE matrices, it requires larger refolding and thus introduces high energy barriers, the longer the transcript becomes. This is due to the fact, that the substructures obtained from backtracking in the MFE matrices tend to be structurally far away from the current state of the transient structure, after all, they only represent the most probable ones, and do not need to merge well into a non-equilibrium state.

To avoid large structural changes and to allow a more fine grained prediction of the cotranscriptional folding trajectory, alternative, non-equilibrium structure prediction needs to be applied. For this purpose, the use of hard constraints in structure prediction may be a promising technique. Therefore, (parts of) the current transient structure can be used as hard constraint for conditional structure prediction through subsequent MFE forward recursions. Backtracking in the resulting constrained MFE matrices then yields substructures that are more closely related to the current transient state, and therefore more likely to be adopted. This heuristic also makes sense from a molecular biology point of view. An RNA structure is more prone to undergo a transition into

a more or less closely related state, than to refold a major part at once, as the original `Kinwalker` heuristic assumes.

In contrast to that, advances in computational RNA folding kinetics also contribute largely to the problem of cotranscriptional folding prediction, of course. Although to date only applicable to relatively small sequence lengths of about 100 nt, more efficient methods, such as *ab init io* coarse graining, enable the analysis of RNA sequences with sequence lengths of up to 400 nt already. This was shown in Chapter 12. The current drawback of this method is the involved sampling strategy. It requires a large amount of computations, after all stochastic backtracking from the 2D distance classes requires an additional factor of n^2 compared to regular stochastic backtracking from the ensemble. However, the ansatz to successively distort the Boltzmann ensemble with elaborate pseudo-energy functions and draw samples from the deformed secondary structure free energy landscape, as described above, shows promise that even more efficient methods can be developed.

BIBLIOGRAPHY

[1] Jan Pieter Abrahams, Mijam van den Berg, Eke Van Batenburg, and Cornelis Pleij. Prediction of RNA secondary structure, including pseudoknotting, by computer simulation. *Nucleic Acids Research*, 18(10):3035, 1990.

[2] 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*, 40(4):1407–1423, 2012. doi: 10.1093/nar/gkr810.

[3] Andrés Aguilera and Tatiana García-Muse. R loops: from transcription byproducts to threats to genome stability. *Molecular cell*, 46(2):115–124, 2012.

[4] T Akutsu. Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, 104:45–62, 2000.

[5] Can Alkan, Emre Karakoc, Joseph H Nadeau, S Cenk Sahinalp, and Kaizhong Zhang. RNA-RNA interaction prediction and antisense RNA target search. *Journal of Computational Biology*, 13(2):267–282, 2006.

[6] D Alkema, P A Hader, R A Bell, and T Neilson. Effects of flanking g . c base pairs on internal watson-crick, g . u, and nonbonded base pairs within a short ribonucleic acid duplex. *Biochemistry*, 21(9):2109–2117, Apr 1982.

[7] MP Allen and DJ Tildesley. Computer simulation of liquids. *Physics Today*, 1987.

[8] Fabian Amman, Stephan H Bernhart, Gero Doose, Ivo L Hofacker, Jing Qin, Peter F Stadler, and Sebastian Will. The Trouble with Long-Range Base Pairs in RNA Folding. In *Advances in Bioinformatics and Computational Biology*, pages 1–11. Springer, 2013.

[9] Vincent P Antao and Ignacio Tinoco. Thermodynamic parameters for loop formation in RNA and DNA hairpin tetraloops. *Nucleic acids research*, 20(4):819–824, 1992.

[10] Amit Arora and Beatrix Suess. An RNA G-quadruplex in the 3′ UTR of the proto-oncogene PIM1 represses translation. *RNA Biology*, 8:802–805, 2011.

[11] D B Arter, G C Walker, O C Uhlenbeck, and P G Schmidt. Pmr or the self-complementary oligoribonucleotide cpcpgpg. *Biochem Biophys Res Commun*, 61 (4):1089–1094, Dec 1974.

[12] David Auber, Maylis Delest, Jean-Philippe Domenger, and Serge Dulucq. Efficient drawing of RNA secondary structure. *J. Graph Algorithms Appl.*, 10(2): 329–351, 2006.

[13] Jean-Pierre Bachellerie, Jérôme Cavaillé, and Alexander Hüttenhofer. The expanding snoRNA world. *Biochimie*, 84(8):775–790, 2002.

[14] Robert T Batey, Robert P Rambo, and Jennifer A Doudna. Tertiary motifs in RNA structure and folding. *Angewandte Chemie International Edition*, 38(16):2326–2343, 1999.

[15] Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, USA, 1 edition, 1957.

[16] Richard Bellman. *Introduction to Matrix Analysis*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1970. ISBN 9780898713992.

[17] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, and Philip E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 2000. doi: 10.1093/nar/28.1.235.

[18] Stephan H Bernhart, Ivo L Hofacker, and Peter F Stadler. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–615, 2006.

[19] 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.

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

[21] Stephan H Bernhart, Ulrike Mückstein, and Ivo L Hofacker. RNA accessibility in cubic time. *Algorithms for Molecular Biology*, 6(1), 2011.

[22] Christof K Biebricher and Rudiger Luce. In vitro recombination and terminal elongation of RNA by Q beta replicase. *The EMBO journal*, 11(13):5129, 1992.

[23] E. Biondi, S. Branciamore, L. Fusi, S. Gago, and E. Gallori. Catalytic activity of hammerhead ribozymes in a clay mineral environment: Implications for the RNA world. *Gene*, 389:10–18, 2007.

[24] Elisa Biondi, Adam WR Maxwell, and Donald H Burke. A small ribozyme with dual-site kinase activity. *Nucleic acids research*, 40(15):7528–7540, 2012.

[25] Elizabeth H Blackburn and Kathleen Collins. Telomerase: an RNP enzyme synthesizes DNA. *Cold Spring Harbor perspectives in biology*, 3(5):a003558, 2011.

[26] Kenneth F Blount and Olke C Uhlenbeck. The structure-function dilemma of the hammerhead ribozyme. *Annu. Rev. Biophys. Biomol. Struct.*, 34:415–440, 2005.

[27] Sebastian Bonhoeffer, John S McCaskill, Peter F Stadler, and Peter Schuster. RNA multi-structure landscapes. *European biophysics journal*, 22(1):13–24, 1993.

[28] Philip N Borer, Barbara Dengler, Ignacio Tinoco Jr, and Olke C Uhlenbeck. Stability of ribonucleic acid double-stranded helices. *Journal of molecular biology*, 86 (4):843–853, 1974.

[29] Philippe Brion and Eric Westhof. Hierarchy and dynamics of RNA folding. *Annual review of biophysics and biomolecular structure*, 26(1):113–137, 1997.

