universität wien

Masterarbeit

Titel der Masterarbeit

„Computational Refinement of SHAPE – RNA probing Experiments“

verfasst von

Roman Wilhelm Ochsenreiter, BSc

angestrebter akademischer Grad

Master of Science (MSc)

Wien, 2015

Studienkennzahl lt. Studienblatt:	A 066 863
Studienrichtung lt. Studienblatt:	Masterstudium Biologische Chemie
Betreut von:	Univ.-Prof Dipl.-Phys. Dr. Ivo L. Hofacker

CONTENTS

ACKNOWLEDGEMENTS

Most of all I would like to thank Prof. Dr. Dipl.-Phys. Ivo Hofacker for supervising this work and as well providing me with the necessary equipment. Special thanks also go to Prof. Dr. Christoph Flamm and Dr. Sven Findeiss who both provided helpful input numerous times. Thanks to the whole TBI-team, as everybody here *always* offered helpful advice, no matter how absurd the question or how pressing time was. On this occasion I would also like to express my gratitude to Mag. Stefan Hammer, who was a amicable roommate and introduced me to numerous Unix-related tools that turned out extremely helpful for work.

At last, I would like to express many thanks to my parents, Maria and Wilhelm and to my brother Julian, for their ongoing support which has been a foundation for many past, and hopefully future successes.

1

MOTIVATION

Chemical probing, also called structural mapping methods, have for decades provided a useful tool for RNA structure modeling[1], as they help to identify accessible positions in an RNA molecule. In recent time, the method has gained new dynamics, driven by the development of new probing techniques such as *SHAPE* (selective 2'hydroxyl acylation analyzed by primer extension)[2]. In addition to the "traditional" chemical probing experiments, which are always about probing one sequence ("1-Dimensional"), recently a new 2-Dimensional approach was developed by Kladwang et al.[3]. This approach expands the classical setup from one experiment to a series of experiments where every point-mutated variant of the sequence of interest is probed and analyzed, consequently called Mutate&Map. By analysis of the local perturbations in each experiment, valuable information is obtained about base-base interactions *in vitro*, which can guide the inference of basepairs in the wild-type sequence, thus allowing the building of high-confidence secondary structures.

Despite its usefulness, the experimental protocol of the Mutate&Map approach has some practical constraints. For the complete exploration of all possible point mutations of a given RNA sequence, each position has to be mutated once to every other base, resulting in a large array of necessary experiments. In the wetlab, every Sequence of this mutant library has to be bought or synthesized. While this is no problem for short sequences, it quickly becomes obvious that with increasing RNA length, practicability as well as feasibility are not given even for well-funded labs. Eventhough the prices for synthesis services are coninuously dropping, the cheap availability of long sequences is still in distant future, restricting the rapid application of the Mutate&Map approach to only small RNAs.

Moreover, the analysis of Mutate&Map Data showed that a considerable amount of mutations resulted in no gain of additional information [36]. The major share of those mutations was located in largely unpaired regions of the molecule, as predicted by computational methods. A lot of computational information, such as MFE structures, structural ensembles or

basepair probabilities can be calculated in a fairly quickly and resource-efficient manner for long sequences (>1kb). With all this information available, the question arises whether there is a possibility to preprocess the experiment with suitable computationally derived constraints in order that less mutant sequences need to be probed.

In order to achieve any computational optimization in a meaningful way, it is also necessary to characterize in detail the *SHAPE* -reaction's interaction with RNA. To date, many aspects concerning basic characteristics of the *SHAPE* -reaction are little understood. This is mainly because until recently, the amount of available probing data was small, hindering effective statistical analysis. However, since the advent of endeavours as the EteRNA-project, which typically generate enormous quantities of data with the help of high-throughput methods, scarcity of data is no longer an issue.

Hence, it is of vital interest to investigate unanswered questions, as to what extent *SHAPE* -reactivity is influenced by an RNA's structural features and how well experimental data correlates with respective computed parameters. Thorough understanding of the *SHAPE* -reaction will furthermore be helpful for the simulation of *SHAPE* data, which will be a crucial step in the aforementioned optimization of the Mutate&Map-procedure.

2

INTRODUCTION

This chapter provides the necessary background knowledge, especially about Nucleic acids, RNA probing methods and secondary structure prediction algorithms.

2.1 RNA

Nucleic acids are involved at the most basic mechanisms that underlay life. While Desoxy-ribonucleic acid (DNA) is generally involved in the storage of genetic information, Ribonucleic acid (RNA) was long thought to serve just as an intermediate information carrier, called messenger RNA *mRNA*, serving as a template for the production (called *translation*) of proteins. This model is also known as the central *dogma of molecular biology*.

However, it has become clear that RNA has assumed many other diverse roles in life. According to the dogma, every RNA would function should have a protein which it serves as translation template. This does not hold true for all RNAs, as shown by the discovery of transfer RNA (*tRNA*), ribosomal RNA (*rRNA*), and more generally non coding RNA (*ncRNA*), which all are transcribed from the DNA, but do not encode any proteins.

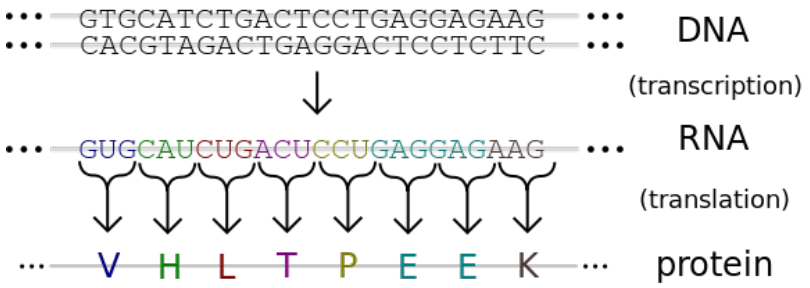


Figure 1: The central Dogma of Molecular Biology. DNA serves as template for mRNA synthesis, the Ribosome translates the encoded information into proteins.

