Albert-Ludwigs-Universtität
Freiburg im Breisgau/Germany

Ant Colony Optimization based Inverse Folding of mono- and bistable RNA Macromolecules

Dissertation

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat der Technischen Fakultät
der Albert-Ludwigs Universtät Freiburg

von

Diplom Bioinformatiker
Robert Kleinkauf

2016

Dekan
Prof. Dr. Georg Lausen

Gutachter
Prof. Dr. Rolf Backofen

Zweitgutachter
Prof. Dr. Ivo Hofacker

Kommissionsvorsitz
Prof. Dr. Christoph Scholl

Kommissionsbeisitz
Prof. Dr. Gerald Urban

Datum der Promotion
27.06.2016

I will take the Ring, though I do not know the way.

Frodo Baggins
The Council of Elrond

ACKNOWLEDGEMENT

...I salute Mr. Frodo Baggins for carrying the Ring and ultimatively destroying it in the fires of Mount Doom...

Well, yes, in that case, it was me who 'handed in' the Ring! But it is very selfish to start with oneself, however, on the way to Mount Doom, Frodo received support, help and company through direct and indirect open and secret support by friends and benefactors. Without those fundamental prerequisites, he would not have gotten any further than to the borders of the Shire.

Therefore,

... I want to thank Rolf Backofen: Thank you very much for giving me the opportunity to be part of your chair in Freiburg and let me have my software tinkering workbench in the lab.

...I want to thank Ivo Hofacker for the kindness of accepting my thesis in the role as the second supervisor.

...I want to thank Martin Mann for his very patient style of mentorship, support and care.

...Especially I want to thank Monika for being the good sprite all the time! Without you some peculiarities would have been unnecessary burdens! I also want to thank Stefan, the ghost in the machine. Without you, life would have been even more complicated, and I would have been stuck.

...I want to thank the ever changing composition of the Bioinformatics hair working alongside during the time. You have been the cause of the constant puzzle and the partially collaborative possibility to get back on track! Without you, the whole chair would be just another dusty place filled with nothing but streamlined smoothies and other strange characters! It was a nice experience with you!

Last but not least I want to thank my family and my friends for their lively support, their constructive encouragement to diverse concerns and their being themselves in diverse situations. Without you, this whole endeavor would have been impossible.

Thank You!

ABSTRACT

Since the discovery of structural conformations of DNA in the middle of the 20^{th} century not only technologies that can elucidate structures of biologically relevant molecules have become more sophisticated, but the understanding of biological processes on the molecular level in general has grown tremendously in the last 60 years of research. In this same time, the early and dogmatic statement, according to which proteins are the only entities in molecular biological perception that can perform and provide necessary biological, e.g. enzymatic, functions within organisms, has undergone major revision. Of course, proteins do still perform the functions, which have been annotated, but in addition to the level of control of the proteins, specific RNA molecules, namely the non-coding RNAs, have been accounted to the executing functional level so far exclusive to proteins. As in the case of the proteins, a functional RNA receives its specific function from a biologically active structure conformation, which strongly correlates with the respective RNA sequence. A comparatively large number of sequences can fold into a similar structural RNA conformation. However, small perturbations as for example point mutations or sequestering of parts of the RNA sequence through other interacting entities can be key for the disruption of the functional structure of the RNA. Alongside with the exploration of new RNA functionalities, RNA based technologies have been derived from single RNA based functionalities and corresponding mechanisms. Their analytical and creative potential in combination with hereof derived computer programs, e.g. predicting structures from RNA sequences or vice versa, predicting RNA sequences from structures, extend the classical biological approach beyond its investigative origin by adding a progressive engineering spirit to the former purely research character.
In this dissertation, a computational RNA design tool and its application performance are presented. The tool is conceptually based on a relatively long heritage of tools, which can solve the 'RNA inverse fold' problem: Given a structure (mostly secondary RNA structure), the programs pursuit different strategies to produce a sequence, which can fold into the specified structure input. Classically a single structure was given as input. With the presented tool and its several capabilities of solving different levels of structural complexity based on RNA secondary structures, not only a new way of solving the problem with the heuristic approach of the ant-colony optimization technique was introduced. Furthermore, new constraints such as the regulation of a very precise GC content of the solution sequence has been given major concern in the concept as well as new structural constraint possibilities of pseudoknots and bistable RNA entities. The new introduced features are benchmarked and tested on structure complexity specific data sets, which have been gained from online data bases and corresponding literature efforts. Also comparative representations with other state of the art computer programs are given.

ZUSAMMENFASSUNG

Seit der Entdeckung der Strukturkonformation von DNA in der Mitte des 20. Jahrhunderts wurden nicht nur die strukturellen Aufklärungsmethoden für biologisch relevante Mikro- und Makromoleküle weiterentwickelt, vielmehr wurde auch das Wissen um biologische Prozesse auf molekularer Ebene enorm erweitert. Im selben Zeitrahmen wurde aber auch das sehr lange vorherrschende Dogma, nach welchem ausschließlich Proteine die einzigen Entitäten innerhalb von Organismen seien, die biologisch relevante, wie z.B. enzymatische Funktionen innehaben, durch Weiterentwicklungen und Neuentdeckungen wiederlegt, beziehungsweise überholt. Selbstverständlich vollziehen die Proteine weiterhin deren annotierten charakteristischen Funktionen innerhalb der Organismen, dennoch wurde die Welt der Proteine funktional um die Klasse der nicht codierenden RNA erweitert. Diese sind in der Lage, ähnliche Aufgaben zu bewerkstelligen, wie die Proteine selbst. Wie bei den Proteinen leitet sich die Funktion einer RNA von deren Strukturkonformation ab, die wiederum von der jeweiligen RNA Sequenz abhängt. Obwohl eine im Vergleich zu den Proteinen größere Anzahl an Sequenzen in immernoch die selbe RNA Struktur falten kann, können durch kleine Störungen, wie zum Beispiel Punktmutationen oder auch durch Interaktion von anderen Faktoren an bestimmten Stellen der RNA die Faltungskompetenz einer RNA Sequenz drastisch verändern, so dass die eigentliche biologische Funktion unterbrochen werden kann. Mit den Entdeckungen neuer RNA Funktionen, wurden RNA basierende Techologien aus einzelnen Funktionen und deren Mechanismen abgleitet. Das hierbei innewohnende analytische und kreative Potential der jeweiligen Technogien gekoppelt und einhergehend mit theoretisch-biologischen Computerprogrammen, die z.B. durch Faltungsvorhersagen für Sequenzen, d.h. gegeben einer RNA Sequenz, deren Structur berechnen können, bzw. umgekehrt, durch inverse Faltungsvorhersagen von Strukturen, d.h. gegeben einer Struktur, die jeweilige Sequenz berechnen können, lässt es zu, bzw. fordert es gerade zu heraus, dass der klassische biologische Forschungsansatz über seinen investigativen Charakter hinauswächst und sich zu einer Ingeneursdisziplin weiterentwickelt.
Innerhalb der vorliegenden Dissertation wird ein Konzept zum Lösen des Problems des RNA Inverse Folding vorgestellt. Die Anwendungsperformanz der daraus hervorgegangenen Implementierung wird vergleichend zu bisherigen Programmen dargestellt. Das Problem des RNA inverse foldings wurde in einem relativ langem Erbe von Programmen schon zu vor gelöst. Die jeweiligen Programme gehen hierbei das Progblem des RNA inverse folding auf unterschiedliche Arten und Weisen an und erlauben es mit einer jeweilig spezifischen Lösungskomplexität das Problem zu lösen. Auch die in der Arbeit verwendete Methode der Ameisen-Kolonie Optimirung hat eine Historie und kam in unterschiedlichsten Bereichen bereits zum Einsatz. Alle RNA inverse folding Programme liefern RNA Sequenzen, die die Eigenschaft besitzen sollen, in eine als Eingabe definier-

te Sekundärstruktur der RNA falten zu können. Hierbei wurde in klassischen Ansätzen eine einzelne Struktur als Eingabe zugelassen. Mit dem hier präsentierten Programm und seinen unterschiedlichen Möglichkeiten der Problemlösung des RNA inverse folding auf unterschiedlichen strukturellen Komplexitätsebenen der RNA SekundÃdrstruktur, wurde nicht nur ein neuartiger Weg zur Lösung des Problems mit Hilfe der heuristischen ant-colony optimization (Ameisen Kolonie Optimierung) dargestellt, vielmehr wurden neuartige Problemrandbedingungen (Constraints) mit berücksichtigt und zur Lösung des Problems herangezogen, so dass es mit nun Hilfe des dargestellten Weges möglich ist, sehr präzise den GC-Gehalt der zu designenden RNA Sequenz einstellen zu können. Zusätzlich können durch die Entwicklung geeigneter Darstellungsmöglichkeiten komplexe Struktureingaben gemacht werden, so dass es mit dem vorgestellten Ansatz auch möglich ist, Pseudoknoten und bistabile Strukturkonformationen zu modellieren. Nach einer generellen EinfÃijhrung in die Themen nicht codierende RNA und deren Faltung, wird die Adaption der Ameisen Kolonie Optimierung hinsichtlich des RNA inverse foldings afgebaut. Die jeweilig benutzten Konzepte und deren Zusammenspiel werden dargestellt. Letztlich werden die Fähigkeiten des Programms in Vergleichs- und Leistungstests auf der Basis von verschiedenen Strukturkomplexitäten mit Hilfe von Daten ermittelt, die aus Online-Datenbanken entnommen wurden, sowie aus Literturrecherchen hervorgegangen sind. Vergleichende Ergebnisse zu bisherigen 'state-of-the-art' Programmen werden präsentiert.

CONTENTS

CHAPTER 1

INTRODUCTION

Survey

In this introductory chapter of the thesis, a brief overview of the essential threads of topics is presented. The different threads contribute primal to the understanding of the different topics, which are mentioned and considered throughout the course of this thesis and its underlaying work. Even though each topic has its own origin and forms its own 'discipline' in academia, their synergistic liaison is not negligible and arranges the basis of computational biology and its bioinformatic approaches within the mentioned fields.

Each of the involved topics is compendiously presented and characterized in a more or less condensed manner. They were sorted and grouped into sections, which allow to introduce the biological foundation of the endeavor. Subsequently they allow to introduce formal definitions, which are needed in order to fortify the presented algorithmic work, which was performed within the frame of this thesis. In addition, basic facts and concepts are described, such that the biological but as well the algorithmic principles can be used to explain and motivate the performed project. Alongside, substantiating examples of different already existent approaches on the part of biotechnological technologies but as well on the part of preceding algorithmic work in the field of computer science are listed to show their importance within the respective research areas and their applications and their impact on society.

1.1 Macromolecular Biopolymers

During a nearly 5 trillion year long evolutionary process, highly specific complex organizational patterns emerged on Earth. Let's call it the life, as we know it. In the beginning, present molecules reacted and in an ongoing process formed noticeably more complex states until organisms evolved. Over the course of time, several earth ages saw the spawning and the extinction of various species. Nowadays, organisms show a high diversity, ranging from minimalistic virus organization over unicellular procaryotes to multicellular 'higher' life forms. Though the present variety is tremendous, Life (on Earth) shares cellular key principles, which are similar to all organisms. Those key principles are mediated by several classes of micro and macro molecules. Hereby, the major concerted process of each organism is to propagate its heritage information about its organization and functionality. Dependent on the organism, this specific information is stored and maintained as molecules of Desoxy-Ribonucleic Acid (DNA) in a redundant/non-redundant form (polyploid/haploid) in each cell.

1.1.1 DNA

DNA is a macromolecule, which consists of polymerized nucleotides (Watson and Crick, 1953). The stored information is encoded with the help of four different nucleotides: Adenine (A), Cytosine (C), Guanine (G) and Tymine (T). Due to its chemical features, namely the complementary character of nucleobases and their ability of forming hydrogen bonds, DNA is able to adopt a double helical structure conformation consisting of a plus and a minus strand. The two strands of DNA which form the double helix, are reverse complements in the composition of nucleotides in each position. In higher cells, more information is saved within the cell especially in the DNA, such that DNA of higher organisms is usually much longer than those of lower organisms, therefore and in order to be able to maintain regulatory control over the assured information, that is actually used, the DNA is compacted by the use of nucleosomes. They allow on-demand de-/coiling of the DNA, such that temporarily unneeded sections can be stored in a space saving manner. Nucleosomes are complexes of histone proteins and represent the basic unit, which is specialized to bind to DNA, such that 147 DNA nucleotides can be wound around one histone complex. With the DNA, the nucleosomes form the basic layer within the chromosomal organization. Presumably further histone proteins attach to this complexes and condense the DNA/histone complex further. The condensed form of DNA is called heterochromatin, whereas the loose form of DNA is called euchromatin.

1.1.2 From DNA to RNA

In order to execute and gain usage from the information, which is stored within the DNA, temporary copies of DNA are made. The copying process is termed 'transcription'. Transcripts show similar characteristics as the DNA and also are member of the chemical class of polynucleotides, but show an additional alcohol group at the single nucleotides' sugar rings (2'C). Nevertheless this 2'C sugar oxidized form of DNA is called Ribonucleic Acid (RNA). Still, this slight change in chemical property allows to use almost the same material for transmitting information to various locations and for various reasons.

DNA is read and transcribed to RNA by complexes called RNA polymerase. Mechanistically there is a difference between procaryotic and eucaryotic systems, on how the DNA is recruited to the RNA polymerase and how the transcription is terminated. The difference is not only manifested by the mechanism of recruitment of RNA polymerases but is detectable by the transcription rates (nucleotides per seconds nts^{-1}) in the different types of organisms. They range from 10-20 nts^{-1} in human cells, to 20-80 nts^{-1} in bacterial systems and up to 200 nts^{-1} in phages (Pan and Sosnick, 2006).

Procaryotes have one type of RNA polymerase. The RNA polymerase is guided by its σ-factor to the promotor region of a transcription start site. When the RNA polymerase is attached correctly to the start site, it will start to transcribe the template DNA until it encounters a termination signal within the DNA.

In eucaryotic systems, three different types of RNApolmerases exists. While type I and III are involved in the transcription of ribosomal, transfer and other small RNAs, RNA polymerase II (RNApolII) plays the major role in eucaryotic general RNA polymerization. In contrast to the procaryotic RNApolmerase, which is almost capable on its own to initiate and execute transcription, eucaryotic RNApolII requires a multitude of different factors in order to be able to transcribe a stretch of DNA into RNA. This increase in complexity of reading DNA in eucaryotes is due to the dense packing of the DNA and its more sensitive transcription rate adjustability. RNApolII requires a set of general transcription factors, which makes the whole process of unwinding and replicating the DNA possible by recruiting the RNApolII to correct positions on the DNA. The factors are termed general, due to their ability to bind to all promotors within the respective organism. Roughly, the eucaryotic transcription factors act in a σ-factor-like manner. Before a RNApolII molecule can be recruited to the DNA it might be necessary to bring the DNA into a 'readable' form. This is achieved by proteins called transcriptional activators. Furthermore, they support the recruiting of the transcription factors and ultimately lead to the recruitment of RNApolII to a specific position. Histone acetylases and chromatine remodeling complexes are required to remodel the state of the chromatin such that the transcription initiation complex can form.

Once the transcription started, elongation factors are required to keep the RNApolII on the DNA while transcribing it to RNA.

DNA regions, which can be transcribed into biological active RNA molecules are termed genes and occur in different forms of organization of regulation. Operons control the 'readability' of a gene. In that way it is possible to precisely orchester the activities of all genes. A gene can be arranged, such that is is transcribed individually. Nevertheless, genes can be organized in groups of genes, which are subordinated under the control of one operon. This often happens, for example in gene cassettes, whose gene products are enzymes of a certain metabolic pathway and are needed all at the same time and quantity.

The transcript RNA, in respect to the organism it was produced in, has to be further processed or can directly be used in its function itself. In general, a procaryotic transcript does not require further processing in order to apply its function. Eucaryotic transcripts, however, needs to be further processed in order to become functional.

In general, there are two types of RNA: There are RNAs which are translated into protein in a procedure called 'translation'. In this process, the nucleotide code is translated into the amino acid code. Amino acids are the monomers of proteins, therefore, this class of RNA is labeled protein-coding RNA. All RNA which is transcribed but not translated into any protein is called non-coding RNA (ncRNA). ncRNA, however, is functional on the RNA level itself and administers a wide spectrum of functionalities. The functionality of a respective ncRNA derives from its structure, which derives from its sequence. The focus of this thesis is to obtain certain RNA sequence, which can fold into specified structure, therefore more detail on RNA but specifically on ncRNA is introduced and dealt with throughout the thesis.

1.2 RNA

In this section, RNA as such is introduced and described in more detail. Build upon biochemical and formal RNA definitions RNA structure is introduced as such, so that the reader can get an initial sense for the importance of RNA structure and its affiliated biological functionality. Based on the initial knowledge, relevant contemporary RNA technology is introduced in the subsequent parts of the section. In the algorithmic part of the introduction, a brief insight to the RNA structural folding problem and its counterpart, the RNA inverse folding problem is presented. Within, a certain survey of the different approaches and strategies are outlined. By mentioning and motivating the ant colony optimization in the last part of the introduction, the circle of conceptual argumentation is closed, such that in the recapitulation of the introduction, a clear purpose of the thesis is stated.

1.2.1 Biochemical Definition

Chemically, a sequence of RNA consists of condensed nucleotides. Each nucleotide consists of a phosphate residue, a pentose sugar ring, and a nucleobase, which is a nitrogeneous base. The phosphate group and the pentose sugar ring form the structural backbone of the RNA molecule. The character of the respective nucleobase residue specifies the nucleotidic character. Within RNA polymers, four different nucleotides can be found: Adenine (A), Cytosine (C), Guanine (G) and Uracil (U). In Figure 1.1 the basic nucleotides are depicted. Adenine and Guanine are categorized into the chemical class of purine molecules, whereas Cytosine and Uracil are members of the chemical class of pyrimidines. A nucleotide is chemically a nucleoside, which has a (mono-, di- or tri) phosphate residue bound to its 5'-Carbon atom. The nucleoside is the pentose sugar ring covalently bound with its 2'-Carbon atom to the 9-Nitrogen atom of affected purines or to the 1-Nitrogen atom of pyrimidines.

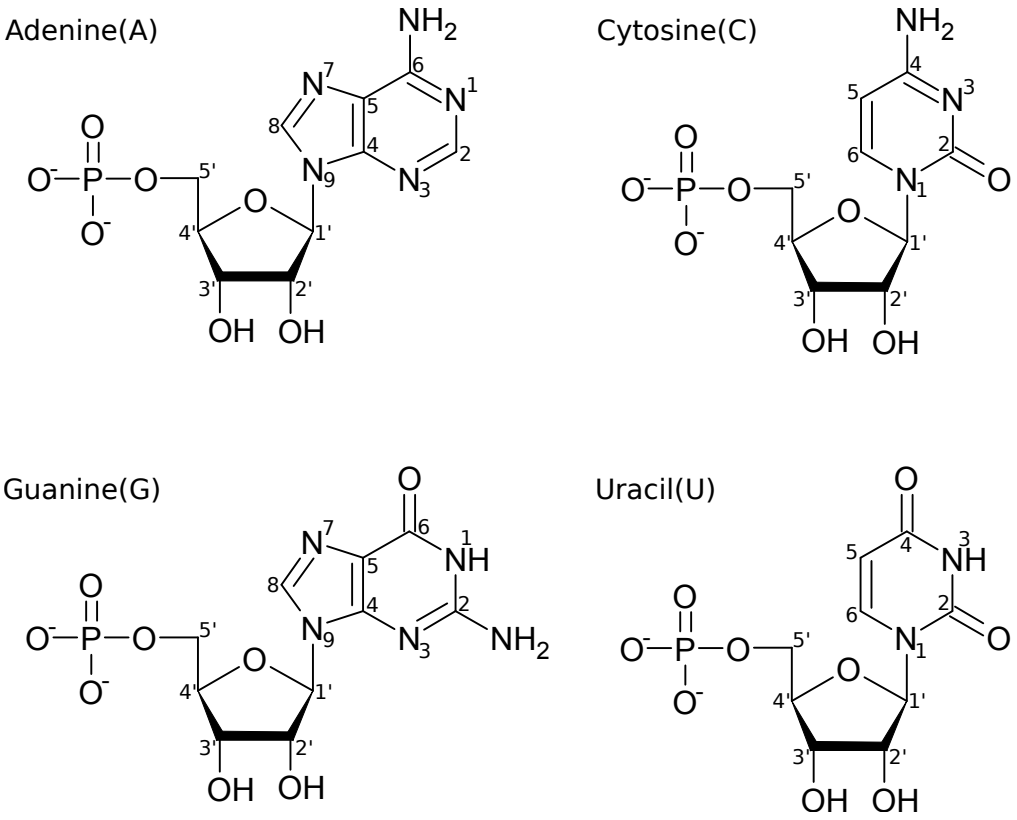

Figure 1.1: RNA Nucleotides
Canonic nucleotides (Nucleoside-Monophosphate) present in RNA molecules according to IUPAC. Modified templates from ChemSketch/ACD Labs.

The RNA polymerization reaction is an addition reaction of a nucleosidtriphosphate and an RNA molecule, which already consists of n added nucleotides. The product is a diphosphate and the extended RNA molecule with $n + 1$ nucleotides.

$$\text{Nucleotidetriphosphate} + \text{RNA}_n \rightarrow \text{Diphosphate} + \text{RNA}_{n+1}$$

This polymerization reaction is facilitated by the molecular complex of RNA polymerase which is classified as nucleotidyltransferase. The reaction itself is propelled by the dissociation of the diphosphate group of the Ribonucleotidetriphosphate (RTP).
The nucleobases of the polymerized nucleotides have the chemical property to be able to form hydrogen bonds. According to IUPAC, a hydrogen bond is defined as: The hydrogen bond is an attractive interaction between a hydrogen atom from a molecule or a molecular fragment $X-H$ in which X is more electronegative than H, and an atom or a group of atoms in the same or a different molecule, in which there is evidence of bond formation (Arunan *et al.*, 2011b,a).
Hydrogen bonds, in comparison with covalent bonds, are roughly double in length. In RNA, there are basic two types of hydrogen bonds, which are responsible for the base pair formation: $N - H \cdots : N$, 3.1 kcal/mol and $N - H \cdots : O$, 1.9 kcal/mol. If two nucleobases of two distinct but compatible nucleotides of one RNA molecule form hydrogen bonds, they form a base pair, the smallest 'building block' of RNA structure. The 'canonical' base pairs between either A and U or G and C and the 'non-canonical' base pair between G and U. The base pair GC can form three hydrogen bonds and is therefore stronger in comparison with the base pair AU and GU, which respectively only form two hydrogen bonds. However, due to the composition of the hydrogen bonds within the different base pairs, the strength of th AU base pair is stronger than the energy of GU.
Based on the nucleotides' capability of forming base pairs with complementary nucleotides or being incompatible to each other, each sequence of RNA can engage a unique set of potential structures. A certain RNA sequence is termed primary structure, a concrete self-interaction of an RNA sequence, which is based on base pairs is called secondary structure. Subsets of base pairs of a secondary structure, which are in specific neighborhood relations, are pooled together and labeled as a secondary structure loop elements. In Figure 1.4, a selection of basic secondary structure elements is displayed. Dependent on which structure complexity is considered, a structure of an RNA can be labeled nested or non-nested pseudoknot structure (Rietveld *et al.*, 1982). Pseudoknot structures mark the beginning of the transition from the structure categories 'secondary structure' to 'tertiary structure', since their complexity exceeds the definition of single nested secondary structure elements, but still can be noted as a secondary structure. Pseudoknots are found to be very important structural motifs in various processes, such as in ribozymes (Rastogi *et al.*, 1996; Ke *et al.*, 2004), self-splicing introns (Adams *et al.*, 2004), and telomerase activity (Theimer *et al.*, 2005).

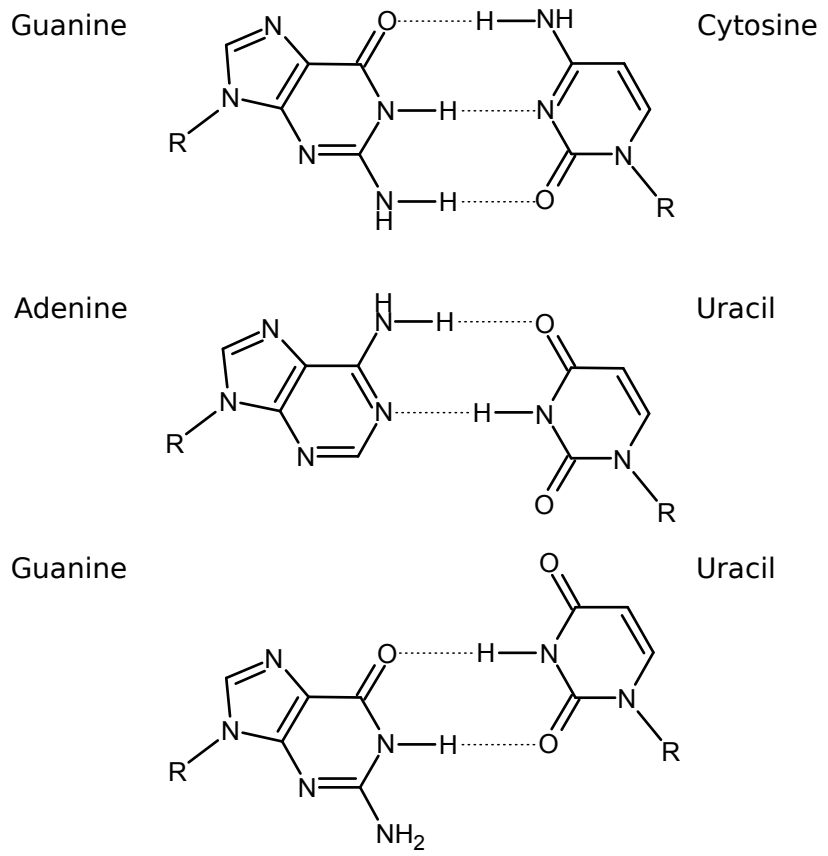

Figure 1.2: RNA Base Pairs
Including the 'canonical' AU and GC and in addition the less strong GU base pair. Modified templates from ChemSketch/ACD Labs.

Additionally, pseudoknots play critical roles in altering gene expression by inducing ribosomal frameshifting in many viruses (Shen and Tinoco, 1995; Nixon *et al.*, 2002; Michiels *et al.*, 2001). Exemplary basis pseudoknot structures are depicted in Figure 1.3. The motifs are very wide spread among pseudoknot structures in nature.

However, the secondary structure representation of RNA 'only' is a reduction of the three dimensional tertiary structure conformation, which provides more interaction information sourcing from the involved nucleotides such as intercalation effects of the sugar rings . The secondary structure interactions, namely the base pairs, remain established within the tertiary structure and comprise to a large extend the 3D structure, such that the secondary structure information is a very good but not complete approximation of the 3D structure. Besides the increased complexity of RNA tertiary structure, it is computationally very expensive. Thus, the secondary structure representation can be a cheap calculative means to explore and investigate RNA structure in a substitute

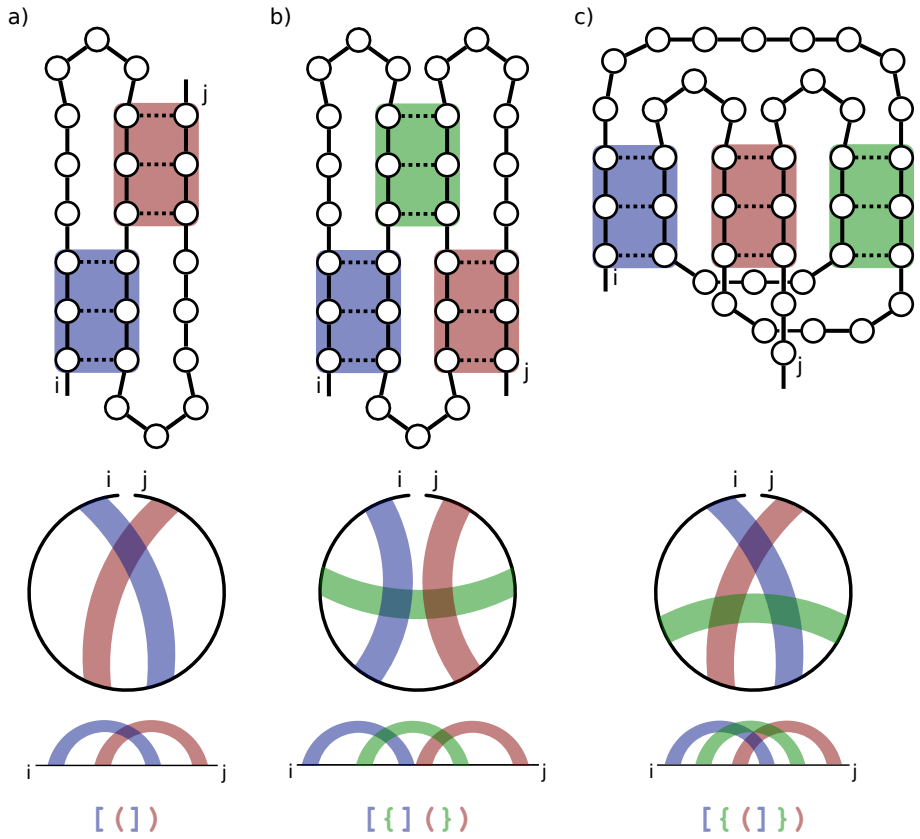

Figure 1.3: RNA characteristic pseudoknot crossing base pairs
Column a) *hairpin-*, Column b) *kissing-hairpin-* and Column c) *threeknot*-pseudoknot examples. The different rows represent different representations of the same structure. R1: secondary structure graph, R2: Circular Plot, R3: Arc Plot, R4: DotBracket Shape Representation (Giegerich *et al.*, 2004). The respective necessary base pairs (stacks) are indicated by color (red, blue and green). *hairpin-* and *kissing-hairpin*-pseudoknots are the most frequent pseudoknots, but also more complex but less frequent pseudoknots are possible.

manner. However, one should keep in mind, that the secondary structure is just a prestep to the tertiary structure interaction. Tertiary structure interaction is less stable than secondary structure interaction. Therefore, it is assumed that RNA secondary structure can be mostly independent from tertiary structure influence (Crothers *et al.*, 1974; Banerjee *et al.*, 1993), but as already seen above, tertiary interaction provides biological functionality (Pleij *et al.*, 1985).

Formal Definition

In order to computationally support the analysis of RNA, an appropriate integration of a suitable RNA representation is required. Throughout the thesis, the following formal definitions of RNA sequence and structure are used. They were employed in different stages of the project respectively.

Definition 1 (RNA Nucleotide Sequence)**:** Let $\mathcal{S}$ be an RNA sequence of length n, such that $\mathcal{S} = \mathcal{S}_1, ..., \mathcal{S}_n$. Each sequence position $\mathcal{S}_i$ consists of one nucleotide of the alphabet $\Sigma_{RNA} = \{A, C, G, U\}$, such that $\mathcal{S}_i \in \Sigma_{RNA}$. A sequence is called gapped, if at least one character of the sequence is substituted by the gap symbol $'\_'$.

Definition 2 (GC Value)**:** Let $GC(\mathcal{S})$ be the GC content value of a RNA nucleotide sequence $\mathcal{S}$, such that $GC(\mathcal{S}) = \frac{|\{\ i \mid \mathcal{S}_i \in \{G,C\}\}|}{n}$.

Definition 3 (Secondary Structure)**:** Let $P = \{(i,j) \mid i,j : i < j - 3\}$ be s set of position pairs over the RNA sequence $\mathcal{S}$. P is called a secondary structure, if all positions interact only once, i.e. $\forall (i,j),(k,l) \in P : i \neq j \neq k \neq l$. If $\exists (i,j),(k,l) \in P : i < k < j < l$, the structure is labeled non-nested or called a pseudoknot. If $\forall (i,j),(k,l) \in P : i < k < l < j \vee k < i < j < l$ holds, the structure is labeled nested.

Definition 4 (Secondary Structure String Representation)**:** Let w^P be a Dot-Bracket representation of a secondary structure P for an RNA sequence $\mathcal{S}$ of length n, such that each position within the string is of the alphabet $\Sigma_P = \{(,),[,],\{,\},<,>,.\}$. Within the string $w^P \in (\Sigma_P)^n$, each base pair $(i,j) \in P$ is indicated by a pair of corresponding brackets on the respective positions w_i^P and w_j^P. A single stranded position is indicated by the literal $'.'$.

Definition 5 (Lonely Base Pairs)**:** Let $LP(P)$ be a set of lonely base pairs of a secondary structure P, such that

$$
\begin{aligned}
LP(P) = \quad & \{\ (i,j) \mid \nexists_{(k,l) \in P} : |i-k| + |j-l| = 2 \\
& \quad \vee ((i+1, j-1) \in P \wedge (i-1, j+1) \notin P \wedge (i+2, j-2) \notin P) \\
& \quad \vee ((i-1, j+1) \in P \wedge (i+1, j-1) \notin P \wedge (i-2, j+2) \notin P) \\
& \} \qquad (1.1)
\end{aligned}
$$

Note: The set $LP(P)$ contains both, truly lonely base pairs (first part in Equation 1.1), that have no immediate neighbors, as well as lonely stacks of two base pairs (second and third part).

Definition 6 (Single Stranded Nucleotides)**:** Let $SS(P) = \{i \mid i : \nexists(i,j),(j,i) \in P\}$ be the set of all sequence positions, which are not involved in any base pair in the structure P.

Definition 7 (Secondary Structure Loops)**:** Let $\phi_{(i,j)} = (L_{BP}, L_{SS})$ be a secondary structure loop of a nested structure P of a given sequence $\mathcal{S}$, which consists of a set of bases $L_{SS} \subseteq SS(P)$ and a set of base pairs $L_{BP} \subseteq P$, which all are enclosed by and accessible from a base pair $(i,j) \in P$, which itself is not part of the loop it encloses, $(i,j) \notin \phi_{(i,j)}$, i.e.:

1. $L_{SS} = \{\, i' \mid i < i' < j \wedge \nexists(k,l) \in P : i < k < i' < l < j\}$
2. $L_{BP} = \{(i',j') \mid i < i' < j' < j \wedge \nexists(k,l) \in P : i < k < i' < j' < l < j\}$

Depending on the context and the content of the structure loop $\phi_{(i,j)}$, the loop is categorized into one of the following structure elements:

1. **Hairpin Loop**: the base pair $(i,j) \in P$ encloses a hairpin loop, if:
   $\forall i < i' < j' < j : (i',j') \notin L_{BP}$.
2. **Base Pair Stack**: the base pair $(i,j) \in P$ is called stacking base pair, if:
   $(i+1, j-1) \in L_{BP}$
3. **Internal Loop**: Two base pairs $(i,j),(i',j') \in P$ form an internal loop (i,j,i',j'), if:
   (a) $(i',j') \in L_{BP}$
   (b) $(i'-i) + (j-j') > 2$ (no stack)
4. **Bulge Loop**: A Bulge loop is a special case of an internal loop. In addition one of the following statements must be true:
   $j = j' + 1$ or $i' = i + 1$
5. **k-Multiloop**: The base pair (i,j) closes a k-multiloop, which is enclosed by k additional internal base pairs $L_{BP} = \{(i_1,j_1),\ldots,(i_k,j_k)\}$, such that:
   $\forall (i_x,j_x),(i_{x+1},j_{x+1}) \in L_{BP} : j_x < i_{x+1}$
6. **External Bases**: The bases L_E, in which each base ist not accessible from any loop closing base pair (i,j), are called external bases. $L_E = \{k \mid \forall (i,j) \in P : k < i \vee k > j\}$

Since each base pair denotes uniquely one loop, we can decompose each secondary structure uniquely into loops. With $\Phi_P = \{\phi_{(i,j)} \mid (i,j) \in P\}$ we denote the unique decomposition of P into loops.
For illustrations of instances of secondary structure elements, see Figure 1.4.

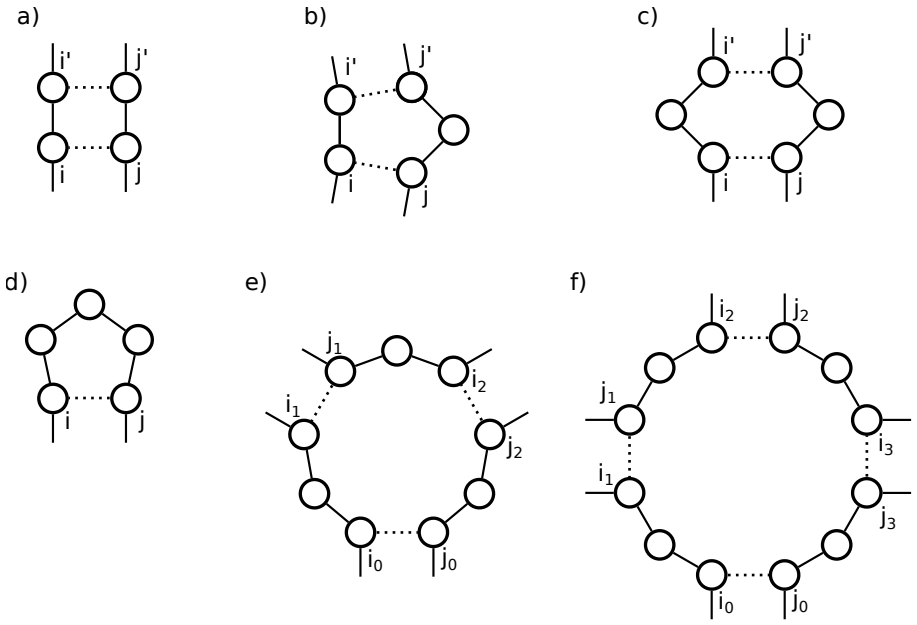

Figure 1.4: RNA secondary structure loop decomposition elements:
a) Stack, b) Bulge, c) Internal Loop, d) Haipin Loop, e) 2-Multiloop, f) 3-Multiloop

Depending on its nucleotide composition in the sequence, each RNA secondary structure has an affiliated energy value. A common measure to determine an RNA secondary structure's energy $E(P, \mathcal{S})$ for a given secondary structure P and a sequence $\mathcal{S}$, is the nearest-neighbor energy model (NNEM) (Turner and Mathews, 2002), which is based on the Gibbs free energy ΔG, compare to Section 1.2.4. The NNEM predicts the change of free energy ΔG which occurs, if a certain structural conformation P is folded from a unfolded structure-less RNA conformation. In order to compute the free energy of a structure, the nearest-neighbor model decomposes the whole structure P in secondary structure loops, such that a unique allocation of all bases and base pairs of the structure into secondary structure loops is accomplished. For each secondary structure element, a specific free energy value can be calculated, such that the overall free energy of the structure P can be predicted. The free energy ΔG of a secondary structure P of a sequence $\mathcal{S}$ is approximated by $E(P, \mathcal{S})$ according to the energy of its secondary structure loops Φ_P and the corresponding nucleotides within those loops. Turner and Mathews (2002) provide a database, where around 7500 specific energy values $E(\phi_{(i,j)}, \mathcal{S})$ for the various loops and their sizes respecting certain nucleotide compositions are available.

The free energy of structure P, given a sequence $\mathcal{S}$ and the corresponding loop decomposition Φ_P can be predicted by Equation 1.2.

$$\Delta G(P) = E(P, \mathcal{S}) = \sum_{\phi_{(i,j)} \in \Phi_P} E(\phi_{(i,j)}, \mathcal{S}) \tag{1.2}$$

Since an RNA nucleotide molecule is versatile and flexible in forming structures and does not only form one single structure until its degradation, the plenitude of structure a single RNA can fold into has to be covered and described as well. The collection of all structures an RNA molecule can fold into is called structural ensemble $\mathcal{P}_\mathcal{S}$.

Definition 8 (Secondary Structure Ensemble)**:** Let

$$\mathcal{P}_\mathcal{S} = \{P \mid P \text{ is secondary structure of } \mathcal{S}\}$$

be the secondary structure ensemble of the sequence $\mathcal{S}$.

According to the nearest-neighbor energy model, structures of a single RNA entity do not all have the same affiliated energy. Each structure has its own energy value; while two different structures can have the same energy value. All structures of a sequence can be represented as a Boltzmann weighted description of the corresponding equilibrium state of all structures, namely the partition function $Z(\mathcal{P}_\mathcal{S})$. In the partition funktion each structure of the ensemble gets weighted according to its free energy.

Definition 9 (Partition Function of RNA)**:** Let $Z(\mathcal{P}_\mathcal{S}) = \sum_{P \in \mathcal{P}_\mathcal{S}} e^{-\frac{E(P,\mathcal{S})}{RT}}$ be the partition function over the structure ensemble $\mathcal{P}_\mathcal{S}$ of a given RNA sequence $\mathcal{S}$, the gas constant R and a temperature T of the investigated system.