[30] Robert E Bruccoleri and Gerhard Heinrich. An improved algorithm for nucleic acid secondary structure display. *Computer applications in the biosciences: CABIOS*, 4(1):167–173, 1988.

[31] Christine Brunel, Roland Marquet, Pascale Romby, and Chantal Ehresmann. RNA loop–loop interactions as dynamic functional motifs. *Biochimie*, 84(9):925–944, 2002.

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

[33] Samuel E. Butcher and Anna Marie Pyle. The Molecular Interactions That Stabilize RNA Tertiary Structure: RNA Motifs, Patterns, and Networks. *Accounts of Chemical Research*, 44(12):1302–1311, 2011. doi: 10.1021/ar200098t.

[34] Song Cao and Shi-Jie Chen. Predicting RNA pseudoknot folding thermodynamics. *Nucleic Acids Research*, 34(9):2634–2652, 2006. doi: 10.1093/nar/gkl346.

[35] Thomas R Cech. The RNA worlds in context. *Cold Spring Harbor perspectives in biology*, 4(7):a006742, 2012.

[36] T.R Cech. A model for the RNA-catalyzed replication of RNA. *Proc. Natl. Acad. Sci. USA*, 83:4360–4363, 1986.

[37] Vladimír Černỳ. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of optimization theory and applications*, 45(1):41–51, 1985.

[38] E. Chargaff. Chemical specificity of nucleic acids and mechanism of their enzymatic degradation. *Experientia*, 6:201–240, 1950.

[39] Gang Chen and Douglas H Turner. Consecutive GA pairs stabilize medium-size RNA internal loops. *Biochemistry*, 45(12):4025–4043, 2006.

[40] Jonathan L Chen, Abigael L Dishler, Scott D Kennedy, Ilyas Yildirim, Biao Liu, Douglas H Turner, and Martin J Serra. Testing the nearest neighbor model for canonical RNA base pairs: revision of GU parameters. *Biochemistry*, 51(16):3508–3522, 2012.

[41] Hamidreza Chitsaz, Raheleh Salari, S Cenk Sahinalp, and Rolf Backofen. A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, 25(12):i365–i373, 2009.

[42] Christine S Chow and Felicia M Bogdan. A structural basis for RNA-ligand interactions. *Chemical reviews*, 97(5):1489–1514, 1997.

[43] F.M. Codoner, J.-A. Daros, R.V. Sole, and S.F. Elena. The fittest versus the flattest: Experimental confirmation of the quasispecies effect with subviral pathogens. *PLoS Pathogens*, 2(12):1187–1193, December 2006.

[44] Francis Crick et al. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

[45] Francis H Crick. On protein synthesis. In *Symposia of the Society for Experimental Biology*, volume 12, page 138, 1958.

[46] Jan Cupal, Ivo L. Hofacker, and Peter F. Stadler. Dynamic programming algorithm for the density of states of rna secondary structures. In *German Conference on Bioinformatics*, pages 184–186, 1996.

[47] Kévin Darty, Alain Denise, and Yann Ponty. Varna: Interactive drawing and editing of the RNA secondary structure. *Bioinformatics*, 25(15):1974, 2009.

[48] Peter De Rijk, Jan Wuyts, and Rupert De Wachter. RnaViz 2: an improved representation of RNA secondary structure. *Bioinformatics*, 19(2):299–300, 2003.

[49] Katherine E Deigan, Tian W Li, David H Mathews, and Kevin M Weeks. Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences*, pages pnas–0806929106, 2008.

[50] J. Demongeot and A. Moreira. A possible circular RNA at the origin of life. *Journal of Theoretical Biology*, 2007. doi: 10.1016/j.jtbi.2007.07.010. in press.

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

[52] Biao Ding. The Biology of Viroid-Host Interactions. *Annual Review of Phytopathology*, 47(1):105–131, 2009. doi: 10.1146/annurev-phyto-080508-081927. PMID: 19400635.

[53] Biao Ding and Asuka Itaya. Viroid: a useful model for studying the basic principles of infection and RNA biology. *Molecular plant-microbe interactions*, 20(1): 7–20, 2007.

[54] Ye Ding and Charles E. Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Research*, 31(24):7280–7301, 2003. doi: 10.1093/nar/gkg938.

[55] R.M. Dirks and N.A. Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of Computational Chemistry*, 24(13):1664–1677, 2003.

[56] 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.

[57] Chuong B Do, Daniel A Woods, and Serafim Batzoglou. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22 (14):e90–e98, 2006.

[58] Ivan Dotu, William A Lorenz, Pascal Van Hentenryck, and Peter Clote. Computing folding pathways between RNA secondary structures. *Nucleic acids research*, 38(5):1711–1722, 2010.

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

[60] David E Draper, Dan Grilley, and Ana Maria Soto. Ions and RNA folding. *Annu. Rev. Biophys. Biomol. Struct.*, 34:221–243, 2005.

[61] Sanaz Farajollahi and Stefan Maas. Molecular diversity through RNA editing: a balancing act. *Trends in Genetics*, 26(5):221–230, 2010.

[62] Junli Feng, Walter D Funk, Sy-Shi Wang, Scott L Weinrich, Ariel A Avilion, Choy-Pik Chiu, Robert R Adams, Edwin Chang, Richard C Allsopp, Jinghua Yu, et al. The RNA component of human telomerase. *Science*, 269(5228):1236–1241, 1995.

[63] Witold Filipowicz, Lukasz Jaskiewicz, Fabrice A Kolb, and Ramesh S Pillai. Post-transcriptional gene silencing by siRNAs and miRNAs. *Current opinion in structural biology*, 15(3):331–341, 2005.

[64] Thomas R Fink and Donald M Crothers. Free energy of imperfect nucleic acid helices: I. The bulge defect. *Journal of molecular biology*, 66(1):1–12, 1972.

[65] Christoph Flamm and L. Ivo Hofacker. Beyond Energy Minimization: Approaches to the kinetic folding of RNA. *Chemical Monthly*, 139(4):447–457, 2008. doi: 10.1007/s00706-008-0895-3.

[66] Christoph Flamm, Walter Fontana, Ivo L Hofacker, and Peter Schuster. RNA folding at elementary step resolution. *RNA*, 6(3):325–338, 2000.

[67] Christoph Flamm, Ivo L Hofacker, Sebastian Maurer-Stroh, Peter F Stadler, and Martin Zehl. Design of multistable RNA molecules. *Rna*, 7(2):254–265, 2001.

[68] Christoph Flamm, Ivo L Hofacker, Peter F Stadler, and Michael T Wolfinger. Barrier trees of degenerate landscapes. *Zeitschrift für Physikalische Chemie*, 216 (2/2002):155, 2002.