Compared to DNA, the most striking striking difference of RNA is its capability to fold into catalytically active conformations, also called *ribozymes*, which assume various roles as RNA-cleavage[4] or even self replication[5]. Especially the self-

replicating RNAs are of particular interest, as it is hypothesized that RNA accounted for a large portion of all biomolecules in prebiotic times *RNA-world*[6].

2.1.1 *Chemical basics*

Nucleic acids are highly polymeric biomolecules and represent one of the most important family of molecules found in nature. Their monomers consist of a 5-carbon sugar, a nucleobase and a phosphate group and are called nucleotides, or nucleosides in case there is no phosphate group. Common *sugars* are D-Ribose for Ribonucleic acid (RNA), and 2-Desoxy-Ribose for Desoxyribonucleic acid (DNA). Usually, the positions in the (Desoxy-) Ribose molecule are numbered (Figure 2), to facilitate orientation. The *Phosphate group* is connected to the sugar at position 3' via a phosphoester bond. The *Nucleobase* is connected by a glycosidic bond to the carbon atom at position 1'. All canonical Nucleobases found in nature are pyrimidine or purine derivatives. For DNA, there are the four bases Adenine (A), Guanine (G), Cytosine (C) and Thymine (U). The same bases are found in RNA, except for Thymin, which is substituted by the chemically almost identical Uracil (U). Nucleotides condense into the polymeric form, often referred to as *strands*. In the polymer, the phosphate connects the 5' and the 3' carbon atoms via a phosphodiester bond, forming the so called *sugar-phosphate-backbone*.

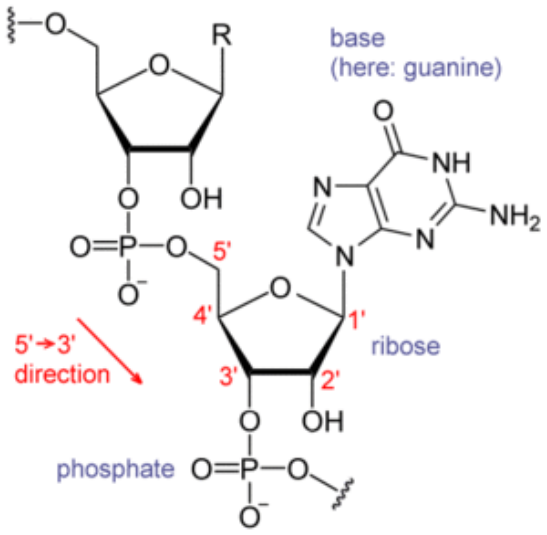


Figure 2: Schematic depiction of the backbone of an RNA molecule. The carbon atoms in the sugar are labeled as it is common with nucleic acids. In DNA the 2'-Hydroxyl group is replaced by a single hydrogen atom

Basepairs

Early on in nucleic acids research, it was observed that the bases Adenine/Thymine and Cytosine/Guanine are always present in equal quantities in cellular extracts. The exact origin of this relationship remained unclear until the work of Franklin, Watson and Crick [7], who discovered the DNAs structure and its tendency to arrange itself in a double-helical conformation of two antiparallel strands. Central for the *hybridization* of two strands are the bases' ability to engage in mutual noncovalent interactions via hydrogen bonds. Among all possible pairwise interactions, the most stable and universally occuring are the so called *Watson-Crick*-basepairs between Adenine-Thymine or Cytosine-Guanine. Ribonucleic acids have similar abilities to form basepairs, with the minor difference that Thymine is substituted by Uracil [8] and the occurrence of the less stable G-U basepairs, also called *Wobble*-basepairs [9]. Also, since RNA predominantly occurs as a single stranded molecule, it shows an even stronger predisposition for the establishment of intramolecular basepairs in contrary to DNAs preference for intermolecular base pairings.

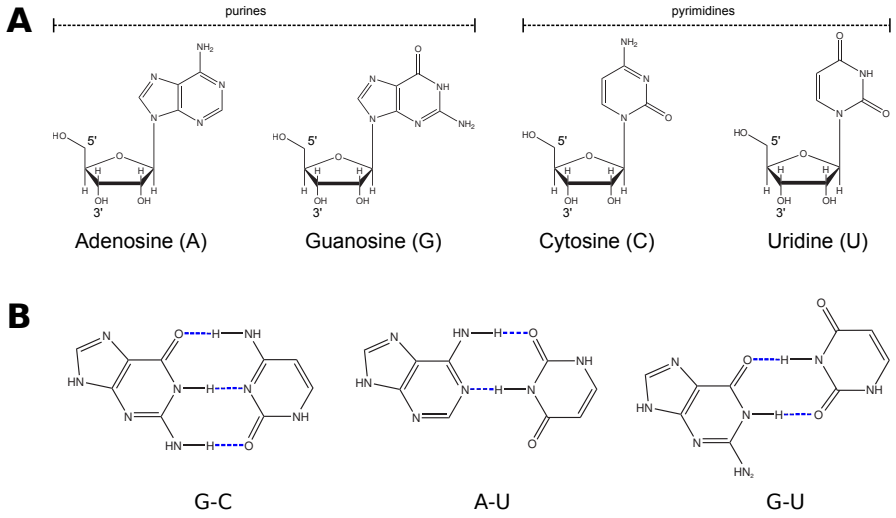


Figure 3: A - Nucleosides found in RNA, the 3′ and 5′ sugar carbon atoms are labeled. B - geometry of the two Watson-Crick basepairs (G-C, A-U) and the Wobble base pair (G-U)

If the bases face each other in the proper geometry, A-T (A-U) and G-U basepairs form two hydrogen bonds, whereas G-C basepairs establish 3 bonds, leading to higher stability of G-C rich structures. Even though base complementarity is a prerequisite for RNA/DNA folding, the involved hydrogen bonds just account for a fraction of the overall stabilizing energies involved in the formation of stable structures, despite their stabilizing interactions [10]. When two basepairs are adjacent to each other, as it is commonly the case in helices, stacking interactions between the pi electrons of the bases greatly stabilize the structure. A single base pair can stabilize a structure up to -3.4 kcal mol^{-1}, slightly above the thermal energy at room temperature RT=0.6 kcal mol^{-1}.