According to the weight of a structure within the Boltzmann equilibrium distribution, a specific probability can be calculated for each structure. Based on the probabilities assigned to each single structure, single base pair probabilities can be computed as well.

Definition 10 (Probability of a Structure)**:** Let $Pr(P|\mathcal{S}) = \frac{e^{-\frac{E(P,\mathcal{S})}{RT}}}{Z(\mathcal{P}_\mathcal{S})}$ be the probability of the structure P of a given sequence $\mathcal{S}$ in the corresponding structure ensemble $\mathcal{P}_\mathcal{S}$ in equilibrium.

Since the energy of a structure dominates largely the probability of a structure in the Boltzmann weighted structure ensemble, a structure, which has the highest probability, or equivalently the lowest free energy, is termed minimum-free-energy (MFE) structure.

Definition 11 (Minimum Free Energy MFE-Structure)**:** Let the minimum free energy MFE structure P_{MFE} be a structure which holds $\forall P \in \mathcal{P}_{\mathcal{S}} : E(P_{\text{MFE}}, \mathcal{S}) \leqslant E(P, \mathcal{S})$.

The probability of two nucleotides i and j of an RNA molecule of being in a base pair (i, j) depends on the Boltzmann weighted energies of all (i, j) containing structures, normalized by the whole structure ensemble.

Definition 12 (Probability of a Base Pair)**:** Let $Pr(i, j) = \sum_{P \ni (i,j)} Pr(P|\mathcal{S})$ be the probability of a certain base pair (i, j) within a secondary structure ensemble $\mathcal{P}_{\mathcal{S}}$ of a sequence $\mathcal{S}$.

Often we want to enforce some certain base pairs in the structure folding (e.g., since they are now to be conserved). For that purpose, we introduce a folding constraint as used by *RNAfold*, which is a structure string w of a secondary structure P as given in Definition 4. However, *RNAfold* internally interprets dots within the string differently, namely they are free to pair or not to pair with other positions. This corresponds to a limited ensemble defined as follows.

Definition 13 (Limited Structure Ensemble)**:** Let w^P be a representation of a folding constraint that encodes an associated secondary structure P. A limited structure ensemble $\mathcal{P}_{\mathcal{S}|w^P} \subseteq \mathcal{P}_{\mathcal{S}}$ is defined as

$$\mathcal{P}_{\mathcal{S}|w^P} = \{P' \mid P' \cap P = P\}.$$

That is, the structures in $\mathcal{P}_{\mathcal{S}|w^P}$ contain the base pairs $(i, j) \in P$, which implies there is no base pair overlapping with P ($\nexists (k, l) \in P' : \exists (i, j) \in P : \{i, j\} \cap \{k, l\} = \varnothing$).

To perceive a global overview on the probability situation of an RNA entity, the base pair probability matrix $M_{\mathcal{P}_{\mathcal{S}}}$ can be derived from a structure ensemble $\mathcal{P}_{\mathcal{S}}$ of an RNA sequence $\mathcal{S}$ of length n, as stated in the following definition.

Definition 14 (Base Pair Probability Matrix)**:** Let $M_{\mathcal{P}_{\mathcal{S}}}$ be a base pair probability matrix $n \times n$ of the structure ensemble $\mathcal{P}_{\mathcal{S}}$ of a sequence S of length n, such that $M_{\mathcal{P}_{\mathcal{S}}}(i, j) = Pr(i, j)$.

Definition 15 (Accuracy)**:** Given a nested secondary structure P and a base pair probability matrix $M_{\mathcal{P}_\mathcal{S}}$ then let the accuracy of the structure P within the base pair probability matrix $M_{\mathcal{P}_\mathcal{S}}$ constitute as

$$\text{Accur}(P, M_{\mathcal{P}_\mathcal{S}}) = \frac{1}{|P|} \sum_{(i,j)\in P} M_{\mathcal{P}_\mathcal{S}}(i,j)$$

Definition 16 (Accessibility Structure)**:** Let $P_{access} \subseteq [1,n]^2$ be a set of position pairs over an RNA sequence $\mathcal{S}$ of length n. P_{access} is called an accessibility structure, if all position pairs interact only within a delimiting position pair $(k,l) \in P_{access}$, such that

$$P_{access} = \{(i,j) \mid (i,j) \in (\{k \cdots l\} \times \{k \cdots n\} \cup \{1 \cdots l\} \times \{k \cdots l\}) \wedge i < j-3\}.$$

Definition 17 (Accessibility Structure String Representation)**:** Let $w^{P_{access}}$ be an accessibility structure string representation of length n of an accessibility structure P_{access}, such that each position within the string is of the alphabet $\Sigma_D = \{.,x\}$, i.e. $w^{P_{access}} \in (\Sigma_D)^n$. Within the string $w^{P_{access}}$, the delimiting pair (k,l) for the definition of the accessibility structure P_{access} is indicated by a substring of length $l-k+1$ consisting of consecutive literals of type $x \in \Sigma_D$. All literals at the string positions in $[1,k[$ and $]l,n]$ of $w^{P_{access}}$ are of type $. \in \Sigma_D$.

Definition 18 (Accessibility)**:** Given an accessibility structure P_{access} and a base pair probability matrix $M_{\mathcal{P}_\mathcal{S}}$ then let the accessibility of that structure P_{access} within the base pair probability matrix $M_{\mathcal{P}_\mathcal{S}}$ constitute as

$$\text{Access}(P_{access}, M_{\mathcal{P}_\mathcal{S}}) = 1 - \text{Accur}(P_{access}, M_{\mathcal{P}_\mathcal{S}})$$

.

Definition 19 (Fuzzy Structure)**:** Let $P_{fuzzy} \subseteq [1,n]^2$ be a set of position pairs over an RNA sequence $\mathcal{S}$ of length n. P_{fuzzy} is called a fuzzy structure, if all position pairs interact only within delimiting position pairs $(k,l),(o,p) \in P_{fuzzy}$, such that $\forall (i,j) \in P_{fuzzy} : (k \leqslant i < j \leqslant l) \vee (o \leqslant i < j \leqslant p) \vee (k \leqslant i \leqslant l \wedge o \leqslant j \leqslant p)$. The base pairs (k,l) and (o,p) are part of the structure.

Definition 20 (Fuzzy Structure String Representation)**:** Let $w^{P_{fuzzy}}$ be a fuzzy structure string representation of a fuzzy structure P_{fuzzy}, such that each position within the string is of the alphabet $\Sigma_D = \{.,x\}$, i.e. $w^{P_{fuzzy}} \in (\Sigma_D)^n$. Within the string $w^{P_{fuzzy}}$, the P_{fuzzy} delimiting position pairs $(k,l),(o,p)$ are indicated by the substrings $w^{P_{fuzzy}}_{[k,l]}$

and $w_{[o,p]}^{P_{fuzzy}}$, which consist of literals of type $x \in \Sigma_D$ only. The literals of all other string positions are of type $. \in \Sigma_D$. The delimiting position pairs $(k,l),(o,p)$ enforce the structure delimiting positions of the resulting fuzzy structure.

Definition 21 (Fuzzy Structuredness)**:** Given a fuzzy structure P_{fuzzy} and a base pair probability matrix $M_{\mathcal{P}_\mathcal{S}}$ then let the fuzzy structuredness of a fuzzy structure P_{fuzzy} within the base pair probability matrix $M_{\mathcal{P}_\mathcal{S}}$ constitute as

$$\text{Fuzzy}(P_{fuzzy}, M_{\mathcal{P}_\mathcal{S}}) = \text{Accur}(P_{fuzzy}, M_{\mathcal{P}_\mathcal{S}}).$$

1.2.2 ncRNA – non-coding RNA

With a basic knowledge on where RNA is to be sorted and some insight on fundamental structural conditions, the class of non-coding (nc)RNA is presented with more emphasis in order to introduce the underlaying biological significance and consequently as well the functional potential of biotechnological and medicinal applications derived from ncRNA.

Junk DNA/RNA

The class of non-coding RNA and its acceptance in the scientific world is a relatively new development and 'renunciation' from the previously dogmatically believed doctrine of the impossibility of RNA functioning as something else than messenger RNA (mRNA) in the process of protein biosynthesis. With some 'necessary' exceptions of course, namely transfer RNA (tRNA) and ribosomal RNA (rRNA), which are key functionalities in the protein biosynthesis. To that extend, the peole back then did not get bothered by the fact that roughly $80-90\%$ of the DNA were somehow transcribed. It was a convenient to omit this fact and the ominous category of 'junk-DNA/RNA' within scientific discourse. Nowadays it is known with evidence, that this 'junk DNA/RNA' has actually a different purpose than just being discarded. 'junk DNA/RNA' shows to have a plenitude of functions within the cell environment, and most astonishingly it is mostly involved in processes, which are not directly connected with the protein biosynthesis, but in multiple other regulative processes connected to more different events.

ncRNA Diversity

A good overview over the vastness of differently categorized non-coding RNA families can be found in the database 'Rfam' (RNA families) (Nawrocki *et al.*, 2015). In its current version (Rfam 12.0) the Rfam database discerns 2450 different families of ncRNA. Although it categorizes subtypes of certain RNA families as own RNA families, and

integrates similar families into RNA clans, the diversity of RNA is still quite large in this database.
Simply the large amount of known and reported ncRNA families allows to assume and expect a functionally wide spread abundance among the different families. In fact, the functional range of RNA not only complements the known functionality of proteins but provides quite own functional entities and their accompanied pathways.

Prominent ncRNA Technologies

The trend reversal in the comprehension of the role of non-coding RNA as a regulative means within cells has been fortified. However, although, or maybe because of the native ncRNA situation being so diverse and nowadays also in a more accepted role, the acquaintance of knowledge and insight to the world of ncRNA resulted in the discoveries and developments of certain ncRNA techniques. Their core concepts have been abstracted from natively occuring processes, which involve non-coding RNA entities as major players. They have been advanced into very powerful molecular tools in the field of biotechnology. Most prominently appreciable are the technologies of RNA interference and the very recently revealed and implemented system of CRISPR/CAS9. Alongside, but prominent on another field within the ncRNA based technology, is the Aptamer class. They are promising players in the field of RNA based anti-bodies (Keefe *et al.*, 2010).

RNA interference The findings of RNA interference (RNAi) first had been published by Fire *et al.* (1998), describes a pathway, whose main feature is to selectively target specific genes by disrupting a respective messenger RNA (mRNA) with small but complementary strands of RNA, which are embedded in a specific protein machinery. The targeted RNA is disintegrated into smaller fragments, such that other nucleases can terminate the digestion of affected targeted RNA. This mechanism can be found both in plants as well as in other eucaryotes. The detailed mechanism is constituted differently in plants, such that the corresponding small fragments of RNA, which are necessary to knock-down targeted RNA is called small interfering RNA (siRNA) (Hamilton and Baulcombe, 1999) whereas in the other cases the corresponding class of RNA is called micro RNA (miRNA) (Elbashir *et al.*, 2001). With this means, the competent organisms can regulate their RNA levels, but are, as in the case of the plants, as well able to encounter intruding genetic material of viruses by complementary permanent disruption. The technology itself was adapted and is mainly used in basic research experimental setups. In those, the functionality of single genes can be tested by knocking them out selectively. Also the identification of single genes, who are responsible for a certain phenotype feature, can be facilitated by this technique. Furthermore, RNAi-based therapeutic studies and developments are, despite of some fails in clinical

stage (H, 2010), still under investigation (Tiemann and Rossi, 2009; Mari and Bardoni, 2014), and therefore are quite promising.

CRISPR The CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) system itself was described in 2007 (Barrangou *et al.*, 2007). It constitutes an adaptive defense system used against viruses and plasmid intrusions within bacteria and archea who are host to the CRISPR system. Each time a successful defense event has taken place, newly gained information of the intruder genetic material is included into the system, such that the host of the system then 'knows' its enemy already, making it more robust and durable in the future. The information about the intruder's genetic material is stored in short repeating stretches of RNA, which can in the case of a new intrusion be incorporated into a carrier protein(CAS). With the obtained repeats, the targeted genetic material can be cleaved into smaller fragments and subsequently be digested by nucleases. The capabilities of the CRISPR/CAS9 (Li *et al.*, 2016) system of selectively destroying foreign DNA/RNA and editing the genome was identified and turned into methods allowing to selectively alter and edit single genes within genomes. This technology was successfully applied to animals cell lines (Wang *et al.*, 2013; Hwang *et al.*, 2013; Jinek *et al.*, 2012), but also to industrial plants (Svitashev *et al.*, 2015; Li *et al.*, 2015) in order to produce 'design' food. In the process of CRISPR/CAS9 characterization bioinformtic approaches have been involved as well (Makarova *et al.*, 2015). The topic is currently strongly in the focus of scientific and ethical discussion, since this technology is seeming open to facilitate genome editing on members of the human race (Mali *et al.*, 2013; Jinek *et al.*, 2013; Wang *et al.*, 2014) as well. Thus, if wrongly applied or misused, it bears the potential of providing a dystopia world perspective formerly only known to science-fiction plots.

SELEX Although Aptamers, specific RNA structures which can sense the presence of ligands, also occur in nature on a native basis within organism, the method of SELEX (Systematic Evolution of Ligands by EXponential Enrichment) introduced the possibility of enriching stretches of RNA, that can bind towards a certain target. This target can be a small low-molecular ligand, such as a metabolite, but the target can also be of larger nature, up to and including whole cell surfaces. The process of SELEX was introduced in 1990 by Tuerk (Tuerk and Gold, 1990) and Ellington (Ellington and Szostak, 1990) in independent reports on the matter. Starting from a certain RNA library, SELEX enriches RNA entities, which demonstrate competitiveness in an iteratively applied three step procedure. Within one iteration, in a 'binding' step, the respectively participating RNA entities get the possibility to interact with an immobilized target. In a subsequent washing step, low affinity binders get washed away, such that only sufficiently high affine entities towards the presented target survive the iteration. In the third step, all entities, whose composition of nucleotides allowed them

to interact in a desired way get amplified by the application of the polymerase chain reaction. Due to the naturally occurring 'mistake rate' of introducing and prolonging a wrong nucleotide on some positions, the formerly strongly narrowed in sequence and conformational space of the surviving library gets expanded again by close sequence neighbors of the current high affine entities, such that in a next round of enrichment, even better binders potentially had been introduced. The methodology stipulates a gradual increase in the stringency of the dilution step, such that in the absolute beginning of the process, a quite broad collection of RNA is enriched, which will be thinned out in subsequent rounds, when the dilution and washing gets more stringent.

Aptamers & Riboswitches

Aptamers are a general category of RNA, which can bind to other specified targets, whose nature is highly diverse. They range from small molecules, such as metabolites and other low-molecular molecules, to larger molecules such as other RNA or even proteins. In nature, aptamers occur mostly in association with regulative mechanisms, in which the Aptamer's function is to sense the presence of a certain associated metabolite in order to cause a specific *cis*-reaction and/or *cis*-regulation of subordinated functional pathways (Winkler and Breaker, 2003). In this context they were termed riboswitches: A resemblance of their switching character on the basis of RNA.
In its application SELEX has given raise to multiple synthetic aptamers with different targets. They have been subject to further extension towards inclusion into regulative RNA entities.
As early as in 1999, Soukup *et. al.* (Soukup and Breaker, 1999) reported on the design principles of riboswitches, which back then had been based on adenosine-triphosphat (ATP) and flavin mononucleotide (FMN). Within their study, they do not only report the enrichment of suitable RNA sequences wihtin a SELEX application, which bind the mentioned molecules as substrate, but also report on their constructs, which in fact, had a regulative impact on the respective system, where they had been introduced to.

A theophilline aptamer was developed using libraries of high affine RNA pools allowing high-resolution molecular discrimination towards their target (Jenison *et al.*, 1994). Furthermore the theophylline aptamer was extended towards a riboswitch system, extending the aptamer towards a helix slippage mechanisms, sequestering a gene expression functional site (Suess *et al.*, 2004).

Also tetracycline (tc) (Chopra and Roberts, 2001), an antibioticum, was target to an application of SELEX, in which a suitable RNA sequence capable of interacting with tc was found, which was sufficiently characterized in order to be categorized as suitable

and functional tc-aptamer (Berens *et al.*, 2001). Subsequent studies and setups introduced the tc aptamer into different artificial systems. Suess *et. al.* (Suess *et al.*, 2003) extended the aptamer into a conditional translation control system, which is validated and further characterized (Hanson *et al.*, 2005; Müller *et al.*, 2006).

These depicted examples should serve as supportive arguments in order to highlight the impact of SELEX to the artificial design world of riboswitches and its development. Of course there are many more examples, in which single molecules and even single ions had been target to a SELEX enrichment in order to identify a suitable RNA sequence for respective interaction. Besides the aptamer/riboswitch part and their fusion to develop new design principles for genetic control within biotechnological, and also maybe in medicinal and therapeutic integration, other aptamers already have been released as medical players and one has been approved by the United States Food and Drug Administration (FDA) in 2004 (Lee *et al.*, 2005; Nimjee and Rusconi, 2005; Ulrich *et al.*, 2006).

Riboswitch architecture Riboswitches base their functionality mainly on the ability of the respectively involved aptamers. A riboswitch at least contains an aptamer part, which is responsible to interact highly specific with certain metabolites, respectively dependent on the particular riboswitches' aptamer domain. Riboswitches are natively located in the 5' regions of RNA which need to be under their control. In most cases those are transcriptional and translational events on which the presence of riboswitches are having an effect on. Based on their architecture, they can administer their regulative functionality to different situations. In most cases, riboswitches govern genetic regulation and its dependent functionality in bacterial 5' mRNA (Barric and Breaker, 2007), but as well in plants (Sudarsan *et al.*, 2003) and fungi (Kubodera *et al.*, 2003). The binding of the ligand to the aptamer is facilitated by an induced-fit procedure (Noeske *et al.*, 2007; Lutz *et al.*, 2014), in which the ligand and the aptamer engage their bound structures step by step.

Road Block - Functionality In some cases, the regulative functionality of a riboswitch is simply realized without any further required structures or sequence. Based on the binding event of the ligand to the aptamer, the structure of the aptamer itself gets stabilized enough, such that complexes which should actually read through this sequence situation, are now hindered to pass and are not able to break the reinforced aptamer structure open (Link and Breaker, 2009). When the ligand is not present to stabilize the respective structure, small minimalistic structures will form, which are so weak that sequence accessing machineries can simply 'open' them. In many cases, those minimalistic structure elements build some kind of leastwise structure, which can

interact with the suitable ligand and therefore can step by step accomplish the stable and durable aptamer ligand complex. This principle is called 'road blocking' event.

Switch - Functionality In some cases, the aptamer platform is coupled with an expression platform, which is some sequence and structure extension to the aptamer. This expression platform carries information, which is necessary to execute some controlled function, such as translation start signals, for example. When a corresponding ligand is bound to the aptamer part of the riboswitch, it induces a certain ligand bound structure conformation, which allows a downstream part of the expression platform to form and structure as a hairpin, such that the contained important information, e.g. some binding site for subordinated functionality, is integrated into the formed hairpin. The emerge of the inaccessibility of that specific stretch now impedes the actual function. The hairpin sequesters the function, and is therefore also called sequester hairpin. If, on the other hand the ligand is absent or its concentration if too low to initiate interaction with the aptamer, the bound structure of the aptamer is not situated, which allows a hybrid hairpin to form. This hairpin consists partially of the aptamer potion and partially of the expression platform, but the functional binding site is not affected of this formation, such that subsequently, due to the formation of the anti-sequester hairpin, the sequestering function of the sequester hairpin is not governed to the binding site, so that the binding site is found to be accessible in this conformation. The principle is called hairpin shift or slip system. Such a system is very good described by Rentmeister *et al.* (2007) within the thiamine pyrophosphate riboswitch.

Riboswitch Structure Folding Pathway Regulation To this extend, the mechanisms how an riboswitch actually works are still under discussions and new findings are made frequently. The 'fuse' model (Wickiser *et al.*, 2005), how ever describes a plausible mechanisms, on how the atpamers, resp. the riboswitches come into their function. It contains the assumption that the conformational decision is made during a transcription event. As soon as the aptamer ligand sensing portion exits the respective machinery, it needs to fold quickly into structures which allow for sensing for its specific ligand (Pan and Sosnick, 1997). Otherwise, wrong structures get trapped kinetically, such that a ligand could not be recognized anymore (Treiber and Williamson, 1999). Dependent on the concentration of the ligand, the aptamer folds into a certain conformation and influences downstream, but not yet transcribed information processing, allowing or disallowing respective signals to be read or processed. The basis on this decision is based on kinetic behavior rather than on a thermodynamic equilibrium properties of the conformation. In this way, a situation dependent decision is made in each copy of the construct being processed. As soon as the concentration changes its concentration towards a interaction concentration, the structural decision is made accordingly and allows a fast reaction towards altering ligand concentrations. This model is supported

by the fact, that in some cases, the respective transcription event attenuates just after the aptamer potion was produced (Perdrizet *et al.*, 2012). This grants some extra time to actually sense the ligand information and then undergo certain structure dependent decision, when the elongation process of the controlled entity continues accordingly.

Conformational Bistable RNA Entities

On the way to artificial riboswitches, supportive RNA molecules which have the feature of having two (and more) major prominent conformations, but lack a clear inducer ligand, that forces them into a specific conformation, have to be mentioned as well (Fürtig *et al.*, 2001). Within the category of bistable RNA, actually the riboswitches constitute one large subgroup, but they are ligand induced bistable molecules. Besides the induced switching molecules, also intrinsically bistable RNA molecules exists, which consist of at least two concurrent structures to each other, such that either structure can get folded likewise.

Exemplary, the hammerhead (Hammann *et al.*, 2012) and the hairpin ribozyme (Müller *et al.*, 2012), which are involved in cleaving and ligating of circular RNA genomes (Flores *et al.*, 2011), are two RNA entities, whose major conformation have the ability to either cleave some RNA strand or fuse some RNA strand together within the underlaying RNA genome machinery. They are layout in a way, in which only one specific conformation, which is in concurrence to other conformation, is the actively cleaving/ligating functional conformation. In the other cases the functionality is set silent.

In the case of the Qβ-replicase, RNA switches are involved, namely SV11 (Biebricher and Luce, 1992), that , as soon, as they are transcribed can fold into their functional substructure, but get governed by a larger and more stable alternative hairpin, that disrupts that initial functionality.

Due to the constant unfolding or refolding, both structures can be found likewise, even though that one structure has slightly increased frequency of being folded. The folding pathway(s) and their intermediate structures depend largely on refolding rate-limiting base stacks (Xu and Chen, 2012). As seen in the riboswitch 'fuse' model, bistable molecules can, at a certain point of time in their folding ensemble be in their thermodynamic equilibrium. Induced refolding by disturbances to that equilibrium allow to get insight on the folding behavior on the structural level (Wenter *et al.*, 2005; Cao *et al.*, 2010). Based energy landscape explorations, the kinetic behavior of a single entity can be computationally simulated as well to get insight on which structure(s) dominate the course of folding of a certain sequence (Mann and Klemm, 2011).

1.2.3 RNA Experimental Structural Probing

In order to be able to derive algorithms, which are able to predict structure of a certain RNA molecule, the underlaying principles need to be measured and explored, such that specific derivations of the folding principles can be made.
The experiments of Turner and Mathews (2002), which are based on the energetic elucidation of single fragments of RNA, have been executed under very precise controlled environmental influence and provide a basis on how an important part of the structure of RNA can be accessed algorithmically. Most of the present RNA secondary structure prediction algorithms are based on the experimentally explored energy parameters. Also knowing that those conditions are almost never met in real cells, methods of single RNA structure determination have been advanced to methods, that can bring elucidation to the whole RNA structurome of a cell. If well evaluated and combined and finally merged into a new definition of an algorithm, which can therein predict structure given a certain cellular context, new structure prediction tools could be generated, based on NGS data.
In order to determine a functional structure of RNA, different methods have been developed over time. Besides nuclear magnetic resonance (NMR) and X-ray crystallography protocols, which resolve 3D structural information from crystallized molecules and from solutions, chemical and enzymatic probing are methods, that deliver information about the structural conformation of an RNA entity. This is done by incubating the RNA with different chemicals and/or enzymes. A specific interaction of a reagent with its specific position within a certain context within the RNA allows to indicate certain nucleotide interactions by different combinations of reagents or RNase as cleaving enzyme. With this approach, different layers of structural information are extracted from RNA molecules in independent reactions.
In chemical probing, four different but specific reaction types are included. On each tier, different combinations of reagents have to be used in order to determine the true character of the involved interactions. Base-specific reagents make it possible to gain insight about base stacking, hydrogen bonding and electrostatic environment directly adjacent to probed bases (Ehresmann *et al.*, 1987; Lavery and Pullmann, 1984). With this approach, all nucleotides of a sequence can be addressed simultaneously in one experiment. Structural parts, which are solvent accessible in the conformation, can be probed by hydroxyl radicals. They induce cleavage of the RNA backbone by proton abstraction in ribose C4' or C5' atoms (Tullius and Greenbaum, 2005). The protocols of in-line probing (Regulski and Breaker, 2008) and 'selective 2'-hydroxyl acylation analyzed by primer extension' (SHAPE)-probing (Merino *et al.*, 2005) introduce the possibility to investigate local nucleotide flexibility or dynamics. The latter is in concordance with NMR based relaxation measurements and therefore indicate the potential as an quantitative measurement of nucleotide flexibility (Gherghe *et al.*, 2008). Furthermore this method is nucleotide independent (Wilkinson *et al.*, 2009). Tertiary

structure interaction can be probed with bifunctional reagents, that interact only with nucleotides, which are in vicinity through structural scaffolding. In addition, UV cross-linking experiments can be applied as substitute. RNAse activity manifests in the cleavage of solution exposed single stranded RNA. This ability is used for enzymatic probing (Daou-Chabo and Condon, 2009). The incubation experiments are strictly simple in comparison to the attached analysis pipeline of the experiments, where all produced fragments and interactions have to be quantified and set into correlation. The information of the fragments is extracted either from a gel or capillary electrophoresis. After all, the fragment quantification analysis (Mitra *et al.*, 2009) and the subsequent already presented structure prediction are done with computational aid to integrate the gained structure model. Advanced hydroxyl footprinting probing methods are capable of resolving the time-dependent character of RNA structure folding. With that means, structure folding observations in a time scale from sub-seconds to several minutes are possible (Shcherbakova and Brenowith, 2008).

By coupling 'classical' chemical probing protocols with next-generation sequencing techniques, SHAPE-seq (Lucks *et al.*, 2011), multiplexed accessibility probing sequencing (MAP-seq) (Seetin *et al.*, 2014), hydroxyl radical footprinting sequencing (HRF-seq) (Kielpinski and Vinther, 2014) and chemical interference of RNA structures sequencing (CIRS-seq) (Incarnato *et al.*, 2014), which perform transcriptome wide RNA structural probing, the focus shifts from single molecular entities to whole RNA libraries, such that RNA folding tendencies within certain molecular context can be elucidated.

After a classical incubation of the probes with the respective reagent, the introduced interactions between the chemical(s) and respective RNA are detected using reverse transcriptase produced libraries of cDNA where the 'mutations' are encoded by rather shorter fragments, since the introduced chemical adducts on the RNA induce a stop in the reverse reading process. In addition, the methods of RNA interaction groups by mutational profiling (RING-MaP) (Homan *et al.*, 2014) and of SHAPE-MaP (Siegfried *et al.*, 2014) provide subsequent association studies leading to detailed maps of nucleotide tertiary interaction and supplemental structure stabilizing interactions. Furthermore the detection of distinct conformations within the solution ensemble is feasible. Large-scale reconfigurations of the structure and intermediate 'hidden' structures can be discovered using this methods.

Classically the experiments for chemical probing of RNA structure have been performed in *in vitro* experiments under non-physiological conditions. With this methods and approaches the foundation for the chemical probing itself has been laid out. Although the experiments *per se* have been correct, they did not reflect the true nature of the situation of the investigated RNA molecules. Only within *in vivo* experiments under 'correct' circumstances of the cell and its provided environment of all influencing particles and molecules, the correct and therefore functional nature of RNA molecules can

be arranged. In recent review (Kwok *et al.*, 2015) transcriptome-wide probing of *in vivo* RNA structure is summarized. A good comparison between *in vitro* and *in vivo* approaches and an overview on how to combine them in order to obtain insight on how cell internal parameters influence on the RNA sequence is given in (Tyrrell *et al.*, 2013). However, the part of RNA structural probing is only mentioned and described in such detail in this thesis for prospect reasons, since it provides a powerful method in order to understand the principles which govern the molecules in their behavior. The mechanisms of those tools are not specifically used within this work, nevertheless, they provide, if developed further into the direction of whole cell RNA structurome analysis, a promising way of consolidation of structure elucidation methods, out of which new prediction algorithms can be derived. Eventually they can be included in methods such as *antaRNA* in order to produce suitable RNA, which function under specific circumstances and therefore are highly specialized RNA designs. For the moment being, the general derivations of Turner and Mathews (2002) in their manifold presence have to be a good base camp for such endeavors and therefore are used as well in this thesis.

1.2.4 RNA Folding

Knowing, that the importance of a biologically functional RNA is exclusively dependent on its correct structure conformation, it is of great interest to derive the structural functional knowledge of an RNA molecule from it. To predict the functional structure from the sequence have been implemented in computational approaches. However, the underlaying principles to those algorithms has been observed and derived from very detailed experiments. The whole process, in which an RNA forms its structure is called folding. Given a certain length and composition of nucleotides, the intrinsic potential of the RNA to fold into a plenitude of structures is enormous. In most cases, only one structure conformation is able to function biologically, which makes it a challenging task to forecast and predict this functional conformation based only on the RNA sequence information. The process in which an RNA sequence adopts its structural conformations is called 'folding' and is on the one hand driven by the RNA intrinsic physico-chemical forces of the atoms from which the RNA is made of. On the other hand, this process is driven by the biochemical environment, the RNA molecule is situated in. Under specific medium and environment influence, the configuration of an RNA molecule therefore contributes to a large extend to its folding. Employing *in vivo* studies and experiments provides insight into principal folding mechanisms of RNA and its dependencies. However, a situation as it is *in vivo* can not be reproduced and represented by folding algorithms. The knowledge of *in vitro* experiments only approximates the *in vivo* situation due to the experimental setup, which does not cover all contributing factors entirely.

Within cells, however, the process of folding an RNA is very complex, since it is highly orchestered and has to be very precise in order to be functional on that level. This is necessary to prevent misfolding and its implicated malfunctioning of RNA entities. To provide its natural functionality, each RNA molecule needs to be in a specific conformation, that is required in order to perform that functionality. RNA structure is achieved through a folding pathway, i.e. that the whole molecule is not folded into its structure within one big folding step, but is build up in hierarchically organized steps into its functional conformation. Co-transcriptional folding describes the process, by which nascent RNA, as soon as it leaves the RNA polymerase, can start to fold and build initial secondary structure elements it is capable of folding into. As the transcription continues, more and more structural elements emerge. Co-transcriptional folding is influenced by several factors (Zemora and Waldsich, 2010; Neugebauer, 2002; Lai D, 2013). The velocity of transcription itself determines and dictates the time an RNA molecule is granted to undergo certain folding steps in its pathway and influence the 'correctness' of a current fold. Furthermore, in some genes, pause sites have been identified, which induce the RNApolmerases to halt their action for a certain amount of time. During those translation pauses, the nascent RNA has more time to under go certain folding steps which define the later conformation of the RNA or the general functionality at all (Landick, 1997; Mooney *et al.*, 1998). On its way from being only a 'linear' sequence to become a fully functional entity, the RNA either undergoes a folding path or experiences refolding into an energetically favorable biologically functional molecule. Through wrong folding, conformations can be engaged, which are also energetically favorable, but are not functional. In order to prevent from misfolding, factors, such as small molecules (Roth and Breaker, 2009; Serganov, 2009), proteins or regulative or guiding RNA, which can interact with and even modify the nascent RNA during its folding (Morris and Mattick, 2014; Bachellerie *et al.*, 2002). Due to this influence, certain potential structure elements can not fold anymore due to sequestration of potential structure, but other alternative conformations become more likely to be folded into. A misfolded and therefore energetically trapped RNA molecule can either be recycled, or can be partially unfolded through the influence of certain proteins in a process called RNA chaperone-guided refolding (Jackson *et al.*, 2006). Temperature dependent bi-stability of RNA directly shifts the success in undergoing a correct folding path in dependence of a 'correct' temperature (Cimdins *et al.*, 2014; Narberhaus, 2010). With such means of control, it is feasible to achieve specific structures and therefore specific functionality, e.g. in order to control transcription or translation (Nahvi *et al.*, 2002).

Gibbs Free Energy

The underlaying physico-chemical forces within cells and cell-free experiments induce the sequence to fold into energetically favored structures. Intrinsically an energy minimization is performed when an RNA molecule folds into its structure. A gaugeable entity in this belonging is the Gibbs free energy. The energy difference between the unfolded structure and a certain conformation allows a first indication on how stable this structure is.

The thermodynamical entity of 'free' energy was firstly introduced and termed by J. W. Gibbs (Gibbs, 1873). Formally the Gibbs' free energy is defined as $G = H - TS$, where H is the enthalpy, T the temperature and S the entropy. The enthalpy $H = U + pV$ describes the internal energy U of a system under a certain pressure p and volume V. The entropy S describes the level of order within a certain system. The entities temperature pressure and volume describe the macrostatic character of a system, whereas the entropy is used to describe the remaining microstatic character. The higher the entropy, the more disorder/mixedupness is within the system.

The formation of the RNA's hydrogen bonds in its structural conformation sets energy free, in such a way that a folded structure's energy is lower/more stable than the completely unfolded RNA molecule. In the following, only the energy difference ΔG to the unfolded molecule is measured and discussed. A negative ΔG indicates hydrogen bonding within a folded structure. It follows, that there is a need for external energy to unfold a structure of RNA into its unfolded structure conformation. The respectively required amount of energy is dependent of the character of the involved hydrogen bonding situation. During the process of folding, the emerging RNA structure excludes water molecules from the hydrophilic interaction of hydrogen bond donors and acceptors from within the RNA. This as well introduces a higher degree of order into the system, such that the entropy is lowered.

Folding Algorithms

RNA Folding algorithms are constructed, such that their respective methods are focused to find a biologically functional structure solution for a given RNA sequence. However, it is very difficult to predict the biologically functional structure of a sequence just out of the knowledge of the sequence itself. Alternatively most algorithms predict the MFE structure of a sequence, even though it might not be the functional structure. In this study, secondary structures of respective RNAs are target of investigation. In order to find such a solution structure of a sequence, different optimization objectives have been employed. Most of them are calculated by applying the principle of dynamic programming. With this means, it is possible to only compute a suitably sized subset of all solutions within the respective objective in order to find an optimal result to the

problem. Hereby, two kinds of problem classes within the secondary structure prediction discipline emerged: Mainly programs focus on the more simple solution of the nested structure problem. The non-nested or pseudoknot structure problem is tackled by less attempts.

In the following, an overview on the historic development on the field of RNA secondary structure prediction algorithms is given.

The Nussinov algorithm (Nussinov *et al.*, 1978) tries to maximize the amount of base pair interactions within a certain nucleotide interaction range and given specific sequence complementarities. Nussinov's algorithm performs in $\mathcal{O}(n^2)/\mathcal{O}(n^3)$ memory/-time complexity.

The Zuker algorithm (Zuker and Stiegler, 1981) firstly introduced the usage of biophysical properties and its objective is to minimize ΔG, such that energetically very stable MFE structures are predicted. Instead of evaluating a structure according to its base pairs, it assumes secondary structural elements of bases within the nearest-neighbor energy model and affiliates corresponding energetic values of Turner and Mathews (2002) to respective structures. Zuker's algorithm performs in $\mathcal{O}(n^2)/\mathcal{O}(n^4)$ memory/-time complexity. With a subtle setting of internal calculation depth of certain structure elements, the time complexity can be reduced to $\mathcal{O}(n^3)$. The concept of an extended nearest-neighbor energy model of Dotu *et al.* (2014) develops the idea of the 2-stack energies of Turner *et. al.* towards a triplet model based on initial experimental values of Gray *et al.* (2010).

The McCaskill algorithm (McCaskill, 1990) firstly provided the possibility to calculate the partition function and the corresponding probabilities of structures and single base pairs given a certain sequence. The memory/time complexity is the same as for the Zuker algorithm. Established implementations of the Zuker and McCaskill algorithms are *Mfold* (Zuker and Stiegler, 1981) and *RNAfold* (Hofacker *et al.*, 1994; Lorenz *et al.*, 2011).

The implementations of *pKiss* (Janssen and Giegerich, 2014) allow the computation and prediction of pseudoknot structures. *pKiss* is as well based on thermodynamics and uses extended Zuker-style recursions, but can only provide for canonical pseudoknot structures, which are effectively the two simplest types of pseudoknot structures (hairpin loops and kissing hairpins). A $\mathcal{O}(n^2)/\mathcal{O}(n^4)$ memory/time complexity is required. Even though it has $\mathcal{O}(n^2)$ space complexity, which is less than other algorithms being able to calculate hairpin pseudoknots, like Akutsu (2000) in $\mathcal{O}(n^3)$, *pKiss* still requires $\mathcal{O}(n^4)$ in time. This is no improvement over others. On the other hand, it still has the same complexity when it comes to kissing hairpin calculation. In comparison, the formulation of Chen *et al.* (2009) shows a $\mathcal{O}(n^3)/\mathcal{O}(n^5)$ memory/time complexity. Exemplary, more complex pseudoknot structures, e.g. the Threeknot, can be computed with the algorithm of Rivas and Eddy (2000) in $\mathcal{O}(n^4)/\mathcal{O}(n^6)$ memory/time complex-

ity. All nested structure predictions in this work are performed with *RNAfold* of the *ViennaRNA*-package v2.1.3 (Lorenz *et al.*, 2011). The pseudoknot predictions in this work have been done with *pKiss* due to its space efficiency and its output compatibility within the established pipeline.
HotKnots (Ren *et al.*, 2005) is a heuristic approach for the solution of the pseudoknot folding problem. It uses multiple parallel structures, which are extended by small building blocks, until a suitable result is derived. However, *HotKnots* is comparatively slow in comparison to *IPknot* (Sato *et al.*, 2011) and according to the same studies, it is not competitive to other predictive programs, in terms of runtime. *IPknot* uses integer programming to tackle pseudoknot decomposition into pseudoknot free substructures, in order to maximize expected accuracies within pseudoknot considering base-pairing probability distributions. Even though it is fast in its unrefined modus, it takes longer than competitive programs to return solution proposals. Alas, no explicit complexity account is given within the publications.

1.2.5 RNA Kinetics

The process of structure folding of an RNA molecule is governed by the constant formation and/or dissolving of base pairs, such that a structure of an RNA molecule is not a static but a highly flexible entity (Shen, 2008). Folding kinetics describe the dynamic change of structures of an RNA over time. In the following, a brief description of the underlaying principles is given.
Folding kinetics can be modeled based on the concept of RNA energy landscapes, which encode all possible structures of an RNA, their structural energies and their transitions into each other. Based on this, one can compute the probability of each structure at any time point for a given start probability distribution using a Markov process model as done by the tool *Treekin* (Wolfinger *et al.*, 2004). An example is given in Figure 1.5, which describes the kinetics of a bistable RNA molecule, that even acts as a switch. A more recent approach, namely *RNAHeliCes* (Huang *et al.*, 2012; Huang and Voss, 2014), circumvents the energy landscape enumerating complexity by abstracting and estimating plausible folding pathways. Since RNA kinetics are only investigated briefly within this work, no detailed discussion is provided in the following. The tool *Treekin* is later used for kinetics studies, thus the rest of the section focuses on the basis of this tool.
The topology considering all structures of the energy landscape of a sequence under the rule of a certain neighborhood transition relation and a given energy function resembles all microstates of the system and is quite complex in order to perform adequate computations on it. To reduce the complexity, two steps can be performed.
The first step is to use an energy cut-off on the energy landscape and to only consider structures and their transitions that range energetically below a certain energetic

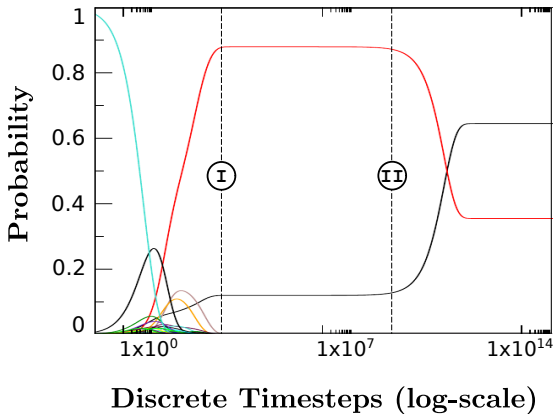

Figure 1.5: Example kinetics plot of a bistable switch of (Mann and Klemm, 2011) The initial open-chain structure entity (turquoise) depletes and transits into several different structure entities in the beginning of the consideration. From that early pool of intermediate structures, two prevailing structures emerge (black and red) and enrich towards a stable situation in terms of the respective structure probability (time point I). At time point II, the quasi-equilibrium changes dramatically, such that the previously underrepresented structure (black) gets promoted over the previously dominant structure (red). The two entities, after they crossed their probability levels, converge into a equilibrium situation and finally stabilize according to their structure probabilities (see Definition 10).

threshold (Wolfinger *et al.*, 2006). In this case, only the connected component is used, which contains the initial open-chain structure.