[69] Walter Fontana, Peter F Stadler, Erich G Bornberg-Bauer, Thomas Griesmacher, Ivo L Hofacker, Manfred Tacker, Pedro Tarazona, Edward D Weinberger, and Peter Schuster. RNA folding and combinatory landscapes. *Physical review E*, 47 (3):2083, 1993.

[70] Susan M Freier, Ryszard Kierzek, John A Jaeger, Naoki Sugimoto, Marvin H Caruthers, Thomas Neilson, and Douglas H Turner. Improved free-energy parameters for predictions of RNA duplex stability. *Proceedings of the National Academy of Sciences*, 83(24):9373–9377, 1986.

[71] J.R. Fresco, B.M. Alberts, and P. Doty. Some molecular details of the secondary structure of ribonucleic acid. *Nature*, 188, 1960.

[72] Eva Freyhult, Vincent Moulton, and Peter Clote. Boltzmann probability of RNA structural neighbors and riboswitch detection. *Bioinformatics*, 23(16):2054–2062, 2007.

[73] Peter Friebe, Julien Boudet, Jean-Pierre Simorre, and Ralf Bartenschlager. Kissing-loop interaction in the 3' end of the hepatitis C virus genome essential for RNA replication. *Journal of virology*, 79(1):380–392, 2005.

[74] Crispin W Gardiner. Handbook of stochastic methods for physics, chemistry and the natural sciences. *Springer ser. in synergetics*, 1985.

[75] Paul P Gardner and Robert Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC bioinformatics*, 5(1):140, 2004.

[76] Junhui Ge and Yi-Tao Yu. RNA pseudouridylation: new insights into an old modification. *Trends in biochemical sciences*, 38(4):210–218, 2013.

[77] 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.

[78] Kenn Gerdes and E Gerhart H Wagner. RNA antitoxins. *Current opinion in microbiology*, 10(2):117–124, 2007.

[79] Wolfgang Gerlach and Robert Giegerich. GUUGle: a utility for fast exact matching under RNA complementary rules including G–U base pairing. *Bioinformatics*, 22(6):762–764, 2006.

[80] Robert Giegerich and Carsten Meyer. Algebraic dynamic programming. In *Algebraic Methodology And Software Technology*, pages 349–364. Springer, 2002.

[81] Robert Giegerich, Björn Voß, and Marc Rehmsmeier. Abstract shapes of RNA. *Nucleic acids research*, 32(16):4843–4851, 2004.

[82] Sunny D Gilbert, Colby D Stoddard, Sarah J Wise, and Robert T Batey. Thermodynamic and kinetic characterization of ligand binding to the purine riboswitch aptamer domain. *Journal of molecular biology*, 359(3):754–768, 2006.

[83] W. Gilbert. Origin of life: The RNA world. *Nature*, 319:618, 1986.

[84] Daniel T Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22 (4):403–434, 1976.

[85] Michael L Gleghorn and Lynne E Maquat. 'black sheep' that don't leave the double-stranded RNA-binding domain fold. *Trends in biochemical sciences*, 39: 328–340, 2014.

[86] 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.

[87] Jay Gralla and Donald M Crothers. Free energy of imperfect nucleic acid helices: II. small hairpin loops. *Journal of molecular biology*, 73(4):497–511, 1973.

[88] Jay Gralla and Donald M Crothers. Free energy of imperfect nucleic acid helices: III. Small internal loops resulting from mismatches. *Journal of molecular biology*, 78(2):301–319, 1973.

[89] HERMAN Groeneveld, KAFIINE Thimon, and J Van Duin. Translational control of maturation-protein synthesis in phage MS2: a role for the kinetics of RNA folding? *Rna*, 1(1):79, 1995.

[90] Cecilia Guerrier-Takada, Katheleen Gardiner, Terry Marsh, Norman Pace, and Sidney Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35(3):849–857, 1983.

[91] Alexander P Gultyaev, FHD Van Batenburg, and Cornelis WA Pleij. The computer simulation of RNA folding pathways using a genetic algorithm. *Journal of molecular biology*, 250(1):37–51, 1995.

[92] Alexander P Gultyaev, FHD Van Batenburg, and Cornelis WA Pleij. Dynamic competition between alternative structures in viroid RNAs simulated by an rna folding algorithm. *Journal of molecular biology*, 276(1):43–55, 1998.

[93] AP Gultyaev. The computer simulation of RNA folding involving pseudoknot formation. *Nucleic acids research*, 19(9):2489–2494, 1991.

[94] Christine E Hajdin, Feng Ding, Nikolay V Dokholyan, and Kevin M Weeks. On the significance of an RNA tertiary structure prediction. *Rna*, 16(7):1340–1349, 2010.

[95] Eliane Hajnsdorf and Irina V Boni. Multiple activities of RNA-binding proteins S1 and Hfq. *Biochimie*, 94(7):1544–1553, 2012.

[96] Michiaki Hamada, Hisanori Kiryu, Kengo Sato, Toutai Mituyama, and Kiyoshi Asai. Prediction of RNA secondary structure using generalized centroid estimators. *Bioinformatics*, 25(4):465–473, 2009.

[97] Kyungsook Han and Yanga Byun. PseudoViewer2: visualization of RNA pseudoknots of any type. *Nucleic Acids Research*, 31(13):3432–3440, 2003.

[98] I.F. Hassen, S. Massart, J. Motard, S. Roussel, O. Parisi, J. Kummert, M. Fakhfakh, H. Marrakchi, J.P. Perreault, and M.H. Jijakli. Molecular features of new peach latent mosaic viroid variants suggest that recombination may have contributed to the evolution of this infectious RNA. *Virology*, 360:50–57, 2007.

[99] W Keith Hastings. Monte Carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[100] Nina Z Hausmann and Brent M Znosko. Thermodynamic Characterization of RNA 2× 3 Nucleotide Internal Loops. *Biochemistry*, 51(26):5359–5368, 2012.

[101] R.M. Hazen, P.L. Griffin, J.M. Carothers, and J.W. Szostak. Functional information and the emergence of biocomplexity. *PNAS*, 104:8574–8581, 2007.

[102] Thomas Hermann and Dinshaw J Patel. Adaptive recognition by nucleic acid aptamers. *Science*, 287(5454):820–825, 2000.

[103] Paul G Higgs. RNA secondary structure: a comparison of real and random sequences. *Journal de Physique I*, 3(1):43–59, 1993.

[104] Ivo L Hofacker, Walter Fontana, Peter F Stadler, L Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188, 1994.

[105] Ivo L Hofacker, Peter Schuster, and Peter F Stadler. Combinatorics of RNA secondary structures. *Discrete Applied Mathematics*, 88(1):207–237, 1998.

[106] 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(7):1308–1316, Jul 2010. doi: 10.1261/rna.2093310.

[107] P. Hogeweg and B. Hesper. The alignment of sets of sequences and the construction of phyletic trees: an integrated method. *Journal of Molecular Evolution*, 20: 175–186, 1984.