	CG	GC	GU	UG	AU	UA
CG	-2.4	-3.3	-2.1	1.4	-2.1	-2.1
GC	-3.3	-3.4	-2.5	-1.5	-2.2	-2.4
GU	-2.1	-2.5	1.3	-0.5	-1.4	-1.3
UG	-1.4	-1.5	-1.4	0.3	-0.6	-1.0
AU	-2.1	-2.2	-1.4	0.6	-1.1	-0.9
UA	-2.1	-2.4	-1.3	-1.0	-0.9	-1.3

Table 1: Free energies in kcal mol^{-1} for stacked basepairs

2.1.2 *RNA Structure*

Primary structure

The primary structure of an RNA molecule is the sequence of nucleotides in the polymer chain, written down as a sequence of the letters A,C,G and U. Conventionally, when written from left to right, the first letter corresponds to the sequence's 5'-end, whereas the last to its 3'-end. Despite its name, the primary structure holds no information on the spatial arrangement of the RNA molecule. Its main advantage and usage is the unambiguous and simple representation of nucleotide sequences, which are the basis for secondary/tertiary structure prediction algorithms. When working with RNA computationally, primary structures are commonly saved in the FASTA-format.

```
>Sequence_Header
AACGUAACGCGUACUGCAUGCAUGCA
```

Secondary structure

Due to their natural ability to form basepairs between complimentary bases, single stranded RNA molecules typically fold onto themselves. The resulting base pairing pattern forms stretches of paired bases, called helices or stems, and unpaired regions, called loops. This pattern of stems and loops is called the secondary structure of an RNA molecule. The secondary structure reflects the local geometry of specific domains of the molecule, but carries no information on how the various domains are organized to each other. Secondary structures are a useful way of describing RNA structure on a coarse grained level and still retain enough accuracy since the energetic contributions of basepairs usually outweigh higher-order interactions. In order to qualify as a valid secondary structure, each base has to fulfill the following criteria:

1. A base cannot participate in more than one base pair.
2. There must be at least three unpaired bases between two bases paired with each other.
3. There can be no two basepairs (i, j) and (k, l) for which $i < k < j < l$ holds true.

Condition 1 excludes tertiary structure motifs like G-quadruplexes and base triplets; condition 2 takes into consideration that in

such small loops, the RNA backbone would have to bend with an sterically impossible angle.

If two basepairs violate condition 3, they are said to form a *Pseudoknot*. These structural elements occur in nature, but are forbidden in the computational model for practical purposes, since the inclusion of pseudoknots in secondary structure prediction algorithms increases the complexity of the problem dramatically. Also, the knowledge about the energetics of pseudoknotted structures is too small to allow the meaningful parametrization of their energies. Therefore, the accurate prediction of pseudoknots becomes important not until dealing with the tertiary structure.

In addition to the separation of bases in basepairs and and unpaired bases, a RNA structure can be further decomposed into loops. A loop consists of the *closing base pair* (i, j) and all positions *k immediately interior* of the pair (i, j). A base k is called immediately interior to the base pair (i, j), if $i < k < j$ and there is no other base pair (p, q) which satisfies $i < p < k < q < j$. Counting the number of basepairs delimiting a loop, including the closing basepairs, allows to assign a *degree* to various loops. Loops of degree 1 are called *hairpin loops*, loops of degree 2 are called *Interior loops* or *Bulges* in case there are unpaired bases just on one side of the interior loop. All loops of higher degrees are referred to as *Multiloops*.

The loop decomposition is especially important for secondary structure prediction algorithms, since the overall free energy of an RNA molecule is calculated as the sums over all loop energies.

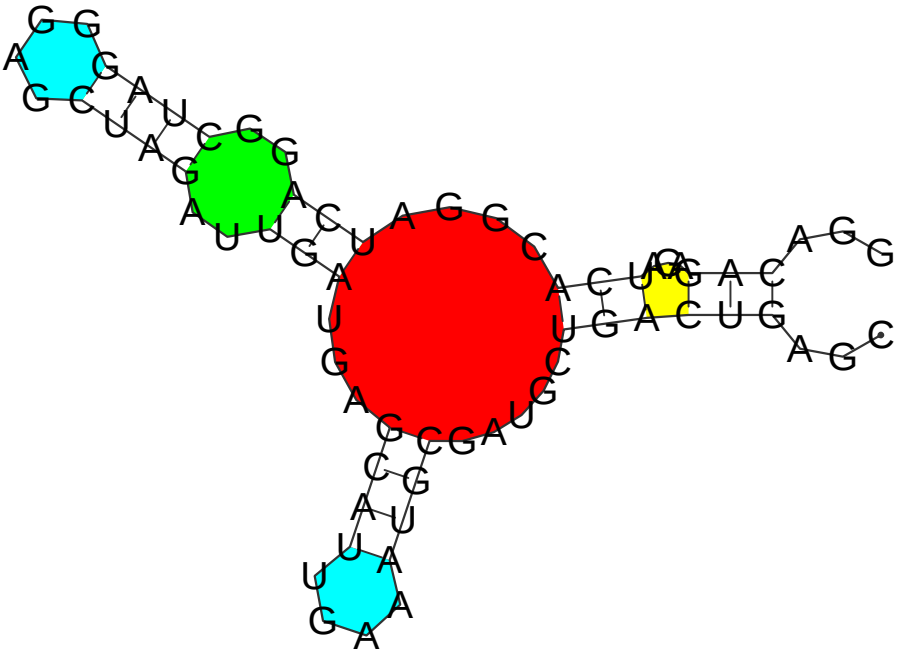


Figure 4: Depiction of a Secondary structure, special loop types are coloured - light blue (Hairpin), green (Interior loop), yellow (Bulge), red (Multiloop)

Tertiary structure

The term tertiary structure describes the exact positioning of a molecule in 3-dimensional space. In analogy to proteinic enzymes, many biochemically active RNAs gain their function not before they have a distinct tertiary structure. Even though there exists a fair amount of experimentally obtained RNA tertiary structures, the computational prediction is still a hard problem due to the large amount of degrees of freedom involved in 3D-folding.