The second step clusters the energy landscape into connected macrostates and derives a new coarse grained energy landscape with according macroscopic transitions based on the underlaying microscopic energy landscape. One possible clustering is based on gradient basins, as summarized in the review of Flamm and Hofacker (2008): In the employed model of RNA kinetics, all structures that transit into the same local minimum due to a gradient walk, that always selects the neighbor of a structure with the lowest energy, are summarized as one basin, or macrostate, of the local minimum. All microscopic transitions from one macrostate into another are fused into an according macroscopic transition.

The transition information combined with the energy value for each state is used to compute transition rates according to the Arrhenius equation. The rate matrix in combination with an initial probability distribution can be used to constitute a time continuous Markov Process. The computed RNA kinetics provide the probability Pr_i^t for a given time point t with which an folding RNA molecule resides within a certain macrostate i (see Figure 1.5).

1.2.6 RNA Inverse Folding

The question of RNA inverse folding describes the algorithmic problem of finding a sequence, whose main feature is to fold into a specific structure conformation, given a folding prediction algorithm according to which the folding hypothesis is built upon. There is no immediate counterpart in nature but the evolution of molecules itself. RNA is quite robust in its structure folding, such that several thousand sequences of a certain length can fold into the same structure (Grüner *et al.*, 1996; Grüner *et al.*, 1996). RNA is therefore quite robust against single pint mutations. In order to solve the issue of obtaining a sequence, which can fold into a user defined structure, several approaches have been presented in the past: The now classical tool *RNAinverse* (Hofacker *et al.*, 1994) opened the field of algorithmical RNA inverse folding. The program generates seed sequences, which are further optimized by applying local search within a second processing step. With *RNA-SSD* (Andronescu *et al.*, 2004) the optimizing step was improved by applying stochastic local search. In *InFoRNA* (Busch and Backofen, 2006, 2007) the seeding procedure was improved by generating a sequence that is maximally stable for the target structure and thus has high probability to fold into that structure. *Inv* (Gao *et al.*, 2010) uses the principles of loop decomposition. They as well follow the principle of generating a seed sequence with subsequent local optimization, such that the sequence is modified towards requested structure. With this approach inverse folding of pseudoknotted structures is feasible.
In the tool *NUPACK* (Zadeh *et al.*, 2011), a dynamic programming approach is applied to hierarchically decompose the given structure for compiling solution sequences. The sequences are evaluated according to the introduced and characteristic ensemble defect structural distance measure. *MODENA* (Taneda, 2011) introduces a genetic algorithm to solve the problem. *MODENA* also provides a platform on which also pseudoknot structures can be solved. Within the evaluative step, *IPknot* (Sato *et al.*, 2011) or *HotKnots* (Ren *et al.*, 2005) are used to fold a current solution sequence. As objective target, *MODENA* aims to maximize the closeness of a solution to the target structure and simultaneously minimize the solution's free energy. In *fRNAkenstein* (Lyngso *et al.*, 2012), as in contrast to *MODENA*, a genetic algorithm to solve a multiple structure constraint inverse folding problem is presented. The evolutionary steps include 'point mutation' of undesired positions and promotion of regions by 'recombination and propagation' to the next generation of sequences. The fitness of a solution is evaluated on an average level according to the multi structure constraint.
Coming from the 'seed and optimize' strategy, in *RNA-ensign* (Levin *et al.*, 2012) the local optimization step is extended towards a k-mutants global sampling of an energetic ensemble model of the RNA mutation landscape. Iteratively the neighborhood of a seed sequence is increased by incorporating all neighbor sequences having k positional nucleotide mutations until a correct solution is found. Also *RNAexinv* (Avihoo *et al.*,

2011) employs this strategy; in addition *RNAinverse* is employed within this strategy in order to generate the used seed sequence and focus on the improvement of the optimization step. For this purpose a simulated annealing strategy is applied, which iteratively mutates and evaluates current sequences. The evaluative objective caters the targeted structure. Newly the thermodynamic stability and mutational robustness of the sequence are considered as well. In addition *RNAfbinv* is an extension of *RNAexinv* and rewards and highlights the presence of user defined substructural motifs within the objective function. *IncaRNAtion* (Reinharz *et al.*, 2013) also uses *RNAinverse* as provider for the seed sequence and performs a weighted sampling procedure in order to do an optimizing step towards the user specified constraints, among which also a GC content can be entered. *RNAiFold* (Garcia-Martin *et al.*, 2013) approaches the problem by providing a constraint programming solution, which employs large neighborhood searches. With this approach, additional side constraints can be introduced. A very important development within *RNAiFold* is the fact, that ranges of target GC content can be specified to restrict a computation of a solution sequence. The quite recent approach *ERD* (Esmaili-Taheri *et al.*, 2014) employs an evolutionary algorithm which is based on a structural decomposition and genetic operations on whole substructures instead on single nucleotide substitution.

In a nutshell, early inverse fold tools pursued the 'seed and optimize' technique, whereas newer tools employ genetic/evolutionary algorithms in oder to heuristically approach a solution which satisfies constraints. It appears as if the use of heuristic approaches allows growing itemization of the underlaying constraint sets. Besides the structural constraint, new tools, in order to provide an increased flexibility towards the constraints should have the full IUPAC nucleotide sequence constraint implemented and also should have competitive capabilities on the field of GC-content constraint fulfillment. The latter point is very crucial, since specific RNA families and especially distinct organisms have particular GC values in order to function in an evolutionary well-rehearsed interplay. Given the fact that classical RNA inverse fold programs have intrinsic specific GC values (as exemplary shown in Figure 1.6), and knowing that the GC content of different RNA entities is spreading widely among various organisms, providing a precise GC content value within their solution sequences seems a very important feature to new RNA inverse folding tools.

1.3 Ant Colony Optimization

Meta-heuristic Ant Colony Optimization (ACO) is a nature inspired algorithm, which belongs to the class of swarm algorithms. Swarm algorithms are derived from biological models and patterns in behavioral observations of animals and mostly aim at the solution of complex problems, to which an explicit solution calculation is expensive due to

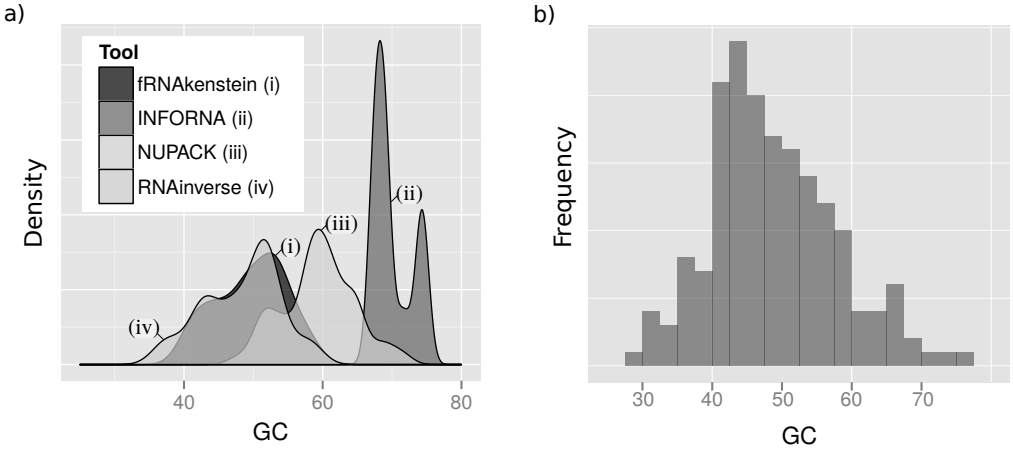

Figure 1.6: Intrinsic GC values of an exemplary selection of programs vs. Rfam seed sequence actual GC values of various RNA entities.
The direct comparison between the intrinsically achieved GC values of available RNA inverse folding programs with the actual GC ranges within Rfam RNA entities allows the assumption, that more precise and adjustable GC control in RNA inverse folding is useful in order to design GC value compatible RNA. As indicated in **a)**, the energy minimizing approach of *InFoRNA* is clearly biased towards a very high GC content value by the incorporation of energetically more stable G-C base pairs. The other approaches are more sane in this respect, but seem limited in the question of obtaining the full potential of the possible GC range. The exemplary sampling was performed on the 3-multiloop tRNA structure of length 74 nt. **b)** Wide spread GC values within the GC content histogram derived from each sequence of each Rfam family seed alignment. The bold majority of instances populates between 40 and 60 % GC, with more emphasis towards the 40% scope of the distribution. Nevertheless, also the even more extreme GC values below 40% and above 60% are quite frequently represented. Drafted using ggplot2/R and VRNA software.

the complexity of the problem itself. The principle behind the class of swarm algorithms is, that 'agents' construct and evaluate instances of the respective complex problem and transmit this information to the 'swarm', such that the 'mind' of the underlaying 'swarm of agents' can be influenced by prior solutions and their outcomes. Using this constant reinforcement of information within the decision making process, the selection of an increasing number of suitable parts converges towards an approximated solution of the problem. Hitherto, different approaches have been presented, exemplary other methods in the class of swarms algorithms are the 'particle swarm optimization' (Kennedy and Eberhart, 1995) and the 'bees algorithm' (Pham *et al.*, 2005).

The underlaying algorithm of this thesis is derived from the 'ant colony optimization' approach (Dorigo *et al.*, 1999), which is an advancement of the 'ant system' (Dorigo *et al.*, 1996). In the ACO approach, the food foraging behavior of ants is mimicked. In an initial differentiated study, Deneubourg *et al.* (1983) described different patterns of ants exploring their environment searching for food. Deneubourg *et al.* (1990) as

well describe and elucidate the trail laying and trail following behavior of ant in simple experiments, which is a key principle within ACO.

Ant-colony optimization has been applied to a plethora of optimization tasks showing its versatile and flexible character as robust problem solver for complex questions under static and dynamic settings. Whereas the initial tests of ACO have been made on the traveling salesman problem, ACO was shown to perform very good is several areas, as summarized in Appendix B. Preeminent applications, such as Bayesian Networks (de Campos *et al.*, 2002), partitioning/clustering (Blum and Blesa, 2005) and classification/data mining (Martens *et al.*, 2011, 2007; Parpinelli *et al.*, 2002a,b) in addition to promising results of ACO adaptations concerning molecular biological questions such as protein folding (Shmygelska and Hoos, 2005; Hu *et al.*, 2008; Nardelli *et al.*, 2013), ligand docking (Korb *et al.*, 2006) and RNA folding (McMillan, 2006), motivated and encouraged to apply and adapt ACO to the RNA inverse folding problem.

Its basis is the ants' behavior to evaluate the quality of a 'food source' and feed that information to the other ants by applying pheromone trails on the underlaying terrain. Here, the quality is a not further specified entity, but can be interpreted as a summary of questions such as, 'How good is the food?', 'How long is the path leading to the food?', 'How difficult/dangerous is the path?'. The pheromone influences the walking and decision making behavior of other ants of the ant colony, such that more ants choose to walk a 'good' segment of the terrain, or avoid walking on a 'bad' one (Goss *et al.*, 1989). The pheromone itself is evaporated with a certain rate due to the environmental influence to the terrain. As illustrated in Figure 1.7, over time, due to the continuous marking of the terrain with pheromone and it's partial evaporational removal, segments and paths, which are not good, disappear due to no refreshing by the ants. Good segments, on the other hand are endorsed by this principle (Deneubourg *et al.*, 1990). In the end, an optimal path through the terrain is established.

1.4 Synopsis of the Introduction and Incentive of the Thesis

Given the fact that an increasing number of RNA structures and their functions are revealed in contemporary research attempts in addition to more and powerful, maybe even fate critical biological methods and computational tools, which have been developed and will be investigated and applied within artificial RNA designs, the impact on research and society in the future will increase as well. Researching personnel is introducing artificial constructs into a increasingly bio-x driven engineered world, therefore it might be of use to provide computational tools, with which the design process could be eased and facilitated on a broad scale.

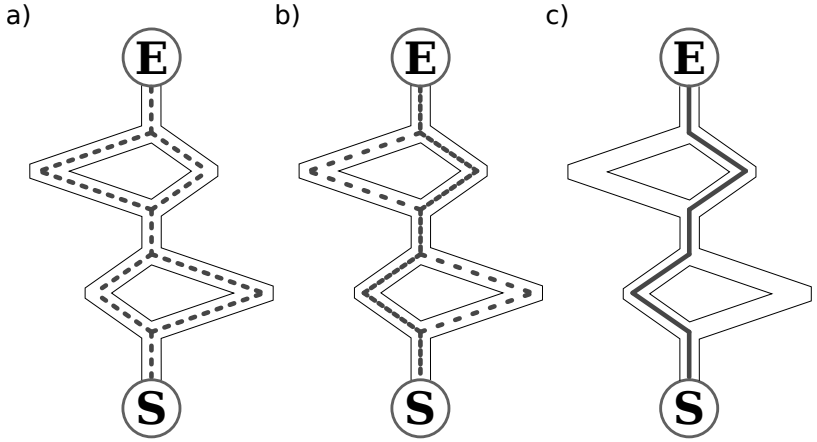

Figure 1.7: Exemplary pheromone marking convergence towards 'optimal' solution.
Ants make a walk from point **S** to point **E**, evaluate the solution, e.g. here the path length, and mark the corresponding path with a quality dependent amount of artificial pheromone. In regular episodes, pheromone evaporates from the paths with a certain rate, such that over some ants' walks (a → c), a prominent solution emerges from all possible selectable paths, which represents an optimal solution to a problem. Setup inspired by the Deneubourg *et al.* (1990) experiments.

By the introductive depiction of RNA and its context, its chemical, biological and also algorithmic definitions, a general survey on the matter is given. This allows to constitute the aim of the thesis.
Knowing the biological and technological impact of RNA based technology, this work presents an attempt, in which the RNA inverse folding problem is approached by the application of an ant-colony optimization heuristic algorithm. Within the algorithm itself, RNA folding algorithms serve as basis for the evaluation of sequences, which are produced by the internal data structures and procedures. In its progress the program was encountering several developmental stages. Each stage was meant to solve increasingly difficult folding hypothesis and produce according sequences as solutions. Therefore increasingly complex generative and evaluative functions have been integrated into the pipeline of *antaRNA*, the resulting algorithm.
Show casing the basic functionality with regular folding hypothesis provided by *RNAfold* or *pKiss*, the ant colony optimization is adopted to provide sequences which can last but not least encounter the RNA inverse folding problem of bistable sequences with user defined constraints on structural, sequence and also newly on GC target value constraints. The thesis is based on two *antaRNA* publications (see Appendix). The yet unpublished extension of *antaRNA* towards more complex bistable RNA behavior is presented as well.

CHAPTER 2

THE *ANTARNA* ALGORITHM

In this section of the thesis the adaptation of the ant-colony optimization (ACO) principles to the RNA inverse folding problem is described. The used constraints to the folding problem are depicted, such that their incorporation into *antaRNA* is explained. The underlaying model of the ACO adaptation to the RNA inverse fold is problem is portrayed and a brief description of the modified and applied functions is build up. The resulting algorithm is presented to the extend, that its general mechanics are presented. The program *antaRNA*, which is the acronym for 'ants assembled RNA', is the resulting implementation. Its functionality will be described within this section.

The adaptation of ACO to the RNA inverse folding problem is aimed to produce sequences, which fold into a predefined structure, given an RNA folding algorithm, which is also used within the *antaRNA* framework during the optimization. Besides the regular structural constraint, additional sequence and GC constraint where implemented and provided to extended the primary structure constraint in a way, such that enhanced definitions of sequence designs can be computed.

This chapter covers the technical aspects of *antaRNA*. The resulting description of *antaRNA* should not be seen as a single monolithic product. It rather should be seen in the specific perspective that *antaRNA* was successively improved and made ready for more complex questions and models in each phase of development of the program. The division into the different levels of structure complexity is especially important and discussed in Chapters 3 and 4, where the different stages of *antaRNA* are presented. In this chapter, the respective adaptations of the code and concept sections to different categories of structure complexity and its resulting extensions of code and concepts used in *antaRNA* are accented.

2.1 Survey of *antaRNA*

The presented implementation of *antaRNA* follows the procedure which is depicted and summarized in Algorithm 1 and is an adaptation to the ant colony optimization principles described by Dorigo *et al.* (1999). In this section, an initial survey of the major principles is given without making any definitions nor more explicit descriptions of the employed mechanism. This first survey is meant to illustrate the overall flow of the presented RNA inverse folding approach using ant colony optimization. The specific definitions and particular detailing of certain used data structures and their principles are given in the trailing sections within this chapter.

Given a set of structure $\mathbb{C}^{\text{str}}$, sequence $\mathbb{C}^{\text{seq}}$ and GC $\mathbb{C}^{\text{GC}}$ constraints for a certain RNA design problem, *antaRNA* successfully generates solution sequences $\mathcal{S}^{\text{sol}}$ to the respective RNA inverse problem. In its course, the constraint set is translated into a terrain graph T, on which virtual ants perform subsequent walks, during which sequences $\mathcal{S}$ were assembled. During a single walk, an ant follows edges within the graph, leading to nucleotide emitting nodes. Each visited node $v_{j.\sigma}$ in the terrain T emits a certain nucleotide σ from the alphabet Σ_{RNA} to its respective sequence position j.

Algorithm 1: Survey of the Ant Colony Optimization Principle in its RNA inverse fold Adaptation: *antaRNA*. In general, the algorithm takes structural, sequence and GC constraint as input and modulates a terrain accordingly. While a suitable solution is not found, the algorithm iteratively produces a solution, performs its evaluation and subsequently updates the artificial pheromone (after a evaporation step) within the terrain according to the quality of the solution in a specific way. If a suitable solution was generated, it will terminate the algorithm, and the solution sequence is returned.

Data: $\mathbb{C}^{\text{str}}$, $\mathbb{C}^{\text{seq}}$, $\mathbb{C}^{\text{GC}}$
Result: $\mathcal{S}^{\text{sol}}$ satisfying $\mathbb{C}^{\text{str}}$, $\mathbb{C}^{\text{seq}}$, $\mathbb{C}^{\text{GC}}$
$T \leftarrow$ intitializeTerrain($\mathbb{C}^{\text{str}}$, $\mathbb{C}^{\text{seq}}$, $\mathbb{C}^{\text{GC}}$); $\mathcal{S}^{\text{sol}} \leftarrow \epsilon$;
while *termination criterion not met* **do**
  $\mathcal{S} \leftarrow$ produceSolution(T);
  $\mathcal{Q} \leftarrow$ evaluateSolution($\mathcal{S}$);
  $T \leftarrow$ evaporatePheromone(T, ρ);
  $T \leftarrow$ updateTerrain($T, \mathcal{S}, \mathcal{Q}$);
  if $\mathcal{S}$ *superior* $\mathcal{S}^{sol}$ **then**
    $\mathcal{S}^{\text{sol}} \leftarrow \mathcal{S}$
  end
end
return $\mathcal{S}^{sol}$;

After a sequence assembly walk, each produced sequence $\mathcal{S}$ is evaluated according to its structural distance d_{str}, its sequence distance d_{seq} and according to its GC distance d_{GC}. An overall distance quality is concentrated in a score $\mathcal{Q}$. The measures are made against the issued constraint set of the design. In a bonification procedure, the terrain graph T is updated according to the score $\mathcal{Q}$ on positions which contributed successfully

to the compliance of requested structural constraints within the current sequence under evaluation. For this, each edge in the terrain which was involved in the emission of a nucleotide and ultimately lead to the compliance of the constraints, is highlighted by the application of virtual 'pheromone'. The amount of applied pheromone is dependent on the overall quality of the current sequence. Additionally, all edges are exposed to an simulated pheromone evaporation event, such that all edges loose a certain percentage of their pheromone in this step.
The procedure of generating and evaluating a sequence and rewarding certain successfully contributing elements in the terrain graph is repeated until one of the terminating criteria is fulfilled: Termination can occur due to solution quality convergence after which the respectively best solution so far is returned upon this event. If a sequence completely complies with the demanded constraints, this respective sequence is returned as an immediate solution of the process.
The constraint input varieties, the respectively derived data structures and each processing step of the algorithm is described in detail in the following sections of this chapter.

2.2 The Constraints $\mathbb{C}$

The constraints $\mathbb{C}$ are the only means to formulate a structural problem and encode it into a design problem formulation within *antaRNA*. Dependent on the level of detail at hand, the design problem can be outlined by the careful specification of the different types of constraints operating within *antaRNA*. In this section, the different types of constraints are portrayed and illustrated. The highlight in this juncture is focused on the conceptual principles of the respective constraints. Their transformation and their usage within the algorithm are explained in later sections within this chapter. The constraints $\mathbb{C}$ comprise structural, sequence and GC constraint.

2.2.1 The Structure Constraint $\mathbb{C}^{\mathbf{str}}$

The structure constraint $\mathbb{C}^{\text{str}}$ provides the ability to formulate structural input for *antaRNA*, which is taken as optimization target template within the procedure. Dependent on the operated modus of *antaRNA*, different types of structural constraint definitions can be employed in order to define the respective objective target structure. The different definitions are presented hereafter.

The MFE Single Structure Constraint

The main and classical constraint to RNA inverse folding is the structure constraint $\mathbb{C}^{\text{str}}$ as in Definition 22. Its secondary structure P can be encoded by a secondary structure representation w^P in dot-bracket notation. Thus, the given constraint structure is taken as objective during the process. A solution sequence will show the structural constraint as its MFE structure.

Definition 22 (Single Structure Constraint)**:** Let $\mathbb{C}^{\text{str}}$ be a single structure constraint, that is a secondary structure P.

The DotPlot Structure Constraint Features

With the extension of the single structure constraint towards the DotPlot (DP) modus, it is possible to request structural properties within a structure ensemble $\mathcal{P}_{\mathcal{S}}$ of a sequence $\mathcal{S}$. These are encoded by multiple constraint features as objective. Within one design effort, it is furthermore possible to target up to two structure ensembles, i.e. one unconstrained ensemble $\mathcal{P}_{\mathcal{S}}$ and one that is a limited structure ensemble $\mathcal{P}_{\mathcal{S}|w^P}$, e.g. due to ligand interaction. Such interaction enforces a certain ligand-dependent substructure within the RNA, which is encoded by a folding constraint w^P on the structure ensemble $\mathcal{P}_{\mathcal{S}|w^P}$.
The structural constraints, that can basically be stated in DP mode are accuracy constraint features $\mathbb{C}^{\text{str}}_{\text{Accur}}$ (Definition 23) and accessibility constraint features $\mathbb{C}^{\text{str}}_{\text{Access}}$ (Definition 24). Each feature can be bound to one of the above structure ensemble, which is encoded by a label $\lambda \in \Lambda = \{\lambda_U, \lambda_C\}$, where λ_U labels the unconstrained ensemble $\mathcal{P}_{\mathcal{S}}$ and λ_C the constrained ensemble $\mathcal{P}_{\mathcal{S}|w^P}$. In addition, a folding constraint feature $\mathbb{C}^{\text{str}}_{\text{Fold}}$ as in Definition 25 can be stated for λ_C. All features form the overall set of structural constraint Ω as defined in Definition 26.

Definition 23 (Accuracy Constraint Feature)**:** Let $\mathbb{C}^{\text{str}}_{\text{Accur}} = (P, \lambda, \gamma)$ be an accuracy constraint feature, that consists of an objective nested target structure P, an ensemble label $\lambda \in \Lambda$ and an objective accuracy target value $\gamma \in [0, 1]$.

Definition 24 (Accessibility Constraint Feature)**:** Let $\mathbb{C}^{\text{str}}_{\text{Access}} = (P_{access}, \lambda, \gamma)$ be an accessibility constraint feature, that consists of an objective accessibility structure P_{access}, an ensemble label $\lambda \in \Lambda$ and an objective accessibility target value $\gamma \in [0, 1]$.

Definition 25 (Folding Constraint Feature)**:** Let $\mathbb{C}^{\text{str}}_{\text{Fold}} = (P, \lambda, \gamma)$ be a folding constraint feature, that is an instance of an accuracy feature, which consists of a nested secondary structure P of a folding constraint w^P as in Definition 3, an ensemble label

$\lambda \in \Lambda$ and an objective accuracy target value $\gamma \in [0, 1]$.
If no limited structure ensemble is to be optimized, the structure is set to $P = \varnothing$, the ensemble label to $\lambda = \lambda_U$ and the target value to $\gamma = 0$.
If a folding constraint P for a limited ensemble $\mathcal{P}_{\mathcal{S}|w^P}$ is present, the ensemble label is set to $\lambda = \lambda_C$ and the objective value to $\gamma = 1$.

Definition 26 (Structure Constraint Feature Set)**:** Let $\mathbb{C}^{\text{Accur}} = \{\mathbb{C}^{\text{str}}_{\text{Fold}}\} \cup \bigcup_{t=1}^{T} \mathbb{C}^{\text{str}}_{\text{Accur}_t}$ be the set of the folding constraint feature and T requested accuracy constraint features and let $\mathbb{C}^{\text{Access}} = \bigcup_{u=1}^{U} \mathbb{C}^{\text{str}}_{\text{Access}_u}$ be the set of U requested accessibility constraint features. Furthermore, let Ω be the set of all constraint structure features, i.e. $\Omega = \mathbb{C}^{\text{Accur}} \cup \mathbb{C}^{\text{Access}}$.

With the use of the structure constraint features, specific base pair enrichment/depletion within the according base pair probability matrices is pursued. Each type of constraint feature pursues a different type of structure, that covers its specific base pair enrichment/depletion. Each feature can be defined several times. The accuracy constraint feature is used for the enrichment of a nested secondary structure, whereas the accessibility constraint feature can be used to enforce base pair depletion in whole areas within the respective probability matrix that belong to the used accessibility structure.

The Fuzzy Structure Constraint

The fuzzy structure constraint is a newly developed constraint concept feature to the RNA inverse folding problem which is provided by *antaRNA* and which does not specify base pairing interactions explicitly. It rather allows to render input, which vaguely defines areas within the design object which do not impose penalties within the evaluation of those structure domains, if they interact. It is an extension to both the MFE modus and the structure constraint features within DP-modus. The general difference between the fuzzy constraint definition within the MFE modus and the DP modus is that in the MFE modus, the fuzzy constraint serves as a wildcard structure for certain regions of the sequence, whereas the fuzzy constraint in the DP modus can model an objective degree of structuredness among certain stretches of the sequence.

MFE Fuzzy Constraint Within the MFE modus, a MFE structure constraint can be used according to the Definition 27 as an extension of the usable underlaying secondary structure of the single structure constraint definition (Definition 22).

Definition 27 (Single Structure Fuzzy Constraint Extension)**:** Let a single structure fuzzy constraint extension be an alphabetical extension within a single structure constraint $\mathbb{C}^{\text{str}}$, such that the structure defining alphabet Σ_P is extended by the alphabet

literals of the standard latin alphabet $\Sigma_L = \{a, b, c, ..., x, y, z, A, B, C, ..., X, Y, Z\}$.

Let $P^{\mathbb{C}^{\text{str}}_B}$ be the set of base pairs, that comply with the alphabetical extension of a single structure fuzzy constraint $\mathbb{C}^{\text{str}}$, that are in the alphabet Σ_L, such that

$$P^{\mathbb{C}^{\text{str}}_B} = \bigcup_{\mathcal{L} \in \Sigma_L} \{(i, j) \mid \mathbb{C}^{\text{str}}_i = \mathcal{L} = \mathbb{C}^{\text{str}}_j \wedge j - i > 3\}$$

Still, the secondary structure $P^{\mathbb{C}^{\text{str}}}$ of the constraint $\mathbb{C}^{\text{str}}$ only consists of the regularly allowed base pair interactions which result from the alphabet Σ_P as stated in Definition 4.

Constraint positions, declared with the same latin character, are allowed to interact within the solution's MFE structure. In the current adaptations, this constraint type is presented and can be used as 'soft' (lower case letters) and/or 'hard' (upper case letters) constraint. In the 'soft' modus, base pairing is not a necessity within a solution, but might occur in such without provoking a constraint violation. In contrast, if the 'hard' constraint variant is chosen as input, there must be at least one positively predicted base pair interaction in the solution sequence among the specified fuzzy constraint region, in order to prevent a constraint violation. An example setup of a MFE fuzzy structure constraint feature can be found in in Figure 2.1.

DP Fuzzy Constraint Within the DP modus, the fuzzy constraint $\mathbb{C}^{\text{str}}_{\text{Fuzzy}}$ (Definition 28) conceptually represents a means, that can enrich/deplete average structuredness among certain stretches of RNA within the underlaying structure ensemble of that sequence. If fuzzy structural constraint is defined within the DP modus, the overall set of structure constraints Ω is updated according to Definition 29.

Definition 28 (Fuzzy Constraint Feature)**:** Let $\mathbb{C}^{\text{str}}_{\text{Fuzzy}} = (P_{fuzzy}, \lambda, \gamma)$ be a fuzzy constraint feature, that consists of a fuzzy structure P_{fuzzy}, an ensemble label $\lambda \in \Lambda$ and an objective fuzzy target value $\gamma \in [0, 1]$.

Definition 29 (Updated Overall Structural Constraint)**:** Let $\mathbb{C}^{\text{Fuzzy}} = \bigcup_{v=1}^{V} \mathbb{C}^{\text{str}}_{\text{Fuzzy}_v}$ be the set of V requested fuzzy feature constraints. Ω is extended, such that the updated set of overall structural constraints is given by $\Omega \cup \mathbb{C}^{\text{Fuzzy}}$.

As depicted in Figure 2.2, the fuzzy constraint features in DP modus pursue the concept of enriching/depleting base pair probabilities within certain areas of the objective base pair probability matrices according to the objective target value γ of the respective

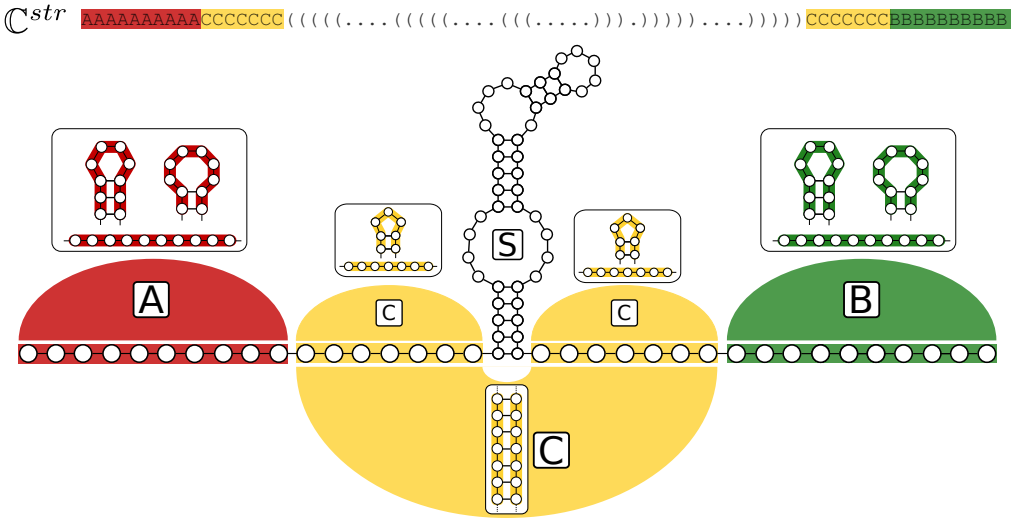

Figure 2.1: Illustrative Schema of $\mathbb{C}^{\mathbf{str}}$ in MFE modus
Single structure constraints: Regular dot-bracket and Fuzzy variations of $\mathbb{C}^{\text{str}}$ within the MFE modus. Only one target structure is imputable in the case of explicit structure request in dot-bracket notation (structure S), whereas different structure peculiarities are allowed in the fuzzy defined regions A, B and C. Each of the substructures are legal in terms of the constraint. Region C can form an basal hairpin extension to S or can fold into singular structures flanking structure S, whereas regions A and B can solely fold into the 'singular' structures, but are not allowed to interact. Taken and extended from Kleinkauf *et al.* (2015a) using Inkscape and ViennaRNA Software.

fuzzy constraint feature. The use of variants of the underlaying fuzzy structure allows to determine fuzzy structure in every area of the base pair probability matrices.

Structure Dependency Graph $G_{\mathbb{C}}$

In order to provide insight into the dependencies between the single positions within a complex structure constraint environment within the DP modus, we use a dependency graph, as depicted in Definition 30. In addition, also its connected components (Definition 31) and singleton nodes (Definition 32) are derived.

Definition 30 (Dependency Graph)**:** Let $G_{\mathbb{C}} = (\mathcal{V}_{\mathbb{C}}, \mathcal{E}_{\mathbb{C}})$ be a dependency graph of the structural accuracy constraints, in which the nodes $\mathcal{V}_{\mathbb{C}}$ represent the single positions of a nucleotide sequence of length n, such that $\mathcal{V}_{\mathbb{C}} = \{v_1, v_2 \dots v_n\}$. The edges $\mathcal{E}_{\mathbb{C}}$ represent base pairs that occur in some defined accuracy constraint feature structures within $\mathbb{C}^{\text{Accur}}$, such that $\mathcal{E}_{\mathbb{C}} = \bigcup_{(P,\lambda,\gamma)\in\mathbb{C}^{\text{Accur}}} \{(v_i, v_j) \mid \exists v_i, v_j \in \mathcal{V}_{\mathbb{C}} : (i,j) \in P\}$.

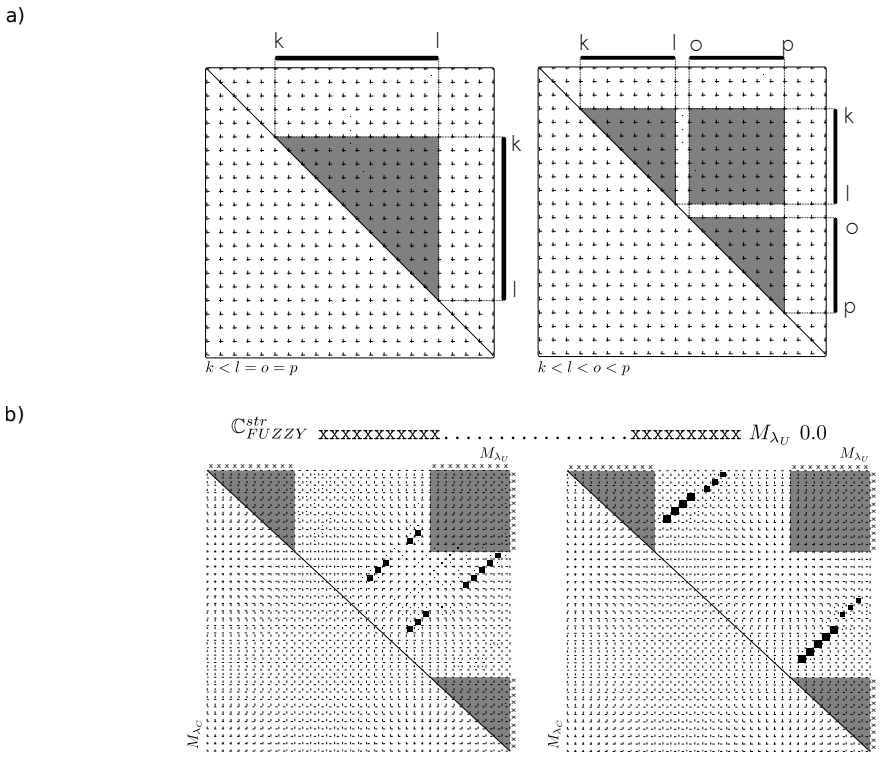

Figure 2.2: Illustrative Example of a fuzzy constraint feature $\mathbb{C}^{\mathbf{str}}_{\mathbf{Fuzzy}}$ in DP modus **a)** Example of the fuzzy constraint feature within the DP modus. The structure boundary is dependent on the specifications of k, l, o, p. The left example covers the special case $k < l = o = p$, whereas the right example truly has varying variables, i.e. $k < l < o < p$. **b)** Exemplary results of a negative design principle, that was realized by the specified fuzzy constraint feature. The objective focuses on not folding into structure elements within the specified regions of 'x' (gray areas). Structure is only allowed within the uncolored regions in the upper triangle of the base pair probability matrices. The indicative examples show, that differently characterized structures can result from the principle of fuzzy constraint within *antaRNA* DP modus. Inkscape and ViennaRNA Software.

Definition 31 (Set of Connected Components)**:** Let CC be the set of connected components within the dependency graph $G_{\mathbb{C}}$, that is

$$\mathsf{CC} = \{\, cc \mid cc \text{ is a connected component in } G_{\mathbb{C}} \ \wedge |\mathcal{V}_{\mathbb{C}|cc}| \geqslant 2\}.$$

A connected component cc consists of nodes $\mathcal{V}_{\mathbb{C}|cc}$ and the set of their connecting edges $\mathcal{E}_{\mathbb{C}|cc}$. Let $id(cc) = \arg\min_i(v_i \in \mathcal{V}_{\mathbb{C}|cc})$ be a function that returns the sequence index i of the node v_i, which has the the smallest sequence index among the nodes of the connected component $\mathcal{V}_{\mathbb{C}|cc}$.

Note, a connected component in the dependency graph thus groups all directly or indirectly dependent sequence positions, which is of importance to combine sequence and structure constraints.

Definition 32 (Set of Singletons)**:** Let S be the set of singletons of the dependency graph $G_{\mathbb{C}}$, that is $\mathsf{S} = \{v \mid \nexists v' : (v, v') \in \mathcal{E}_{\mathbb{C}} \vee (v', v) \in \mathcal{E}_{\mathbb{C}}\}$

2.2.2 The Sequence Constraint $\mathbb{C}^{\mathbf{seq}}$

Besides the structure constraint, adjusting specific sequence positions to certain nucleotides or ambiguous groups of nucleotides is very important, since certain nucleotides coupled with a specific underlaying structure is often key to enable a biological function. In *antaRNA* two different sequence constraint possibilities are available. The realization of the sequence constraint is in both variants of the program regulated within the construction layout of the specific terrain of a design but follow Definition 33.

Definition 33 (Sequence Constraint)**:** Let $\mathbb{C}^{\text{seq}}$ be the sequence constraint of length n, such that each position in $\mathbb{C}^{\text{seq}}$ constraints a position i in the sighted solution sequence $\mathcal{S}$. The sequence constraint is defined from the ambiguous RNA nucleotide alphabet $\Sigma_S = \{A, C, G, U, R, Y, S, W, K, M, B, D, H, V, N\}$, i.e. $\mathbb{C}^{\text{seq}} = (\Sigma_S)^n$. The decoding of each ambiguous alphabet letter to the according set of nucleotides is done by $\text{NT} : \Sigma_S \to \wp(\Sigma_{RNA})$, where $\wp$ denotes the power set of a set. The function $\text{COMPL} : \Sigma_S \to \wp(\Sigma_{RNA})$ provides the set of possible base pairing partners. Both are given in Table 2.1.

Both constraint variants are based on all allowed RNA IUPAC nucleotides, as listed in Table 2.1. During the construction of the terrain, all sequence constrained nodes within the terrain setup, which do not agree with the respective sequence constraint on a specific position, are not allocated, such that they can not be visited by a virtual ant, i.e. a not allowed nucleotide can not be emitted into a result sequence.

Explicit Sequence Constraint By the use of the sequence constraint $\mathbb{C}^{\text{seq}}$, the target sequence can be specified in a way, that certain nucleotide positions are instanced by an explicit or ambiguous nucleotide representations. Apart from the N case that, all other IUPAC ambiguous RNA nucleotide definitions can be used in order to constrain positions within the sequence construct.