[108] Stephen R Holbrook and Sung-Hou Kim. RNA crystallography. *Biopolymers*, 44 (1):3–21, 1997.

[109] Christian Höner zu Siederdissen. Sneaking around concatMap: Efficient combinators for dynamic programming. In *ACM SIGPLAN Notices*, volume 47, pages 215–226. ACM, 2012.

[110] Fenix WD Huang, Jing Qin, Christian M Reidys, and Peter F Stadler. Partition function and base pairing probabilities for RNA–RNA interaction prediction. *Bioinformatics*, 25(20):2646–2654, 2009.

[111] Jiabin Huang, Rolf Backofen, and Björn Voß. Abstract folding space analysis based on helices. *RNA*, 18(12):2135–2147, 2012.

[112] Graham A. Hudson, Richard J. Bloomingdale, and Brent M. Znosko. Thermodynamic contribution and nearest-neighbor parameters of pseudouridine-adenosine base pairs in oligoribonucleotides. *RNA*, 19(11):1474–1482, 2013. doi: 10.1261/rna.039610.113.

[113] J L Huppert, A Bugaut, S Kumari, and S. Balasubramanian. G-quadruplexes: the beginning and end of UTRs. *Nucleic Acids Res.*, 36:6260–6268, 2008.

[114] 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.

[115] Homer Jacobson and Walter H Stockmayer. Intramolecular reaction in polycondensations. I. The theory of linear systems. *The Journal of Chemical Physics*, 18(12): 1600–1606, 1950.

[116] John a. Jaeger, Douglas H. Turner, and Michael Zuker. Improved predictions of secondary structures for RNA. *Proc. Natl. Acad. Sci., USA, Biochemistry*, 86: 7706–7710, 1989.

[117] Stefan Janssen and Robert Giegerich. Faster computation of exact RNA shape probabilities. *Bioinformatics*, 26(5):632–639, 2010.

[118] Michael F Jantsch. Reaching complexity through RNA-editing. *RNA biology*, 7 (2):191–191, 2010.

[119] A. Joachimi, A. Benz, and J.S. Hartig. A comparison of DNA and RNA quadruplex structures and stabilities. *Bioorganic & medicinal chemistry*, 17(19):6811–6815, 2009. energy parameters.

[120] Gerald F Joyce. The antiquity of RNA-based evolution. *Nature*, 418(6894):214–221, 2002.

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

[122] Elzbieta Kierzek, Magdalena Malgowska, Jolanta Lisowiec, Douglas H Turner, Zofia Gdaniec, and Ryszard Kierzek. The contribution of pseudouridine to stabilities and structure of RNAs. *Nucleic acids research*, 42(5):3492–3501, 2014.

[123] Scott Kirkpatrick, C Daniel Gelatt, Mario P Vecchi, et al. Optimization by simmulated annealing. *science*, 220(4598):671–680, 1983.

[124] Bjarne Knudsen and Jotun Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic acids research*, 31(13):3423–3428, 2003.

[125] D Lai, J R Proctor, and I M Meyer. On the importance of cotranscriptional rna structure formation. *RNA*, 19(11):1461–73, Nov 2013. doi: 10.1261/rna.037390.112.

[126] G Lapalme, RJ Cedergren, and D Sankoff. An algorithm for the display of nucleic acid secondary structure. *Nucleic acids research*, 10(24):8351–8356, 1982.

[127] Sébastien Lemieux and François Major. Automated extraction and classification of RNA tertiary structure cyclic motifs. *Nucleic acids research*, 34(8):2340–2346, 2006.

[128] Neocles B Leontis and Eric Westhof. Conserved geometrical base-pairing patterns in RNA. *Quarterly reviews of biophysics*, 31(04):399–455, 1998.

[129] Neocles B Leontis and Eric Westhof. Geometric nomenclature and classification of RNA base pairs. *Rna*, 7(4):499–512, 2001.

[130] Neocles B Leontis and Eric Westhof. The annotation of RNA motifs. *Comparative and functional genomics*, 3(6):518–524, 2002.

[131] Neocles B Leontis, Aurelie Lescoute, and Eric Westhof. The building blocks and motifs of RNA architecture. *Current opinion in structural biology*, 16(3):279–287, 2006.

[132] Josef Leydold and Peter F Stadler. Minimal cycle bases of outerplanar graphs. *JOURNAL OF COMBINATORICS*, 5:209–222, 1998.

[133] Yuan Li and Shaojie Zhang. Predicting folding pathways between RNA conformational structures guided by RNA stacks. *BMC bioinformatics*, 13(Suppl 3):S5, 2012.

[134] Jan Lipfert and Sebastian Doniach. Small-angle X-ray scattering from RNA, proteins, and protein complexes. *Annu. Rev. Biophys. Biomol. Struct.*, 36:307–327, 2007.

[135] R. Lorenz, S. Bernhart, J. Qin, C. Höner zu Siederdissen, A. Tanzer, F. Amman, I. Hofacker, and P. Stadler. 2D meets 4G: G-Quadruplexes in RNA Secondary Structure Prediction. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, PP(99):1, 2013. ISSN 1545-5963. doi: 10.1109/TCBB.2013.7.

[136] Ronny Lorenz, Christoph Flamm, and Ivo L. Hofacker. 2D projections of RNA folding landscapes. In Ivo Grosse, Steffen Neumann, Stefan Posch, Falk Schreiber, and Peter F. Stadler, editors, *German Conference on Bioinformatics 2009*, volume 157 of *Lecture Notes in Informatics*, pages 11–20, Bonn, September 2009. Gesellschaft f. Informatik. ISBN 978-3-88579-251-2.

[137] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA Package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011. doi: 10.1186/1748-7188-6-26.

[138] Ronny Lorenz, Ivo L. Hofacker, and Stephan H. Bernhart. Folding RNA/DNA hybrid duplexes. *Bioinformatics*, 2012. doi: 10.1093/bioinformatics/bts466.

[139] Lauren Lui and Todd Lowe. Small nucleolar RNAs and RNA-guided post-transcriptional modification. *Essays in biochemistry*, 54(1):53–77, 2013.

[140] R B Lyngsø and C N Pedersen. RNA pseudoknot prediction in energy based models. *Journal of Computational Biology*, 7:409–428, 2001.

[141] Rune B Lyngsø, Michael Zuker, and CN Pedersen. Fast evaluation of internal loops in RNA secondary structure prediction. *Bioinformatics*, 15(6):440–445, 1999.

[142] CH Mak and Paul S Henke. Ions and RNAs: Free Energies of Counterion-Mediated RNA Fold Stabilities. *Journal of Chemical Theory and Computation*, 9(1):621–639, 2012.

[143] Arka Mallela and Kazuko Nishikura. A-to-I editing of protein coding and non-coding RNAs. *Critical reviews in biochemistry and molecular biology*, 47(6):493–501, 2012.