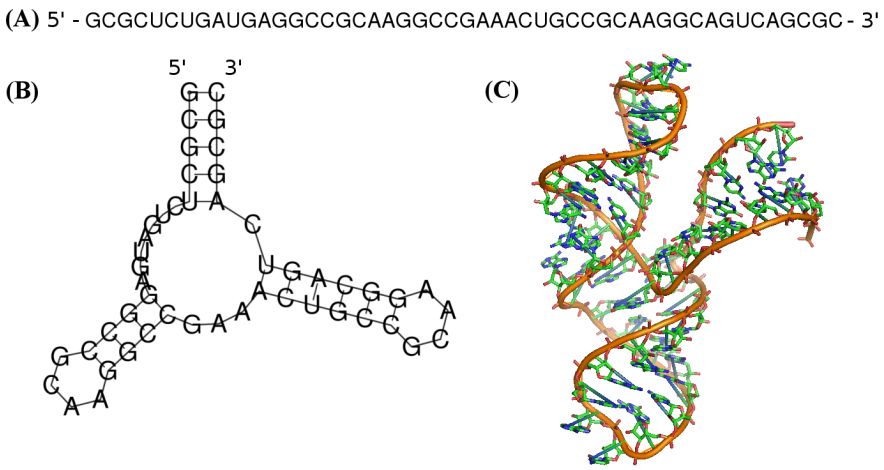


Figure 5: Representations of the Hammerhead Ribozyme. (A) Sequence or Primary structure, (B) Secondary structure, (C) Tertiary structure

2.1.3 *RNA structure Representations*

From a formal point of view, every secondary structure can be viewed as a graph, whose nodes represent nucleotides at position $i = 1, 2.., n$ of a sequence with length n. For each valid secondary structure a sequence can form, two separate sets of edges can be drawn. The first set of edges represents the covalent phosphate backbone connecting each node i with node $i + 1, i = 1, 2...n - 1$. Since the backbone does not change upon folding, this set of edges is the same for all secondary structures. Base pairs account for the second set of edges, and can be drawn independently from each other, as long as they do not violate any of the constraints for valid secondary structures.

Graph notation

Structure graphs are a convenient way to represent a secondary structures in a graph-like way. For this purpose, the backbone is drawn first, and the basepairs are subsequently inserted without crossing each other. Since pseudoknots are not allowed, the resulting resulting graph will always be planar.

Dot-Bracket notation

For a more condensed storage, secondary structures can be described as a string s of length n, where $s_{[1..n]}$ corresponds to the respective nucleotide in the backbone. All positions which

do not participate in any base pair are labeled with a dot ".", then for each base pair $(i, j)\{j|j < i\}$ an opening bracket "(" is inserted at s_i and a respective closing bracket ")" at s_j. Despite its simplicity, the Dot-Bracket string still remains unambiguous, since basepairs are not allowed to cross.

...((((((((((.....))))))))))...

Figure 6: Dot-bracket notation of a short model hairpin.

Coarse grained RNA-Shapes

When comparing RNA structures, the most interesting differences often do not lie in the addition or removal of single basepairs or the change in helix lengths, but in the large scale rearrangement of complete folding domains. With common secondary structure representations as the dot-bracket string, it is necessary to investigate and compare every single base pair of all sample structures, when investigating a large amount of RNA structures generated by stochastic backtracking or suboptimal folding. The RNA-shape approach [11] heavily facilitates this analysis by reducing any secondary structure to its *nesting pattern* through ignoring loop and helix lengths, highlighting only the most important structural features. In such patterns, that can be regarded as abstract representations of secondary structures, loops are represented by a pair of square brackets and unpaired regions by an underscore.

In the current implementation the program RNAshapes [12] distinguishes between 5 levels of abstraction, also called "Shape-Types". Common to all levels is that they abstract from loop and stack lengths, and represent unpaired regions are by an underscore and stacking regions by a pair of squared brackets. Since multiple different secondary structures can map into the same shape, it should be noted that the process of shape abstraction is not reversible.

Type	Description	Example
1	Least abstract - The shape represents all loops and all unpaired regions	`[_[_[]]_[_[]_]]_`
2	Nesting patterns for all loop types and unpaired regions in external loops and multiloops	`[[_[]][_[]_]]`
3	Nesting patterns for all loop types but without unpaired regions	`[[[]][[]]]`
4	Helix nesting patterns in external loops and multiloops	`[[][[]]]`
5	Most abstract - only the nesting pattern of helices, no unpaired regions	`[[][]]`

Table 2: Shape abstraction levels and examples for the structure `((((((...(((..(((...))))))...(((..((.....))..)))))))))`.

In addition to the nesting pattern, there exists a variety of different parameters that can be determined for each Shape. Out of all structures falling into one shape, the so called Shape-Representative (*Shrep*) is the secondary structure with the lowest free energy. As the probability of a structure increases with its (negative) free energy, the Shrep will always be the most probable structure and serves as an appropriate representative for the whole shape.

In case the partition function of the RNA sequence is calculated, it is possible to determine the probability of a structure in the ensemble. Therefore, the probability to encounter a shape, termed *shape probability*, is the sum of the probability of all structures mapping into respective shape. The accuracy of the resulting shape probabilities depends on the method used for the generation of the input structures. Exact shape probabilities can only be calculated when the structure space sampled exhaustively, but since this usually is not an option due to the enormous computational effort, the input structures are generated by boltzmann sampling. Since structures in boltzmann sampling are picked according to their weight in the ensemble, it is assured that even with small samples (1000 structures), the shape probabilities approximate the exhaustive approach.