Implicit Temporary Sequence Constraint Modification Given two positions i and j are constrained by the structure constraint $\mathbb{C}^{\text{str}}$ and are forming the base pair within a current structure folding $(i, j) \in P$, furthermore given that sequence position $\mathcal{S}_i$ is constrained with a specific nucleotide $\mathbb{C}^{\text{seq}}_i = \sigma \in \Sigma_{RNA}$ and $\mathbb{C}^{\text{seq}}_j$ is constrained by N,

then the sequence position $\mathbb{C}_j^{\text{seq}}$ gets implicitely constrained by the possible complement nucleotide(s) of σ, $\mathbb{C}_j^{\text{seq}} = \text{NT}^{-1}(\text{COMPL}(\sigma))$. The implicit sequence constraint is only available in the MFE modus, since it provides a specific MFE structure, on which this principle can be applied. The application of this principle in the DP modus was not pursued in this respect.

$s \in \Sigma_S$	NT(s)	COMPL(s)\ 'GU'	COMPL(s)
A	{A}	{U}	{U}
C	{C}	{G}	{G}
G	{G}	{C}	{C, U}
U	{U}	{A}	{A, G}
R	{A, G}	{U,C}	{U,C}
Y	{C, U}	{A, G}	{A, G}
S	{G, C}	{G, C}	{U, G, C}
W	{A, U}	{U, A}	{U, A, G}
K	{G, U}	{C, A}	{U, C, A, G}
M	{A, C}	{U, G}	{U, G}
B	{C, G, U}	{A, G, C}	{A, G, C, U}
D	{A, G, U}	{A, C, U}	{A, G, C, U}
H	{A, C, U}	{U, G, A}	{U, G, A}
V	{A, C, G}	{U, G, C}	{U, G, C}
N	{A, C, G, U}	{A, C, G, U}	{A, C, G, U}

Table 2.1: IUPAC RNA Nucleotide Ambiguity Code
The IUPAC RNA nucleotide codes $s \in \Sigma_S$ resemble nucleotides or groups of them that are decoded by NT(s). COMPL(s) represents the complementing set, and COMPL(s)\'GC' the same set without G::U forming base pairs.

2.2.3 The GC Value Constraint $\mathbb{C}^{\text{GC}}$

Since the GC content of RNA sequences varies not only among the different entities of RNA, but as well among the different organisms, it is of interest to control the GC content of an designed RNA sequence to a very precise level, such that organism- and RNA specific GC values can be adjusted. With this means, the designed entities can be optimized towards the respective RNA system, but as well can be adjusted to the organism it is going to be introduced into.

Definition 34 (GC Content Feature)**:** Let $c^{\text{GC}} = (\gamma, i, j)$ be a GC content feature constraining a construct stretch limited by the positions i and j ($i, j \in \mathbb{N} \wedge 1 \leqslant i < j \leqslant n$) with an objective GC value of $\gamma \in [0, 1]$.

Definition 35 (GC Constraint)**:** Let $\mathbb{C}^{\text{GC}} = \bigcup_{l=1}^{L} c_l^{\text{GC}}$ be the set of L GC content features. Each sequence postion must be covered by exactly one GC content feature, i.e. $\forall k \in [1, n] : \exists(\gamma, i, j) \in \mathbb{C}^{\text{GC}} : k \in [i, j] \wedge \nexists(\gamma', i', j') \neq (\gamma, i, j) \in \mathbb{C}^{\text{GC}} : k \in [i', j']$.

Even though the primary function of the GC constraint $\mathbb{C}^{\mathrm{GC}}$ (Definition 35) is to constrain the design object to one general or several stretch specific, very precise GC value(s), the GC value constraint $\mathbb{C}^{\mathrm{GC}}$ in *antaRNA* can be used in different ways. Once generated as mandatory input, the adaptation will produce a sequence, which is closest to the specification. As an advancement of the very precise single value GC constraint, also more complex constraint definitions are possible: On the one hand, the input can be split up into several non-overlapping sections requesting different GC values within one construct, and on the other hand, a Gauss- or Normal-distribution dependent GC value can be specified as target value, such that populations of sequences generated under the picked constraint comply the respectively selected underlaying distribution. Since the concept of GC constraint is rendered within the terrain graph as its attribute, please consult the terrain graph section for further specifications of the adaptation of the GC constraint.

2.3 The Terrain

As hinted at in the survey functionality description of *antaRNA*, the algorithm emulates a virtual Terrain on which virtual ants will walk in order to assemble an RNA sequence. By the way the Terrain is constructed and composed, the assembly of sequences will be guided into suitable directions. This is done by representing the Terrain using a graph data structure, which allows an ant to walk on the edges and simultaneously invoke the emission of sequence position specific nucleotides with a certain probability at the respective nodes of the graph. In this section, the used graph and its composition are described in detail. For each modus of *antaRNA*, a different type of graph layout is initialized in order to reflect the respective situation of the differently defined and used structural constraints.

2.3.1 The Terrain Graph T

The Terrain is initially represented as a graph as in Definition 38. Dependent on the modus of the program, the Terrain is initialized differently. The main reason for this are the alternatively characterized structure constraints that are used in the different modi. The structure constraint in the MFE modus can only state base pairs, i.e. the dependency graph shows only connected components of size 2 (the base pairs) and singletons (unpaired positions). In the DP modus, the structure constraint representation is more complex than in the MFE modus since structure can be defined and constrained via various independent but yet position-wise overlapping constraint features. Therefore, connected components of the resulting structure dependency graph can grow larger. During the construction of the Terrain the topology of the dependency graph is transformed into suitable subgraphs of the Terrain. Based on that, dependent positions,

which are grouped in a connected component, can be assigned with nucleotides in a way, such that they fulfill both the sequence and structure constraints.
The edges and nodes of the Terrain, independent of the modus, allow an ant to start walking through the terrain passing nucleotide emission nodes. The type specifications for nodes (Definition 36) and edges (Definition 37) provide fundamental characteristics, such that an ant can select an edge based on its distinct features. As soon as a nucleotide emission node is visited, its distinct nucleotide is emitted and assigned to the specific position, which is given by the node as well. A more detailed description of the initialization of the Terrain T in terms of its specific layout is given after the initial definitions for the cases of the MFE and the DP modus respectively. The sequence assembly and its evaluation are explained in more detail after the Terrain layout part.

Definition 36 (Emission Vertex $v_{i.\sigma}$)**:** Let $v_{i.\sigma}$ be an emission node, that emits, if visited, a specific nucleotide symbol $\sigma \in \Sigma_{\text{RNA}}$ to the i^{th} position of a currently assembles solution sequence $\mathcal{S}$.

Definition 37 (Transition Edge $e_{i,j}$)**:** Let a transition edge $e_{i,j}$ be a directed edge, that connects two emission nodes v_i and v_j. Let τ be the pheromonic weight of the edge, such that $\tau(e_{i,j}) \in \mathbb{R}$ and let η be the length of the edge, such that $\eta(e_{i,j}) \in \mathbb{R}$.

Definition 38 (Terrain T)**:** Let the Terrain T be a directed graph $G = (\mathcal{V}, \mathcal{E})$ with a set of vertices $\mathcal{V}$, which is comprised of a set of emitting vertices $\mathcal{V}_e$ and a set of non-emitting vertices $\mathcal{V}_\bullet$, i.e. $\mathcal{V} = \mathcal{V}_e \cup \mathcal{V}_\bullet$, such that $\mathcal{V}_e \subseteq \mathcal{V}_E = \{v_{i.\sigma} \mid 1 \leqslant i \leqslant n \wedge \sigma \in \Sigma_{\text{RNA}}\}$ and $\mathcal{V}_\bullet \subseteq \{v_\bullet, v_\bullet^1, \cdots, v_\bullet^n\}$. The connecting edges constitute as $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$. Let $\mathcal{V}^i = \{v_{i.\sigma} \mid v_{i.\sigma} \in \mathcal{V}_e\}$ be the set of nodes, that can emit to a sequence position i.

Imposed sequence and structure constraints combined can, due to their overlapping nature, contradict themselves, such that a suitable solution production is not possible at all. The contradictions, that can occur upon such situation can be identified as sequence/structure constraint clashes, probability overload and odd cycles within connected components.
In both modi of the program, sequence/structure constraint clashes, that emerge, if for example, a base pair is allocated with two explicit nucleotides, that are not complementary to each other, can be identified due to the allocated layout of the Terrain T, i.e. $\forall \mathcal{V}^i : \mathcal{V}^i \neq \{\}$ must hold for all positions, that are assigned to emitting nodes. Otherwise, there exists an allocation flaw due to this problem category.
In DP modus, imposed objective structure feature values can overlap on certain positions to that extent, such that their requested probabilities all together exceed a maximally possible probability load of 1 on affected positions. Therefore a check on

all constrained positions and their resulting base pairs must be performed, in order to exclude this contradiction.
And furthermore, requested base pairs among different accuracy constraints features can produce situations, such that odd cycles within their affected connected components of the dependency graph $G_{\mathbb{C}}$ occur. In those cases, due to the complementarity of nucleotides, no appropriate nucleotide allocation to the terrain and subsequently to the solution sequence would be possible. A cycle in a graph, that is a trail in a graph G that visits the same node v_i twice. A trail is a sequence of edges, in which a single edge is not allowed to be used twice. An odd cycle therefore requires, that the number of edges in the trail, which has been used to revisit v_i twice, is odd.

Minimum Free Energy (MFE) - Modus

Within the MFE modus, the set of non-emitting nodes $\mathcal{V}_\bullet$ comprises only one node $\mathcal{V}_\bullet = \{v_\bullet\}$ and defines, together with the set of emitting nodes $\mathcal{V}_e = \mathcal{V}_E \backslash \{v_{i,\sigma} \mid \sigma \notin \mathrm{NT}(\mathbb{C}_i^{\mathrm{seq}}) \vee \exists (k,i) \in \mathbb{C}^{\mathrm{str}} : \sigma \notin \textsc{compl}(\mathbb{C}_k^{\mathrm{seq}})\}$, the set of vertexes $\mathcal{V}$. The connecting edges constitute as $\mathcal{E} = \mathcal{V}_\bullet \times \mathcal{V}_e^1 \cup \bigcup_{x=2}^{n} \mathcal{V}_e^{x-1} \times \mathcal{V}_e^x$. An initial setup of the terrain is depicted in Figure 2.3.a.

Dot Plot (DP) - Modus

As indicated in the beginning of the Terrain description, it is paramount to capture the dependencies of all positions, which result from structural constraints, and translate them into suitable sets of emission nodes and their connecting edges, such that contradiction free sequence assembly can be performed. For that reason, the Terrain is set up based on the connectivity information of the connected components of the dependency graph that are given by the structural constraints.
Therefore, within the DP modus, the set of non-emitting nodes $\mathcal{V}_\bullet$ comprises a root non-emitting node $v\bullet$ and one non-emitting node for each detected connected component and each singleton in the dependency graph $G_{\mathbb{C}}$, such that $\mathcal{V}\bullet = \{v_\bullet\} \cup \bigcup_{cc \in \mathsf{CC}} v_\bullet^{id(cc)} \cup \bigcup_{v_i \in \mathsf{S}} v_\bullet^i$. Each connected component and each singleton of the deoendency graph get their own starting node, since the probability of the edges depend on all edges that leave a node, i.e. if all emitting components were directly connected to the root, an appropriate selection of edges would result in fail. However, the emitting nodes are set up analogously to the MFE modus, with the only difference that now the structural dependency check is based on edges within the dependency graph $G_{\mathbb{C}}$ and not based on base pair information from the structure constraint: $\mathcal{V}_e = \mathcal{V}_E \backslash \{v_{i,\sigma} \mid \sigma \notin \mathrm{NT}(\mathbb{C}_i^{\mathrm{seq}}) \vee \exists (k,i) \in \mathcal{E}_{\mathbb{C}} : \sigma \notin \textsc{compl}(\mathbb{C}_k^{\mathrm{seq}})\}$.
Within the Terrain T, the non-emitting root node $v_\bullet$ is connected to the remaining non-emitting nodes in $\mathcal{V}_\bullet$. From each intermediate non-emitting node $v_\bullet^i \in \mathcal{V}_\bullet$,

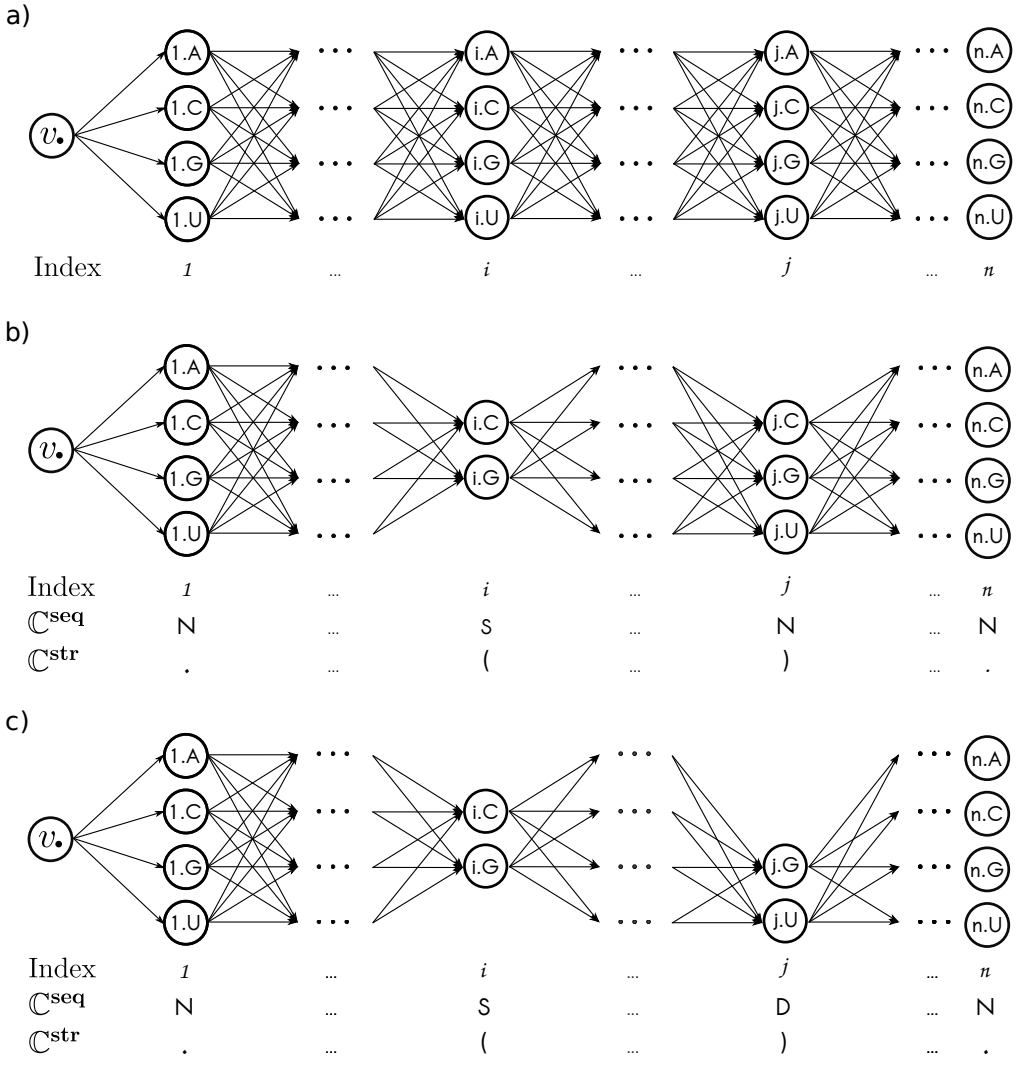

Figure 2.3: Exemplary MFE Modus Terrain Layout
a) Plain terrain without the influence of any constraint. For all positions, all four possible nucleotide emitting nodes are initialized. **b)** Influence of structure and sequence constraint. A requested base pair between positions i and j in accordance with sequence constraint "S" on position i induces sequence complementary constraint on to position j, considering GU base pairs. **c)** Sequence complementary constraint of position i on to position j is limited by the sequence constraint of position j itself. Taken and extended from Kleinkauf *et al.* (2015b) using Inkscape.

edges are drawn to its emitting group, which is either representing a connected component $cc \in \mathsf{CC}$ or a singleton $s \in \mathsf{S}$ of the dependency graph $G_{\mathbb{C}}$. For each connected component cc, a pre-order traversal of the nodes is used to get a deterministic layout of the according Terrain's subgraph. The pre-order is generated by a

depth-first search DFS(cc, v) starting from a node $v \in \mathcal{V}_{\mathbb{C}|cc}$. Within the following, the DFS(cc,v) returns the per-order traversal information in form of visited edges $(v_i, v_j) \in \mathcal{E}_{\mathbb{C}|cc}$ instead of visited nodes only. Based on the obtained pre-order of a connected component, the according set of edges $\mathcal{E}_{cc,v} \subset \mathcal{E}$ in the Terrain T is defined by $\mathcal{E}_{cc,v} = \bigcup_{(v_i,v_j)\in \mathrm{DFS}(cc,v)} \{(v_{i,\sigma}, v_{j,\sigma'}) \in \mathcal{V}_e^i \times \mathcal{V}_e^j \mid \sigma' \in \textsc{compl}(\sigma)\}$. In addition to the actual structure constraint, the set of edges in $\mathcal{E}_{cc,v}$ considers nucleotide complementarity between neighboring nodes within the connected component as well.
The overall set of edges $\mathcal{E}$ in the DP modus is therefore given by

$$\mathcal{E} = \bigcup \begin{cases} \bigcup_{cc\in \mathsf{CC}} (v_\bullet, v_\bullet^{id(cc)}) \cup \bigcup_{v_i \in \mathsf{S}} (v_\bullet, v_\bullet^i) & \text{Root} \rightarrow \text{Non-emitting Nodes} \\ \bigcup_{cc\in \mathsf{CC}} \left\{ \{v_\bullet^{id(cc)}\} \times \mathcal{V}_e^{id(cc)} \right\} \cup \mathcal{E}_{cc,v_{id(cc)}} & \text{Connected Components Internally} \\ \bigcup_{v_i \in \mathsf{S}} \{v_\bullet^i\} \times \mathcal{V}_e^i & \text{Non-emitting Nodes} \rightarrow \text{Singletons} \end{cases}$$

2.4 Sequence Assembly

In each iteration of *antaRNA*, k sequences are compiled under the same terrain condition. The sequence with the best evaluation gets promoted and wins the underlaying tournament of the iteration. In order to assemble a single sequence of length n, an ant performs a walk through the modus dependent layout of the terrain graph T, that guides the ant to assemble the sequence based on the respective conditions $(\tau(e), \eta(e))$ of the edges $e \in \mathcal{E}$ within the terrain. For each sequence position $j \in \{1, \cdots, n\}$ the ant selects an edge $e_{(i.\sigma,j.\sigma')} \in \mathcal{E}$ leading from a topologically precurser vertex $v_i.\sigma$ to a vertex $v_j.\sigma'$, which is then emitting its nucleotide σ' to the sequence position j. The ant selects the respective edge e according to its probability p. The probability p of an edge $e_{(i.\sigma,j.\sigma')}$ solely depends on its intrinsic heuristic information $\eta(e_{(i.\sigma,j.\sigma')})$, its respective pheromone value $\tau(e_{(i.\sigma,j.\sigma')})$ and on the context of all edges leading away from $v_{i.\sigma}$, namely $\mathcal{E}_{(i.\sigma,j.\sigma*)} \subset \mathcal{E}$. The parameters α and β are weight factors to the edge features $\tau(e)$ and $\eta(e)$ respectively. Equation 2.1 summarizes the single influences of α on $\tau(e)$ and β on $\eta(e)$, and depicts the overall relation of a probability of an edge to be selected in dependence of its competing edges.

$$p(e_{(i.\sigma,j.\sigma')}) = \frac{\alpha \cdot \tau(e_{(i.\sigma,j.\sigma')}) + \beta \cdot \eta(e_{(i.\sigma,j.\sigma')})}{\sum_{\sigma* \in \mathbb{C}_j^{\mathrm{seq}}} \left(\alpha \cdot \tau(e_{(i.\sigma,j.\sigma*)}) + \beta \cdot \eta(e_{(i.\sigma,j.\sigma*)}) \right)} \tag{2.1}$$

Note, the above procedure covers only edges that connect two emitting nodes from $\mathcal{V}_e$. However, edges $e \in \mathcal{E} \cap (\mathcal{V}_\bullet \times \mathcal{V}_e)$ that start in a non-emitting node $v_\bullet \in \mathcal{V}_\bullet$ but lead to an emitting node are dealt equally in their probability based selection.
Within the MFE modus, the sequence is assembled linearly from the start to the end of the sequence. The only exceptions are nucleotides on sequence positions which have closing base character. They are solely dependent on their already assigned base pair

partner nucleotide. In some cases of allowed 'GU' base pairs, there can be a probability dependent selection for that case.
The sequence assembly within the DP modus iterates over each emitting subgraph of the Terrain T in order to fill every position in the sequence. Since the topology of the Terrain T is assigned according to pre-order of the underlaying depth-first search on the basis of the connected components of the dependency graph, the assembling ant uses a respective walk order which informs the ant which position of the sequence has to be detailed next. Each selected edge is stored within solution specific sequence assembly trail, as described in Definition 39.

Definition 39 (Sequence Assembly Trail)**:** Let a sequence assembly trail be $\mathcal{E}_{\text{walk}} \subset \mathcal{E}$, which is the set of edges that are used during a Terrain T traversal in order to assemble a sequence. For each sequence position $i \in [1, n]$ there exists exactly one edge in $\mathcal{E}_{\text{walk}}$ which ends in an emitting $v_{i.\sigma} \in \mathcal{V}$.

2.5 Sequence Quality Evaluation

Within *antaRNA*, the evaluation of a sequence includes the measurement of three distances d between the sequence intrinsic features and the constraints. All distances are normalized, such that they represent a percentage measurement of the respective deviation from the specific feature request and therefor can be compared to other calculations and do not stay a relative measurement to the current design concept.

2.5.1 GC-Distance $d_{\mathbf{GC}}$

The GC distance adapts and resembles the deviation of an actual GC value of a sequence $\mathcal{S}$ to all respectively imposed GC objective target values, that are provided by GC constraint features $c^{\text{GC}} \in \mathbb{C}^{\text{GC}}$. Based on the single GC value difference between a constraint feature c^{GC} and the sequence $\mathcal{S}$ (Definition 40), the an intrinsic GC error compensated (Definition 41) based GC feature distance (Definition 42) is provided, which in the end cumulatively is combined into a global GC distance between the sequence $\mathcal{S}$ and all imposed GC constraint features $c^{\text{GC}} \in \mathbb{C}^{\text{GC}}$ (Definition 43).

Definition 40 (GC Difference)**:** Let Δ_{GC} be the difference between the measured GC value of a subsequence $\mathcal{S}_{i..j}$ that is constrained by a GC constraint feature $(\gamma, i, j) \in \mathbb{C}^{\text{GC}}$ with its objective value γ, i.e.

$$\Delta_{GC}(\mathcal{S}_{i..j}, \gamma) = |GC(\mathcal{S}_{i..j}) - \gamma|.$$

Definition 41 (GC Compensation)**:** Given a GC constraint feature $(\gamma, i, j) \in \mathbb{C}^{\mathrm{GC}}$ that delimits a subsequence $\mathcal{S}_{i..j}$ of length $m = j-i+1$, let $\delta^{+}_{GC}(\mathcal{S}_{i..j}, \gamma)$ and $\delta^{-}_{GC}(\mathcal{S}_{i..j}, \gamma)$ be the GC compensation values, i.e.

$$\delta^{+}_{GC}(\mathcal{S}_{i..j}, \gamma) = \frac{|(\gamma \cdot m) - \lceil \gamma \cdot m \rceil|}{m} \ , \ \delta^{-}_{GC}(\mathcal{S}_{i..j}, \gamma) = \frac{|(\gamma \cdot m) - \lfloor \gamma \cdot m \rfloor|}{m}.$$

Both are GC compensatory values that describe, if any, the intrinsic GC error within a sequence, which arises from nucleotide discretization.

Definition 42 (GC Feature Distance)**:** Let $d_{\mathrm{gc}}(\mathcal{S}_{i..j}, \gamma)$ be the GC feature distance between a sequence $\mathcal{S}_{i..j}$ and a GC objective value γ that is given by a GC constraint feature $(\gamma, i, j) \in \mathbb{C}^{\mathrm{GC}}$, i.e.

$$d_{\mathrm{gc}}(\mathcal{S}_{i..j}, \gamma) = \begin{cases} \Delta_{GC}(\mathcal{S}_{i..j}, \gamma) - \delta^{+}_{GC}(\mathcal{S}_{i..j}, \gamma) & \text{if } \Delta_{GC}(\mathcal{S}_{i..j}, \gamma) > 0 \\ \Delta_{GC}(\mathcal{S}_{i..j}, \gamma) + \delta^{-}_{GC}(\mathcal{S}_{i..j}, \gamma) & \text{if } \Delta_{GC}(\mathcal{S}_{i..j}, \gamma) < 0 \\ \Delta_{GC}(\mathcal{S}_{i..j}, \gamma) & \text{else} \end{cases} \tag{2.2}$$

Definition 43 (GC Distance)**:** Let d_{GC} be the GC distance between a sequence $\mathcal{S}$ and all imposed GC constraint features $\mathbb{C}^{\mathrm{GC}}$. Each GC constraint feature $(\gamma, i, j) \in \mathbb{C}^{\mathrm{GC}}$ evaluates the GC content of a substring of the sequence $\mathcal{S}_{i..j}$, delimited by i and j with its specific objective value γ, i.e.

$$d_{\mathrm{GC}} = \sum_{(\gamma,i,j) \in \mathbb{C}^{\mathrm{GC}}} d_{\mathrm{gc}}(\mathcal{S}_{i..j}, \gamma) \tag{2.3}$$

2.5.2 Sequence Distance d_{seq}

The sequence distance d_{seq} measures the distance between the requested sequence constraint and the actual nucleotide which was emitted to the respective position of the solution sequence $\mathcal{S}$. Any mismatched character not satisfying the requested constraint is dealt equally by a raise of a penalty score of 1. The total penalty score of the sequence is normalized by the sequence length n to transform it into a percentage value, as shown in Definition 44.

Definition 44 (Sequence Distance)**:** Let d_{seq} be the length normalized distance between a sequence $\mathcal{S}$ and its sequence constraint $\mathbb{C}^{\mathrm{seq}}$, such that

$$d_{\mathrm{seq}}(\mathcal{S}, \mathbb{C}^{\mathrm{seq}}) = \frac{|\{\ i \mid \mathcal{S}_i \notin \mathrm{NT}(\mathbb{C}^{\mathrm{seq}}_i)\ \}|}{n} \tag{2.4}$$

2.5.3 Structure Distance $d_{\mathbf{str}}$

The structural distance d_{str} is calculated according to the structure constraint type and the underlaying modus, in which *antaRNA* is operated. Basically there are two different mode: MFE modus allows to define one target structure constraint in dot-bracket notation. The structural distance to the requested structure is based on the binary presence of the respective base pairs within the request and the minimum free energy structure of the current solution sequence. Within the DP modus, however, the underlaying idea is to provide a basis for the modeling and definition of bistable RNA molecules. Therefore the structural constraint can be provided by multiple structural requests. The distance of a current sequence's structure to the requested structural constraint is provided by a distance function based on the deviation of the current base pair probability matrices.

Minimum Free Energy (MFE) - Modus

Within the MFE evaluation of the structural distance the congruence of the MFE structure P^{sol} of the current solution sequence to the requested structure of the structure constraint $\mathbb{C}^{\text{str}}$ is examined. Since the target structure $P^{\mathbb{C}}$ is not only comprised of the structure constraint $\mathbb{C}^{\text{str}}$, it needs to be fused with other contributing factors as depicted in Definition 46, in which the lonely base pairs of a structure, if enabled, are accounted according to Definition 5 as well. In *antaRNA*, the classical lonely base pairs are considered upon user request. To furthermore promote lonely base pairs within MFE structures, the concept of lonely base pairs has been extended to two base lonely base pairs in general.

Definition 45 (Sequence Constraint Induced Base Pairs)**:** Let $P^{SI}(P|\mathbb{C}^{\text{seq}}, \mathbb{C}^{\text{str}})$ be the set of sequence constraint induced base pairs within a structure P, that have been induced by a sequence constraint $\mathbb{C}^{\text{seq}}$ and the single structure constraint $\mathbb{C}^{\text{str}}$, such that $P^{SI}(P|\mathbb{C}^{\text{seq}}, \mathbb{C}^{\text{str}}) = \{(i,j) \mid (i,j) \in (P \backslash \mathbb{C}^{\text{str}}) \wedge \mathbb{C}^{\text{seq}}_i \in \Sigma_{RNA} \wedge \mathbb{C}^{\text{seq}}_j \in \Sigma_{RNA}\}$.

Definition 46 (Effective Target Structure)**:** Let $P^{\mathbb{C}}$ be the effective target structure, in which the constraint structure $\mathbb{C}^{\text{str}}$, the lonely base pairs of that structure $LP(\mathbb{C}^{\text{str}})$, if any, the fuzzy constraint pairs $P^{\mathbb{C}^{\text{str}}_B}$, if any, and the sequence constraint $\mathbb{C}^{\text{seq}}$ are used together with a solution structure P^{sol}, such that

$$P^{\mathbb{C}} = \bigcup \begin{cases} \mathbb{C}^{\text{str}} \backslash LP(\mathbb{C}^{\text{str}}) & \text{lonely base pair free structure constraint} \\ P^{\mathbb{C}^{\text{str}}_B} \cap P^{\text{sol}} & \text{validated fuzzy structure in solution} \\ LP(\mathbb{C}^{\text{str}}) \cap P^{\text{sol}} & \text{validated lonely base pairs in solution} \\ P^{SI}(P^{\text{sol}}|\mathbb{C}^{\text{seq}}, \mathbb{C}^{\text{str}}) & \text{sequence constraint induced structure} \end{cases} \tag{2.5}$$

Incompletely produced lonely base pairs, that are failed to be produced in a solution sequence are dealt in a special way. They are allowed to not specifically be considered within the evaluation of the structure prediction. Instead they need to have their requested complementary bases at the respective sequence positions within the solution sequence, such that they are not penalized by an additional term within the MFE distance calculation, as in Definition 47.

Definition 47 (Structural Distance MFE Modus)**:** Let $d_{\text{str}}(P^{\text{sol}}, P^{\mathbb{C}})$ be the structural distance between the structure of a solution sequence P^{sol} and a constraint structure $P^{\mathbb{C}}$, such that the structural distance is a symmetric difference (Equation 2.6) between the two resulting sets of base pairs of the given structures. The distance is normalized by the length of the underlaying constraint system n and extended by the penalty term of unsatisfied lonely base pairs.

$$d_{\text{str}}(P^{\text{sol}}, P^{\mathbb{C}}) = \frac{|(P^{\text{sol}} \backslash P^{\mathbb{C}}) \cup (P^{\mathbb{C}} \backslash P^{\text{sol}})| + \frac{|LP(\mathbb{C}^{\text{str}}) \backslash P^{\text{sol}}|}{2}}{n} \tag{2.6}$$

Dot Plot (DP) - Modus

The overall structural distance in the DP modus gauges the question, how well all structural constraints are met by a solution sequence within the its base pair probability matrix/matrices. The base pair probability matrices of a sequence $\mathcal{S}$ are computed by *RNAfold* and depend on the initially made structure constraint features. They consist either only of the unconstrained structure ensemble, or comprises both, the unconstrained and the limited structure ensemble, that was constrained with the secondary structure folding constraint $\mathbb{C}^{\text{str}}_{\text{Fold}}$.

Definition 48 (Solution Matrices)**:** For the structure ensemble label λ_U, let $M_{\lambda_U} = M_{\mathcal{P}_{\mathcal{S}}}$ be the base pair probability matrix of the unconstrained structure ensemble $\mathcal{P}_{\mathcal{S}}$ of a sequence $\mathcal{S}$. For the structure ensemble label λ_C, let $M_{\lambda_C} = M_{\mathcal{P}_{\mathcal{S}|w^P}}$ be the base pair probability matrix of the limited structure ensemble $\mathcal{P}_{\mathcal{S}|w^P}$, that is constrained by the folding constraint w^P of the structure, that is denoted within the folding constraint feature $\mathbb{C}^{\text{str}}_{\text{Fold}}$.

For each requested constraint structure feature $(P, \lambda, \gamma) \in \Omega$, their feature deviations $\delta(P, \lambda, \gamma)$ are calculated on the basis of the underlaying solution matrices. As depicted in Definition 49, each constraint feature is evaluated according to its type.

Definition 49 (Structure Constraint Feature Deviation)**:** Given a structure constraint feature $(P, \lambda, \gamma) \in \Omega$ then let $\delta(P, \lambda, \gamma)$ be the structure distance between the objective

value γ and an actual feature structure measurement of the structure P within a base pair probability matrix M_λ. The structure measurement depends on the type of the structure constraint feature (P, λ, γ), i.e.

$$\delta(P,\lambda,\gamma) = \left| \gamma - \begin{cases} Accur(P, M_\lambda) & \text{iff } (P,\lambda,\gamma) \in \mathbb{C}^{\text{Accur}} \\ Access(P, M_\lambda) & \text{iff } (P,\lambda,\gamma) \in \mathbb{C}^{\text{Access}} \\ Fuzzy(P, M_\lambda) & \text{iff } (P,\lambda,\gamma) \in \mathbb{C}^{\text{Fuzzy}} \end{cases} \right| \tag{2.7}$$

In order to retrieve normalization within the measurement, for each constraint structure feature, its maximally possible deviation value $\theta(P, \lambda, \gamma)$ is determined as in Definition 50. They are combined within the maximum deviation value Θ_Ω (Definition 51).

Definition 50 (Maximum Deviation of a Structure Constraint Feature)**:** Let $\theta(P, \lambda, \gamma)$ be the maximum deviation of a structure constraint feature $(P, \lambda, \gamma) \in \Omega$, that can be adopted according to its objective value γ, i.e.

$$\theta(P,\lambda,\gamma) = \begin{cases} 1-\gamma & \text{iff } \gamma < 0.5 \\ \gamma & \text{else} \end{cases} \tag{2.8}$$

Definition 51 (Maximum Deviation of a Structure Constraint Feature Set)**:** Let Θ_Ω be the maximum deviation, that can be adopted by all features of the structure constraint feature set Ω, i.e.

$$\Theta_\Omega = \sum_{(P,\lambda,\gamma)\in\Omega} \theta(P,\lambda,\gamma) \tag{2.9}$$

Given all the previous measurements, within the solution matrices dependent on the set of constraint structure features, the structural distance constitutes as in Definition 52.

Definition 52 (Structure Distance DP Modus)**:** Let $d_{\text{str}}(\Omega)$ be the structural distance between the set of base pair probability matrices and a structure constraint feature set Ω, from which the maximally achievable deviation Θ_Ω normalizes the single deviations of the constraint structure features $(P, \lambda, \gamma) \in \Omega$, such that

$$d_{\text{str}}(\Omega) = \frac{1}{\Theta_\Omega} \sum_{(P,\lambda,\gamma)\in\Omega} \delta(P,\lambda,\gamma). \tag{2.10}$$

2.6 Quality Dependent Terrain Update

The generally pursued principle of the Terrain update is that virtual pheromone values promote elements within the Terrain T, which contributed to a solution. The pheromone intensity depends on the quality of a solution. To determine the quality of a solution, measured feature distances are transfered into specific feature scores as depicted in Definition 53. The overall quality of a solution constitutes according to all stated distances resulting in the score $\mathcal{Q}$, as in Definition 54.

Definition 53 (Distance Score)**:** Let $q(d)$ be a distance score, that is dependent on a distance d and a limiting distance value $\vartheta \in \mathbb{R}$, $(\vartheta > 0)$, such that

$$q(d) = \frac{1}{\max(d, \vartheta)} \tag{2.11}$$

Definition 54 (Quality Score)**:** Let $\mathcal{Q}$ be the quality score of the measured features (structure distance d_{str}, sequence distance d_{seq} and GC-value distance d_{GC}) of a sequence. In dependence of the modus, the structure distance is calculated differently, i.e. in MFE modus, the distance is calculated on the basis of the MFE structure P^{sol} of the solution sequence and the effective target structure $P^{\mathbb{C}}$, where as in the DP modus, the structure distance is based on the base pair matrices of the solution and all imposed structure constraint features Ω. Each measured feature can be weighted according to a feature specific weight $\kappa \in \mathbb{R}$, such that $\mathcal{Q}$ constitutes as

$$\mathcal{Q} = \kappa_{GC} \cdot q(d_{\text{GC}}) + \kappa_{seq} \cdot q(d_{\text{seq}}) + \begin{cases} \kappa_{str} \cdot q(d_{\text{str}}(P^{\text{sol}}, P^{\mathbb{C}})) & \text{MFE Modus} \\ \kappa_{str} \cdot q(d_{\text{str}}(\Omega)) & \text{DP Modus} \end{cases} \tag{2.12}$$

Before a sequence quality dependent update of the successfully contributing edges is conducted in the Terrain T, all edges are exposed to a simulated evaporation event (as in Definition 55) in which a uniformly applied rate of pheromone is removed from the edges.

Definition 55 (Edge Pheromone Evaporation)**:** Each edge in the Terrain graph T is object of virtual evaporation, that is dependent on the parameter $\rho \in [0, 1]$. That results in a pheromone reduction on each edge, such that

$$\forall_{e \in \mathcal{E}} : \tau(e) = (1 - \rho)\tau^{\text{old}}(e). \tag{2.13}$$

The interplay of evaporation and the events of pheromone application by the ants effectuates the convergence of the walks. That is due to a selection of edges, that, over

time, perceived and accumulated a superior amount of pheromone, such that they are predominantly selected over the other edges in the terrain.

2.6.1 Minimum Free Energy (MFE) - Modus Pheromone Update

Within the MFE modus, the updating of successfully contributing edges follows a relatively easy pattern and has a binary character according to the question: 'Did you have success in forming the requested structural component within the solution $\mathcal{S}$ or not?' the corresponding edges are marked. Only edges of bases are promoted and get a quality dependent pheromone update, if they successfully accomplished to form a base pair interaction with the respective corresponding base within the solution, as it was requested by the constraint. Also single stranded bases get promoted, if they comply with the structure constraint.
For each edge $e \in \mathcal{E}_{\text{walk}}$ that contributed to a solution sequence $\mathcal{S}$, a check is performed, that determines, whether the edge is dignified to receive a pheromone value update. According to Definition 56, the value update follows the principle as stated in Equation 2.14. A supportive indicator function Equation 2.15 allows to decide, whether an edge gets promoted or not.

Definition 56 (Pheromone Update MFE Modus)**:** The pheromone update rule is applied to all edges, that have been participating in the generation of a sequence, that is the sequence assembly trail $\mathcal{E}_{\text{walk}}$. The contribution is according to the quality score $\mathcal{Q}$.

$$\forall e_{(i.\sigma,j.\sigma)} \in \mathcal{E}_{\text{walk}} : \tau(e_{(i.\sigma,j.\sigma)}) = \tau^{\text{old}}(e_{(i.\sigma,j.\sigma)}) + m(j)\mathcal{Q} \tag{2.14}$$

$$m(j) \quad = \quad \begin{cases} 1 & \exists(j,k) \vee (k,j) \in (P^{\text{sol}} \cap P^{\mathbb{C}}) \quad \text{base pair} \\ 1 & \nexists(j,k) \wedge (k,j) \in (P^{\text{sol}} \cup P^{\mathbb{C}}) \quad \text{single stranded} \\ 0 & \text{else} \end{cases} \tag{2.15}$$

Dot Plot (DP) - Modus Pheromone Update

Within the DP modus, the binary character of structure features within an MFE structure is not given in the structure ensemble anymore, but is replaced by a continuous range in $[0, 1]$ for each structure element. The concept of highlighting successful edges is extended and is based on the relative successfulness of single structure constraint features or groups of them.
Within the current solution, the structure constraint features of accuracy and accessibility are used to promote the relative success of each single edge such that the edge is

highlighted by the average quality of the structure features, whose requested structure is affected by the node to which the edge is leading to. This principle is only applied to nodes, that have been allocated by a connection component $cc \in \mathsf{CC}$. The affected structure constraint features, which are considered for a single pheromone contribution of a connected component in relation to an emission position i, are obtained as in Definition 57.

The edge pheromone dependent update is based on all fuzzy structure features, that affect a position i, which is covered by an edge leading to a corresponding emission node. Definition 58 indicates the relation of all fuzzy constraint features that are allocated for a position i.

In both cases, however, the respectively affected edges only receive a pheromone update, if the structural distance of the underlaying structure constraint features is below a maximum deviation threshold ξ. The overall pheromone update rule in the DP modus is given in Definition 59.