[144] S.C. Manrubia and C. Briones. Modular evolution and increase of functional complexity in replicating RNA molecules. *RNA*, 13:97–107, 2007.

[145] Ján Maňuch, Chris Thachuk, Ladislav Stacho, and Anne Condon. NP-completeness of the direct energy barrier problem without pseudoknots. In *DNA Computing and Molecular Programming*, pages 106–115. Springer, 2009.

[146] Ján 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.

[147] Nicholas R Markham and Michael Zuker. UNAFold. In *Bioinformatics*, pages 3–31. Springer, 2008.

[148] Hugo M Martinez. An RNA folding rule. *Nucleic acids research*, 12(1Part1):323–334, 1984.

[149] Encarnación Martínez-Salas, Gloria Lozano, Javier Fernandez-Chamorro, Rosario Francisco-Velilla, Alfonso Galan, and Rosa Diaz. RNA-Binding Proteins Impacting on Internal Initiation of Translation. *International journal of molecular sciences*, 14(11):21705–21726, 2013.

[150] David H Mathews and Douglas H Turner. Experimentally derived nearest-neighbor parameters for the stability of RNA three-and four-way multibranch loops. *Biochemistry*, 41(3):869–880, 2002.

[151] David H Mathews, Matthew D Disney, Jessica L Childs, Susan J Schroeder, Michael Zuker, and Douglas H Turner. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proceedings of the National Academy of Sciences of the United States of America*, 101(19):7287–7292, 2004.

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

[153] John S Mattick. The functional genomics of noncoding RNA. *Science*, 309(5740): 1527–1528, 2005.

[154] Marjori Matzke, Werner Aufsatz, Tatsuo Kanno, Lucia Daxinger, Istvan Papp, M Florian Mette, and Antonius JM Matzke. Genetic analysis of RNA-mediated transcriptional gene silencing. *Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression*, 1677(1):129–141, 2004.

[155] John S McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990.

[156] J L Mergny and L Lacroix. Uv melting of g-quadruplexes. *Curr Protoc Nucleic Acid Chem.*, Unit 17.1., 2009.

[157] Edward J Merino, Kevin A Wilkinson, Jennifer L Coughlan, and Kevin M Weeks. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *Journal of the American Chemical Society*, 127(12):4223–4231, 2005.

[158] 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*, 21:1087, 1953.

[159] Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix. *SIAM review*, 20(4):801–836, 1978.

[160] 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.

[161] Attila Molnar, Charles W Melnyk, Andrew Bassett, Thomas J Hardcastle, Ruth Dunn, and David C Baulcombe. Small silencing RNAs in plants are mobile and direct epigenetic modification in recipient cells. *science*, 328(5980):872–875, 2010.

[162] David Monchaud and Marie-Paule Teulade-Fichou. A hitchhiker's guide to G-quadruplex ligands. *Organic & biomolecular chemistry*, 6(4):627–636, 2008.

[163] Peter B Moore and Thomas A Steitz. The roles of RNA in the synthesis of protein. *Cold Spring Harbor perspectives in biology*, 3(11):a003780, 2011.

[164] 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.

[165] Kevin V Morris and John S Mattick. The rise of regulatory RNA. *Nature Reviews Genetics*, 15(6):423–437, 2014.

[166] Stefanie A Mortimer and Kevin M Weeks. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *Journal of the American Chemical Society*, 129(14):4144–4145, 2007.

[167] Ulrike Mückstein, Hakim Tafer, Jörg Hackermüller, Stephan H Bernhart, Peter F Stadler, and Ivo L Hofacker. Thermodynamics of RNA–RNA binding. *Bioinformatics*, 22(10):1177–1182, 2006.

[168] Franz Narberhaus, Torsten Waldminghaus, and Saheli Chowdhury. RNA thermometers. *FEMS microbiology reviews*, 30(1):3–16, 2006.

[169] T Neilson, P J Romaniuk, D Alkema, D W Hughes, J R Everett, and R A Bell. The effects of base sequence and dangling bases on the stability of short ribonucleic acid duplexes. *Nucleic Acids Symp Ser*, (7):293–311, 1980.

[170] Beth L Nicholson and K Andrew White. Functional long-range RNA-RNA interactions in positive-strand RNA viruses. *Nature Reviews Microbiology*, 12(7): 493–504, 2014.

[171] Poul Nissen, Jeffrey Hansen, Nenad Ban, Peter B Moore, and Thomas A Steitz. The structural basis of ribosome activity in peptide bond synthesis. *Science*, 289 (5481):920–930, 2000.

[172] Ruth Nussinov and George Pieczenik. Structural and combinatorial constraints on base pairing in large nucleotide sequences. *Journal of theoretical biology*, 106(3): 245–259, 1984.

[173] Ruth Nussinov, George Pieczenik, Jerrold R. Griggs, and Daniel J. Kleitman. Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, 35(1):68–82, 1978. doi: 10.1137/0135006.

[174] Tao Pan and Tobin Sosnick. RNA folding during transcription. *Annu. Rev. Biophys. Biomol. Struct.*, 35:161–175, 2006.

[175] Marc Parisien and Francois Major. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. *Nature*, 452(7183):51–55, 2008.

[176] Dinshaw J Patel, Lawrence Shapiro, and Dennis Hare. DNA and RNA: NMR studies of conformations and dynamics in solution. *Quarterly reviews of biophysics*, 20(1-2):35–112, 1987.

[177] Dmitri D Pervouchine. IRIS: intermolecular RNA interaction search. *Genome Informatics Series*, 15(2):92, 2004.

[178] Alla Peselis and Alexander Serganov. Themes and variations in riboswitch structure and function. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 2014. doi: 10.1016/j.bbagrm.2014.02.012.

[179] Matthew Petersheim and Douglas H Turner. Base-stacking and base-pairing contributions to helix stability: thermodynamics of double-helix formation with ccgg, ccggp, ccggap, accggp, ccggup, and accggup. *Biochemistry*, 22(2):256–263, 1983.

[180] Anton I Petrov, Craig L Zirbel, and Neocles B Leontis. Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA*, 19(10):1327–1340, 2013.

[181] Ian R Price, Jason C Grigg, and Ailong Ke. Common themes and differences in SAM recognition among SAM riboswitches. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 2014.

[182] Jeff R Proctor and Irmtraud M Meyer. CoFold: an RNA secondary structure prediction method that takes co-transcriptional folding into account. *Nucleic acids research*, 41(9):e102–e102, 2013.

[183] Arati Ramesh and Wade C Winkler. Metabolite-binding ribozymes. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 2014. doi: 10.1016/j.bbagrm.2014.04.015.

[184] J. Reeder and R. Giegerich. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. *BMC bioinformatics*, 5 (1):104, 2004.