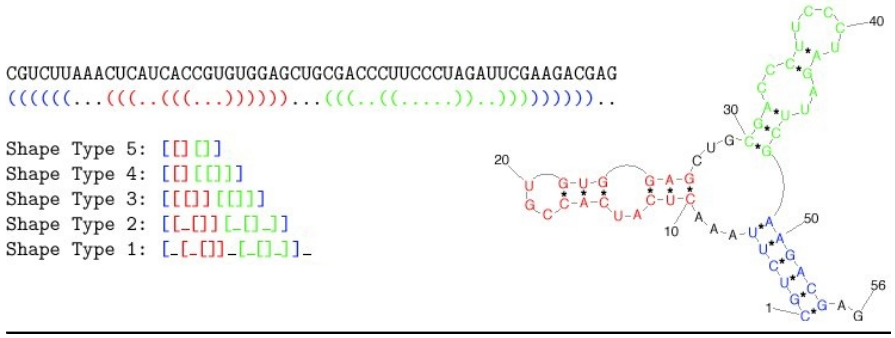


Figure 7: RNA secondary structure and its various shape abstraction levels

2.2 RNA-PROBING AND SHAPE

RNA is a molecule that has been intensively studied by chemical probing for decades. So long as this technique has been around, as diversely it has evolved, comprising many methodologically different approaches. In all approaches the target RNA is treated with some form of reagent, either small molecules [13], enzymes [14] or metal ions [15], that either cleave the RNA directly or form a covalent adducts with the RNAs' bases or backbones. Since the crucial reactions as adduct formation or backbone cleavage are governed by the local environment at each nucleotide, the resulting signals can be, depending on the method, interpreted to quantitatively or qualitatively give a signal about local nucleotide parameters as flexibility or base-pairings.

2'-Single Hydroxyl Acylation analyzed by Primer Extension (*SHAPE*) is a chemical mapping technique which makes it possible to quantitatively analyze RNA conformations at nucleotide-level resolution [16]. In contrary to other probing techniques, SHAPE offers strong advantages over previous techniques as: invariance to Base identity [17], independence of solvent accessibility [18], and overall strong correlation of modification rates with backbone flexibility [19]. Moreover, recent advances in capillary electrophoresis[20] and electropherogram analysis tools[21] have opened up the possibility to perform *SHAPE* -probing in a high-throughput manner.

Due to its inherent utility *SHAPE* has not only been used for exact structure determination [22], but also as a tool for the investigation of folding kinetics [23] or the assistance in sequence design [24].

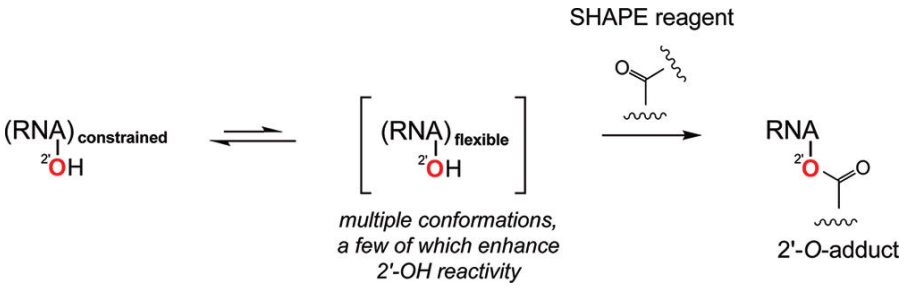


Figure 8: Basic principle of the SHAPE reaction. Figure taken from [18]

2.2.1 *Mechanism*

Where other chemical probes specifically modify nucleobases, SHAPE interacts with RNA's backbone. Commonly used reagents are 1-methyl-7-nitroisatoic anhydride (1M7) or N-methylisatoic anhydride (NMIA), both anhydrid species that preferentially acylate 2'OH groups. After modification, the RNA molecules are transcribed by a reverse transcriptase with fluorescently labeled primers. Since 2'-acylated nucleotides are impassable to reverse transcriptases, the transcribed library will not only consist of full length RNA transcripts, but also of smaller fragments in different quantities depending on how likely the molecule was to be modified at a specific residue. The fragments then are separated by capillary gel electrophoresis and their quantities are determined by readout of band fluorescence. As a last step, (absolute) band intensities are converted to a normalized scale which not only facilitates comparison of intramolecular reactivities, but also enables cross-experiment comparison.

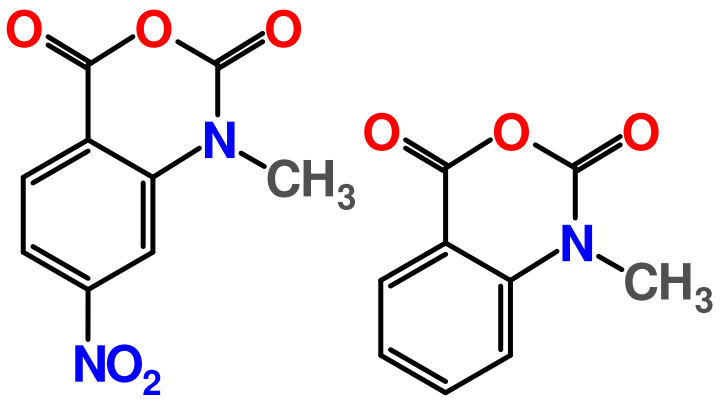


Figure 9: SHAPE-reagents 1M7 (left) and NMIA (right)

2.2.2 *The Mutate&Map strategy*

Even though the inclusion of SHAPE data into structure secondary prediction algorithms greatly improves accuracy [22], it still suffers from intrinsic inaccuracies [25]. Moreover, standard *SHAPE* experiments just investigate the general conformational flexibility of a base, whereas knowing the exact interaction partner would yield much more valuable information.

In order to investigate not only nucleotide flexibility, but actual basepairs, the **Mutate-and-Map** approach takes the whole probing approach to a new level. It refines the standard *SHAPE* -approach by not only probing the wild-type sequence, but by systematically point-mutating each base into its complement, followed by a probing experiment. Since the conversion of a base into its complement eliminates a base pair, an increase in modification is not only visible at the site of mutation, but also becomes visible at the site of the former base pairing partner, the experimental data can be used for the inference of base pairings.

In contrary to the standard *SHAPE* -approach, which yields only information on each base's reactivity and is therefore labeled as "linear" or 1-dimensional, the Mutate&Map-approach consists of 2 "Dimensions", being reactivity and site of mutation, hence the name "2-D"*SHAPE* .