Definition 57 (Position Dependent Regular Structure Constraint Features)**:** Let $\Omega_{cc,i} \in \mathbb{C}^{\text{Accur}} \cup \mathbb{C}^{\text{Access}}$ be the set of accuracy and accessibility structure constraint features that get affected by a connected component $cc \in \mathsf{CC}$ in accordance with a position i in sequence $\mathcal{S}$, i.e.

$$\Omega_{cc,i} = \{(P, \lambda, \gamma) \in \mathbb{C}^{\text{Accur}} \cup \mathbb{C}^{\text{Access}} \mid \forall v_k \in \mathcal{V}_{\mathbb{C}|cc} : (k,i) \vee (i,k) \in P\}.$$

Definition 58 (Position Dependent Fuzzy Constraint Features)**:** Let $\Omega_{\text{Fuzzy},i}$ be the set of fuzzy constraint features $\Omega_{\text{Fuzzy},i} \subseteq \mathbb{C}^{\text{Fuzzy}}$, which cover the position i of a sequence $\mathcal{S}$, i.e.

$$\Omega_{\text{Fuzzy},i} = \{(P, \lambda, \gamma) \in \mathbb{C}^{\text{Fuzzy}} \mid ((i,j) \vee (j,i) \in P)\}.$$

Definition 59 (Pheromone Update DP Modus)**:** The pheromone update is applied to each edge in sequence assembly trail $e \in \mathcal{E}_{\text{walk}}$. A minimum quality threshold ξ delimits respective position contributions of the regular and fuzzy constraint feature evaluation. Therefore, let $s : \mathcal{E} \to \mathbb{R}$ be a function, that returns the pheromone update value, that results from the respective quality of the structure features, i.e.

$$s(e_{(i.\sigma,j.\sigma')}) = \sum \begin{cases} \sum_{\substack{cc \in \mathsf{CC} \\ v_j \in \mathcal{V}_{\mathbb{C}|cc}}} \begin{cases} \mathcal{Q} \cdot (1 - d_{\text{str}}(\Omega_{cc,j})) & d_{\text{str}}(\Omega_{cc,j}) \leqslant \xi \\ 0 & \text{else} \end{cases} \\ \begin{cases} \mathcal{Q} \cdot (1 - d_{\text{str}}(\Omega_{\text{Fuzzy},j})) & d_{\text{str}}(\Omega_{\text{Fuzzy},j}) \leqslant \xi \\ 0 & \text{else} \end{cases} \end{cases} \tag{2.16}$$

The resulting pheromone update for each edge in the sequence assembly trail $e \in \mathcal{E}_{\text{walk}}$ follows

$$\forall e \in \mathcal{E}_{\text{walk}} : \tau(e) = \tau(e)^{\text{old}} + s(e). \tag{2.17}$$

2.7 Termination of the Algorithm

With the above described mechanisms working together, the ants produce better and better sequences over the course of time, observed from a global perspective. However, the program has several internal conditions on when to terminate the procedure.
If the features of a sequence comply and satisfy the imposed requests of the constraints, the program immediately halts and returns the current solution. No further computations are made hereafter.
However, not all imposed constraint situations allow to produce a perfect compliance of the constraints by the sequence, such that the program would run for ever pursuing to find a better solution, which is improved in comparison to the currently best, but not perfect one. Still, the features of a nearly perfect solution trigger the pheromone bonification of the contributing edges in the Terrain T. If this situation happens at a specific point during the execution of the algorithm, the corresponding edges of the solution get promoted in such a way, that the resulting ratio between the just promoted edge and its potential competitors is shifted to a (much) higher chance of the winning edge of being reselected in subsequent ant walks. This introduces situations, in which the Terrain T and sequence generating process are trapped in a local minimum. To circumvent this situation of local minimum traps of the Terrain T, and to provide a possibility of reinitializing it and other contributing factors, *antaRNA* keeps track of the best solution of each iteration during the design process. A basis of 50 last solutions is taken into account in order to fit a linear curve on the overall qualities, which have been gained so far. If a slope threshold is exceeded by the slope of the curve, the program triggers the reset procedure, based on the assumption, that the current terrain situation does not produce any better solutions than the current ones. It is considered more suitable to reset and give raise to subsequent tries, rather than sticking to one local minimum hoping to encounter an improvement of the rather stuck situation. If the situation is such, that the algorithm is restarting the Terrain T over and over again, because it does not provide a precise solution to the constraints, the overall termination of the program is provided by a maximally allowed reset number, after which the so far best solution of all resets is returned.
In addition, and in concurrence to a high number of allowed resets, *antaRNA* provides termination upon a maximally allowed time boundary, after which the so far best solution is returned before the program terminates.

CHAPTER 3

EXPERIMENTS

antaRNA was developed incrementally, such that in a step wise fashion more and more complex structural situations have been able to be represented, modeled and finally computed. Each complexity level induced the alteration of specific parts of the primary algorithmic setup of *antaRNA* and covers different structure complexity problem classes. On each stage of *antaRNA*, its performance was parametrized and benchmarked according to different datasets, which have been derived from their respective sources using different extracting procedures. For each dataset that was used to parametrize *antaRNA*, the underlaying dataset was split up into a training and a test set, such that both processes use different data and avoids over fitting to that extent.

3.1 Nested Structures - MFE Modus

As introduced, nested structures represent the class of lowest complexity among all classes of secondary structures among RNA. However, nested structure elements represent the basic structural 'building blocks' of RNA. Every more complex class contains elements which are member of this class. They are the foundation to higher structures. In a first step, *antaRNA* was implemented to solve the inverse folding problem to this structure class only.

3.1.1 *antaRNA* - nested MFE Setup

In the nested MFE modus setup,*antaRNA* takes exactly one single nested structure constraint, a sequence constraint in IUPAC nomenclature, and a single target GC value as input. The structure distance calculations of intermediate solution sequences within the MFE modus are based on the structure predictions of *RNAfold* from the

ViennaRNATools, such that the predicted structure of *RNAfold* represents the minimum free energy structure (MFE) of a sequence. The resulting structure constraint objective of *antaRNA* in this stage always represents an MFE structure. In this stage of the program, only one global target GC constraint value was definable. *antaRNA* in the MFE modus allows the distribution definition of uniform and Gauss distributions of GC values in sampling situations, but is not displayed at this point. Within the structural distance calculations, the use of lonely base pairs is available, as well as the potential consideration of GU base pairs. An example call and result constitute as displayed in Appendix C.1.

3.1.2 *Rfam* Dataset

The nested structure dataset was derived from the *Rfam* data bank Version 11.0. To that date, the database contained 2208 different classified RNA families. For each family, the seed alignment was obtained, from which a representing sequence and its structure was derived, if some requirements had been fulfilled, which are to explained in the following. If an entry was obtained, it was either grouped into the *Rfam* train set $Rfam_{\text{Train}}$ or the *Rfam* test set $Rfam_{\text{Test}}$. Only seed alignments have been considered that are larger than 20 entries. For a considered alignment, one entry was selected, that was the shortest entry among all entries of the alignment in their ungapped variant. The consensus structure, together with the winning entry in its gapped form are reduced by those positions, which are indicated by gap symbols in the alignment entry. Like that, both the entry and the consensus structure shrink to the size of the initially winning ungapped sequence. After the structure derivation, each position of the derived sequence is conditionally set to N. For that two conditions must hold, that are: First, if the ratio of positions within the sequence, which are involved in base pairs within the obtained structure in comparison to the total amount of sequence positions in the sequence exceeds 0.2 and second, if the amount of nucleotides within the affected alignment column that are of the same character as the indicated nucleotide at the position under investigation in the ungapped sequence does not exceed a majority threshold. To make sure that not too much sequence constraint is derived from the alignment, the majority threshold is set individually for each family, such that an amount of explicit sequence constraint positions between 20% and 30% is realized. Furthermore, the GC content of the produced sequence constraint should not exceed 15%. If the entry can be used at all, i.e. if the mentioned prerequisites are met, it is either grouped into the training dataset, if there is no other entry in the training set, whose length is in 5nt range of the currently introduced entry. Otherwise it will be grouped to the test data set. The used thresholds and boundaries within the procedure have been determined manually. The resulting training dataset comprises 20, the test data set counts 63 different constraint entities, ranging from 34 to 274 nucleotides in length. The training set is basis

Parameter	Description	Value Range
α	Edge Pheromone Weight	$\{0.2, 0.5, 1.0, 2.0, 4.0\}$
β	Edge Heuristic Weight	$\{0.2, 0.5, 1.0, 2.0, 4.0\}$
ρ	Terrain Evaporation Factor	$\{0.05, 0.1, 0.2\}$
κ_{str}	$\mathbb{C}^{\text{str}}$ Scoring Weight	$\{0.5, 1.0, 5.0\}$
κ_{GC}	$\mathbb{C}^{\text{GC}}$ Scoring Weight	$\{0.5, 1.0, 5.0\}$
κ_{seq}	$\mathbb{C}^{\text{seq}}$ Scoring Weight	$\{0.5, 1.0, 5.0\}$

Table 3.1: Used Parameter Ranges within Nested Structure Parametrization For each varied program parameter, the respective range of used variable instances is reported.

for a parametrization of *antaRNA* for the nested structure MFE modus, whereas the test dataset is used to validate and benchmark the performance of *antaRNA* in that modus.

3.1.3 Parametrization Setup

The parametrization of *antaRNA* in the nested structure MFE modus was performed on the $Rfam_{\text{Train}}$ dataset. In order to find appropriate parameters for *antaRNA* for later default usage in this modus, an array of varying parameter values was the base for a grid search, in which the best parameter set for the MFE modus was determined. As summarized in Table 3.1, a preselected range of values has been object within the performance tests in order to determine a winning set of parameters. The winning parameters gained from $Rfam_{\text{Train}}$ dataset, have been tested on subsequent benchmark runs on the $Rfam_{\text{Test}}$ dataset. Each parameters combination has been tested on the $Rfam_{\text{Train}}$ dataset. Herein, not only three different target GC constraint values $\mathbb{C}^{\text{GC}} \in \{25\%, 50\%, 75\%\}$ have been tested, but in addition each entity was tested under two different sequence constraint situations: Without specific sequence constraint on the one side, with entity specific sequence constraint on the other side. For each single setup 10 samples have been calculated. In total, the performance of each parameter configuration was tested on a resulting set of (20 constraint structures $*$ 2 sequence constraint situations $*$ 3 GC objectives $*$ 10 samples) $=$ 1200 constraint situations (CS). For each CS a time limitation was set to 600 seconds. As depicted in Equation 3.1, a winning parameter set was computed from that set.

$$PC_{win} = \underset{pc \in PC}{\text{argmin}} \frac{1}{|CS_{pc}|} \sum_{cs \in CS_{pc}} d_{\text{str}}(cs) + d_{\text{GC}}(cs) + d_{\text{seq}}(cs) + Time(cs) \tag{3.1}$$

3.1.4 Benchmark Setup

Quality Comparison Probing

The benchmark on the winning parameter configuration was performed on the $Rfam_{\text{Test}}$ dataset. For that purpose, three different $\mathbb{C}^{\text{GC}}$ situations and respectively two $\mathbb{C}^{\text{seq}}$ constraint situations have been applied to each of the 63 entities of the underlaying dataset. Again, for each constraint situation 10 samples have been computed. A maximum time of 1200 seconds was allowed to be used in the benchmark set.

As a comparative means, the calculated sequence data sets of standard calls of the tools *IncaRNAtion* and *RNAiFold* have been employed. As a primary comparison of the quality of the sequences, that have been produced by the particular programs, the structural distance d_{str} and the GC deviation d_{GC} off the made target GC constraints are taken into consideration. Supportive measurements such as the design success rate and the respectively achieved sequence diversity among the produced sequences are also respected in order to evaluate the particular quality of the employed software implementations. The diversity of the resulting sequences is measured according to the Shannon-Entropy H as described in Equation 3.2. Therefore, for each sequence unconstrained position within a batch is evaluated for its entropy.

$$H = -log_2 \sum_{i=1}^{m} p_i log_2(p_i) \tag{3.2}$$

where m is the sequences' alphabet size and p_i denotes the frequency of the according alphabet letter in all sequences at the position of interest.

Here the mean values over all sequence positions which are sequence unconstrained in combination with all designed sequence sets are taken as basis for the comparative value calculation. In order to also tell, if a program produces longer stretches of alike nucleotides, the entropy of dinucleotides was calculated in addition. In the case of mononucleic evaluation a maximum bit score of 2 and in the case of dinucleotides a maximum bit sore of 4 can be achieved, if a position within the sequence block is of maximally diverse composition.

Since *IncaRNAtion* is a preprocessor to *RNAinverse*, the results of *IncaRNAtion* have been post-processed by *RNAinverse*. The resulting sequences have been subject to evaluation. Each of the calls produced a sequence, whose features were compared towards the features of th obtained *antaRNA* sequences. The calls where as much as possible fed with the existent constraints at hand, especially the constraint of the GC target value was set specifically. The execution of *IncaRNAtion* was performed according to the displayed calls. The first one was used to perform executions, which have been subject to a sequence constraint. The latter calls have been executed for the calculation of sequence constraint free calculations of solution sequences.

After each invocation of *IncaRNAtion*, *RNAinverse* is called on each seed sequence to compute the actual outcome.

***IncaRNAtion* Calls**

```
$ IncaRNAtion -d [file_name]
              -a 1 -no_profile
              -s_gc [tgc] 100
              -c [Constraint Sequence]
```

```
$ IncaRNAtion -d [file_name]
              -a 1
              -no_profile
              -s_gc [tgc] 100
```

The calculation of the solution sequences of *RNAiFold* could not be done in a local attempt, since a proprietary dependent program was not obtainable anymore. Therefore, the used sequences in this study have been kindly provided by Dr. Ivan Dotu, who is collaborating author of the *RNAiFold* software.
The constraints used in the made *RNAiFold* executions summarize as follows:

- Since *RNAiFold* only returns precise results, if any, the allowed GC target value constraint interval within the execution was was set to an interval of the respective target GC value with a allowed variance k. The resulting and used interval $[x - k\%, x + k\%]$, where x was chosen from the values $[25, 50, 75]$ and k was set to 2%.
- A time out of one hour was applied.
- The internal option 'LNS' was used.
- Per execution one sequence, if any, was returned.
- The underlaying used *ViennaRNATools* Tools are in version 1.8.5. During its usage the '-d2' option was used.

'Sample and Filter' versus 'Direct Computation'

To exemplary show that the heuristic approach of 'Direct Computation' of *antaRNA* is legit and competitive in comparison to already existing tools, a 'Sample and Filter' pipeline was set up and the resulting average times where compared to each other.
Therefore, the winning parameter configuration was compared in its average time which was used to compute 10 respective sequences, which complied the requested constraints. The respective performance was evaluated under the usage of varying target GC values ranging from 20% GC to 80% GC with step size of 10%.

The test was exemplary performed by the application of three different *Rfam* derived constraints sets, namely the constraint IDs RF00480, RF00007 and RF00563 which can be categorized substitutively for different length categories L1, L2 and L3 with $L1 = 1 \leqslant x < 100, L2 = 100 \leqslant x < 200, L3 = 200 \leqslant x < 300$.
The exemplary 'Sample and Filter' approach was performed with three different programs: *NUPACK*, *ERD* and *RNAinverse*. In each attempt, a set maximum of 1000 allowed sampling trials was used in order to probe time categories in which it would be possible for the respective program to obtain 10 suitable sequences for each $\mathbb{C}^{GC}$ category. If the given number of 10 solutions was not achieved after the 1000 trials, a forecast time requirement was stated. If none were existent, the attempt was evaluated as fail. In comparison, *antaRNA* was executed 100 times for each constraint setup. The average time was computed and compared to the respective values of the competitive pipeline results.
The programs *NUPACK*, *ERD* and *RNAinverse* have been executed in a wrapper script, which called the program as shown exemplary hereafter. The programs were run in standard settings, such that they could perceive the respectively desired structure request. The configuration of *RNAinverse* allows to set it up as a 'listener program', such that subsequent requests can be piped to it.

```
1 $ RNAinverse
2 RNAinverse Prompt $ [Constraint Structure]
```

NUPACK only accepts an input file, which contains the respective constraints. For each execution it was called as

```
1 $ /NUPACK/bin/design [Constraint Structure File]
```

ERD has also a straight forward execution call, which takes the constraint structure on the command line.

```
1 $ ./erd -x [Constraint Structure]
```

3.2 Pseudoknot Structures - MFE Modus

3.2.1 *antaRNA* - Pseudoknot MFE Setup

antaRNA was extended by the possibility to accept pseudoknot structure input. Furthermore, the structural comparison of *antaRNA* was adapted. For that the structure folding capacities were extended by *pKiss*, a pseudoknot prediction software, which can fold both nested and pseudoknot structures. In addition to *pKiss*, also *IPknot* and *HotKnots* have been included into the framework, such that a user can select those methods for specific pseudoknot folding situations. *pKiss* was chosen to be the standard

program for this purpose, since it provides a regular command line output and is also tailored for the common pseudoknot classes H (hairpin)- and K (kissing-hairpin)-type. Due to the more complex structure format, some internal data structure representation had to be adapted as well, such that the previously established functionality was still functional as before. Also the respective integrity checks were improved and adapted to the new situation to be able to deal with pseudoknot structures.
The execution of *antaRNA* in order to compute sequences folding in a certain pseudoknot is still operated in the MFE modus. Example calls for the usage of pseudoknots can be found in Appendix C.2.

3.2.2 *Pseudobase++* Dataset

The pseudoknot structure dataset was derived from a *Pseudobase++* database download as of 2014/12. The 'raw' *Pseudobase++* dataset consists of 304 entries. Starting from that, several filtering steps were applied to keep only a qualitatively suitable dataset. Detected cases, where non-canonical base pairs (different from AU, GC or GU) have been identified, were removed. The remaining pseudoknot structures are classified into four complexity categories, representing simple hairpin pseudoknots (H), bulge hairpin pseudoknots (B), complex hairpin pseudoknots (cH) and kissing hairpin pseudoknots (K). Pseudoknots of higher categories were excluded from the dataset. The dataset was split into a training Pseudo++$_{\text{Train}}$ and a benchmarking Pseudo++$_{\text{Test}}$ dataset. This was done by adding entities to the pool of training data Pseudo++$_{\text{Train}}$, such that each member constraint was at least 5 nucleotides difference in length to all entries. For each constraint, the respective explicit sequence was transformed into a sequence constraint only holding 25% of explicit sequence constraint positions. The rest of the position was transformed into an ambiguous sequence constraint. The transformation process was repeated until a sequence constraint was found, that only had a GC content of maximally 15%. With the latter requirement it was made sure, that any given GC objective that was used in the parametrization and the subsequent benchmark was potentially achievable by the constraint.
The training set Pseudo++$_{\text{Train}}$ is build up from 7 H-type, 3 B-type and 6 cH-type constraint systems and comprises 16 entities in total. The test set Pseudo++$_{\text{Test}}$ is built from 209 H-type, 29 B-type, 8 cH-type and 3 K-type structure constraints resulting in a total number of 249 the instances.

Parameter	Description	Value Range
α	Edge Pheromone Weight	$\{0.1, 1.0, 2.0, 5.0\}$
β	Edge Heuristic Weight	$\{0.1, 1.0, 2.0, 5.0\}$
ρ	Terrain Evaporation Factor	$\{0.1, 0.2\}$
κ_{str}	$\mathbb{C}^{\text{str}}$ Scoring Weight	$\{0.1, 0.5, 1.0\}$
κ_{GC}	$\mathbb{C}^{\text{GC}}$ Scoring Weight	$\{0.1, 0.5, 1.0\}$
κ_{seq}	$\mathbb{C}^{\text{seq}}$ Scoring Weight	$\{0.1, 0.5, 1.0\}$

Table 3.2: Used Parameter Ranges within the Pseudoknot Structure Parametrization For each varied program parameter, the respective array of respectively used variable instanciation is reported. The winning parameters are listed in the last column.

3.2.3 Parametrization and Benchmark Setup

Parametrization

The Pseudo++$_{\text{Train}}$ dataset was used to carry out a parametrization procedure, in which a grid search on a defined parameter configuration set was performed to find suitable parameter values for the execution of the MFE mode under the usage of *pKiss*. The varied parameters were given in Table 3.2. All possible combinations of parameter instances within the specified arrays have been considered to set up the parameter configuration grid. Per configuration within the grid search, a performance evaluation based on the 16 entities of the Pseudo++$_{\text{Train}}$ dataset was executed, in which a winning configuration was identified. For each entity in the data set three different target GC values $\mathbb{C}^{\text{GC}}$ of 25%, 50%, 75% and two different sequence constraints (ambiguous and explicit) were tested. For each resulting constraint setup 10 sequences have been produced. For each parameter configuration the respective set of produced sequences has been subject to determine an overall performance value 16 structure constraints $*$ 3 GC constraints $*$ 10 samples $=$ 480 constraint situations (CS) have been used to evaluate a single parameter configuration.

3.2.4 Benchmark of *antaRNA* against *MODENA*

To benchmark the quality of *antaRNA*, the resulting sequences have been compared to the sequences produced by *MODENA*. For that purpose the Pseudo++$_{\text{Test}}$ dataset was used as foundation. Each entity was subject to three target GC values $\mathbb{C}^{\text{GC}} \in \{25\%, 50\%, 75\%\}$, for each constraint setup 10 sequences have been calculated. For the benchmark the parametrized values for the respective pseudoknot configuration have been applied, as in Table 3.2 in the last column. A maximum time of 1200 seconds was allowed to be used in the benchmark set.
Since *MODENA* uses *HotKnots* or *IPknot* as underlaying folding mechanism for pseudoknot prediction, a comparison towards the performance of two versions of *MODENA*

was possible. The solution sequences of *MODENA* have kindly been provided by Mr. Dr. A. Taneda, main contributor to *MODENA*. However, only the solutions of ambiguous sequence constraint have been calculated and therefore only this aspect can be subject to comparison between *antaRNA* and *MODENA*.

3.2.5 Benchmark of Pseudoknot Predictions within *antaRNA*

In order to evaluate which tool is well suited for the internal use as an appropriate pseudoknot sequence folding hypothesis tool within *antaRNA*, the respective wrappers for the internal usage of *pKiss*, *HotKnots* and *IPknot* have been implemented to provide their functionality internally within *antaRNA* and complement *RNAfold* for the pseudoknot MFE modus. To evaluate a comparison between the tools the Pseudo++$_{\text{Test}}$ dataset is employed as a benchmark scaffold. For each present entity, three different target GC constraint values 25%, 50% and 75%, as well as the two sequence constraint situations using ambiguous or explicit sequence constraint have been applied. From each setup 10 solutions have been commissioned for execution.
pKiss provides a folding hypothesis for H- and K- type pseudoknots and a good integration into the *antaRNA* pipeline, since there is no file I/O when using *pKiss*. *HotKnots* and *IPknot* do not provide a good interface for the usage within a pipeline: The drawback of using *IPknot* in the pipeline is that it needs to have input files, which need constantly to be written as *antaRNA* constantly produces suboptimal solutions, which need to be evaluated. This produces a massive overhead file I/O, which is not easily absorbed within the general performance. Furthermore *HotKnots* has been shown to be much slower (Sato *et al.*, 2011), even though it does not need special input files management. So the expected run times are much higher than the runtime of *antaRNA* under the usage of *pKiss*. All that was a legitimation for an early usage of *pKiss* within *antaRNA*. For integrity reasons, the underlaying performance comparison was elaborated in an afterthought for completeness in that respect. So the made comparison should highlight the categorical superiority of *pKiss* to underline its usage within *antaRNA*.

3.3 Multistable Structures - DP Modus

3.3.1 *antaRNA* - Multistable DP Setup

The aim of this stage of *antaRNA* is to be able to indicate and represent multistable RNA conformations, such that bistable RNA sequences can be object to design attempts. For that reason the already presented structure descriptors have been introduced to *antaRNA*. Based on the provided flexible structure constraint input format, at

least two main alternative structure conformations can be modeled and defined within one design RNA attempt. Alongside with the necessary structure constraint input format revisit, the way of parsing the input and representing it internally was altered, such that the whole terrain graph architecture, the sequence assembly procedure, the evaluative structural distance and the terrain rewarding system have been subject of reconsideration in the development phase. As well internal structure input correctness checks have been adjusted to the new situation. Especially, the true graph representation was realized to parse and represent the input structures appropriately. For the calculation of the base pair probability matrices, *antaRNA* was extended with the python based functions of the *ViennaRNATools*. Appropriate functions have been implemented, which allow to compute the respective ensemble base pair probability matrices of investigated sequences.

3.3.2 Multistable Dataset

The data collection of the multistable structure section is directly dependent on basic literature research. Within the multistable structure dataset, only a handful of structure elucidated entities have been collected. The collection of bistable RNA entities is rather small in comparison to the *Rfam* and the *Pseudobase++* collections of previous steps of the project. In addition, the data set has to be split up into 'intrinsically bistable' and 'induced bistable' structure constraints, resulting in even less instances of structural constraint for each type of switch. This narrows down the possibility of configuring a sufficient data split among each category, such that no suitable parametrization attempts of *antaRNA* could be conducted. Therefore the parameters of the initial MFE calculations of the nested structures are employed to that extent. Since *antaRNA* in its multistable version is still in a developmental phase, in which further details within the algorithm might be subject to deeper revision, all information gathered in the multistable program configuration are of initial indicative character. Their not optimal character is arguable.

Intrinsically Bistable RNA

To this extent the intrinsically bistable data set consists of 6 RNA entities, which have been described in their respective manuscripts. They are of different lengths and they all have two alternative structures, which are used as constraint during the execution of *antaRNA* in its initial parameters. The constructs are derived from the publications of Mann and Klemm (2011)(d33), Bogomolov *et al.* (2010)(Bog1-2) and Hörbartner and Micura (2003)(HB1-3). In Figure 3.1, the alternative secondary structures of the constructs are displayed.

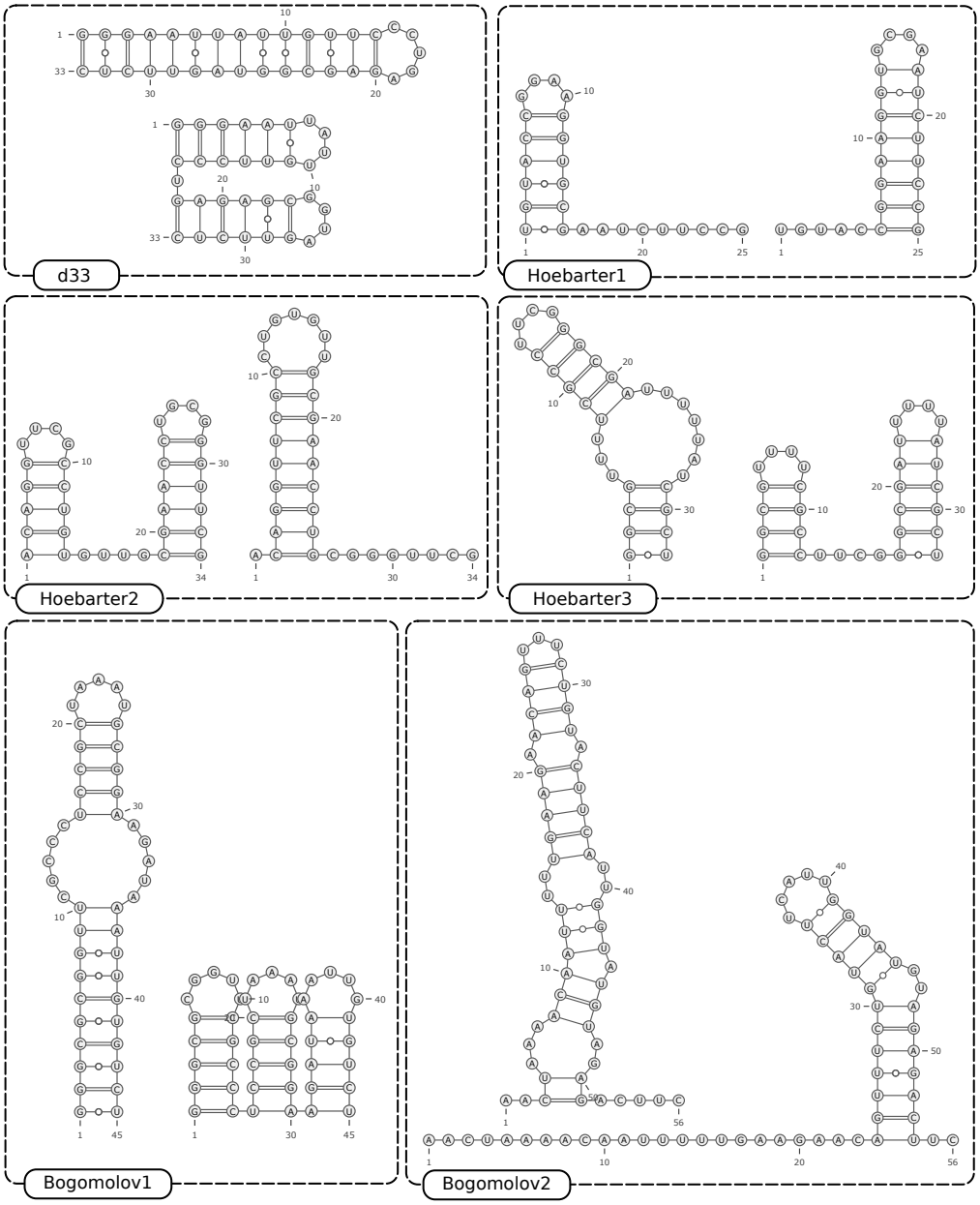

Figure 3.1: Intrinsically Bistable RNA Molecules Alternative secondary structures of the investigated intrinsically bistable RNAs. In each inlet, two alternative secondary structures are displayed, which are respectively used to request structural constraint within the DP modus of *antaRNA* by using accuracy structure features.

Ligand-induced Bistable RNA (Riboswitches)

Among the ligand-induced riboswitches and their designs, different entities have been used as a basis for the investigation of the design behavior of *antaRNA* within this class.

Among them are structure elucidated full riboswitches, such as the thiamine pyrophosphate (TPP) riboswitch thiM of E.*coli* (Rentmeister *et al.*, 2007), already computationally designed and validated riboswitches such as the theophylline riboswitch (Wachsmuth *et al.*, 2013) and a proposed own design of a riboswitch using the tetracycline aptamer (Suess *et al.*, 2003).

3.3.3 Benchmark

antaRNA

In order to evaluate the performance of *antaRNA* in the DP modus with its new setups of the terrain and other functions, each available constraint set of the dataset was subject to a design attempt of 250 entities in a benchmark setup, such that initial characteristics of the resulting sequences could be observed and described. A revisit and readjustment of internal basic principles was necessary in this stage of the program was required in order to adopt to the new situation of far more complex systems of multiple structural constraint and due to the newly used principle of probability enrichment within structure ensemble base pair probability matrices.
For each structure/sequence system, a dissection of the respective system into the newly introduced suitable smaller, but still meaningful basic structure constraint features was performed. Each setup included an objective target GC value of 50%. In some cases, a specific sequence constraint was applicable. The respectively maximally allowed times for the different systems was chosen to be the standard 600 seconds or 5 terrain resets. To this extent, no other tools, which would be able to design RNA sequences on that level, are hitherto tested and compared with *antaRNA*. Besides the standard evaluation of structural distance to the constraints, also again the GC content of the respective sequences is of interest.

RNA Kinetics

Whether an RNA sequence has the capability to fold into alternative structures can be verified by investigating the resulting dotplots of their folding. RNA kinetics allow to further investigate, if a sequence folds into certain specific structural entities at the same time, i.e. has bistable and even switching character.
In this study, the program *Treekin* was employed to compute RNA kinetics for the d33 and the Hoebarter constructs. Before the execution of *Treekin* is possible, the rate matrix for an underlaying model has to be calculated. The rate matrix in this setup is obtained by the program *barriers* (Flamm *et al.*, 2002). The input to *barriers* in turn is produced by *RNAsubopt* (Lorenz *et al.*, 2011). Given an RNA sequence, *RNAsubopt* is able to enumerate all (suboptimal) structures, which the sequence is able to fold into. In order to reduce the considered enumeration space, *RNAsubopt* has an

argument, that allows to restrict the energetic range of structural consideration during the enumeration. Within the applied setup, a range between the energy of the MFE-structure of a sequence $E(P_{MFE})$ and the value $7 - E(P_{MFE})$ was used to cutoff the energy landscape height to reduce the complexity of the computation.

CHAPTER 4

RESULTS AND EVALUATION

The resulting sequences and their features have been put into contest with other sequences from programs, which are direct competitors on the specific complexity level within the field of RNA inverse folding. The comparison is based on the sequences' features, namely $d_{\mathrm{GC}}, d_{\mathrm{seq}}$ and d_{str}. Furthermore, if applicable also the position wise entropy of sampled pools of respective sequences are subject to comparison. In addition, and only for suitable cases of multistable molecules, RNA kinetic computations have been made to indicate their behavior.

4.1 Nested Structures Analysis

4.1.1 Parametrization Result

The winning parameter configuration was derived according to a performance score, which is described in Equation 3.1. The score of a parameter set is an average value over the different optimization criteria of an underlaying batch and shows Pareto optimality. The winning parameter configuration has a total score of 0.1995. The least performing configuration scores 2.8132. Figure 4.1 displays the overall performance of all participating configurations. The course of decrease of the scores towards the minimal score can be described as 'flat'. The top 5% values are in the interval $[0.1995, 0.2405]$, which describes a 1.457% score decrease in comparison to the maximal score achieved in the ranking. The values of the winning parameter configuration are listed in Table 4.1. They were taken subsequently for all comparison experiments of *antaRNA* in the nested MFE modus.

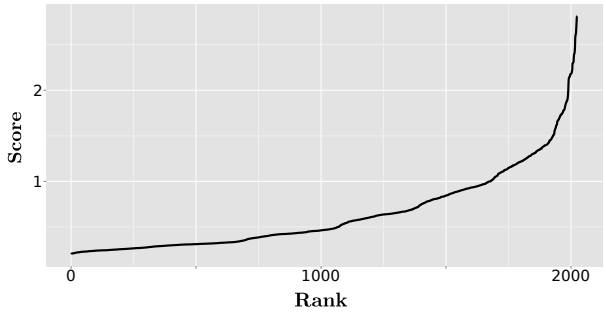

Figure 4.1: Nested Structure Parametrization Results Overall scored ranking of the executed nested structure parametrization of *antaRNA*. The winner configuration is taken for computations among nested structures.

Parameter	Description	Value
α	Edge Pheromone Weight	1.0
β	Edge Heuristic Weight	1.0
ρ	Terrain Evaporation Factor	0.2
κ_{str}	$\mathbb{C}^{\text{str}}$ Scoring Weight	0.5
κ_{GC}	$\mathbb{C}^{\text{GC}}$ Scoring Weight	5.0
κ_{seq}	$\mathbb{C}^{\text{seq}}$ Scoring Weight	1.0
	Score	0.1995

Table 4.1: Winning Parameter Configuration of the nested structure parametrization.

4.1.2 Benchmark Results

In Figure 4.2, the primary results of the benchmark of *antaRNA* on the nested structure *Rfam* dataset are summarized. In the plot, the features of structural distance d_{str} and GC aberration distance d_{GC} of the result sequences of *antaRNA* are compared to the sequence qualities of the sequences produced by the tools *IncaRNAtion* and *RNAiFold*.

Structure Compliance d_{str}

Structure Compliance of *IncaRNAtion* The capabilities of the sequences produced by *IncaRNAtion* to fold into their respective structure constraint can be described into two categories: As long as there is no specific sequence constraint applied during the execution, the structural compliance of all produced sequences is very good. The requested structural constraints are fulfilled to full extend. As soon, as the specific sequence constraint is inferred to the calculation, *IncaRNAtion* produces sequences, whose structural compliance is not met anymore entirely. Instead, besides some still compliant entities, the medians of all length categories and constraint configurations range between 1 and 4% structural deviation from the made constraints.

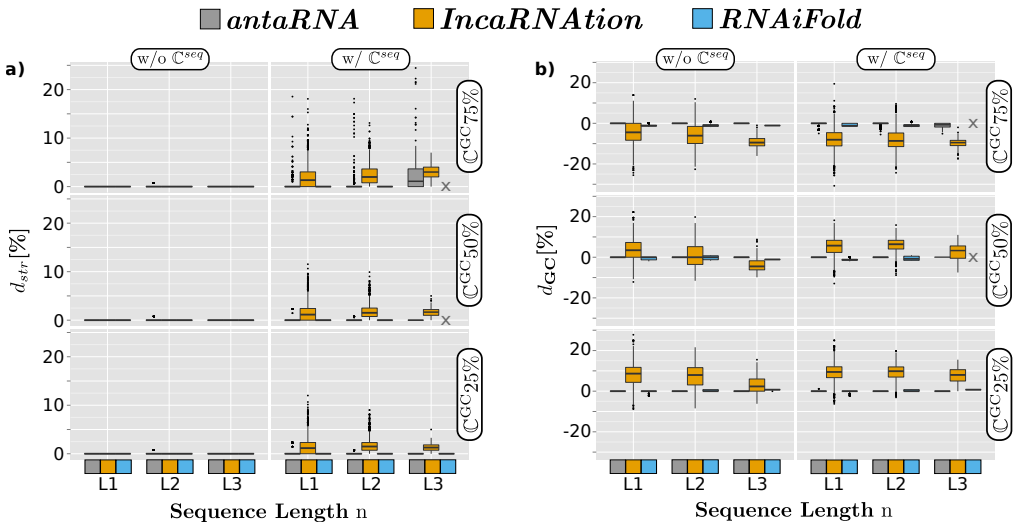

Figure 4.2: Nested Structure Primary Benchmark Results a) Structural distances d_{str} and **b)** GC aberrations d_{GC} of the resulting sequences of *antaRNA*, *IncaRNAtion* and *RNAiFold* of the Rfam nested structure benchmark experiments. Per tool the set of structure constraints $\mathbb{C}^{\text{str}}$ have been subject to calculation under the usage of specific or ambiguous sequence constraint $\mathbb{C}^{\text{seq}}$ and three different target GC constraints $\mathbb{C}^{\text{GC}}$.

Structure Compliance of *RNAiFold* In contrast to the solution sequences of *IncaRNAtion* the characteristics of the produced sequences of *RNAiFold* are more promising when inspecting their sequence features. In terms of the structural compliance of the solution sequences towards the respective constraint, *RNAiFold* appears to be very strict on keeping a high compliance rate to the made structure constraints. As the plot indicates, there is no structural deviation tolerance among the reckoned sequences. The d_{str} is 0 for all reported cases.

Structure Compliance of *antaRNA* The sequences produced by *antaRNA* comply to their particular structural constraint in a very good way. Among the ambiguous sequence constraint-free setups, *antaRNA* manages to construct sequences that show a structural distance of 0 to their constraints. Among the specific sequence constraint governed calculations, the quality still is very good, although under the extreme target GC value constraint situation for length category L3 the median of the solution rises to 2% structural deviation. In general the extreme target GC value constraint situations produce more structural deviation from the sighted constraint.

GC Content Compliance $d_{\mathbf{GC}}$

GC Compliance of *IncaRNAtion* Conspicuous characteristics are observed among the GC distance values of *IncaRNAtion*. Besides a median of 0% deviation in the case of a target GC $\mathbb{C}^{\mathrm{GC}}$ of 50% and the ambiguous application of sequence constraint within the L2 length category, all other observations indicate a decrease in GC compliance. Herein two trends are asserted: The more the respective targeted GC value dissents from an intrinsic GC base value of around 55%, the more deviant the resulting behavior becomes. To categorize the resulting deviation in some scale: Medians deviate up to 10%; upper/lower quartiles differ in extreme cases up to 13%. In addition, the influence of the applied specific sequence constraint seems to have only a small effect on the GC target deviation. At least in the extreme GC target value situations, this effect is small. In the cases of moderate target GC value constraint of 50% the shifts of the affected GC aberrations are more prominent. The influence of the constraint length category does not allow to deduce a clear trend, only that the resulting sequences of the L3 category have a relatively smaller resulting GC value than the respective result sequences of the other length categories. The GC median values of L3 distributions are always $1 - 7\%$ lower than their L1 counterparts.

GC Compliance of *RNAiFold* The GC deviations of the sequences of *RNAiFold*, if any, are remarkably low and well performing and comply all with the specified 2% margins around the targeted GC values in all situations raised by the different constraint setups. Among category L3 of the data of *RNAiFold*, however, under the influence of ambiguous sequence constraint and imposed target GC values of 50 and 75% no results were reported, indicating that during the execution of *RNAiFold* it was not possible to get into the legal solution space of the made constraints within the set time limit of one hour.