[185] Marc Rehmsmeier, Peter Steffen, Matthias Höchsmann, and Robert Giegerich. Fast and effective prediction of microRNA/target duplexes. *Rna*, 10(10):1507–1517, 2004.

[186] Jihong Ren, Baharak Rastegari, Anne Condon, and Holger H Hoos. HotKnots: Heuristic prediction of RNA secondary structures including pseudoknots. *Rna*, 11(10):1494–1504, 2005.

[187] Dirk Repsilber, Sabine Wiese, Marc Rachen, ASTRID W SCHROeDER, Detlev Riesner, and Gerhard Steger. Formation of metastable RNA structures by sequential folding during transcription: Time-resolved structural analysis of potato spindle tuber viroid (-)-stranded RNA by temperature-gradient gel electrophoresis. *Rna*, 5(04):574–584, 1999.

[188] Jessica S Reuter and David H Mathews. RNAstructure: software for RNA secondary structure prediction and analysis. *BMC bioinformatics*, 11(1):129, 2010.

[189] A. Rich and J.D. Watson. Some relations between DNA and RNA. *Proc. Natl. Acad. Sci. USA*, 40(8):759–764, 1954.

[190] Krijn Rietveld, R Van Poelgeest, Cornelis WA Pleij, JH Van Boom, and Leendert Bosch. The tRNA-Uke structure at the 3′ terminus of turnip yellow mosaic virus RNA. differences and similarities with canonical tRNA. *Nucleic acids research*, 10 (6):1929–1946, 1982.

[191] R. F. Rinehart. The equivalence of definitions of a matric function. *The American Mathematical Monthly*, 62(6):pp. 395–414, 1955. ISSN 00029890.

[192] Elena Rivas and Sean R Eddy. A dynamic programming algorithm for RNA structure prediction including pseudoknots. *Journal of molecular biology*, 285(5): 2053–2068, 1999.

[193] Marina V Rodnina, Malte Beringer, and Wolfgang Wintermeyer. How ribosomes make peptide bonds. *Trends in biochemical sciences*, 32(1):20–26, 2007.

[194] Igor B Rogozin, Liran Carmel, Miklos Csuros, and Eugene V Koonin. Origin and evolution of spliceosomal introns. *Biol Direct*, 7(11), 2012.

[195] David Sankoff and Joseph B Kruskal. Time warps, string edits, and macromolecules: the theory and practice of sequence comparison. *Reading: Addison-Wesley Publication, 1983, edited by Sankoff, David; Kruskal, Joseph B.*, 1, 1983.

[196] Georg Sauthoff, Stefan Janssen, and Robert Giegerich. Bellman's GAP: A declarative language for dynamic programming. In *Proceedings of the 13th international ACM SIGPLAN symposium on Principles and practices of declarative programming*, pages 29–40. ACM, 2011.

[197] Georg Sauthoff, Mathias Möhl, Stefan Janssen, and Robert Giegerich. Bellman's GAP - A language and compiler for dynamic programming in sequence analysis. *Bioinformatics*, 29(5):551–560, 2013. doi: 10.1093/bioinformatics/btt022.

[198] Alexander Serganov and Evgeny Nudler. A decade of riboswitches. *Cell*, 152(1): 17–24, 2013.

[199] Claude E Shannon. A note on the concept of entropy. *Bell System Tech. J*, 27: 379–423, 1948.

[200] Bruce A Shapiro. An algorithm for comparing multiple RNA secondary structures. *Computer applications in the biosciences: CABIOS*, 4(3):387–393, 1988.

[201] Bruce A Shapiro and Kaizhong Zhang. Comparing multiple RNA secondary structures using tree comparisons. *Computer applications in the biosciences: CABIOS*, 6(4):309–318, 1990.

[202] Bruce A Shapiro, Jacob Maizel, Lewis E Lipkin, Kathleen Currey, and Carol Whitney. Generating non-overlapping displays of nucleic acid secondary structure. *Nucleic acids research*, 12(1Part1):75–88, 1984.

[203] Shantanu Sharma, Feng Ding, and Nikolay V. Dokholyan. iFoldRNA: three-dimensional RNA structure prediction and folding. *Bioinformatics*, 24(17):1951–1952, 2008. doi: 10.1093/bioinformatics/btn328.

[204] Ling X Shen and Ignacio Tinoco Jr. The structure of an RNA pseudoknot that causes efficient frameshifting in mouse mammary tumor virus. *Journal of molecular biology*, 247(5):963–978, 1995.

[205] K Snoussi, S Nonin-Lecomte, and J-L Leroy. The RNA i-motif. *Journal of molecular biology*, 309(1):139–153, 2001.

[206] J. Sperschneider and A. Datta. DotKnot: pseudoknot prediction using the probability dot plot under a refined energy model. *Nucleic Acids Research*, 2010.

[207] Peter F Stadler and Christoph Flamm. Barrier trees on poset-valued landscapes. *Genetic Programming and Evolvable Machines*, 4(1):7–20, 2003.

[208] David W Staple and Samuel E Butcher. Pseudoknots: Rna structures with diverse functions. *PLoS biology*, 3(6):e213, 2005.

[209] Peter Steffen, Björn Voß, Marc Rehmsmeier, Jens Reeder, and Robert Giegerich. RNAshapes: An integrated RNA analysis package based on abstract shapes. *Bioinformatics*, 22(4):500–503, 2006.

[210] 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.

[211] Gisela Storz, Jörg Vogel, and Karen M Wassarman. Regulation by small RNAs in bacteria: expanding frontiers. *Molecular cell*, 43(6):880–891, 2011.

[212] Murugan Subramanian, Florence Rage, Ricardos Tabet, Eric Flatter, Jean-Louis Mandel, and Hervé Moine. G–quadruplex RNA structure as a signal for neurite mRNA targeting. *EMBO reports*, 12(7):697–704, 2011.

[213] Hakim Tafer and Ivo L Hofacker. RNAplex: a fast tool for RNA–RNA interaction search. *Bioinformatics*, 24(22):2657–2663, 2008.

[214] Kuo-Chung Tai. The tree-to-tree correction problem. *Journal of the ACM (JACM)*, 26(3):422–433, 1979.

[215] Zhi-Jie Tan and Shi-Jie Chen. Rna Helix stability in mixed na^{+}/mg^{2+} Solution. *Biophysical journal*, 92(10):3615–3632, 2007.

[216] Zhi-Jie Tan and Shi-Jie Chen. Salt dependence of nucleic acid hairpin stability. *Biophysical journal*, 95(2):738–752, 2008.

[217] Xinyu Tang, Bonnie Kirkpatrick, Shawna Thomas, Guang Song, and Nancy M Amato. Using motion planning to study RNA folding kinetics. *Journal of Computational Biology*, 12(6):862–881, 2005.