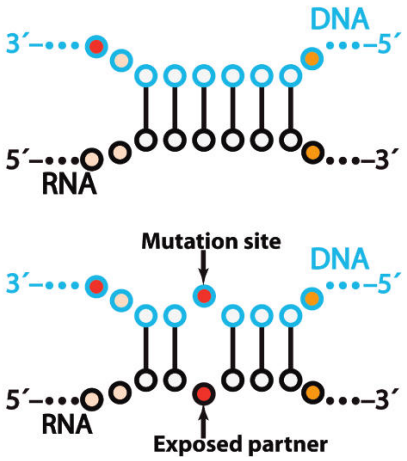


Figure 10: Basic concept of the Mutate&Map approach, explained on a DNA/RNA hybrid. Top: Wild-type DNA/RNA double strand; Bottom: The base pair is eliminated through point-mutation, releasing the other involved bases. Figure adapted from [3]

After proof-of-concept studies on a DNA/RNA hybrid helix [3], which backed up the basic idea behind MM approach with

experimental evidence, its application to a small model hairpin loop was shown to robustly infer most basepairs [36]. Also, it became apparent that for sequence-blind analysis, in order to avoid false positives and isolate true basepairs, the data must be filtered rigorously, since *SHAPE* -signals from mutations not always can be meaningfully interpreted only by manual inspection. In [26], Mutate&Map was successfully used to confirm the secondary and tertiary structure of several small non-coding RNAs.

In the latest work, the Mutate&Map-approach was further improved by not only requiring a base pair to be disrupted by point mutation, but also demanding that the wild type-probing pattern can be recovered after a compensatory mutation is induced at the site of the putative base pairing-partner[27]. Structure models built in that way, termed Mutate-Map-Rescue, were shown to be significantly more accurate than Mutate&Map-experiments alone.

2.2.3 *EteRNA*

The EteRNA-project aims at developing reliable methods for RNA inverse folding[24], where a secondary structure is given, and the sequence which implements this structure in a most stable way has to be found. In contrary to already existing approaches as RNAinverse[28], the idea behind EteRNA is to use machine learning techniques, to extract design rules from a large number of sequence designs, whose quality has been assessed through *SHAPE* -probing. Only design Rules that lead to sequences whose probing pattern is congruent with the target structure, are taken into account.

What makes EteRNAspecial is the source of the numerous sequence designs. All sequences were designed manually by a community of thousands of nonexperts, who build sequences in form of a game according to their own personal developed design rules. Those rules are of purely individual (e.g. heuristic) nature.

The "game" progresses through numerous rounds of sequence design, evaluation by probing, and rule refinement. Rules that consistently led to high-scoring designs are integrated into a new sequence design algorithm. Already after a few rounds, the community's set of rules was able to outperform any existing inverse folding program.

Another major benefit from the EteRNA-project lies in the massive amount of gathered probing data. All probing experiments are made publicly available at the Standford RNA Mapping Database (RMDB - www.rmdb.stanford.edu), acting as a major source for any kind of analysis of *SHAPE* -probing.

2.2.4 *EteRNA-Score*

The quality of a sequence design was summarized as a structure mapping score (EteRNA-score). A nucleotide designed to be unpaired, is assigned a point if its reactivity exceeded 0.25 or if designed to be paired, is less than 0.50. For unpaired nucleotides, this threshold was less stringent to allow for the possibility that a nucleotide could have reduced reactivity from non-Watson–Crick or other interactions. The final score is calculated as the sum of all points divided by the total number of nucleotides.

2.3 RNA SECONDARY STRUCTURE PREDICTION

2.3.1 *Counting structures*

In order to understand the principle of RNA folding algorithms, it is helpful to consider the problem of counting all possible structures for a given sequence:

For n nucleotides, a structure can be formed in two distinct ways from shorter sub-structures: either the first nucleotide is unpaired, and it is followed by another structure on the shorter sequence $x[i+1, ..., j]$, or the first nucleotide is paired with some partner base k. For the latter case, both secondary structures are independent from each other, since basepairs must not cross. The recursion scheme can be visualized as following:

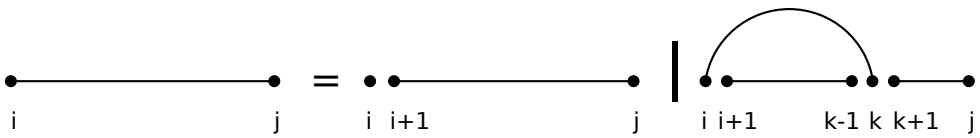


Figure 11: Recursion scheme for the maximum matching problem

The number $N_{i,j}$ of secondary structures that can be formed by the sequence $x[i, ..., j]$ can be recursively formulated as:

$$N_{i,j} = N_{i+1,j} + \sum_{k,(i,k)pairs} (N_{i+1,k-1}N_k + 1, j) \tag{1}$$

2.3.2 *Maximizing Basepairs and Energy minimization*

First attempts of RNA folding revolved around the maximization of basepairs for a given sequence. The Nussinov algorithm [29] can be easily derived from Recursion 1, by substituting N_{ij} by $E_{i,j}$, which represents the maximum number of basepairs on the secondary (sub)structure $x[i, ..., j]$. Also, a weight β_{ik} is assigned to each pair x_i and x_j which is 1 if both form a base pair, otherwise 0.

$$E_{i,j} = \max \left\{ E_{i+1,j}, \max_{k,(i,k)pairs} \left\{ E_{i+1,k-1} + E_{k+1,j} + \beta_{ik} \right\} \right\} \tag{2}$$

Since basepairs differ in energetic contributions, a simple energy model can be established by replacing the weights with the basepairs' free energy. Furthermore, as we want to minimize the free energy, the *max* operation in 2 is replaced by *min*. The structure with the lowest free energy is also called the *Minimum Free Energy* structure (**MFE**).

As the recursions used for energy minimization only compute the minimum energy, but not the structure realizing this energy, the corresponding structure has to be deduced via backtracking. In this process, the path and therefore the list of base pairs for $E(1, n)$ is reconstructed by going backwards through the calculated matrix.