GC Compliance of *antaRNA* The target GC value compliance of *antaRNA* is very good. As a standard compliance situation the assembled sequences perfectly fit the requested constraint. While there is still no deviation from that situation among the solutions of the application of ambiguous sequence constraint, the situation changes slightly, when the specific sequence constraint is applied. In the circumstances of the extreme target GC values of 25% and 75% the overall compliance is not kept up anymore. In the respective situations, the implementation of the constraint is corrupted. Within the L3 category sufficient solutions diverge from a d_{GC} of 0, such that the median of the distribution deviates around 1%. The other deviating entities are categorized as outliers in their respective group.

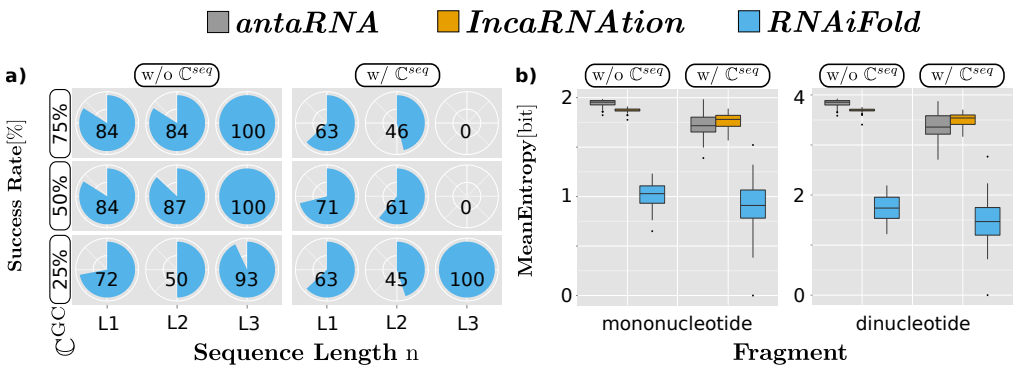

Figure 4.3: Nested Structure Secondary Benchmark Results a) Success Rates of the respective constraint categories of the tool *RNAiFold* report on how many executed program calls returned their result. **b)** Summarizing mono- and dinucleotide position specific entropy evaluation of the respectively sampled sequences within their constraint set as a means to see how diverse the respectively produced sequences are.

Success Rate

Success Rate of *RNAiFold* By the virtue of 'skipping' to return a result after a time out, *RNAiFold* received further evaluation on its success to reach into the legal solution space. As summarized by supportive Figure 4.3 a) the respective fail/success rates of *RNAiFold* among the constraint category derived data partitions. Apparently, it turns out that even though *RNAiFold* achieves high quality sequences in terms of structural and target GC value compliance, if a solution is returned, that it has diverse success rates among the different length categories and applied constraint situations. The success rate range from complete success in the design process to complete fail. Mostly the fail rate increases if the specific sequence constraint is applied. Only the L3 category within 25% applied target GC constraint is able to increase its success rate after the application of sequence constraint. Here, the success rate increases when the respective sequence constraints are operated. On average *RNAiFold* is more successful if no sequence constraint is applied (83.8% successful calculations) in comparison to the sequence constraint cases (49.9% successful calculations).

Entropy Examination

Within the done entropy comparison of the position wise and the tuple wise evaluation of batch positions, a high entropy value indicates a high diversity among the present nucleotides or dinucleotides within the respective column of the investigated batch.

Entropy of *IncaRNAtion* The results of entropy indicate, that sequences produced by *IncaRNAtion* show a high entropy bit score within the mononucleic diversity evaluation. In the cases of ambiguous sequence constraints, the mean bit score reaches 1.87 ± 0.021 while 1.77 ± 0.071 are observed in the cases of specific sequence constraints. Also in the question of consecutive nucleotides diversity, the sequences emitted by *IncaRNAtion* have a high degree of diversity among the respective sample batches. In the cases of ambiguous sequence constraints, the mean bit score reaches 3.69 ± 0.0469 and 3.51 ± 0.131 in the cases of specific sequence constraints.

Entropy of *RNAiFold* The entropy results reveal an internal bias of the sequence composing part of *RNAiFold*. This is manifested in both the mono- and the dinucleotide entropy measurements of the produced sequences of *RNAiFold*. In the mononucleotidic evaluation, the mean entropy of the ambiguous sequence constraint state amount 1.01 ± 0.13 whereas in the specific constraint situation the entropy even reduces to 0.9 ± 0.3. In the dinucleotidic succession evaluation, the entropy in the case of ambiguous sequence constraint scales up to 1.74 ± 0.25 while it is 1.47 ± 0.47 in the specific sequence constraint case.

Entropy of *antaRNA* The overall diversity among the sequences of *antaRNA* populate in the uppermost section of the possible bit score scopes of the respective experiments. According to the single position mononucleotidic diversity measurement, the sequences of *antaRNA* are very high in diversity in the ambiguous sequence constraint case. There the diversity reaches up to 1.95 ± 0.04, where as in the sequence constraint cases the diversity bit score still reaches 1.72 ± 0.12. Also in the nucleotide succession diversity measurement, the sequences of *antaRNA* demonstrate a very competitive performance of 3.83 ± 0.08 among the ambiguous sequence constraints and 3.38 ± 0.25 in the sequence constraint situations.

4.1.3 Appraisal of Generative Strategies

In Figure 4.4 the results of the made strategy comparison 'Sample and Filter' versus 'Direct Computation' are displayed. They are organized in three different length dependent categories, in which the effective run times for the respective experiments of every participating tool are summarized and compared to each other. Within each subplot of Figure 4.4, for each tool the respective accumulated or extrapolated time is plotted, if any results have been able to compute in the specific cases.

The longer a used constraint is, the more time is required among all tools to solve the given task. Moreover, the more a targeted GC constraint deviates from the GC range between 50% and 60%, the more time is required as well. This prominently holds for

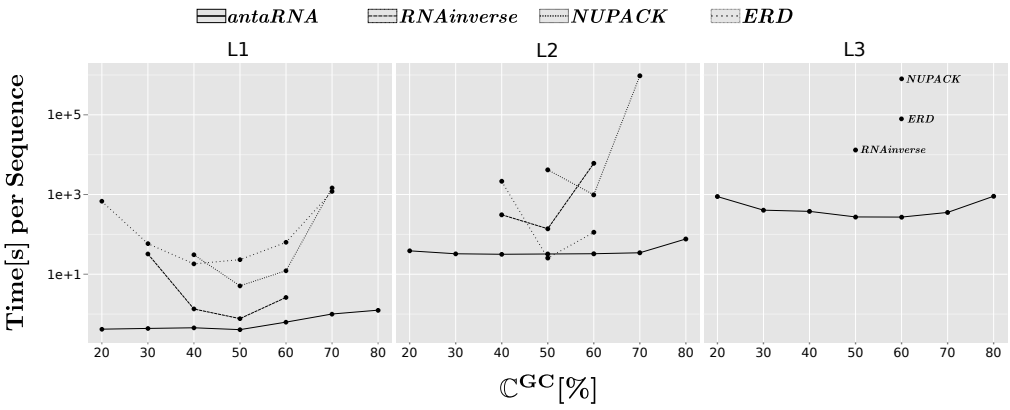

Figure 4.4: Nested Structure Strategy Comparison Results The tools *NUPACK*, *ERD* and *RNAinverse* have been applied in a 'Sample and Filter' pipeline, whereas *antaRNA* was executed in its 'Direct Computation' strategy. For each target GC value, 10 sequences should be enriched in a sampling of 1000 sequences. If 10 sequences have not been enriched within the given sampling extend, but fewer, the reported time is extrapolated based on the resulting average time for one sequence in order to be sampled successfully. If no data point is present, the respective tool in its enrichment attempt did not sample a sequence which was classified successfully within 1000 samples. To have a more comprehensive impression of the behavior among differently long solutions, the categories L1 ($n \in [1, 100]$), L2 ($n \in [101, 200]$) and L3 ($n \in [201, 300]$) differentiate the underlaying sequences.

the executions of *ERD*, *NUPACK* and *RNAinverse*. This effect is strongly moderated among the calls of *antaRNA* in comparison to the other tools.

The failure rate of the tools *ERD*, *NUPACK* and *RNAinverse* increases with an increasing length of the respectively used constraints.While *ERD* allows to obtain results in a GC range of $[20\%, 70\%]$ with a best result in 40% within the L1 panel, the ranges of succes narrow down for *NUPACK* ($[40\%, 70\%]$, best 50%) and *RNAinverse* ($[30\%, 60\%]$, best 50%). In the case of L2, this observation narrows further down for the results of all compared programs: *ERD* ($[40\%, 60\%]$, best 50%), *NUPACK* ($[50\%, 70\%]$, best 60% and *RNAinverse* ($[40\%, 60\%]$, best 50%. Among the L3 category, those programs respectively produce only one valid result batch, namely *ERD* at a GC target value of 60%, *NUPACK* at 60% and *RNAinverse* at 50%. *antaRNA* produces results for all requested cases (including the apparently difficult extreme GC values specifications), it always performs faster in its direct computation approach than the employed tools in their respective 'sample and filter' pipelines. Only *ERD* is faster in one case within category L2, but also only slightly.

4.1.4 Comparison and Assessment

Within the nested structure experiment setups, the quality of the sequences of *antaRNA* is benchmarked against the resulting sequences of *IncaRNAtion* and *RNAiFold*. During the comparative executions of *antaRNA*, the previously gained parametrized parameter set (Table 4.1) is used as standard setting within this benchmark endeavor against the mentioned tools. The protruding results of the benchmark were satisfying enough to skip a cross-validated fine-tuned investigation of the parameter sets on the bases of a re-partitioning of the initial dataset to that extend. The performed strategy comparison between the 'sample and filter' strategy and its representatives *ERD*, *RNAinverse* and *NUPACK* and the 'direct computation' strategy of *antaRNA* is taken into consideration for bolstering legit use of a tool such as *antaRNA*.
The comparison on the level of the structure complexity category of the nested MFE structures is primary performed on the level of the structural distance d_{str} and the GC aberration distance d_{GC} of the resulting sequences. At this point, a sequence distance discussion is not pursued, since the sequence distance of *antaRNA* in the nested structure case is always 0. To get a notion of diversity among the sequences and the respective nucleotides, additional measurements of the mononucleic and dinucleic position-wise entropy within the respectively produced batches of sequences is evaluated.
All sequences of *IncaRNAtion*, *RNAiFold* and as well of *antaRNA*, that have been produced without sequence constraint, despite some outliers, show a perfect structural compliance towards their structural constraints. However, after the application of the uniquely derived sequence constraint on to the same constraint setups as before, *IncaRNAtion* starts to produce solution sequences, whose structural distance deviate stronger from the objective in comparison with the solution sequences of *antaRNA* and *RNAiFold*. However, *antaRNA* and *RNAiFold* also start to produce less good results in the respective direct comparison to the sequence constraint free situation of the same constraint setup. This distinct behavior is prominently present among the extreme GC constraint situations of 25% and 75%. Among the 75% GC value constraint results, the structural quality in general drops the most among all tools. Nonetheless, *RNAiFold* inclines to produce no result if only an imperfect solution is going to be produced, such that it is located outside the legal constraint restrictions, whereas *antaRNA* returns an imperfect result sequence as a 'so-far-so-good' result. This explains the decrease in structural distance compliance among the *antaRNA* solutions within the sequence constraint batch of computations in comparison to *RNAiFold*, whose not compliant results are not present in order to compare them at all.
As in the case of structural deviation of the result sequences of the respective tools, a similar notion towards the respective results can be asserted for the observations of the GC aberrations. In the case of *antaRNA* the GC value compliance can only be compared to the results of *RNAiFold* as a serious competitor within this comparison, since the

result sequences of *IncaRNAtion* deviate so severe from their respective GC constraint, that, even if the structural constraint was successfully complied in the first place, the GC content of the design is not met in most cases. Among the results of *RNAiFold*, still, imperfect solutions are neglected, otherwise, the requested GC constraint is fulfilled among the returned solutions. As discussed for the structure compliance situation, the extreme GC objective of 75% produces result sequences of *antaRNA*, which also do not comply the respective GC constraint. In the rest of the cases with lower GC value objectivity, this is not the case.

Although *IncaRNAtion* does not comply with the given structural and GC constraints that well, it appears to produce quite 'well shuffled' sequences, that means, that subsequent nucleotides show mostly different character. The sequences of *IncaRNAtion*, albeit they fail in other objective respects, show a high bit score entropy in the particularly measured instances of mononucleotide and dinucleotide entropy comparisons. In that respect, the sequence quality of *IncaRNAtion* is even comparable with the sequences of *antaRNA*. The sequences of *antaRNA* show the highest degree of shuffled nucleotide orders among the sequence unconstrained sequence instances, whereas it has to give room to the sequences of *IncaRNAtion* in that respect in the case of sequence constrained computations. This situation holds true in both cases of mononucleic and dinucleic investigation. In terms of the sequences entropy characteristics, the sequences of *RNAiFold* seem to produce monotonic stretches of same characters to a large extend, such that monotonic repetitions occur within the design. This circumstance is expressed through a low entropy among the sequences of *RNAiFold* in all cases of perspective. In addition, the 'success rates' of *RNAiFold* are quite diverse, and spare whole design batches of difficult constraint setups.

Within the results of the strategy comparison between 'sample and filter' and 'direct computation', *antaRNA* as representative of the 'direct computation' strategy, even though it uses heuristic means, can beat the investigated tools of the 'sample and filter' strategy by time and by the circumstance, that it also produces results for the cases of extreme GC objective value ranges, which is not the case for the other investigated programs. A time comparison between *antaRNA* and *IncaRNAtion* was not pursued due to the lack of quality among the sequences of *IncaRNAtion*. Furthermore, a time comparison between *antaRNA* and *RNAiFold* was not realizable, since the sequences of *RNAiFold* have been computed and kindly provided by Dr. Ivan Dotu, Boston College, Massachusetts.

4.2 Pseudoknot Structures Analysis

4.2.1 Parametrization Results

The plot of the parametrization of the pseudoknot configuration (Figure 4.5) employing *pKiss* as an underlaying folding routine allows to select a winning parameter configuration. This process is facilitated by the fact, that the curve describes an increasing steepness among the top 100 configurations. Also among the trailing configurations, the steepness of the curve increases.

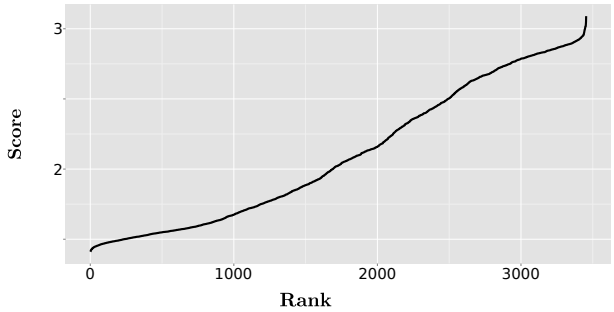

Figure 4.5: Pseudoknot Structure Parametrization Results Overall scored ranking of the executed pseudoknot structure parametrization of *antaRNA*. The winner configuration is taken for computations among pseudoknot structures.

Parameter	Description	Value
α	Edge Pheromone Weight	1.0
β	Edge Heuristic Weight	0.1
ρ	Terrain Evaporation Factor	0.2
κ_{str}	$\mathbb{C}^{\text{str}}$ Scoring Weight	0.1
κ_{GC}	$\mathbb{C}^{\text{GC}}$ Scoring Weight	1.0
κ_{seq}	$\mathbb{C}^{\text{seq}}$ Scoring Weight	0.5
	Score	1.4106

Table 4.2: Winning Parameter Configuration of the pseudoknot structure parametrization.

Between the (in comparison) steep initial and the closing regions of the curve, a broad middle part is populated, which can be more or less be described as a linear increase in score with a lowering in rank. The rank 1 configuration has a score of 1.4106 where in contrast the least performing configuration has a score of 3.089. The winning configuration is taken as standard configuration for the pseudoknot experiments; the values of the winning configuration are displayed in Table 4.2.

4.2.2 Benchmark Results

The results of the the benchmark among the pseudoknot structure base sequence prediction of *antaRNA* and *MODENA* are summarized within Figure 4.6 and comprise the structural distance and the GC aberration respectively the intrinsic GC value of the participating sequences.

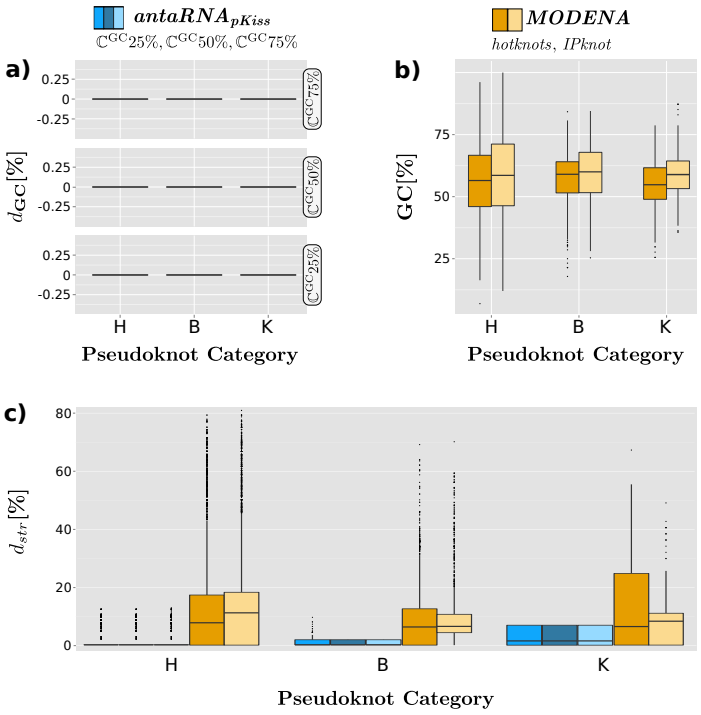

Figure 4.6: Pseudoknot Structure Benchmark Results Within the benchmark setup, *antaRNA* is compared against *MODENA*. The data is categorized into the underlaying structure complexities of pseudoknots, namely hairpin pseudoknots(K), slightly more complex bulge-pseudoknots(B) and kissing hairpin pseudoknots(K): **a)** GC distances of the resulting sequences of *antaRNA* employing *pKiss* towards their GC objective. **b)** Intrinsic GC values of the sequences produced by *MODENA* employing *HotKnots* and *IPknot*. **c)** Structure distances of the resulting sequences of *antaRNA* and *MODENA*. The results of *antaRNA* are grouped into the three differently used objective GC values of 25%, 50% and 75%. The results of *MODENA* respective comprise the sequences of the employments of *HotKnots* and *IPknot*.

Structure Compliance $d_{\mathbf{str}}$

Structural Distance of *MODENA* The sequences of *MODENA*, operating on *IPknot* and *HotKnots*, perform on average the same structural quality among all categories of the done test series. An increase in structural distance among the different

categories is not present. The median values of *HotKnots* indicate a structural distance of around $7 - 8\%$, whereas the upper quartiles range from 18% in the hairpin class to 13% in the bulge hairpin class and ca. 23% in the kissing hairpin class. The distributions of the sequences of *IPknot* show structural distance medians of 12% in the hairpin class, 8% in the bulge hairpin class and 9% in the kissing hairpin class. The quartiles, however, perform differently as the quartiles of the sequences of *HotKnots*. Although the distribution among the hairpin category performs similarly to the *HotKnots* pendent, the lower quartile of the bulge category mounts to 5% and its upper quartile to 11%. In the kissing hairpin category, the upper quartile again indicates a structural distance of 11%.

Structural Distance of *antaRNA* As indicated in Table 4.6.c, the sequences predicted by *antaRNA*, while employing *pKiss* as the underlaying structure prediction tool, comply with the respectively requested structure objective in different quality levels. With increasing pseudoknot structure complexity (from hairpins, over bulge-hairpins to kissing hairpins) the structural distance gets systematically worse. The objective structure compliance of the hairpin pseudoknot class performs best. Over 95% of the investigated sequences successfully reach a structural distance of 0. Although the median of the sequences among the bulge-hairpin class also performs with a structural distance of 0, the upper quartile amounts to 2%. In the kissing hairpin structure category, the median is populated around a structural distance of 1.5%, the distributions upper quartile reaches up to a structural distance of 7.5%.

GC Content Compliance $d_{\mathbf{GC}}$

GC Distance of *MODENA* The tool *MODENA* does not have the intrinsic option of adjusting a certain objective GC value as a constraint in order to influence the design towards that specific GC content. To highlight the differences between the underlaying tools *HotKnots* and *IPknot*, which *MODENA* is using, and furthermore to characterize the intrinsic GC content of resulting sequences of *MODENA*, the GC value of the resulting sequences is portraited. Both underlaying tools do not differ to large extent from each other in their predictive qualities. This statement can be drawn by considering the plot of Figure 4.6.b. The distributions of the measured sequences show a median range between 55 and 60% for all situations. Only the peculiarities of the quartiles allow a diversification among the sequence distributions of the tools. In both cases, the distributions get narrower, the more complex the underlaying constraint structure gets. The widest spread within the quartiles is reached for both tools among the hairpin category, $[48\%, 65\%]$ for *HotKnots* and $[48\%, 70\%]$ for *IPknot*. Among the bulge hairpin and the kissing hairpin pseudoknot categories, this spread effect reduces.

For *HotKnots* the interquartile range constitutes $[52\%, 64\%]$ and in the case of *IPknot* ranges in $[52\%, 66\%]$.

GC Distance of *antaRNA* As illustrated in Figure 4.6.a, the measured GC distances of the sequences generated by *antaRNA* in its usage of *pKiss* under the influence of three different objective GC values $\mathbb{C}^{\mathrm{GC}} \in \{25\%, 50\%, 75\%\}$ over all benchmarked structural difficulty categories, namely the hairpin, bulge-hairpin and kissing-hairpin category of the pseudoknot dataset are 0.0 in all cases of constraint.

4.2.3 Appraisal of Pseudoknot Structure Folding Prediction

The tools *pKiss*, *HotKnots* and *IPknot* potentially represent structure folding routines, which can be employed within *antaRNA*. To include a comparison between the different performances of the tools, their structural, GC and time performance are compared within the usage of *antaRNA*. A rough overview of the comparison between the tools is given in Figure 4.7.

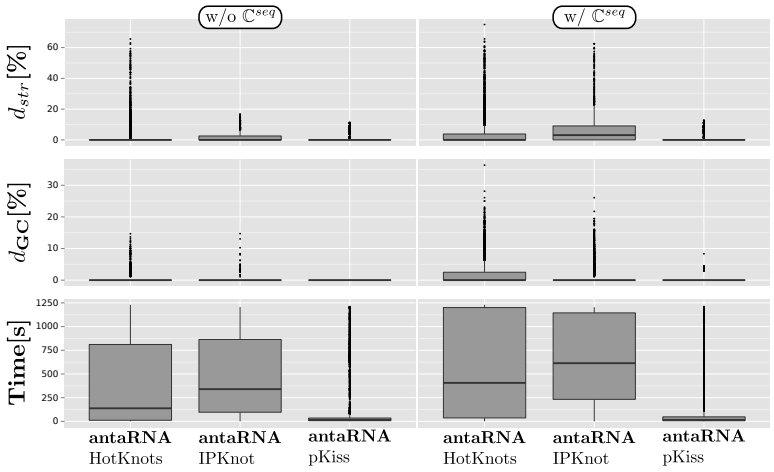

Figure 4.7: Psedoknot Folding Hypothesis Benchmark Results *pKiss*, *HotKnots* and *IPknot* based executions of *antaRNA* are based on the same set of constraints and are competitive against each other, such that a preferential tool could be selected for the default usage within *antaRNA*. The compared features are **a)** Structural Distances and **b)** GC distance of the solution sequences to their respective structural constraint. **c)** Time comparisons of the different executions.

The structural distance of *antaRNA* using *pKiss* performs best among the three choices. This holds true even for the sequence constrained circumstances. With its median and its quartiles and in addition with their 95% whiskers, both distributions indicate a structural distance of 0. When using *HotKnots* or *IPknot*, the structural deviations rise, such that, when using *IPknot* 10% structure deviation can be reached, when sequence

constraint is applied. Even though the median of *HotKnots* in the case of sequence constraint is 0, the upper quartile still hits the 5% level of deviation.

The GC distances of the investigated sequences demonstrates a perfect compliance to that constraint among the sequences of $antaRNA_{pKiss}$. Some outliers blur that result, when the sequence constraints are applied. Among the sequences of *IPknot*, this behavior is observed, though outliers are as well present in the case of no sequence constraint and are extended and enriched in the case of applied sequence constraint. When using *HotKnots* as underlying tool, the GC distances in the case of no applied GC constraint also perform well, although, the outliers are more prominent that in the other cases. Only when sequence constraints are applied, the upper quartile lifts to 3% and the upper 95%-whisker pushes to 6%.

In the category of execution times, the employment of *pKiss* results in ca. maximal 2 minutes runtime for both sequence constraint situations by the indication of the upper quartile of the respective distribution. Some outliers, however, consume the maximum allowed time of 1200 seconds. Both, *HotKnots* and *IPknot*, range with their medians in $[130, 630]$ seconds. Generally, when applying sequence constraint, both tools seem to encounter difficulties, which needs to be compensated by allocating longer computation times.

4.2.4 Comparison and Assessment

Within the pseudoknot structure setup, *antaRNA* is compared solely against *MODENA*, since *MODENA* is the only tool to this extent, with capabilities of performing RNA inverse folding on pseudoknot structures level. For the computations of *antaRNA*, the gained set of parameters of the pseudoknot structure parametrization (Figure 4.2) by using *pKiss* as prediction tool have been employed in the comparative executions against *MODENA* and is used for the comparison setup among the structure prediction tools *pKiss*, *HotKnots* and *IPknot* likewise. As in the case of the nested MFE structure parametrization, a cross-validation of the parameter on the basis of a reshuffled partition of the dataset was omitted in that occasion due to already satisfying performance within the benchmark.

The comparison between the sequence results of the tools is performed by an inspection of the structural distance d_{str} and the GC aberration distance d_{GC}. The comparison on the sequence distance level has not been considered due to the fact, that the sequences of *antaRNA* always fully complied to the respective sequence constraints. Furthermore, the provided sequences of *MODENA* have not been computed under the application of sequence constraint. No additional measurement has been applied to support the basic conclusions of the capabilities of the tools. Since both tools are heuristic approaches to the problem of RNA inverse folding, they both use underlaying folding hypotheses. In

a comparative account, the different folding algorithms were tested and benchmarked against each other as *antaRNA* structure prediction modules.

Even though *MODENA* can solve the RNA inverse problem, it is, however not in the position to pose a GC value constraint objective within its design process. Therefore, it is limited to the depicted GC ranges between ca. 45 and 70% GC content in its generated sequences, no matter which underlying structure prediction algorithm is employed in its course. Opposite to that situation and in dependence on the underlaying folding hypothesis, *antaRNA* not only can incorporate GC constraint objectives into the design of its sequences, but the sequences comply with all given GC value constraints very well. An exception, however, is posed by the sequences of *HotKnots*, which somehow influences the quality of the GC contents of the sequences such that they do not fulfill the constraint to the standard extend of a GC value aberration of $d_{\mathrm{GC}} = 0$.

In comparison to the sequences of *antaRNA*, the generated sequences of *MODENA* show poor structural compliance and are therefore not competing.Even though the quality of the structural compliance is declining with an increasing structural complexity among the solution sequences of *antaRNA* employing *pKiss*, it excels *MODENA* using both, *IPknot* and *HotKnots*, with an median distance difference of $4 - 8\%$, dependent on the complexity of the respective category.

This behavior might be an effect of the fact that *IPknot* and *HotKnots* both also do not perform well when they were used within the setup of *antaRNA*. But they do perform better on the structural compliance level when employed within *antaRNA* in comparison than within *MODENA* (Compare Figures 4.6.c and 4.7 within d_{str} category). Due to the better interface on the program level, and of course of the convincing results, *pKiss* was selected as standard tool for the structure prediction in *antaRNA* and no further benchmarking setup was included into consideration in order to further fortify the statement made.

A time-wise comparison was not possible due to installation issues. However, the used *MODENA* sequences have been computed, published and kindly provided by Dr. Akito Taneda, Hirosaki Univertsity, Japan.

4.3 Multistable Structures Analysis

Sine there currently is no comparative study, which compares the results of *antaRNA* in its DP modus towards other potential competing programs yet, only the properties of the sequences produced by *antaRNA* are presented in this section. The presentation of the results is separated according to the two distinct categories, which have been considered within the dataset: Intrinsically bistable RNA entities and ligand-induced riboswitch RNA sequences.

4.3.1 Intrinsically bistable RNA Molecules

Among the intrinsically bistable RNA molecules, the regular structural distance and the GC distance is reported and evaluated. The sequence distance towards the constraints is 0 in all cases since the constructs are all computed without the application of specific sequence constraints.

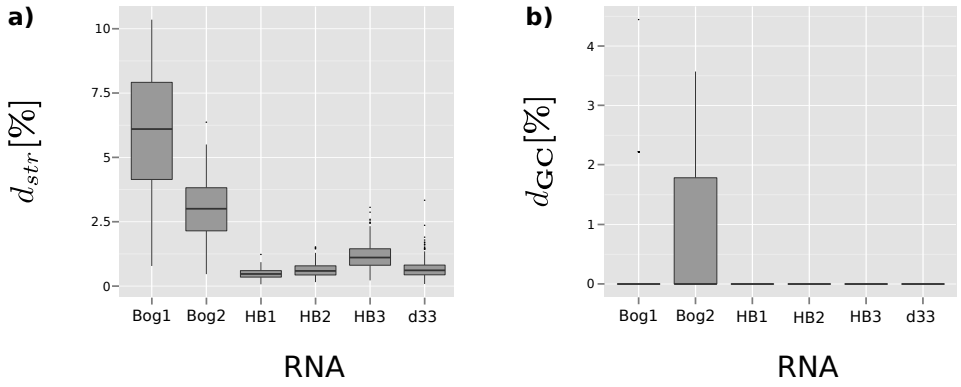

Figure 4.8: Features of the computed intrinsically bistable RNA a) Structural compliance **b)** GC aberration

Structural Compliance $d_{\mathbf{str}}$ As shown in Figure 4.8.a, the structural distances are diverse among the calculated entities. The constructs HB1-3, together with the construct d33 perform in a structural distance range of $0.5-1.5\%$ structural deviation. The medians are located at HB1 0.7%, HB2 0.75%, HB3 1% and d33 at 0.75%. In comparison to the other constructs of the category, they perform with closest compliance to their constraints. The Bog2 constructs spread from $2-3.75\%$ with a median of 6.1%. The largest spread of structural distances, as well as the largest measured values among the tested entities can be found among the Bog1 construct entities. They spread from $4-8\%$ structural distance to their requested structure constraints. The actual best and worst examples from the batch are exemplified in Figure 4.9.

GC Content Compliance $d_{\mathbf{GC}}$ Apart from the constructs of Bog2, all computed constructs perform with a perfect GC content compliance towards their objective target GC value. As shown in Figure 4.8.b, Bog2 deviates from perfect GC compliance with 1.78% GC content in the distribution's upper quartile. In the deviating cases the upper quartile represents one nucleotide, that was allocated in a wrong way, such that the exact GC content was missed.

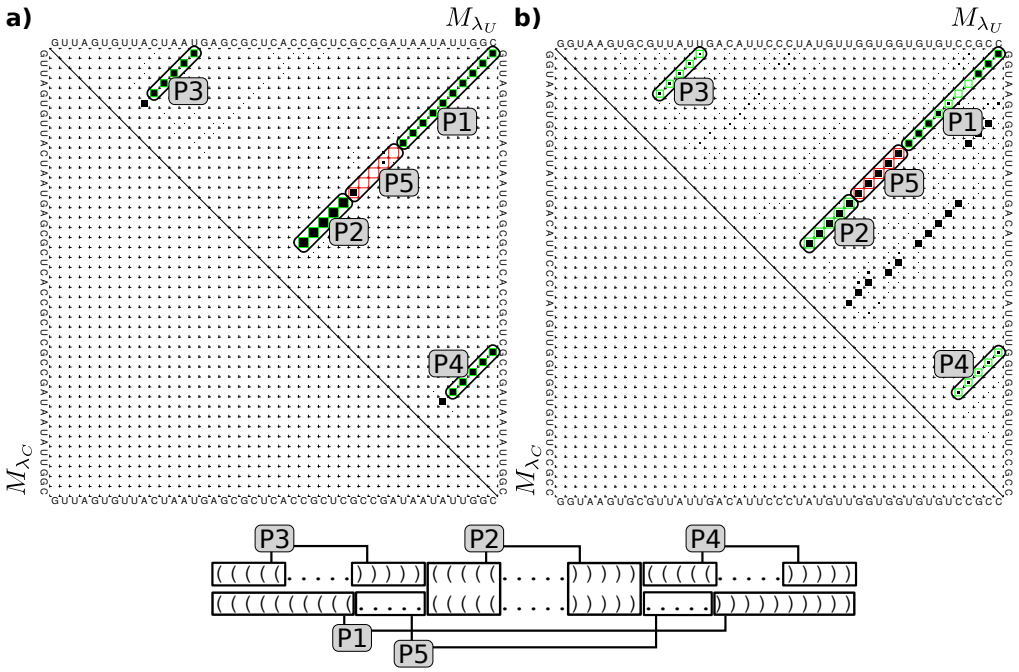

Figure 4.9: Exemplary resulting Dotplots of Bog1 Green colored hairpins within the base pair probability matrix M_{λ_U} (upper trianglar matrices) are actively requested, red ones are tried to be minimized. Both are defined through the usage of accuracy structure features. The alternative secondary structures of the constraint are depicted below the dotplots. Each hairpin is highlighted as in the dotplots. No structure features have been requested within M_{λ_C} (lower triangular matrix). **a)** Best compliance to the structure request with a structural distance of 2.80 **b)** Worst structure compliance towards the made constraints ($d_{\text{str}} = 10.02$), the central hairpin composition allows disruption of all requested alternative structure in the structure ensemble. Instead a third larger hairpin emerges within the ensemble.

RNA Kinetics Analysis The RNA kinetics analysis, as described in the experiment section, was performed with the sequences of the d33 and the HB 1-3 constructs. Each sequence was object to an individual RNA kinetics analysis, while the kinetics were computed up to 10^9 discrete time steps. Each batch of sequences was able to show that it consists to large extent of bistable RNA structure entities. Figure 4.10 indicates the RNA kinetics of the original sequences that have been templates for subsequent design attempts of *antaRNA*. Figure 4.11 indicates categorical behavior that was observed among the investigated sequences, designed by *antaRNA*.

The bistable character among the entities can be categorized based on different probability differences between prominent curves in the plots. For a vast amount of constructs, a clear course is observed: The open-chain structure probability degrades into other structures at the beginning of the simulation. The early peaks emerge within the time step range of 10^0 to 10^3 with a variation of 10^1 time steps. From that early pool

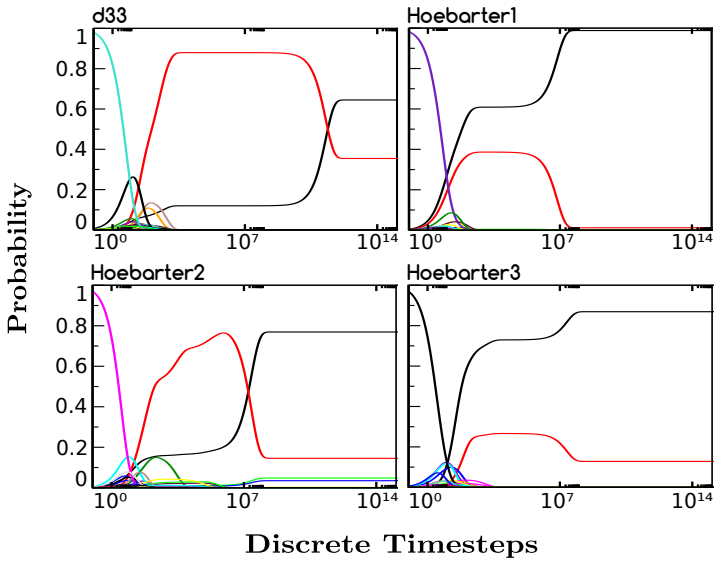

Figure 4.10: RNA kinetics analysis of the original sequences Each kinetics was computed for 10^{14} time steps. The d33 and the HB2 sequences show bistable switching behavior: after a phase of clear dominance of one structure over another, the probabilities of the structures start to reverse and a role change occurs, which endures into the equilibrium. The kinetics of the HB1 and HB3 sequences converge into equilibria that show strong dominance of one structure over the other. The equilibrium situations are amplified from previous alike dominance relations between the alternative structures, such that the roles of the dominant structures do not change. In equilibrium, HB1 seems to lose its bistable character, since one structure is super dominant over the other.

of structures, 2 to 4 more stable structures prevail in the course of the simulations. Those prevailing structures enrich to a certain extent, and mostly keep the respectively achieved probability levels over a certain tranche within the plots. After that plateau phase, in which the levels do not shift severely, a more or less pronounced probability alteration occurs after which the levels of the prevailing structures change again. This changing process takes place at very long intervals of time, as the scales in the kinetics plots are in log-scale. The outcomes of this switching change can not be categorized within one class, but allow to describe the different sequences accordingly. In total, the examined sequences show largely distinct behavior when compared to the kinetic analyses of the originally template sequences. However, the computed entities have all been requested with equally probable structure constraint features in all instances among all constructs. In the most cases, the equally probable structure requests manifest in equally probable alternative structures within the equilibrium of the kinetics.

However, even though the majority of the computed sequences can fold into alternative bistable structures in the equilibrium, they do not all share the same kinetic behavior to get into the equilibrium, such that categorization within the sets need to be introduced.

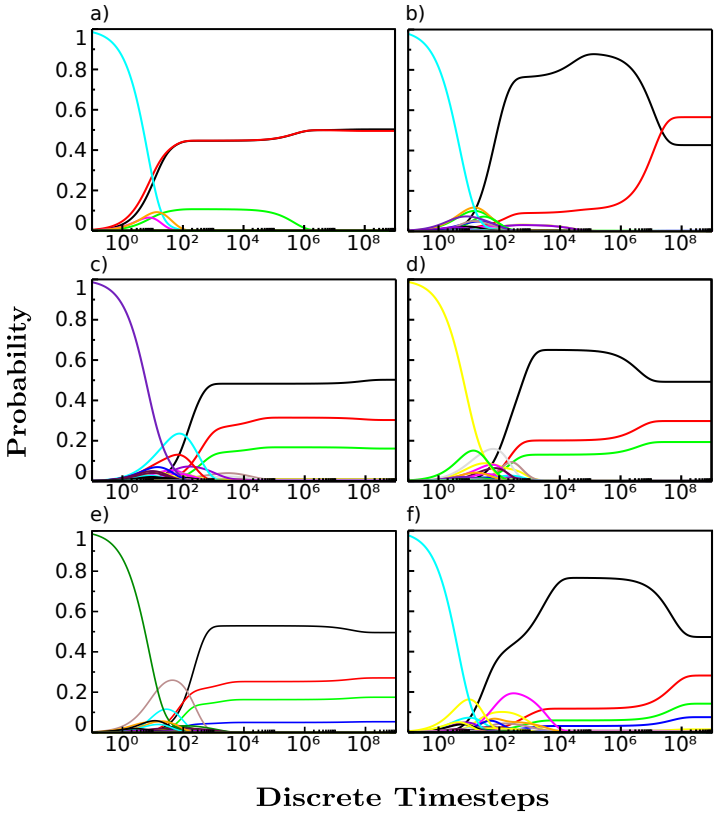

Figure 4.11: Exemplary entities of identified categories of RNA kinetics analysis Among the entities of the RNA kinetics analysis, two major categories can be identified, whose structures show convergence into structurally likewise probable equilibrium. However, in some cases three and also four major structures emerge with probabilities over 5%. In each case, homogeneous achievement of the equilibrium probabilities without differential behavior of major structures (left column) and bistable switching behavior of the entities (right column) can be observed. Each kinetic was calculated within a same time step consideration of $1e^9$ discrete time steps. The displayed categories label as **a)** Bistable homogeneous equilibrium of two major alternative structures, **b)** Bistable switch behavior among two major alternative structures before the equilibrium, **c)** Multistable homogeneous equilibrium of three major structures, **d)** Multistable alternating behavior of three structures, **e)** Multistable homogeneous equilibrium of four major structures and **f)** Multistable alternating behavior of four structures.

For each construct 250 entities have been investigated in separate RNA kinetics analysis. The categorization among each set has been done by visual inspection. The major classification criterion was if the kinetic could show a probability difference of at least 20% among the probability enrichment curves of the two most dominant structures of each plot.

The categories differentiate between entities, which show bistable switching behavior and those which presumably do only show bi- or multistability by having more than one structure enriched in the equilibrium. The solely bistable sequences show a homogeneous convergence behavior into the equilibrium, i.e. two or more major structures more or less directly engage their final probabilities without any specific probability change over time, and more importantly, without showing larger probability differences among the major structures.