[218] Xinyu Tang, Shawna Thomas, Lydia Tapia, David P Giedroc, and Nancy M Amato. Simulating RNA folding kinetics on approximated energy landscapes. *Journal of molecular biology*, 381(4):1055–1067, 2008.

[219] N Kyle Tanner. Ribozymes: the characteristics and properties of catalytic RNAs. *FEMS microbiology reviews*, 23(3):257–275, 1999.

[220] Andrea Tanzer and Peter F Stadler. Evolution of micrornas. In *MicroRNA Protocols*, pages 335–350. Springer, 2006.

[221] Corinna Theis, Christian Höner zu Siederdissen, Ivo L. Hofacker, and Jan Gorodkin. Automated identification of RNA 3D modules with discriminative power in RNA structural alignments. *Nucleic Acids Research*, 41(22):9999–10009, 2013. doi: 10.1093/nar/gkt795.

[222] Thomas Thisted, Nina S Sørensen, and Kenn Gerdes. Mechanism of post-segregational killing: secondary structure analysis of the entire Hok mRNA from plasmid R1 suggests a fold-back structure that prevents translation and antisense RNA binding. *Journal of molecular biology*, 247(5):859–873, 1995.

[223] I. Tinoco, O. C. Uhlenbeck, and M. D. Levine. Estimation of secondary structure in ribonucleic acids. *Nature*, 230(5293):362–367, April 1971. ISSN 0028-0836.

[224] I Tinoco, P N Borer, B Dengler, M D Levin, O C Uhlenbeck, D M Crothers, and J Bralla. Improved estimation of secondary structure in ribonucleic acids. *Nat New Biol*, 246(150):40–41, Nov 1973.

[225] Ignacio Tinoco, Gang Chen, and Xiaohui Qu. RNA reactions one molecule at a time. *Cold Spring Harbor perspectives in biology*, 2(11):a003624, 2010.

[226] Mark Tuckerman. *Statistical Mechanics and Molecular Simulations*. Oxford University Press, 2008.

[227] DH Turner, N Sugimoto, JA Jaeger, CE Longfellow, SM Freier, and R Kierzek. Improved parameters for prediction of RNA structure. In *Cold Spring Harbor symposia on quantitative biology*, volume 52, pages 123–133. Cold Spring Harbor Laboratory Press, 1987.

[228] Douglas H Turner and David H Mathews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic acids research*, page gkp892, 2009.

[229] T. Tuschl, C. Gohlke, T.M. Jovin, E. Westhof, and F. Eckstein. A three-dimensional model for the hammerhead ribozyme based on fluorescence measurements. *Science*, 266:785–789, 1994.

[230] Jason G Underwood, Andrew V Uzilov, Sol Katzman, Courtney S Onodera, Jacob E Mainzer, David H Mathews, Todd M Lowe, Sofie R Salama, and David Haussler. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nature methods*, 7(12):995–1001, 2010.

[231] Ramon Van der Werf, Sybren S Wijmenga, Hans A Heus, and René CL Olsthoorn. Structural and thermodynamic signatures that define pseudotriloop RNA hairpins. *RNA*, 19(12):1833–1839, 2013.

[232] Dico van Meerten, Genevieve Girard, and J Van Duin. Translational control by delayed RNA folding: identification of the kinetic trap. *Rna*, 7(3):483–494, 2001.

[233] G. Varani and W. H. McClain. The G x U wobble base pair. A fundamental building block of RNA structure crucial to RNA function in diverse biological systems. *EMBO Rep*, 1(1):18–23, July 2000. ISSN 1469-221X.

[234] Gabriele Varani and Ignacio Tinoco. RNA structure and NMR spectroscopy. *Quarterly reviews of biophysics*, 24(04):479–532, 1991.

[235] Alexey G Vitreschak, Dimitry A Rodionov, Andrey A Mironov, and Mikhail S Gelfand. Riboswitches: the oldest mechanism for the regulation of gene expression? *TRENDS in Genetics*, 20(1):44–50, 2004.

[236] Björn Voß, Robert Giegerich, and Marc Rehmsmeier. Complete probabilistic analysis of RNA shapes. *BMC biology*, 4(1):5, 2006.

[237] E Gerhart H Wagner and Klas Flärdh. Antisense RNAs everywhere? *TRENDS in Genetics*, 18(5):223–226, 2002.

[238] A E Walter, D H Turner, J Kim, M H Lyttle, P Müller, D H Mathews, and M Zuker. Coaxial stacking of helixes enhances binding of oligoribonucleotides and improves predictions of RNA folding. *Proc Natl Acad Sci U S A*, 91(20):9218–9222, Sep 1994.

[239] Stefan Washietl, Ivo L Hofacker, Peter F Stadler, and Manolis Kellis. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic acids research*, 40(10):4261–4272, 2012.

[240] M. S. Waterman. Secondary structure of single - stranded nucleic acids. *Studies on foundations and combinatorics, Advances in mathematics supplementary studies, Academic Press N.Y.*, 1:167–212, 1978.

[241] M. S. Waterman and T. H. Byers. A dynamic programming algorithm to find all solutions in a neighborhood of the optimum. *Math. Biosci.*, 77:179–188, 1985.

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

[243] Lauren S Waters and Gisela Storz. Regulatory RNAs in bacteria. *Cell*, 136(4): 615–628, 2009.

[244] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; A structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–738, April 1953.

[245] Kevin M Weeks. Advances in RNA structure analysis by chemical probing. *Current opinion in structural biology*, 20(3):295–304, 2010.

[246] Sandra E Wells, John MX Hughes, A Haller Igel, and M Ares. Use of dimethyl sulfate to probe RNA structure in vivo. *Methods in enzymology*, pages 479–492, 2000.

[247] Eric Westhof and Valérie Fritsch. RNA folding: beyond Watson–Crick pairs. *Structure*, 8(3):R55–R65, 2000.

[248] Eric Westhof, Benoît Masquida, and Fabrice Jossinet. Predicting and modeling RNA architecture. *Cold Spring Harbor perspectives in biology*, 3(2):a003632, 2011.

[249] M Wieland and J S Hartig. RNA quadruplex-based modulation of gene expression. *Chem. Biol.*, 14:757–763, 2007.

[250] Christopher J Wilds and Masad J Damha. 2'-Deoxy-2'-fluoro-β-D-arabinonucleosides and oligonucleotides (2' F-ANA): synthesis and physicochemical studies. *Nucleic acids research*, 28(18):3625–3635, 2000.

[251] Ross C Wilson and Jennifer A Doudna. Molecular mechanisms of RNA interference. *Annual review of biophysics*, 42:217–239, 2013.

[252] 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(17):4731, 2004.

[253] Sarah A Woodson. Metal ions and RNA folding: a highly charged topic with a dynamic future. *Current opinion in chemical biology*, 9(2):104–109, 2005.