2.3.3 *Energy minimization with the Loop-based energy model*

According to the loop-based energy model, an RNA molecules' free energy can be seen as the sum of its constituting loops. Generally, a loop's energies depends on its size and type. With the exception of small loops, which have been analyzed exhaustively, and a loop's closing base pair, the loop-based energy model is sequence-independent.

The main difference from the previously discussed model is that now, different types of loops have to distinguished. Thus, given the base pair (i, k), all enclosed substructures have to

be further decompose according to their loop types: hairpin-, interior- and multiloops.

The multi loop case is different, as its energy depends on the number of substructures which emanate from the loop. Therefore, it is necessary to decompose structures within a multiloop in a way that record can be kept of the number of its (sub-)components. This problem is solved by representing a multiloop's substructure as a concatenation of two components: an arbitrary 5' part containing at least one component, and a 3' part starting with a base pair and containing just one component. As a result, both multiloop substructures can now be further decomposed into simpler loop types.

Given the recursive decomposition of the structures, the recursions for the energy minimization algorithm are formulated:

$$
\begin{aligned}
\mathcal{F}_{ij} &= \min\left\{\mathcal{F}_{i+1,j}, \min_{i\le k\le j}\left(\mathcal{C}_{ik}+\mathcal{F}_{k+1,j}\right)\right\}\\
\mathcal{C}_{ij} &= \min\left\{\mathcal{H}_{ij}, \min_{i\le k\le l\le j}\left(\mathcal{C}_{kl}+I(i,j;k,l)\right), \min_{i\le u\le j}\left(\mathcal{M}_{i+1,u}+\mathcal{M}^1_{u+1,j-i}+a\right)\right\}\\
\mathcal{M}_{ij} &= \min\left\{\min_{i\le u\le j}\left((u-i+1)c+\mathcal{C}_{u+1,j}+b\right) \min_{i\le u\le j}\left(\mathcal{M}_{i,u}+\mathcal{C}_{u+1,j}+b\right), \mathcal{M}_{i,j-1}+c\right\}\\
\mathcal{M}^1_{ij} &= \min\left\{\mathcal{M}^1_{i,j-1}+c, \mathcal{C}_{ij}+b\right\}
\end{aligned}
\tag{3}
$$

with the following quantities:

- $\mathcal{F}_{ij}$
  free energy of the optimal substructure on the subsequence $x[i,...,j]$.
- $\mathcal{H}_{ij}$
  free energy of a hairpin loop, closed by the base pair (i,j).
- $\mathcal{I}_{ij}$
  free energy of an interior loop, determined by the two base pairs (i,j) and (k,l).
- $\mathcal{C}_{ik}$
  free energy of the optimal substructure on the subsequence $x[i,...,j]$ subject to the constraint that i and j form a base pair.
- $\mathcal{M}_{ij}$
  free energy of the optimal substructure on the subsequence

$x[i, ..., j]$ subject to the constraint that that $x[i, ..., j]$ is part of a multiloop and has at least one component.

- $\mathcal{M}^1_{ij}$
  free energy of the optimal substructure on the subsequence $x[i, ..., j]$ subject to the constraint that that $x[i, ..., j]$ is part of a multiloop and has exactly one component, which has the closing pair i, h for some h satisfying $i \leq h \leq j$.

The recursions for computing the minimum free energy of an RNA molecule in the loop based energy model were first formulated by Zuker and Stiegler [30].

2.3.4 *Including Probing Data as folding Constraints*

As probing data yields quantitative information on the structural flexibility experienced by a single base, it is of interest to improve computational folding by using the experimental data as constraints. This idea is nearly as old as RNA-probing itself and has been implemented in various ways.

The most straightforward approach is to use probing data in the form of *hard constraints*, meaning that every nucleotide which shows high reactivity, is not allowed to participate in any base pair at all [30]. Despite its simplicity, this approach is of limited usefulness, as it is rarely absolutely that specific nucleotides are unpaired. Beyond that, the basic assumption that unpaired nucleotides always show high reactivity values and vice versa, is not necessarily true (See Section 4.1), making a hard-constrained folding rather error-prone.

In order to overcome the shortfalls of this approach, Deigan et. al[31] included probing data as *Soft constraints*. Here, the *SHAPE* -reactivity of a stacked nucleotide i is added as a pseudo free energy ΔG_{SHAPE} which is calculated as

$$\Delta G_{SHAPE}(i) = m \ln(\text{SHAPE reactivity}(i) + 1) + b \qquad (4)$$

where $m = 2.6$ and $b = -0.8$ are empirically determined parameters. This method has led to remarkable improvements in structure prediction accuracy, as shown with the complete folding HIV-1 genome [32], which is about 10kb in size.

Also other attempts were made to include probing data in a soft-constraint manner, differing in their methodology [33][34], but ultimately leading to comparable results (Luntzer et. al., in preparation).

2.3.5 *The Partition function*

In solution RNA molecules exist not in one, defined ground state but in a distribution of states, where every state is differently populated or thus probable. From thermodynamics it can be derived that in equilibrium, the probability of a structure Ψ is proportional to its Boltzmann factor $\exp(-E(\Psi))/R$, where $E(\Psi)$ is the structure's energy, R the molar gas constant, and T the absolute temperature. The structural *ensemble*, consisting of all possible structures Ψ of a sequence, is defined via its *partition function* Z:

$$Z = \sum_{\Psi} \exp(-E(\Psi)/RT) \tag{5}$$

For an ensemble of RNA structures, the partition function can be computed analogous to Equation 2 in the framework of energy minimization, by replacing $E_{i,j}$ with $Z_{i,j}$ - the partition function over all possible structures on the subsequences $x[i, ..., j]$:

$$Z_{i,j} = Z_{i+1,j} + \sum_{k,(i,k)\text{pairs}} Z_{i+1,k-1} Z_{k+1,j} \exp(-\beta_{ik}/RT) \tag{6}$$

Note that we can transform the recursion for E_{ij} in Equation 2 into the equation for Z_{ij} simply by exchanging maximum operations with sums, sums with multiplications and energies by their corresponding Boltzmann factors.