The bistable switching entities differ in the course of the underlaying probability enrichments of the major structures of the kinetics. During the course of the kinetics, one structure dominates all other structures with a certain probability difference. Only over time, this dominance is altering into different equilibrium situations, which can still be called bistable. Variations of these patterns, in which three or even four major structure are involved have been subject to own categories. The resulting set separations are displayed in Figure 4.11. The sizes of the respective categories are summarized in Table 4.3.

Construct	RNA kinetics Category	Category Size
d33	bistable homogeneous	27
	bistable switch	223
HB1	bistable homogeneous	83
	bistable switch	161
	kinetics with three major structures ($Pr(P) > 0.05$)	6
	Multistable homogeneous	4
	Multistable alternating	2
HB2	bistable homogeneous	23
	bistable switch	227
HB3	bistable homogeneous	57
	bistable switch	107
	kinetics with three major structures ($Pr(P) > 0.05$)	65
	Multistable homogeneous	29
	Multistable alternating	36
	kinetics with four major structures ($Pr(P) > 0.05$)	21
	Multistable homogeneous	9
	Multistable alternating	12

Table 4.3: Categories and their sizes of the different kinetic behaviors, that have been observed among the executed RNA kinetics analysis of the constructs HB1-3 and d33. The categorization of the entities is based on visual classification and differentiates between bistable homogeneous and bistable switching behavior of the two most dominant structures of a plot. A detailing sub-classification considers, if present, three and four major structures and their behavior.

The first category, as in Figure 4.11.a, contains sequences that have two dominating structures in the equilibrium of the kinetics, which have more or less the same equilib-

rium probability, but show homogeneous enrichment of both structure probabilities to the equilibrium levels in the beginning of the kinetic computation.
In the second category, as in Figure 4.11.b, the sequences also have two dominating structures in the equilibrium that are more or less equally probable. However, the equilibrium probabilities of the major structures in this category are achieved on two distinct yet directly dependent folding behaviors. One of the structures dominates the other structure over a period of time, after which the probabilities of both structures approximate each other or even change their role within the kinetics, such that the previously underrepresented structure now is the dominating the previously overrepresented structure. In those cases, we suppose the results to be bistable. However, this category shows a high variation in the individual probability time courses.
Some kinetics indicate the presence of three (Figure 4.11.c and .d) and even four structures (Figure 4.11.e and .f) with an enriched probability over 5% in the last time steps of the kinetics. In some cases this is even the equilibrated situation. Even though those cases only comprise small fractions of the analyzed kinetics, in HB3 those cases enrich more often in comparison to the other construct batches. For each category also the immediate convergence fraction (left column in the plot) and the bistable-like fraction (right column) is present among the entities.
According to the fraction sizes within the single batches in Table 4.3, 71.8% of all investigated entities show bistable or at least bistable-like behavior. Other entities are categorized into immediate convergence into equilibrium (19%), three structures equilibrium (7.1%) and four structures equilibrium (2.1%).

4.3.2 Ligand induced riboswitch-alike RNA

The presumably ligand induced riboswitch-like sequences and their achieved features are evaluated on the level of structural compliance to the respective constraint. The issue of the GC distance is not discussed here explicitly, since all investigated values indicate 0% GC aberration from the applied objective GC values. Although the constructs have been calculated with an applied sequence constraint, their sequence distances to the used constraints is 0 in all cases. This is especially interesting in the case of the theophylline constraints, since they constitute the same structural conformations, only the applied sequence constraint is alternating among the variants. The application of an RNA kinetics evaluation could not be performed due to the length of the respective resulting sequences.

Structural Compliance $d_{\mathbf{str}}$ As displayed in Figure 4.12, the structural distances of the investigated constructs show different levels of structural compliance towards their respective structural constraint. The tetracycline construct (tc), has an interquartile range of $1.03 - 1.29\%$ with an median of 1.15%. The TPP-riboswitch (thiM) has an

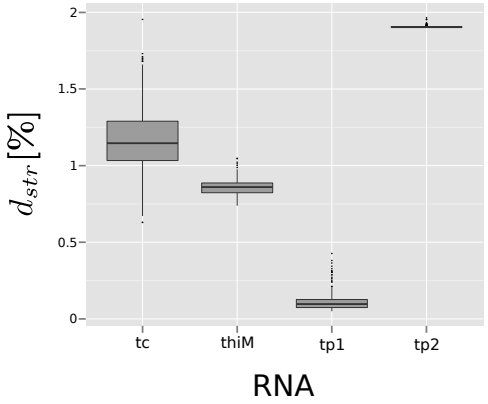

Figure 4.12: Structural compliance features of the computed riboswitch RNA Each of the investigated riboswitch like bistable constructs received different complex structure and sequence constraints during their computation, which might be the explanation for the different qualities in constraint compliance towards the structural constraint. The sequence and objective GC-content distance of the construct is 0 in both cases.

interquartile range between $0.82 - 0.89\%$ with a median of 0.86%. In the cases of the theophylline riboswitch (tp1+2), which are only different in the amount of applied sequence constraint (tp1 has less applied sequence constraint than tp2), remarkable differences in the structural compliance of the constructs are observed. Whereas tp1 faces an interquartile structural compliance of $0.07 - 0.13\%$ and a median of 0.01%, tp2 precisely shows a much worse structural compliance of 1.9% among the sequences.

To further illustrate the structural compliance of the riboswitch-like constraint setups of the computed entities, well performing entities have been selected to be highlighted (tp riboswitch in Figure 4.13, TPP riboswitch in Figure 4.14 and tc riboswitch design in Figure 4.15). The structural deviation observations from the structural constraint results not, as in the MFE structure case, from the binary presence or absence of certain base pairs, but from the slight deviation off the requested probability conditions within the dotplots, such that the deviation is more of a cumulative character. Each of the systems demonstrates a structural compliance situation.
In the case of the theophylline (tp) riboswitch, Figure 4.13, two examples are presented. During the computation, the first one experienced a less stringent, more ambiguous sequence constraint than the second one. However, although in general, the second variant of constraint achieves a less good structural compliance than the first ones, the overall structural picture within th dotplot representation is still compliant with the made constraints. Upon the emulated ligand binding event represented within M_{λ_C}, a clear structural rearrangement is observed, which depicts the requested situation. The increased structural deviation from the constraint features is located in the P2

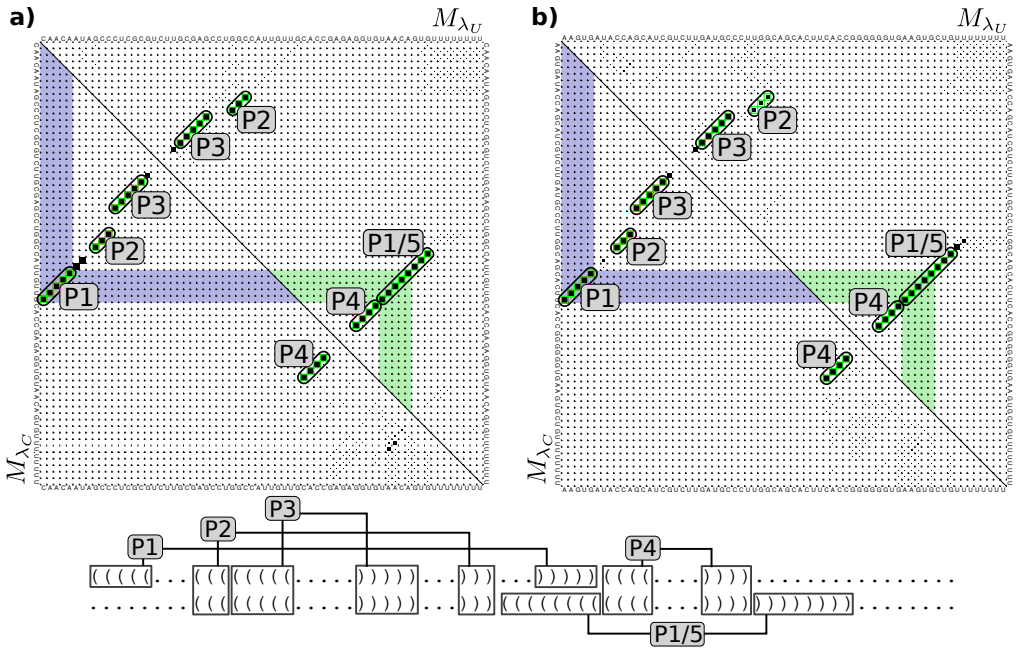

Figure 4.13: Dotplot of a computed Theophylline Riboswitch Constructs Requested alternative structures among the matrices M_{λ_U} (upper triangular matrix) and M_{λ_C} (lower triangular matrix) are indicated by blue and gray shimmed areas. Without the emulated interaction of tetracycline, which induces the formation of the hairpin P1 in M_{λ_C} case, a hybrid hairpin P1/5 can form and basally extend hairpin P4. The hairpins P1 and P2 serve evidently as minimalistic docking station for the ligand to be able to induce the interaction. The 3' stretch of P1/5 can therefore hold some recognition site, which is sequestered by P1/5. **a)** Resulting Dotplot of tp1, which was calculated using more ambiguous sequence constraint as tp2. The structural distance to the constraints is 0.06. **b)** Resulting Dotplot of tp2. Its structural distance to the constraints is 1.903. In comparison, the structural quality loss results from a less probable hairpin P2.

hairpin. In comparison to the imposed request, the resulting base pair probabilities are underrepresented.

In the case of the thiamine pyrophosphate (TPP) riboswitch, Figure 4.14, the structural indications in the dotplots show that the overall compliance towards the made structure feature constraints is presumably high, although the compliance the structure towards the made constraints is not 100%. The structural rearrangement upon binding of the ligand constitutes an expected hairpin-slip, as it is observed and described in the literature (Rentmeister *et al.*, 2007). In the presented example, the slightly increased structural deviation is resulting from an insufficient enrichment of probabilities in the accuracy features of the hairpins P2 and P3 and some present base pair probabilities within the accessibility declarations in the unbound case.

As depicted in Figure 4.15, the artificial construct of a tetracycline (tc) aptamer and its designed hairpin-slip model can sequester a recognition site (red indicated stretch

of sequence), upon the application of the folding constraint emulating the anticipated ligand interaction to the aptamer region. The sequestered region is located 5' to the aptamer, such that, based on the underlaying model of the base pair probability matrices and from a computational point of view, the hairpin-slip system works in upstream of the aptamer.
The riboswitch is extended by constrained sequence to both sides, in order to design the riboswitch within a larger context. Given that difficulty, the achieved sequence, according to the dotplots, is able to produce the exact requested hairpins, which are required to sequester a specific stretch of sequence within the M_{λ_C} m the expected behavior. The structural deviation from the constraints within the presented example only arises from deviation among the accessibility declarations. The requested hairpins seem to be correctly optimized. Nevertheless, as indicated in Figure 4.15, the used constraints do not disallow a lot of substructures to emerge without penalty, such that the used constraint set might be too narrowly selected in order to provide for an adequate sequence design.

4.3.3 Assessment

In the section about the multistable RNA entities, some literature described structures (and their sequences) of bistable RNA entities have been transformed into suitable structure feature constraints of *antaRNA* that depict the alternative structures of the RNA. They have been executed in DP modus of *antaRNA*. The resulting computations have been presented in two categories: Intrinsically bistable and ligand-induced bistable RNA entities. In both categories, the structure, sequence and GC compliance of the resulting sequences have been subject to evaluation. In addition, for suitably small RNA sequences, RNA kinetic analyses have been performed. Since for larger sequences of ligand induced folding pathways, this method in not applicable, only the small intrinsically bistable entities have been subject to that analysis.
Among all entities of the DP modus, the distances of the resulting sequences towards their sequence and GC constraint perform very good in respect to their respective constraints. They show, except for some outlying entities in the GC perspective, the targeted GC value and sequence configuration very precisely.
However, the resulting sequences have to be inspected more careful towards their actual structural compliance than the sequences that have been gained from runs within the MFE modus. The interpretation of the structure constraint compliance has to be rethought in terms of their actual meaning: In comparison to the MFE modus, in which the presence or absence of a base pair impacted in binary character to the structural distance, the structure distance in the DP modus is a gradual measurement of structure compliance. The overall structure deviation of an entity might have been caused by a part of the structure, which completely counteracts the design, even though the rest of

the structure is suited perfectly, or, on the contrary, the measured distance might result from very small deviations from all contributing structure elements.
As the results show, structural deviation in general results mostly from the whole set of all constraints, since they almost all slightly deviate from their sighted objective probability value. In sum, that comprises low valued structural deviation in the regular cases.
The analysis of the RNA kinetics computations revealed that a majority of the computed sequences showed bistable character. Among them are bistable homogeneous entities as well as presumable bistable switches. Among the ligand-induced entities, this functional examination was not possible. If they truly have switching and truly regulative character is another question, which needs to be addressed in biological experiments, since especially the actual switching behavior of the those entities is hard to predict. However, the sequences demonstrated, that the compliance of their respectively posed computational constraints was fulfilled to satisfactory extend.
Nevertheless, a comparison towards other tools has not been subject within this stage of development yet, since the aforementioned components have not been fully integrated into a consistent workflow so far. However, there exist other tools, which produced bistable RNA molecules, e.g. *RNAdesign* (Flamm *et al.*, 2001; Hoener zu Siederdissen *et al.*, 2013). Also the biophysics based models of Espah Borujeni *et al.* (2016) have to be taken into account within such a comparison.

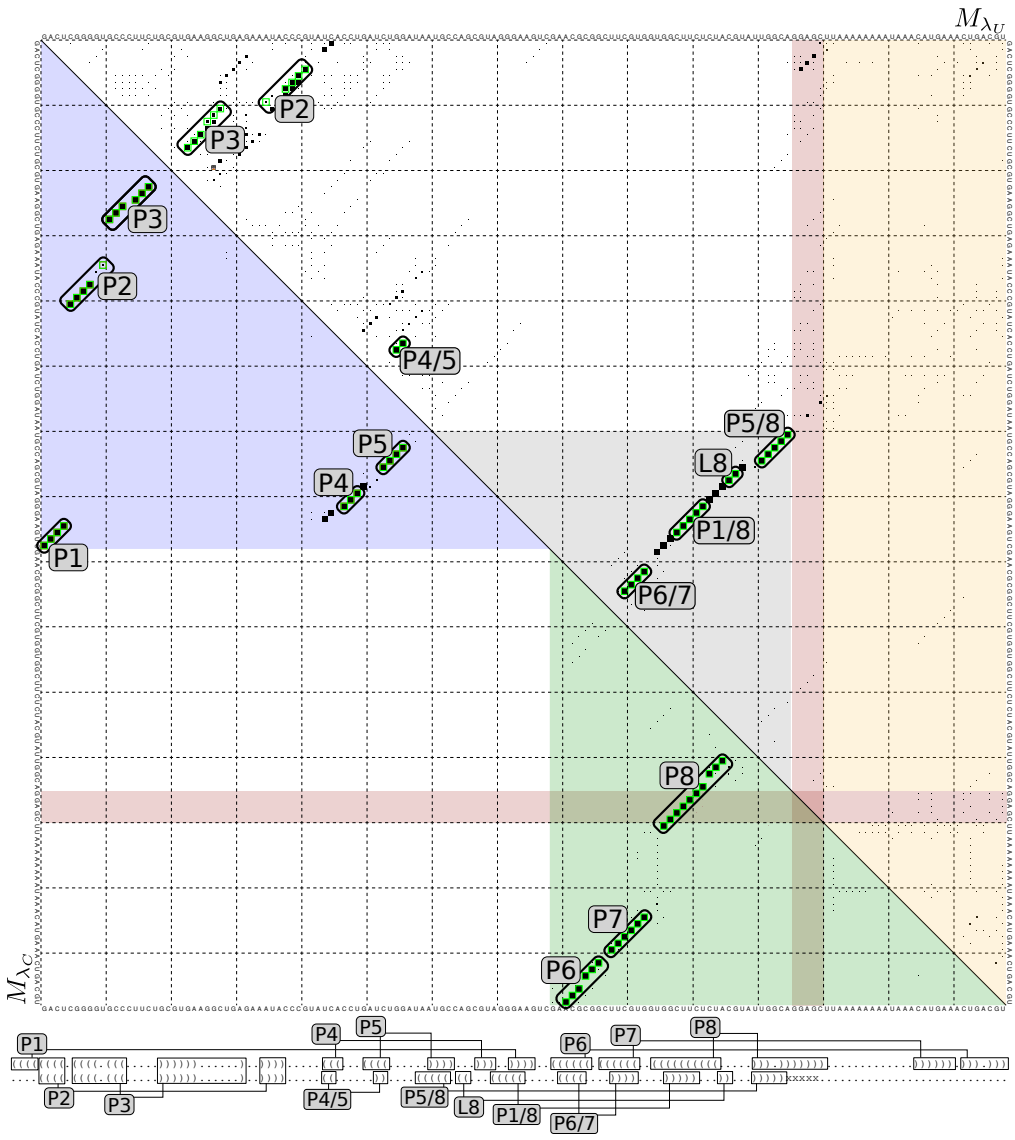

Figure 4.14: Dotplot of a computed Thiamine Pyrophosphate (TPP) Riboswitch Constructs The resulting matrices M_{λ_U} (upper triangular matrix) and M_{λ_C} (lower triangular matrix) indicate, that TPP can be bound to a 'listening' interaction domain, that consists of the hairpins P2, P3 and P4/5 within M_{λ_U} . While the constraint of the ligand-bound conformation of the aptamer is not applied (blue region in M_{λ_C}), the anti-Sequestor hairpin (gray region in M_{λ_U}) is able to encounter structure, which consists of the hybrid hairpins P6/7, P1/8, L and P5/8, such that a recognition site is left single stranded (red stripes in both solution base pair probability matrices). As soon as the ligand interacts with the aptamer stretch, the hybrid hairpin is dissolved, since the ligand interaction 'consumes' the constituting hairpins P1, P4 and P5. The induced restructuring allows the formation of the Sequestor hairpin (green region within M_{λ_C}), which consists of the hairpins P6, P7 and P8. Through the formation of the Sequestor hairpin, the recognition site then is integrated into this hairpin. Potential interaction with external factors is disrupted in that case. The yellow stripe highlights additional constraints that should prevent the introduction of sequence interaction in that area with the rest of the sequence.

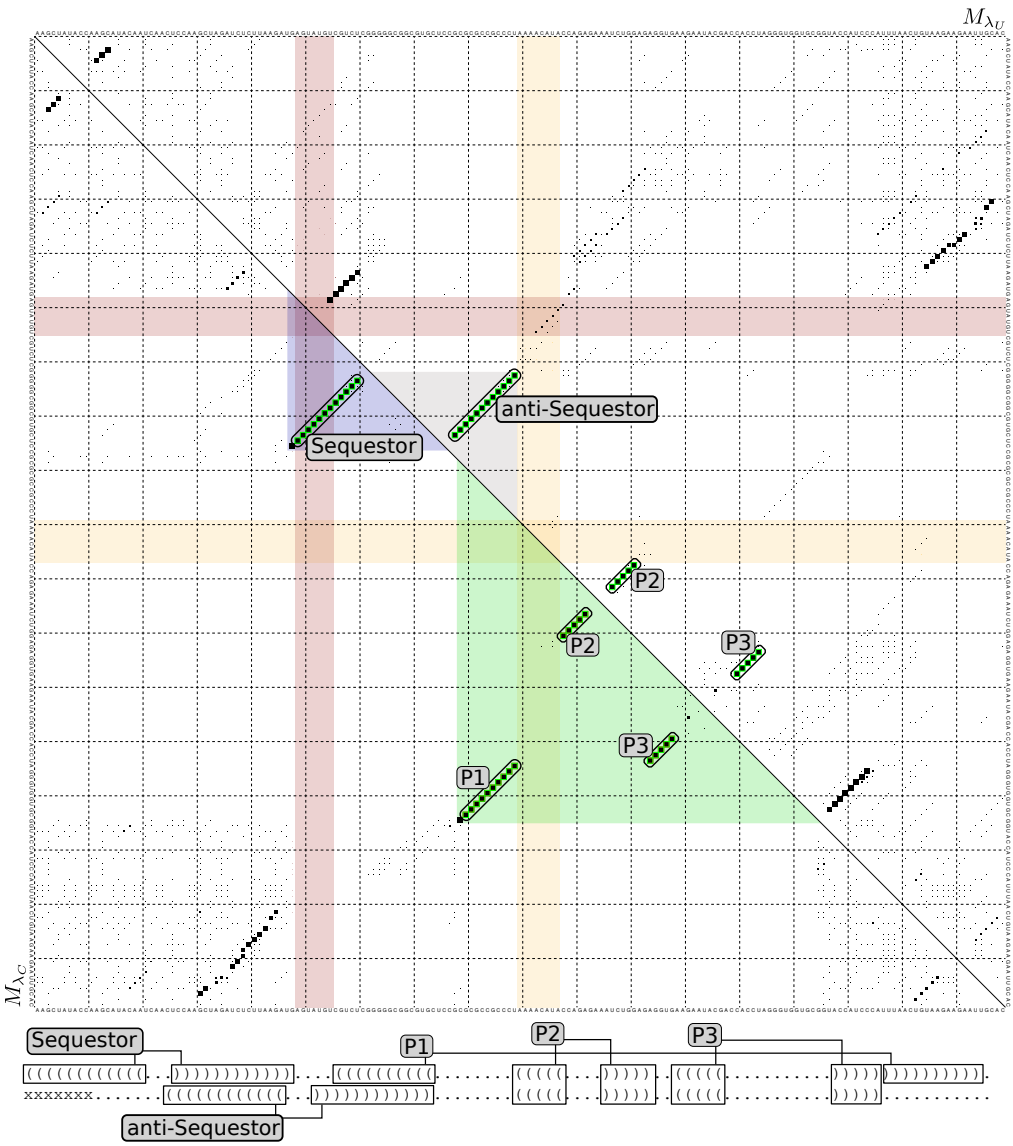

Figure 4.15: Dotplot of a computed Tetracycline (tc) Riboswitch Construct In the conformation within matrix M_{λ_U} (upper triangular matrix), the tc aptamer was constrained in a way, such that the base pairs of the hairpins P2 and P3 were requested as a minimalistic 'ligand-listening' portion of structure of the RNA, that can interact with the ligand. If the ligand would be present, as emulated in matrix M_{λ_C} (lower triangular matrix with folding constraint P1, P2 and P3), an additional hairpin (P1) would emerge due to the interaction with the ligand. The interaction of the sequence regions of the P1 hairpin would interfere with the capability of the anti-Sequestor to form, such that, in the conformation within M_{λ_C}, the Sequestor hairpin would ultimately prevent a recognition site (sequence of red squares in intersection with the Sequestor hairpin) to be accessible by other factors. The additional structure constraints of the sequence of gray squares has supportive character in terms of keeping the loop region between P1 and P2 accessible. However, it does not directly contribute to the actual character of the switching behavior. The secondary structure is displayed without context sequence.

CHAPTER 5

CONCLUSIONS

To conclude the thesis, a brief review of the presented contents is given, such that the ensuing discussion of the obtained benchmark results of the different complexity categories can be integrated into a overall statement about *antaRNA* and its classification within the world of RNA inverse folding.

In the presented thesis, the algorithmic concept of RNA inverse folding was approached by applying the algorithmic multi agent optimization concept of ant colony optimization (ACO). The principle of ACO allowed to set up a flexible environment, in which several multiple constraint definitions to an RNA inverse folding problem have been adapted by the use of objective functions and highly specific constraint types of the problem. Stated solutions of the approach indicate very precise compliance to the imposed constraints.

After a biologically motivated introduction into the context of RNA macromolecules and their folding behavior, methods of structural elucidation and the formal description of RNA, both strands of the respective algorithmic background on the RNA folding/inverse folding and the ant colony optimization are given separately. In the algorithmic section of the thesis, *antaRNA*, the resulting implementation of an adaptation of the ant colony optimization meta heuristic to the RNA inverse problem is presented. In its course, the explanation considers different developmental stages of the project and its affiliated classes of objective structures, which are of increasingly complex character and therefore specific for the respective stage. In the data preparation and benchmark setup sections of the thesis, the derivation of the used data collection is described alongside with the respectively used execution calls of the particularly employed versions of *antaRNA*. In the result section, the obtained characteristics and resulting features of

sequences in different categories are described and compared to other inverse folding programs.

In the pursued course of development and benchmark, *antaRNA* produced suitable sequences that satisfy the setups of varying complex structure constraint classes, GC and sequence constraint. The resulting sequences show high degrees of compliance towards their respectively made constraints. In comparison with different RNA inverse folding tools, the sequences demonstrate superior quality. Additionally, and this is one of the great novelties in *antaRNA*, the newly introduced GC content constraint allows to design RNA entities, that comply with a requested objective GC value with high precision.

The superior quality of the sequences produced by *antaRNA* results from simple facts, i.e. that the sequences did comply better towards the applied structure constraints, and deviated less or not at all from the imposed GC content objectives. Furthermore, the superiority was also hidden in the fact, that the resulting sequences of *antaRNA* showed diverse nucleotide composition, such that a whole batch of results did show comparatively high entropy values for single positions but as well among consecutive arrangements of nucleotide fragments within the solution sequences. The increased nucleotide diversity among the *antaRNA* sequences might results from the initially random process of sequence assembly. Furthermore *antaRNA* does not pursue the strategy of introducing minimalistic alterations to a initially monotonic sequence, such that the inferred constraints are compiled. However, nucleotide diversity within a sequence or a batch of sequences is not stated as explicit objective.

antaRNA was presented as a flexible basis framework for the modularized adaptation and usage of different problem classes within the inverse folding problem. This was done by providing 'simple' tool substitution and subsequent extension of underlaying routines and definitions according to the differently structure complexity classes. In this context, the portation of the inverse folding problem to the level of the base pair probability matrices of a solution sequence has shown its feasibility.

As seen in the cases of the bistable RNA entities, the sequences complied with their encountered structural and GC constraints again very good, but the resulting kinetic behavior was quite diverse among the sequences. However, since the kinetic behavior of a sequence was not an immediate constraint to the ACO inverse folding problem, it was impossible to pursue optimization towards that objective. So in order to truly be able to formulate specific objective bistable behavior of a sequence, more context knowledge about that concern has to be included into the optimizing principle. However, the introduction of an RNA kinetics analysis within *antaRNA* immediately implies

several obstacles: It includes the question on how to formulate and encode kinetic behavior as objective constraint to the program. Furthermore, it rises the question of how to incorporate the knowledge of the kinetic request into the terrain, such that sequences can be compiled accordingly. Moreover, the compliance of a sequence towards its kinetic objective has to be measured and transmitted into the terrain adequately. Another obstacle, which would be introduced to *antaRNA* by the incorporation of RNA kinetics analysis as an objective feature, is the circumstance, that the computation of RNA kinetics requires a lot of time. Since each solution within the design process of *antaRNA* presumably has to be adequately evaluated for its quality, this special issue, if introduced, would result in an massive overall time consumption.

Appendices

APPENDIX A

ANTARNA PUBLICATIONS

The presented work is based on the following published manuscripts. For each, the proper reference is listed here. Additionally a description is given for each manuscript, as well as the abstract from the publication itself. The intention is to give an overview of the underlaying work and its digestion within scientific literature. Also it should highlight that the constitution of *antaRNA* was accomplished in staggered developmental and evaluative phases and was not setup in one leap.

R. Kleinkauf, M. Mann and R. Backofen
[P1] ***antaRNA*: ant colony-based RNA sequence design**
DOI:10.1093/bioinformatics/btv319
Bioinformatics, 2015 Oct 1;31(19):3114-21

Summary
The aim of the projected manuscript was to show that ant-colony optimization is in general adaptable to the RNA inverse folding problem. The demonstrated basic functionality setup, the fundamental design of an underlaying terrain and furthermore the development of an adequate bonification system for the pheromone update phase have been initial target of the implementing work. The proof of concept and the applicability of the introduced target GC value constraint have been parametrized and benchmarked on a data set of nested secondary structures of well known RNA families. The results show that the sequences produced by *antaRNA* not only satisfy particular constraints to the RNA inverse folding problem very good, but indicate superior quality towards competitor tools on the field of RNA inverse fold on the problem class of nested secondary structures.

Abstract
Motivation: RNA sequence design is studied at least as long as the classical folding problem. While for the latter the functional fold of an RNA molecule is to be found, inverse folding tries to identify RNA sequences that fold into a function-specific target structure. In combination with RNA-based biotechnology and synthetic biology, reliable

RNA sequence design becomes a crucial step to generate novel biochemical components. **Results:** In this article, the computational tool *antaRNA* is presented. It is capable of compiling RNA sequences for a given structure that comply in addition with an adjustable full range objective GC- content distribution, specific sequence constraints and additional fuzzy structure constraints. *antaRNA* applies ant colony optimization meta-heuristics and its superior performance is shown on a biological datasets.

R. Kleinkauf, T. Houwaart, R. Backofen, M. Mann
[P2] ***antaRNA* – Multi-objective inverse folding of pseudoknot RNA using ant-colony optimization**
DOI:10.1186/s12859-015-0815-6,
BMC Bioinformatics(2015) 16:389

Summary
The aim of the projected manuscript was to show and apply the flexibility of the ant-colony approach of solving the RNA inverse folding problem. For that reason particular affected functions and data structures have been altered and adapted; and new interface functions have been added to support the usage of different algorithms providing additional pseudoknot folding hypothesis. Furthermore the target GC value has been extended to allow multiple target GC definitions in non-overlapping distinct domains of the design object. The extension of new functionality was parametrized and benchmarked on a data set providing pseudoknoted structures. The results highlight the general flexible exchangeability of underlaying folding hypothesis within *antaRNA* by pointing out the respective qualities among differentiated pseudoknot classes. Furthermore, the results illustrate competitiveness of *antaRNA* among pseudoknot solving RNA inverse folding tools.

Abstract
Background: Many functional RNA molecules fold into pseudoknot structures, which are often essential for the formation of an RNA's 3D structure. Currently the design of RNA molecules, which fold into a specific structure (known as RNA inverse folding) within biotechnological applications, is lacking the feature of incorporating pseudoknot structures into the design. Hairpin-(H)- and kissing hairpin-(K)-type pseudoknots cover a wide range of biologically functional pseudoknots and can be represented on a secondary structure level.
Results: The RNA inverse folding program *antaRNA*, which takes secondary structure, target GC-content and sequence constraints as input, is extended to provide solutions for such H- and K-type pseudoknotted secondary structure constraint. We demonstrate the easy and flexible interchangeability of modules within the *antaRNA* framework by incorporating pKiss as structure prediction tool capable of predicting the mentioned pseudoknot types. The performance of the approach is demonstrated on a subset of the Pseudobase++ dataset.

APPENDIX B

ANT COLONY OPTIMIZATION APPLICATIONS

The ant colony optimization has been applied to several problems. This list is not complete, but allows to have a good overview on which problems the ACO was applied.

B.1 Bioinformatics Related Problems

- Protein Folding (Shmygelska and Hoos, 2005; Hu *et al.*, 2008; Nardelli *et al.*, 2013)
- Ligand Docking (Korb *et al.*, 2006)
- RNA folding (McMillan, 2006)
- Bayesian Networks (de Campos *et al.*, 2002)
- Partitioning/Clustering (Blum and Blesa, 2005)
- Classification and Data Mining (Martens *et al.*, 2011, 2007; Parpinelli *et al.*, 2002a,b)

B.2 Scheduling Problem

- Open-shop scheduling problem (OSP) (Blem, 2003)
- Single machine total tardiness problem (SMTTP) (Baucer *et al.*, 2000)
- Single machine total weighted tardiness problem (SMTWTP) (Merkle and Middendorf, 2000)
- Resource-constrained project scheduling problem (RCPSP) (Merkle *et al.*, 2000)
- Group-shop scheduling problem (GSP) (Blum, 2002)
- Multistage flowshop scheduling problem (MFSP) with sequence dependent setup/changeover times (Donati *et al.*, 2008)

B.3 Vehicle Routing Problem

- Vehicle routing problem (VRP) (Bullnheimer *et al.*, 1997)
- Multi-depot vehicle routing problem (MDVRP) (Yu *et al.*, 2011)
- Vehicle routing problem with time windows (VRPTW) (Gambardella *et al.*, 1999)
- Time dependent vehicle routing problem with time windows (TDVRPTW) (Donati *et al.*, 2008)

B.4 Assignment Problem

- Quadratic assignment problem (QAP) (Stützle, 1997)
- Generalized assignment problem (GAP) (R. and Serra, 2002)
- Frequency assignment problem (FAP) (Maniezzo and Carbonaro, 2000)
- Redundancy allocation problem (RAP) (Liang, 2004)

B.5 Set Cover/Partition Problem

- Set cover problem (SCP) (Leguizamon and Michalewicz, 1999)
- Partition problem (SPP) (Maniezzo and Milandri, 2002)
- Weight constrained graph tree partition problem (WCGTPP) (Cordone and Maffioli, 2001)
- Arc-weighted l-cardinality tree problem (AWlCTP) (Blum and Blesa, 2005)
- Multiple knapsack problem (MKP) (Fidanova, 2003)
- Maximum independent set problem (MIS) (Leguizamon and Michalewicz, 2001)

APPENDIX C

ANTARNA EXAMPLE CALLS

antaRNA is capable of computing the RNA inverse folding problem in different modi for different classes of structural complexity.

C.1 Nested Structure MFE Modus

In order to compute a basic sequence, which fulfills a nested strucutre constraint, *antaRNA* can be used in MFE modus. In the following, an example call is illustrated:

```
1 $ python antaRNA.py -tGC 0.5
2                     -noGU
3                     -noLBP
4                     -ov
5                     MFE
6                     -Cstr "...(((...)))..."
```

Here, a GC target value $\mathbb{C}^{\mathrm{GC}}$ of 50% (`-tGC 0.5`, line 1) is used to constrain a setup, with a nested structure constraint $\mathbb{C}^{\mathrm{str}}$ of `"...(((...)))..."` (line 6). In addition, the usage of the 'GU' base pair is disallowed (`-noGU`, line 2), as well as the usage of lonely base pairs in the sense of *antaRNA* (`-noLBP`, line 3). Furthermore, verbose output is requested to be listed (`-ov`, line 4). The indication, that *antaRNA* should be used in MFE modus, is given in line 5.
A result to this call could look like this:

```
1 >antaRNA0|Cstr:...(((...)))...|Cseq:NNNNNNNNNNNNNNN|Alpha:1.0|Beta:1.0|
2 tGC:0.5|ER:0.2|Struct_CT:0.5|GC_CT:5.0|Seq_CT:1.0|UsedProgram:RNAfold|
3 Modus:MFE|Ants:3|Resets:0/5|AntsTC:50|CC:130|IP:s|BSS:0|ds:0.0|dGC:0.0|
4 GC:46.6666666667|dseq:0.0|L:15|Time:0.0671908855438 ...
5 Rseq:UAUAGCAUUGCUCGG
6 Rstr:...(((...)))...
```

In the first output line (indicated lines 1-4), the identifier with additional verbose output is listed. The verbose output is comprised out of the used constraints, the internally used variable for the computation, internal counters for the computation and finally the features of the retured sequence. In the second and third line (indicated lines 5 and 6), the resulting sequence and the structure which is predicted for the result sequence are listed respectively.

C.2 Pseudoknot Structure MFE Modus

The execution of *antaRNA* has to be extended by various parameters in order to provide the functionality to allow for a pseudoknot specific computation of sequence. New possibilites are described subsequently.
Standard Call:

```
1 $ python antaRNA.py -tGC 0.5
2                     -p
3                     -pkPar
4                     -ov
5                     MFE
6                     -Cstr "(((.[[[.))).]]]"
```

To change the underlaying folding algorithm, for example to use *IPknot*:

```
1 $ python antaRNA.py -tGC 0.5
2                     -p
3                     -pkP "IPKnot"
4                     -ov
5                     MFE
6                     -Cstr "(((.[[[.))).]]]"
```

In both cases the example setting of Section C.1 is picked up and modified by a respective pseudoknotic structure constraint. The most important flag option to set, when operating pseudoknot structures is the `-p` flag (line 2). If not specified differently, *antaRNA* uses'*RNAfold* to predict strcutre. With the `-pkPar` flag (line 3), pseudoknot specific parameters are loaded as default. The pseudoknot parameters derive from the parametrization experiment and only cover the parametrization of *pKiss*. However, the alternative programs can be selected with the `-pkP` flag (line 3). If "HotKnots" is selected instead, a specific path to the respecive installation of *HotKnots* has to be set by `-HKPATH` (not indicated in the example). If the pseudoknot specifications have been set, also a suitable pseudoknot structure can be targeted.

C.3 Multiple Structure DP Modus

The possibility of *antaRNA* to model bistable RNA structure conformation entailed a remodeling of the so far known input format, which was used in the nested and the pseudoknot input format of the structure. The new input format allows the definition for multiple structures, which should be substructures of the overall structure, which are to be modeled within the design.
To explain the input, a small bistable 'oscillator' and a induced bistable 'riboswitch'-like structure example are presented. The examples are reduced examples of the presented designs and discussed results of the bistable RNA design. Also the transformation of the regular structures into the input structures is constituted.

C.3.1 Intrinsically Bistable RNA Molecule Design

The intrinsically bistable RNA example is a borrowed but reduced example taken from (Mann and Klemm, 2011): It consists of 2 concurrent almost equally stable alternative hairpin structures, where the first structure displays two hairpins, the second structure

should have only one large hairpin, covering the whole construct, such that the whole sequence is involved in this hairpin.

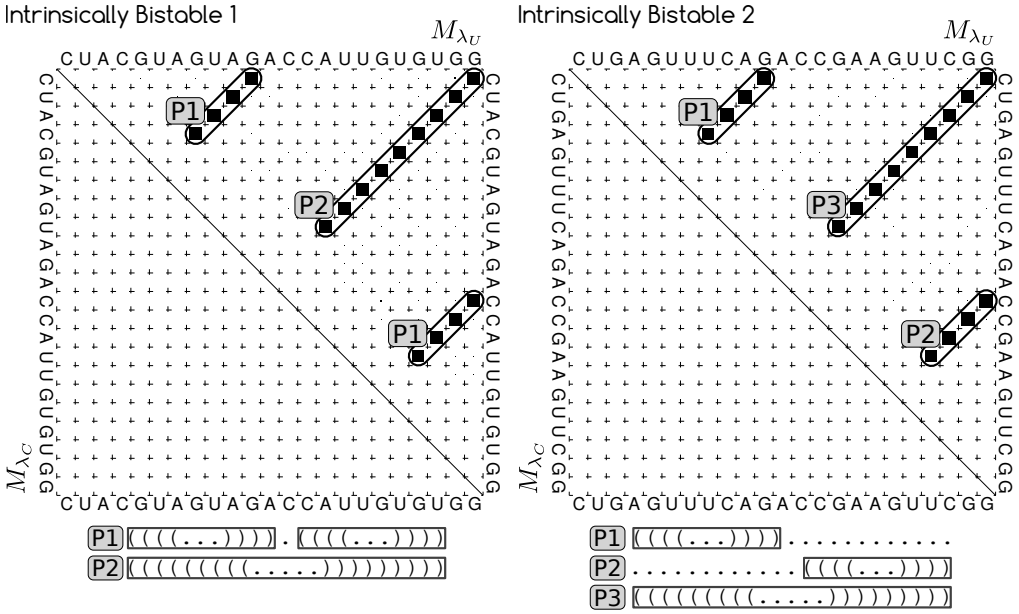

Figure C.1: Dotplot Representation of the intrinsically bistable RNA Molecule Design The folding representations of the solution sequences of both bistale RNA Molecule Designs have been gained from two separate foldings using the partition function in *RNAfold*. The requested constraints and their respective compliance are highlighted in different representations according to the used constraint situations. All accuracies of the respective constraints have been requested with a probability of 0.5 each.

```
1 >IntrinsicallyBistable|d33,KlemmMannConstruct
2 ((((...)))).((((...))))
3 ((((((((((.....))))))))))
```

The transformation into the *antaRNA* input can be modeled in two ways. The first resembles the possibility to declare two accuracy structure features, such that each represents one whole alternative structure: In the call

```
1 $ python antaRNA.py -tGC 0.5
2                     -ov
3                     DP
4                     --accuracy "((((...)))).((((...)))) UB 0.5"
5                     --accuracy "((((((((((.....)))))))))) UB 0.5"
```

line 4 and 5 contain an accuracy structure feature. Both request the their part of the alternative structure with equal probability in the unconstrained (UB) base pair probability matrix M_{λ_U}. The result is composed as in the MFE case.

```
1 >antaRNA|Cseq:NNNNNNNNNNNNNNNNNNNNNN|Alpha:1.0|Beta:1.0|tGC:0-22>0.5|
2 ER:0.2|Struct_CT:0.5|GC_CT:5.0|Seq_CT:1.0|UsedProgram:RNAfold|Modus:DP|
3 Ants:2561|Resets:5/5|AntsTC:50|CC:130|IP:s|BSS:1|ds:8.69937696509|
4 dGC:0.0|GC:47.8260869565|dseq:0
5 .0|L:23
```

```
Time:100.524312019
Rseq:CUACGUAGUAGACCAUUGUGUGG
```

A second way would be to declare three constraint fragments, in which the alternative structure variant is dissected into its smaller hairpins. Still, each one is requested with equal probability. This is possible, since for each base pair only a maximum probability of 1 is allocated:

```
$ python antaRNA.py -tGC 0.5
                    -ov
                    DP
                    --accuracy ".............((((...)))) UB 0.5"
                    --accuracy "((((...))))............. UB 0.5"
                    --accuracy "((((((((((.....)))))))))) UB 0.5"
```

The corresponding output is

```
>antaRNA|Cseq:NNNNNNNNNNNNNNNNNNNNNN|Alpha:1.0|Beta:1.0|tGC:0-22>0.5|
ER:0.2|Struct_CT:0.5|GC_CT:5.0|Seq_CT:1.0|UsedProgram:RNAfold|Modus:DP|
Ants:3809|Resets:5/5|AntsTC:50|CC:130|IP:s|BSS:0|ds:8.19734504617|dGC:0.0|
GC:52.1739130435|dseq:0 .0|L:23
Time:150.652819157
Rseq:CUGAGUUUCAGACCGAAGUUCGG
```

In both cases the respective weights of the single alternative structure declarations of the accuracy constraint features are declared a in a way, according to which the resulting requested probabilities are equally distributed between the alternative conformations. In this situation only constraint free foldings have been produced to highlight the resulting high base pair probabilities and their primal structure feature request accuracies within the structure ensemble diagrams.