[254] Daniel J Wright, Jamie L Rice, Dawn M Yanker, and Brent M Znosko. Nearest Neighbor Parameters for Inosine-Uridine Pairs in RNA Duplexes. *Biochemistry*, 46(15):4625–4634, 2007.

[255] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2): 145–165, February 1999. ISSN 0006-3525.

[256] Jacqueline R Wyatt, Joseph D Puglisi, and Ignacio Tinoco. RNA folding: pseudoknots, loops and bulges. *BioEssays*, 11(4):100–106, 1989.

[257] 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.

[258] A Xayaphoummine, V Viasnoff, S Harlepp, and H Isambert. Encoding folding paths of RNA switches. *Nucleic acids research*, 35(2):614–622, 2007.

[259] Bradley Zamft, Lacramioara Bintu, Toyotaka Ishibashi, and Carlos Bustamante. Nascent RNA structure modulates the transcriptional dynamics of RNA polymerases. *Proceedings of the National Academy of Sciences*, 109(23):8948–8953, 2012.

[260] Nikolay Zenkin. Ancient RNA stems that terminate transcription. *RNA biology*, 11(4):0–1, 2014.

[261] A Y Zhang, A Bugaut, and S Balasubramanian. A sequence-independent analysis of the loop length dependence of intramolecular RNA G-quadruplex stability and topology. *Biochemistry*, 50:7251–7258, 2011.

[262] Cuncong Zhong, Haixu Tang, and Shaojie Zhang. RNAMotifScan: automatic identification of RNA structural motifs using secondary structural alignment. *Nucleic acids research*, page gkq672, 2010.

[263] Christian Höner zu Siederdissen, Stephan H Bernhart, Peter F Stadler, and Ivo L Hofacker. A folding algorithm for extended RNA secondary structures. *Bioinformatics*, 27(13):i129–i136, 2011.

[264] M. Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 244 (4900):48–52, April 1989. ISSN 0036-8075.

[265] M. Zuker and D. Sankoff. RNA secondary structures and their prediction. *Bulletin of Mathematical Biology*, 46(4):591–621, July 1984.

[266] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–148, January 1981. ISSN 0305-1048.

Ronny Lorenz
Institut für Theoretische Chemie
1090 Wien
Austria
Tel: +43-1-4277-52734
Fax: +43-1-4277-52793
ronny@tbi.univie.ac.at
http://www.tbi.univie.ac.at/∼ronny/

Research Interests

- RNA secondary structures
- Riboswitches and ncRNAs
- Classified Dynamic Programming
- RNA landscape
- RNA folding kinetics

Education (University)

October 2013 –	• Research associate with Prof. Dr. Peter F. Stadler at the *University of Leipzig, Department of Computer Science (chair for bioinformatics)*
December 2007 – September 2013	• Phd student with Prof. Dr. Ivo L. Hofacker at the *University of Vienna, Institute for Theoretical Chemistry (Applied Theoretical Biochemistry Group)*
September 2000 – November 2007	• Diploma student with Prof. Dr. Peter F. Stadler at the *University of Leipzig, Department of Computer Science (focus on bioinformatics)* **thesis title**: *Secondary Structure Prediction for circular RNAs*
September 1999 – August 2000	• Diploma student at the *University of Leipzig, Department of Chemistry*

List of Publications

[1] Fabian Amman, Michael T Wolfinger, Ronny Lorenz, Ivo L Hofacker, Peter F Stadler, and Sven Findeiß. TSSAR: TSS annotation regime for dRNA-seq data. *BMC bioinformatics*, 15(1):89, 2014.

[2] Ivo L Hofacker and Ronny Lorenz. Predicting RNA structure: Advances and Limitations. In Christina Waldsich, editor, *RNA Folding*, volume 1086 of *Methods in Molecular Biology*, pages 1–19. Humana Press, 2014.

[3] R. Lorenz, S. Bernhart, J. Qin, C. Höner zu Siederdissen, A. Tanzer, F. Amman, I. Hofacker, and P. Stadler. 2D meets 4G: G-Quadruplexes in RNA secondary structure prediction. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, PP(99):1, 2013.

[4] Ronny Lorenz, Stephan H. Bernhart, Fabian Externbrink, Jing Qin, Christian Höner zu Siederdissen, Fabian Amman, Ivo L. Hofacker, and Peter F. Stadler. RNA folding algorithms with G-Quadruplexes. In Marcilio de Souto and Maricel Kann, editors, *Advances in Bioinformatics and Computational Biology*, volume 7409 of *Lecture Notes in Computer Science*, pages 49–60. Springer Berlin / Heidelberg, 2012.

[5] Ronny Lorenz, Ivo L. Hofacker, and Stephan H. Bernhart. Folding RNA/DNA hybrid duplexes. *Bioinformatics*, 2012.

[6] Ronny Lorenz, Stephan H. Bernhart, Christian Höner zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA package 2.0. *Algorithms for Molecular Biology*, 6(1):26, 2011.

[7] 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, I.L. Hofacker, and P.F. Stadler. Animal snoRNAs and scaRNAs with exceptional structures. *RNA Biology*, 8(6):938–946, 2011.

List of Publications (continued)

[8] I. Boria, A.R. Gruber, A. Tanzer, S.H. Bernhart, R. Lorenz, M.M. Mueller, I.L. Hofacker, and P.F. Stadler. Nematode sbRNAs: homologs of vertebrate Y RNAs. *Journal of Molecular Evolution*, 70(4):346–358, 2010.

[9] Ronny Lorenz, Christoph Flamm, and Ivo L. Hofacker. 2D projections of RNA folding landscapes. In Ivo Grosse, Steffen Neumann, Stefan Posch, Falk Schreiber, and Peter F. Stadler, editors, *German Conference on Bioinformatics 2009*, volume 157 of *Lecture Notes in Informatics*, pages 11–20, Bonn, September 2009. Gesellschaft f. Informatik.

[10] I. Glauche, R. Lorenz, D. Hasenclever, and I. Roeder. A novel view on stem cell development: analysing the shape of cellular genealogies. *Cell proliferation*, 42(2):248–263, 2009.

[11] Andreas R. Gruber, Ronny Lorenz, Stephan H. Bernhart, Richard Neuböck, and Ivo L. Hofacker. The vienna RNA websuite. *Nucleic Acids Research*, April 2008.

[12] I. Roeder, K. Braesel, R. Lorenz, and M. Loeffler. Stem cell fate analysis revisited: interpretation of individual clone dynamics in the light of a new paradigm of stem cell organization. *Journal of Biomedicine and Biotechnology*, 1:84656, 2007.

[13] Ingo Roeder and Ronny Lorenz. Asymmetry of stem cell fate and the potential impact of the niche. *Stem Cell Reviews*, 2(3):171–180, 2006.

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/