C.3.2 Ligand induced riboswitch-like RNA Molecule Design

In this example, the design of a riboswitch is demonstratively exercised. Based on the idea that a ligand is binding to a specific domain, namely the 'aptamer', which is only a part of the total RNA, the rest of the construct is defined such that the ligand-induced structure shifts the regularly present hairpin equilibrium towards another but concurrent structure conformation of the unbound conformation, if the constraint of the ligand-interaction structure is used as folding constraint for the computation of the limited base pair probability matrix.

```
>Riboswitch1
(((((...)))))(((((...))))) B
.......((((((...)))))).... UB
......................xxx. UB
```

The above example structures need to be dissected into suitable constraint partitions, in which the ligand-binding aptamer is having the special role in this setup. The ligand-bound conformation (B), is therefore split into the folding constraint and an accuracy structure feature. The ligand unbound conformation needs no partition into smaller features. In addition, an accessibility constraint feature is listed, which is set to make sure, that the specified region is single stranded in the ligand unbound case and vice versa. The differential behavior of the requested accessibility is realized by the usage of an differential version of the accessibility constraint feature.
The structure constraint is transfered into an adequate input lines with their according structure features.

```
$ python antaRNA.py -tGC 0.5
                    -ov
                    DP
                    --Cstr "(((((...)))))............."
                    --accessibility ".....................xxx. UB 1.0 B 0.0"
                    --accuracy ".............(((((...))))) B 1.0"
                    --accuracy ".......((((((...)))))).... UB 1.0"
```

which outputs the following result to the prompt:

```
>antaRNA|Cstr:(((((...)))))............|Cseq:NNNNNNNNNNNNNNNNNNNNNNNNN|
Alpha:1.0|Beta:1.0|tGC:0-25>0.5|ER:0.2|Struct_CT:0.5|GC_CT:5.0|Seq_CT:1.0|
UsedProgram:RNAfold|Modus:DP|Ants:1557|Resets:5/5|AntsTC:50|CC:130|IP:s|
BSS:3|ds:3.63167808392|dGC:0.0|GC:50.0|dseq:0.0|L:26
Time:114.533756018
Rseq:UCUUCCACGAGGAGUAUCCUCGAUAC
```

indicating a total structural distance d_{str} of 3.63, the respectively made GC-contraint of 50% was perfectly met.
In Figure C.2, the graphical representation of the above example and its solution is displayed. The script processes the designed sequence and uses the applied constraint of $\mathbb{C}^{\text{str}}$ as input. Given the presence of the folding constraint, it produces one unconstrained dotplot and a constrained dotplot, with constraining structure $\mathbb{C}^{\text{str}}$ (mimicking the binding of the ligand to the RNA structure at that positions).

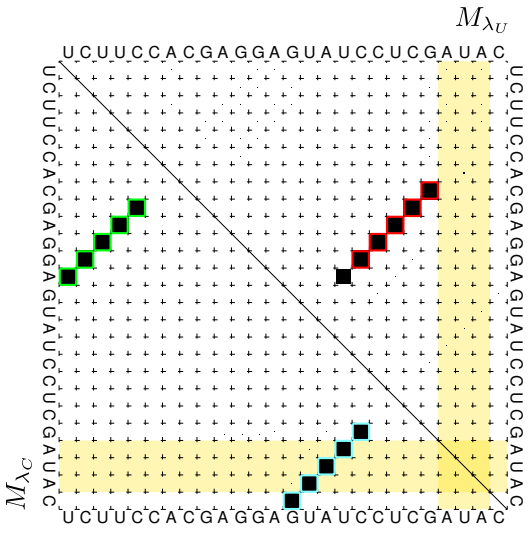

Figure C.2: Dotplot Representation of the 'Riboswitch'-like RNA Molecule Design
The folding representation of the solution sequence of the 'Riboswitch'-like RNA Molecule Design was gained from two foldings using the partition function in *RNAfold*. The M_{λ_U} base pair probability matrix was folded without structural constraint, in contrast to the the limited M_{λ_C} matrix, which was folded using the folding constraint of $\mathbb{C}^{\text{str}}$ of the input. The requested constraints and their respective compliance are highlighted in different colors: $\mathbb{C}^{\text{str}}$:green, 'accessibility': opaque yellow, 'accuracy M_{λ_U}': red and 'accuracy M_{λ_C}':blue. The accessibility, as requested, is high in the M_{λ_C} case, whereas in the M_{λ_U} case it is low (no interfering structure formation).

BIBLIOGRAPHY

Adams, P., Stahley, M., Kosek, A., Wang, J., and Strobel, S. (2004). Crystal structure of a self-splicing group i intron with both exons. *Nature*, **430**, 45–50.

Akutsu, T. (2000). Dynamic programming algorithms for RNA secondary structure prediction with pseudoknots. *Discrete Applied Mathematics*, **104**, 45–62.

Andronescu, M., Fejes, A., Hutter, F., Hoos, H., and Condon, A. (2004). A new algorithm for RNA secondary structure design. **336**(3), 607–624.

Arunan, E., Desiraju, G. R., Klein, R. A., Sadlej, J., Scheiner, S., Alkorta, I., Clary, D. C., Crabtree, R. H., Dannenberg, J. J., Hobza, P., Kjaergaard, H. G., Legon, A. C., Mennucci, B., and Nesbitt, D. J. (2011a). Defining the hydrogen bond: An account. *Pure Appl. Chem.*, **83**(8), 1619–1636.

Arunan, E., Desiraju, G. R., Klein, R. A., Sadlej, J., Scheiner, S., Alkorta, I., Clary, D. C., Crabtree, R. H., Dannenberg, J. J., Hobza, P., Kjaergaard, H. G., Legon, A. C., Mennucci, B., and Nesbitt, D. J. (2011b). Definition of the hydrogen bond. *Pure Appl. Chem.*, **83**(8), 1637–1641.

Avihoo, A., Churkin, A., and Barash, D. (2011). Rnaexinv: An extended inverse rna folding from shape and physical attributes to sequences. *BMC Bioinformatics*, **12**, 319.

Bachellerie, J.-P., CavaillÃĬ, J., and Hüttenhofer, A. (2002). The expanding snorna world. *Biochimie*, **84**(8), 775 – 790.

Banerjee, A., Jaeger, J., and Turner, D. (1993). Thermal unfolding of a group i ribozyme: the low temperature transition is primarily a disruption of tertiary structure. *Biochemistry*, **32**, 153–163.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., and Horvath, P. (2007). Crispr provides acquired resistance against viruses in prokaryotes. *Science*, **315**(5819), 1709–1712.

Barric, J. and Breaker, R. (2007). The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biol*, **8**(11), R239.

Baucer, A., Bullnheimer, B., Hartl, R. F., and Strauss, C. (2000). Minimizing total tardiness on a single machine using ant colony optimization. *Central European Journal for Operations Research and Economics*, **8**(2), 125–141.

Berens, C., Thain, A., and Schroeder, R. (2001). A tetracycline-binding rna aptamer. *Bioorg Med Chem.*, **9**(10), 2549–56.

Biebricher, C. K. and Luce, R. (1992). In vitro recombination and terminal elongation of rna by qβ replicase. *The EMBO Journal*, **11**(13), 5129–5135.

Blem, C. (2003). Beam-aco, hybridizing ant colony optimization with beam search. an application to open shop scheduling. *Technical report TR/IRIDIA*.

Blum, C. (2002). Aco applied to group shop scheduling: a case study on intensification and diversification. *Proceedings of ANTS 2002, Lecture Notes in Computer Science*, **2463**, 14–27.

Blum, C. and Blesa, M. J. (2005). New metaheuristic approaches for the edge-weighted k-cardinality tree problem. *COMPUT OPER RES*, **32**(6), 1355 – 1377.

Bogomolov, S., Mann, M., Voss, B., Podelski, A., and Backofen, R. (2010). Shape-based barrier estimation for RNAs. In *In Proceedings of German Conference on Bioinformatics GCB'10*, volume 173 of *LNI*, pages 42–51. GI. SB and MM contributed equally to this work.

Bullnheimer, B., Hartl, R. F., and Strauss, C. (1997). An improved ant system algorithm for the vehicle routing problem. *Annals of Operations Research*, **89**, 319–328.

Busch, A. and Backofen, R. (2006). INFO-RNA–a fast approach to inverse RNA folding. *Bioinformatics*, **22**(15), 1823–31.

Busch, A. and Backofen, R. (2007). INFO-RNA–a server for fast inverse RNA folding satisfying sequence constraints. **35**(Web Server issue), W310–3.

Cao, S., Fürtig, B., Schwalbe, H., and CHEN, S.-J. (2010). Folding kinetics for the conformational switch between alternative rna structures. *J Phys Chem B.*, **114**(114), 13609âĂŞ13615.

Chen, H.-L., Condon, A., and Jabbari, H. (2009). An O(n(5)) Algorithm for MFE Prediction of Kissing Hairpins and 4-Chains in Nucleic Acids. **16**(6), 803–15.

Chopra, I. and Roberts, M. (2001). Tetracycline antibiotics: mode of action, applications, molecular biology, and epidemiology of bacterial resistance. *Microbiol Mol Biol Rev.*, **65**(2), 232–60.

Cimdins, A., Klinkert, B., Aschke-Sonnenborn, U., Kaiser, F., Kortmann, J., and Narberhaus, F. (2014). Translational control of small heat shock genes in mesophilic and thermophilic cyanobacteria by rna thermometers. *RNA Biology*, **11**(5), 594–608.

Cordone, R. and Maffioli, F. (2001). Colored ant system and local search to design local telecommunication networks. *Applications of Evolutionary Computing: Proceedings of Evo Workshops*, **2037**, 60–69.

Crothers, D., Cole, P., Hilbers, C., and Schulman, R. (1974). The molecular mechanism of thermal unfolding of escherichia coli formylmethionine transfer rna. *J. Mol. Biol*, **87**, 63–88.

Daou-Chabo, R. and Condon, C. (2009). Rnase j1 endonuclease activity as a probe of rna secondary structure. *RNA*, **15**, 1417–1425.

de Campos, L. M., Fernandez-Luna, J. M., Gamez, J. A., and Puerta, J. M. (2002). Ant colony optimization for learning bayesian networks. *INT J APPROX REASON*, **31**(3), 291–311.

Deneubourg, J., Pasteels, J., and Verhaeghe, J. (1983). Probabilistic behaviour in ants : a strategy of errors? *Journal of Theoretical Biology*, **105**, 259–271.

Deneubourg, J.-L., Aron, S., Goss, S., and Pasteels, J. (1990). The self-organizing exploratory pattern of the argentine ant. *J INSECT BEHAV*, **3**(2), 159–168.

Donati, A. V., Darley, V., and Ramachandran, B. (2008). An ant-bidding algorithm for multistage flowshop scheduling problem: Optimization and phase transitions. *Book Chapter in Advances in Metaheuristics for Hard Optimization*, pages 111–138.

Dorigo, M., Maniezzo, V., and Colorni, A. (1996). Ant system: Optimization by a colony of cooperating agents. *IEEE Transactions on Systems, Man, and Cybernetics-Part B*, **26**(1), 29–41.

Dorigo, M., Di Caro, G., and Gambardella, L. (1999). Ant algorithms for discrete optimization. *Artificial Life*, **5**(2), 137–172.

Dotu, I., Mechery, V., and Clote, P. (2014). Energy parameters and novel algorithms for an extended nearest neighbor energy model of rna. *PLoS One*, **9**(2).

Ehresmann, C., Baudin, F., M., M., Romby, P., Ebel, J.-P., and Ehresmann, B. (1987). Probing the structure of rnas in solution. *NAR*, **15**, 9109–9128.

Elbashir, S., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001). Duplexes of 21-nucleotide rnas mediate rna interference in cultured mammalian cells. *Nature*, **411**, 494–498.

Ellington, A. D. and Szostak, J. W. (1990). In vitro selection of rna molecules that bind specific ligands. *Nature*, **346**(6287), 818–822.

Esmaili-Taheri, A., Ganjtabesh, M., and Mohammad-Noori, M. (2014). Evolutionary solution for the RNA design problem. *Bioinformatics*, **30**(9), 1250–8.

Espah Borujeni, A., Mishler, D. M., Wang, J., Huso, W., and Salis, H. M. (2016). Automated physics-based design of synthetic riboswitches from diverse rna aptamers. *Nucleic Acids Research*, **44**(1), 1–13.

Fidanova, S. (2003). Aco algorithm for mkp using various heuristic information. *Numerical Methods and Applications*, **2542**, 438–444.

Fire, A., Xu, S., Montgomery, M., Kostas, S., Driver, S., and Mello, C. (1998). Potent and specific genetic interference by double-stranded rna in caenorhabditis elegans. *Nature*, **391**, 806–811.

Flamm, C. and Hofacker, I. (2008). Beyond energy minimization: approaches to the kinetic folding of rna. *Chemical Monthly*, **137**, 447–457.

Flamm, C., Hofacker, I., Maurer-Stroh, S., Stadler, P., and Zehl, M. (2001). Design of multistable rna molecules. *RNA*, **7**(2), 254–265.

Flamm, C., Hofacker, I. L., Stadler, P. F., and Wolfinger, M. T. (2002). Barrier trees of degenerate landscapes. *Zeitschrift für Physikalische Chemie*, **216**(2), 155.

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

Fürtig, B., Reining, A., Sochor, F., Oberhauser, E. M., Heckel, A., and Schwalbe, H. (2001). *Characterization of Conformational Dynamics of Bistable RNA by Equilibrium and Non-Equilibrium NMR*. John Wiley & Sons, Inc.

Gambardella, L., Taillard, E., and Agazzi, G. (1999). *New Ideas in Optimization*, chapter A multiple ant colony system for vehicle routing problems with time windows, pages 63–76. McGraw Hill, London.

Gao, J., Li, L., and Reidys, C. (2010). Inverse folding of rna pseudoknot structures. *ALGORITHM MOL BIOL*, **5**(27).

Garcia-Martin, J. A., Clote, P., and Dotu, I. (2013). RNAiFold: A constraint programming algorithm for RNA inverse folding and molecular design. *J Bioinform Comput Biol*, **11**(02), 1350001. PMID: 23600819.

Gherghe, C., Shajani, Z., Wilkinson, K., Varani, G., and Weeks, K. (2008). Strong correlaction between shape chemistry and thr generalized nmr order parameter (s^2) in rna. *J Am Chem Soc*, **130**, 12244–12245.

Gibbs, J. (1873). A method of geometrical representation of the thermodynamic properties of substances by means of surfaces. *Transactions of the Connecticut Academy of Arts and Sciences*, **2**, 382–404.

Giegerich, R., Voss, B., and Rehmsmeier, M. (2004). Abstract shapes of RNA. **32**(16), 4843–51.

Goss, S., Aron, S., Deneubourg, J., and Pasteels, J. (1989). Self-organized shortcuts in the argentine ant. *Naturwissenschaften*, **76**(12), 579–581.

Gray, D. M., Gray, C. W., Yoo, B. H., and Lou, T. F. (2010). Antisense dna parameters derived from next-nearest-neighbor analysis of experimental data. *BMC. Bioinformatics*, **11**(252), 382–404.

Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I., Stadler, P., and Schuster, P. (1996). Analysis of rna sequence structure maps by exhaustive enumeration i. neutral networks. *Monatshefte für Chemie, Chemical Monthly*, **127**, 355–374.

Grüner, W., Giegerich, R., Strothmann, D., Reidys, C., Weber, J., Hofacker, I., Stadler, P., and Schuster, P. (1996). Analysis of rna sequence structure maps by exhaustive enumeration ii. structures of neutral networks and shape space covering. *Monatshefte für Chemie, Chemical Monthly*, **127**, 375–389.

H, L. (2010). Drug giants turn their backs on rna interference. *Nature*, **468**(7323), 487.

Hamilton, A. and Baulcombe, D. (1999). A species of small antisense rna in posttranscriptional gene silencing in plants. *Science*, **286**(5441), 950–952.

Hammann, C., Luptak, A., Perreault, J., and de la PeÃśa, M. (2012). The ubiquitous hammerhead ribozyme. *RNA*, **18**, 871–885.

Hanson, S., Bauer, G., Fink, B., and Suess, B. (2005). Molecular analysis of a synthetic tetracycline-binding riboswitch. *RNA*, **11**, 503–511.

Hoener zu Siederdissen, C., Hammer, S., Abfalter, I., Hofacker, I. L., Flamm, C., and Stadler, P. F. (2013). Computational design of rnas with complex energy landscapes. *Biopolymers*, **99**(12), 1124–1136.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte Chemie*, **125**, 167–188.

Homan, J., Favorov, O., Lavender, C., Kursun, O., Ge, X., Busan, S., Dokholyan, N., and Weeks, K. (2014). Single-molecule correlated chemical probing of rna. *PNAS*, **111**(38), 13858–13863.

Hörbartner, C. and Micura, R. (2003). Bistable secondary structures of small rnas and their structural probing by comparative imino proton nmr spectroscopy. *J.Mol. Biol.*, **325**, 421–431.

Hu, X. M., ZHANGïijŇJ., Xiao, J., and Li, Y. (2008). Protein folding in hydrophobic-polar lattice model: A flexible ant- colony optimization approach. *Protein and Peptide Letters*, **15**(5), 469–477.

Huang, J. and Voss, B. (2014). Rna-kinetics based on folding space abstraction. *BMC Bioinformatics*, **15**(1), 60.

Huang, J., Backofen, R., and Voss, B. (2012). Abstract folding space analysis based on helices. *RNA*, **18**(12), 2135–2147.

Hwang, W. Y., Fu, Y., Reyon, D., Maeder, M. L., Tsai, S. Q., Sander, J. D., and Joung, J. K. (2013). Efficient in vivo genome editing using rna-guided nucleases. *Nature Biotechnology*, **31**(3), 227–229.

Incarnato, D., Neri, F., Anselmi, F., and Oliviero, S. (2014). Genome-wide profiling of mouse rna secondary structures reveals key features of the mammalian transcriptome. *Nucleic Acids Res*, **15**(10), 491.

Jackson, S. A., Koduvayur, S., and Woodson, S. A. (2006). Self-splicing of a group i intron reveals partitioning of native and misfolded rna populations in yeast. *RNA*, **12**(12), 2149–2159.

Janssen, S. and Giegerich, R. (2014). The rna shapes studio. *Bioinformatics*.

Jenison, R., Gill, S., Pardi, A., and Poliski, B. (1994). High-resolution molecular discrimination by rna. *Science*, **263**, 1425–1429.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J., and Charpentier, E. (2012). A programmable dual-rna-guided dna endonuclease in adaptive bacterial immunity. *Science*, **17**(337), 816–21.

Jinek, M., East, A., Cheng, A., Lin, S., Ma, E., and Doudna, J. (2013). Rna-programmed genome editing in human cells. *eLife*, **2**.

Ke, A., Zhou, K., Ding, F., Cate, J., and Doudna, J. (2004). A conformational switch controls hepatitis delta virus ribozyme catalysis. *Nature*, **429**, 201–205.

Keefe, A., Pai, S., and Ellington, A. (2010). Aptamers as therapeutics. *Nat Rev Drug Discov*, **9**(7), 537–550.

Kennedy, J. and Eberhart, R. (1995). Particle swarm optimization. *Proceedings of IEEE International Conference on Neural Networks IV*, pages 1942–1948.

Kielpinski, L. and Vinther, J. (2014). Massive parallel-sequencing-based hydroxyl radical probing of rna accessibility. *Nucleic Acids Res*, **42**(8), e70.

Kleinkauf, R., Mann, M., and Backofen, R. (2015a). antaRNA – ant colony based RNA sequence design. *Bioinformatics*, **31**(19), 3114–3121.

Kleinkauf, R., Houwaart, T., Backofen, R., and Mann, M. (2015b). antaRNA - multi-objective inverse folding of pseudoknot RNA using ant-colony optimization. *BMC Bioinformatics*, **16:389**.

Korb, O., Stützle, T., and Exner, T. E. (2006). Application of ant colony optimization to structure-based drug design. In *Ant Colony Optimization and Swarm Intelligence, 5th International Workshop, ANTS 2006*, volume 4150 of *Lecture Notes in Computer Science*, pages 247–258. Springer Verlag.

Kubodera, T., Watanabe, M., Yoshiuchi, K., Yamashita, N., Nishimura, A., Nakai, S., Gomi, K., and Hanamoto, H. (2003). Thiamine-regulated gene expression of aspergillus oryzae thia requires splicing of the intron containing a riboswitch-like domain in the 5'-utr. *FEBS Lett.*, **555**, 516–520.

Kwok, C., Tang, Y., Assmann, S., and Bevilacqua, P. (2015). The rna-structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends in Biochemical Sciences*, **40**(4), 221–232.

Lai D, Proctor JR, M. I. (2013). On the importance of cotranscriptional rna structure formation. *RNA*, **19**(11), 1461–73.

Landick, R. (1997). Rna polymerase slides home: Pause and termination site recognition. *Cell*, **88**(6), 741 – 744.

Lavery, R. and Pullmann, A. (1984). A new theoretical index of biochemical reactivity combining steric and electrostatic factors. *Biophys Chem*, **19**, 171–181.

Lee, J., Canny, M., De Erkenez, A., Krilleke, D., Ng, Y., and Shima, D. (2005). A therapeutic aptamer inhibits angiogenesis by specifically targeting the heparin binding domain of vegf165. *Proc Natl Acad Sci USA*, **102**, 18902–18907.

Leguizamon, G. and Michalewicz, Z.and Schutz, M. (2001). An ant system for the maximum independent set problem. *Proceedings of the 2001 Argentinian Congress on Computer Science*, **2**, 1027–1040.

Leguizamon, G. and Michalewicz, Z. (1999). A new version of ant system for subset problems. *Proceedings of the 1999 Congress on Evolutionary Computation(CEC 99)*, **2**, 1458–1464.

Levin, A., Lis, M., Ponty, Y., O'Donnell, C. W., Devadas, S., Berger, B., and Waldispühl, J. (2012). A global sampling approach to designing and reengineering rna secondary structures. *Nucl. Acids Res.*, **40**(20), 10041–10052.

Li, Y., Pan, S., Zhang, Y., Ren, M., Feng, M., Peng, N., Chen, L., Liang, Y. X., and She, Q. (2016). Harnessing type i and type iii crispr-cas systems for genome editing. *Nucleic Acids Research*, **44**(4), e34.

Li, Z., Liu, Z.-B., Xing, A., Moon, B. P., Koellhoffer, J. P., Huang, L., Ward, R. T., Clifton, E., Falco, S. C., and Cigan, A. M. (2015). Cas9-guide rna directed genome editing in soybean. *Plant Physiology*, **169**(2), 960–970.

Liang, Y. C.and Smith, A. E. (2004). An ant colony optimization algorithm for the redundancy allocation problem (rap). *IEEE Transactions on Reliability*, **53**(3), 417–423.

Link, K. H. and Breaker, R. (2009). Engineering ligand-responsive gene-control elements: lessons learned from natural riboswitches. *Gene Therapy*, **16**, 1189–1201.

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms Mol Biol*, **6**, 26.

Lucks, J., Mortimer, S., and Trapnell, C. (2011). Multiplexed rna structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (shape-seq). *Proceedings of the National Academy of Sciences of the United States of America*, **108**(27), 11063–11068.

Lutz, B., Faber, M., Verma, A., Klumpp, S., and Schug, A. (2014). Differences between cotranscriptional and free riboswitch folding. *Nucleic Acids Research*, **42**(4), 2687–2696.

Lyngso, R., Anderson, J., Sizikova, E., Badugu, A., Hyland, T., and Hein, J. (2012). Frnakenstein: multiple target inverse RNA folding. *BMC Bioinformatics*, **13**(1), 260.

Makarova, K., Wolf, Y., Alkhnbashi, O., Costa, F., Shah, S., Saunders, S., Barrangou, R., Brouns, S., Charpentier, E., Haft, D., Horvath, P., Moineau, S., Mojica, F., Terns, R., Terns, M., White, M., Yakunin, A., Garrett, R., van der Oost, J., Backofen, R., and Koonin, E. (2015). An updated evolutionary classification of crispr-cas systems. *Nat Rev Micro*, **13**(11), 722–736.

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., Norville, J. E., and Church, G. M. (2013). Rna-guided human genome engineering via cas9. *Science*, **339**(6121), 823–826.

Maniezzo, V. and Carbonaro, A. (2000). An ants heuristic for the frequency assignment problem. *Future Generation Computer Systems*, **16**(8), 927–935.

Maniezzo, V. and Milandri, M. (2002). An ant-based framework for very strongly constrained problems. *Proceedings of ANTS2000*, pages 222–227.

Mann, M. and Klemm, K. (2011). Efficient exploration of discrete energy landscapes. *Phys. Rev. E*, **83**, online.

Mari, B. and Bardoni, B. (2014). Applied rnai: from fundamental research to therapeutic applications. *Frontiers in Genetics*, **5**(398).

Martens, D., De Backer, M., Haesen, R., Vanthienen, J., Snoeck, M., and Baesens, B. (2007). Classification with ant colony optimization. *IEEE Transactions on Evolutionary Computation*, **11**(5), 651–665.

Martens, D., Baesens, B., and Fawcett, T. (2011). Editorial survey: Swarm intelligence for data mining. *Machine Learning*, **82**(1), 1–42.

McCaskill, J. (1990). The equilibrium partition function and base pair probabilities for rna secondary structure. *Biopolymers Acids Research*, **29**, 105–1119.

McMillan, N. (2006). *RNA Secondary Structure Prediction using Ant Colony Optimization*. Master's thesis, School of Informatics, University of Edinburgh.

Merino, E., Wilkinson, K., Coughlan, J., and Weeks, K. (2005). Rna structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (shape). *J Am Chem Soc*, **127**, 4223–4231.

Merkle, D. and Middendorf, M. (2000). An ant algorithm with a new pheromone evaluation rule for total tardiness problems. *Real World Applications of Evolutionary Computing*, **1803 of Lecture Notes in Computer Science**, 287–296.

Merkle, D., Middendorf, M., and Schmeck, H. (2000). Ant colony optimization for resource-constrained project scheduling. *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 893–900.

Michiels, P., Versleijen, A., Verlaan, P., Pleij, C., and Hilbers, C. (2001). Solution structure of the pseudoknot of srv-1 rna, involved in ribosomal frameshifting. *J Mol Biol*, **310**, 1109–1123.

Mitra, S., Shcherbakova, I., Altman, R., Brenowith, M., and Laederach, A. (2009). High-throughput single nucleotide structural mapping by capillary automated footprinting analysis. *NAR*, **36**, e36.

Mooney, R. A., Artsimovitch, I., and Landick, R. (1998). Information processing by rna polymerase: Recognition of regulatory signals during rna chain elongation. *Journal of Bacteriology*, **180**(13), 3265–3275.

Morris, K. V. and Mattick, J. S. (2014). The rise of regulatory rna. *Nat Rev Genet*, **15**(6), 423 – 437.

Müller, M., Weigand, J. E., Weichenrieder, O., and Suess, B. (2006). Thermodynamic characterization of an engineered tetracycline-binding riboswitch. *Nucleic Acids Research*, **34**(9), 2607–2617.

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

Nahvi, A., Sudarsan, N., Ebert, M. S., Zou, X., Brown, K. L., and Breaker, R. (2002). Genetic control by a metabolite binding mrna. *Chemistry & Biology*, **9**(9), 1043–1049.

Narberhaus, F. (2010). Translational control of bacterial heat shock and virulence genes by temperature-sensing mrnas. *RNA Biology*, **7**(1), 84–89.

Nardelli, M., Tedesco, L., and Bechini, A. (2013). Cross-lattice behavior of general aco folding for proteins in the hp model. *Proc. of ACM SAC*, pages 1320–1327.

Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and Finn, R. D. (2015). Rfam 12.0: updates to the rna families database. *Nucleic Acids Research*, **43**(D1), D130–D137.

Neugebauer, K. M. (2002). On the importance of being co-transcriptional. *Journal of Cell Science*, **115**, 3865–3871.

Nimjee, S. and Rusconi, C.P.and Sullenger, B. (2005). Aptamers: an emerging class of therapeutics. *Annu. Rev. Med.*, **56**, 555–583.

Nixon, P., Rangan, A., Kim, Y., Rich, A., and Hoffman, D. (2002). Solution structure of a luteoviral p1-p2 frameshifting mrna pseudoknot. *J Mol Biol*, **322**, 621–633.

Noeske, J., Buck, J., Fürtig, B., Nasiri, H. R., Schwalbe, H., and Wöhnert, J. (2007). Interplay of âĂIJinduced fitâĂİ and preorganization in the ligand induced folding of the aptamer domain of the guanine binding riboswitch. *Nucleic Acids Research*, **35**(2), 572–583.

Nussinov, R., Pieczenik, G., Griggs, J. R., and Kleitman, D. J. (1978). Algorithms for loop matchings. *SIAM Journal on Applied Mathematics*, **35**(1), 68–82.

Pan, T. and Sosnick, T. (1997). Intermediates and kinetic traps in the folding of a large ribozyme revealed by circular dichroism and uv absorbance spectroscopies and catalytic activity. *Nat Struct Biol.*, **4**(11), 931–8.

Pan, T. and Sosnick, T. (2006). Rna folding during transcription. *Annual Review of Biophysics and Biomolecular Structure*, **35**, 161–175.

Parpinelli, R. S., Lopes, H. S., and Freitas, A. A. (2002a). An ant colony algorithm for classification rule discovery. *Data Mining: A heuristic Approach*, pages 191–209.

Parpinelli, R. S., Lopes, H. S., and Freitas, A. A. (2002b). Data mining with an ant colony optimization algorithm. *IEEE Transaction on Evolutionary Computation*, **6**(4), 321–332.

Perdrizet, G. A., Artsimovitch, I., Furman, R., Sosnick, T. R., and Pan, T. (2012). Transcriptional pausing coordinates folding of the aptamer domain and the expression platform of a riboswitch. *Proceedings of the National Academy of Sciences of the United States of America*, **109**(9), 3323–3328.

Pham, D., Ghanbarzadeh, A., Koc, E., Otri, S., Rahim, S., and Zaidi, M. (2005). The bees algorithm. *Technical Note, Manufacturing Engineering Centre, Cardiff University, UK*.

Pleij, C., Rietveld, K., and Bosch, L. (1985). A new principle of rna folding based on pseudoknotting. *Nucleic Acids Res*, **13**(5), 1717–31.

R., L. and Serra, D. (2002). Adaptive search heuristics for the generalized assignment problem. *Mathware & soft computing*, **9**(2-3), 417–423.

Rastogi, T., Beattie, T., Olive, J., and Collins, R. (1996). A long-range pseudoknot is required for activity of the neurospora vs ribozyme. *EMBO J*, **15**, 2820–2825.

Regulski, E. and Breaker, R. (2008). In-line probing analysis of riboswitches. *Methods Mol Biol*, **419**, 53–67.

Reinharz, V., Ponty, Y., and Waldispühl, J. (2013). A weighted sampling algorithm for the design of RNA sequences with targeted secondary structure and nucleotide distribution. *Bioinformatics*, **29**(13), i308–i315.

Ren, J., Rastegari, B., Condon, A., and Hoos, H. (2005). Hotknots: Heuristic prediction of rna secondary structures including pseudoknots. *RNA*, **15**.

Rentmeister, A., Mayer, G., Kuhn, N., and Famulok, M. (2007). Conformational changes in the expression domain of the escherichia coli thim riboswitch. *Nucleic Acids Research*, **35**(11), 3713–3722.

Rietveld, K., Van Poelgeest, R., Pleij, C., Van Boom, J., and Bosch, L. (1982). The trna-like structure at the 3'-terminus of turnip yellow mosaic virus rna. differences and similarities with canonical trna. *Nucleic Acids Res*, **10**, 1929–1946.

Rivas, E. and Eddy, S. R. (2000). Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**(7), 583–605.

Roth, A. and Breaker, R. R. (2009). The structural and functional diversity of metabolite-binding riboswitches. *Annual Review of Biochemistry*, **78**(1), 305–334.

Sato, K., Kato, Y., Hamada, M., Akutsu, T., and Asai, K. (2011). Ipknot: fast and accurate prediction of rna secondary structures with pseudoknots using integer programming. *Bioinformatics*, **27**(13), i85–93.

Seetin, M., Kladwang, W., Bida, J., and Das, R. (2014). Massively parallel rna chemical mapping with a reduced bias map-seq protocol. *Methods Mol Biol*, **1086**, 95–117.

Serganov, A. (2009). The long and the short of riboswitches. *Current opinion in structural biology*, **19**(3), 251–259.

Shcherbakova, I. and Brenowith, M. (2008). Monitoring structural changes in nucleic acids with single residue spatial and millisecond time resolution by quantitative hydroxyl radical footprinting. *Nature Protoc*, **3**, 288–302.

Shen, L. and Tinoco, I. (1995). The structure of an rna pseudoknot that causes efficient frameshifting in mouse mammary tumor virus. *J Mol Biol*, **247**, 963–978.

Shen, S. (2008). Rna folding: Conformational statistics, folding kinetics, and ion electrostatics. *Annu. Rev. Biophys*, **37**, 179–214.

Shmygelska, A. and Hoos, H. (2005). An ant colony optimisation algorithm for the 2D and 3D hydrophobic polar protein folding problem. *BMC Bioinformatics*, **6**(1), 30.

Siegfried, N. A., Busan, S., Rice, G. M., Nelson, J. A. E., and Weeks, K. M. (2014). Rna motif discovery by shape and mutational profiling (shape-map). *Nat Meth*, **11**(9), 959–965.

Soukup, G. and Breaker, R. (1999). Engineering precision rna molecular switches. *Proc. Natl. Acad. Sci. USA*, **96**, 3584–3589.

Stützle, T. (1997). An ant approach to the flow shop problem. *Technical report AIDA*.

Sudarsan, N., Barrick, J., and Breaker, R. (2003). Metabolite-binding rna domains are present in the genes of eukaryotes. *RNA*, **9**, 644–647.

Suess, B., Hanson, S., Berens, C., Fink, B., Schroeder, R., and Hillen, W. (2003). Conditional gene expression by controlling translation with tetracycline-binding aptamers. *Nucleic Acids Research*, **31**(7), 1853–1858.

Suess, B., Fink, B., Berens, C., Stentz, R., and Hillen, W. (2004). A theophylline responsive riboswitch based on helix slipping controls gene expression in vivo. *Nucleic Acids Research*, **32**(4), 1610–1614.

Svitashev, S., Young, J. K., Schwartz, C., Gao, H., Falco, S. C., and Cigan, A. M. (2015). Targeted mutagenesis, precise gene editing, and site-specific gene insertion in maize using cas9 and guide rna. *Plant Physiology*, **169**(2), 931–945.

Taneda, A. (2011). MODENA: a multi-objective RNA inverse folding. *Adv Appl Bioinform Chem*, **4**, 1–12.

Theimer, C., Blois, C., and Feigon, J. (2005). Structure of the human telomerase rna pseudoknot reveals conserved tertiary interactions essential for function. *Mol Cell*, **17**, 671–682.

Tiemann, K. and Rossi, J. (2009). Rnai-based therapeutics-current status, challenges and prospects. *EMBO Mol Med*, **1**(3), 142–151.

Treiber, D. and Williamson, J. (1999). Exposing the kinetic traps in rna folding. *Curr Opin Struct Biol.*, **9**(3), 339–45.

Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: Rna ligands to bacteriophage t4 dna polymerase. *Science*, **249**(4968), 505–515.

Tullius, T. and Greenbaum, J. (2005). Mapping nucleic acid structure by hydroxyl radical cleavage. *Curr Opin Struct Biol*, **9**, 127–134.

Turner, D. H. and Mathews, D. H. (2002). Nndb: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research*, **38**, D280–D282.

Tyrrell, J., McGinnis, J. L., Weeks, K. M., and Pielak, G. J. (2013). The cellular environment stabilizes adenine riboswitch rna structure. *Biochemistry*, **52**(48), 8777–8785.

Ulrich, H., Trujillo, C., Nery, A., Alves, J., Majumder, P., Resende, R., and Martins, A. (2006). Dna and rna aptamers: from tools for basic research towards therapeutic applications. *Comb. Chem. High Throughput Screen.*, **9**, 619–632.

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

Wang, H., Yang, H., Shivalila, C. S., Dawlaty, M. M., Cheng, A. W., Zhang, F., and Jaenisch, R. (2013). One-step generation of mice carrying mutations in multiple genes by crispr/cas-mediated genome engineering. *Cell*, **153**(4), 910–918.

Wang, T., Wei, J. J., Sabatini, D. M., and Lander, E. S. (2014). Genetic screens in human cells using the crispr-cas9 system. *Science*, **343**(6166), 80–84.

Watson, J. and Crick, F. (1953). Molecular structure of nucleic acids. *Nature*, **171**, 737–738.

Wenter, P., Fürtig, B., Hainard, A., Schwalbe, H., and Pitsch, S. (2005). Kinetics of photoinduced rna refolding by real-time nmr spectroscopy. *Angewandte Chemie International Edition*, **44**(17), 2600–2603.

Wickiser, J. K., Winkler, W. C., Breaker, R. R., and Crothers, D. M. (2005). The speed of {RNA} transcription and metabolite binding kinetics operate an {FMN} riboswitch. *Molecular Cell*, **18**(1), 49 – 60.

Wilkinson, J., Vasa, S., Deigan, K., Mortimer, S., Giddings, M., and Weeks, K. (2009). Influence of nucleotide identity in ribose 2'hydroxyl reactivity in rna. *RNA*, **15**, 1314–1321.

Winkler, W. and Breaker, R. (2003). Genetic control by metabolite-binding riboswitches. *Chembiochem*, **4**(10), 1024–32.

Wolfinger, M. T., Svrcek-Seiler, W. A., Flamm, C., Hofacker, I. L., and Stadler, P. F. (2004). Efficient computation of rna folding dynamics. *Journal of Physics A: Mathematical and General*, **37**(17), 4731.

Wolfinger, M. T., Will, S., Hofacker, I. L., Backofen, R., and Stadler, P. F. (2006). Exploring the lower part of discrete polymer model energy landscapes. *Europhysics Letters*, **74**(4), 725–732.

Xu, X. and Chen, S.-J. (2012). Kinetic mechanism of conformational switch between bistable rna hairpins. *Comb. Chem. High Throughput Screen.*, **134**(30), 12499–12507.

Yu, B., Yang, Z., and Xie, J.-X. (2011). A parallel improved ant colony optimization for multi-depot vehicle routing problem. *Journal of the Operational Research Society*, **62**(1), 183–188.

Zadeh, J. N., Wolfe, B. R., and Pierce, N. A. (2011). Nucleic acid sequence design via efficient ensemble defect optimization. *J Comb Chem*, **32**(3), 439–452.

Zemora, G. and Waldsich, C. (2010). Rna folding in living cells. *RNA Biology*, **7**(6), 634–641.

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. **9**(1), 133–48.