Beyond that, it is also possible to compute the equilibrium probability of a structure Ψ:

$$p(\Psi) = \exp \frac{-E(\Psi)/RT}{Z} \tag{7}$$

along with the probability p_{ij} for a single base pair (i, j)

$$p_{ij} = \sum_{(i,j)\in\Psi} p(\Psi). \tag{8}$$

The algorithm can be easily derived from the MFE recursion (Eq. 2) by replacing all energies with their corresponding Boltzmann factors, *max*-operations with sums and sums with multiplications.

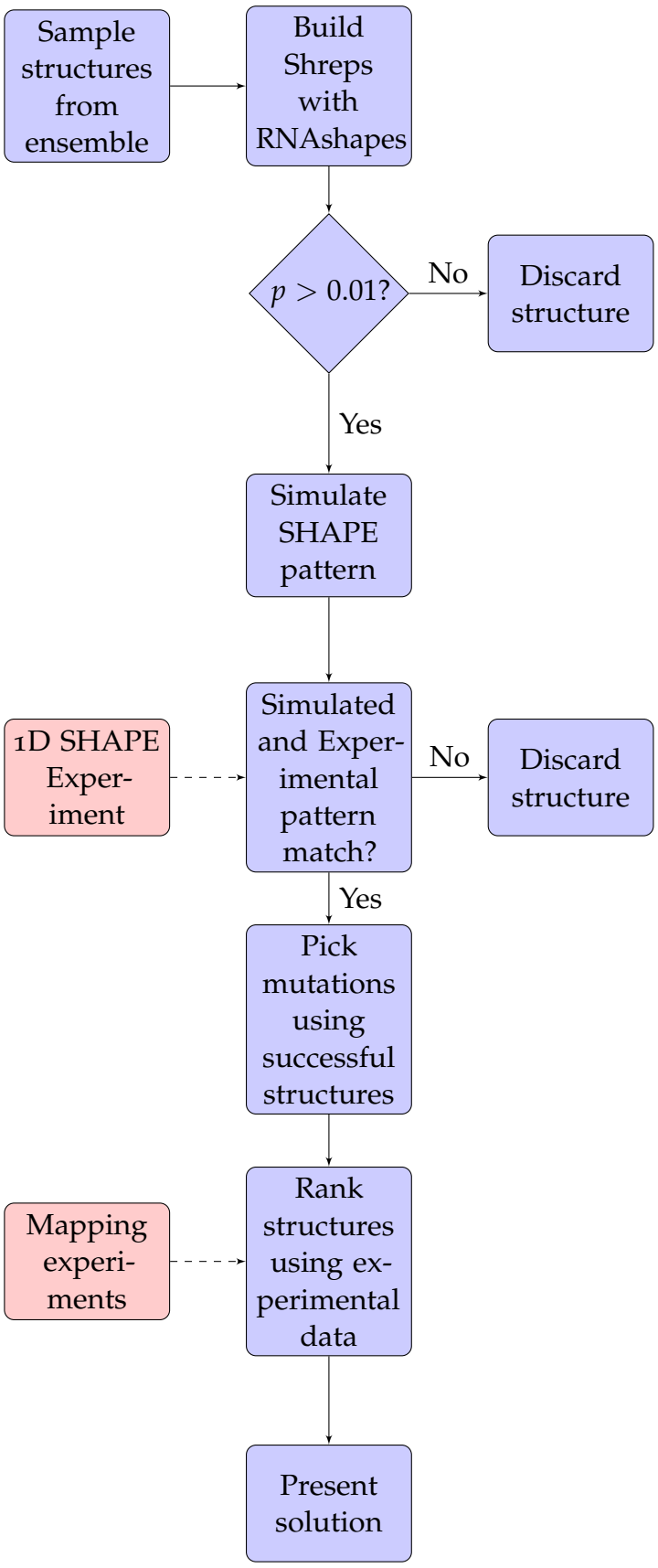
3

METHODS

The following chapter will give an overview over the methods that were developed and applied in this master thesis. All RNA-folding related tools used for this work are part of the Vienna RNA package [28], if not mentioned otherwise. All relevant scripts (parsing Data-files, simulation probability distributions) were written in *Perl 5*, statistics and plots were made with *R*.

3.1 WORKFLOW OVERVIEW

In the following, the steps of the opimisation approach are explained in detail. For a quick overview, the *SHAPE* -optimization procedure designed in this master thesis can be conveniently summarized in a flowchart diagram:



Steps coloured red require work in the wet lab.

1. Structure Sampling
   From a Boltzmann ensemble, secondary structures are obtained by stochastic backtracking with `RNAfold`. This gives a set of structures with thermodynamically favoured (e.g. negative E) secondary structures being likely to be generated.

2. Clustering of Structures - Section 3.2.1
   In a next step, the candidate pool is reduced by clustering all structures with RNAshapes. For each group, a centroid structure (Shrep) is calculated, which form the final pool of candidate structures.

3. Pattern Simulation - Section 3.3
   For each Shrep a *SHAPE* pattern is simulated. This simulation is based on empirical distributions of *SHAPE* - Reactivities of certain structural motives and has been derived from *SHAPE* experiments.

4. 1D-SHAPE Experiment
   An 1D-SHAPE Experiment of the wild type-RNA is performed in the wet lab. This step is not strictly necessary, but additional experimental constraints greatly improve the accuracy of the overall workflow (see next step).

5. Pattern Comparison -Section 3.3.3
   The pool of candidate structures is reduced by comparison of the simulated patterns with the experimental data of the previous step. In case of strong discrepancies between simulated pattern and experiment, the respective predicted structure is deemed inaccurate and subsequently removed from the pool of candidates.

6. Mutation Picking - Section 3.4
   From the pool of candidate structures, a set of mutations is chosen. The set of mutations is constructed in a particular way to maximize the expected information gain from the experiment. Depending on the outcome, it should be possible to unambiguously choose one structure.

7. M&M Experiments
   The mutants are synthesized and the SHAPE-Experiments are performed. This step has to be performed in the wet lab.

8. Basepair inference and model building - Section 3.5
   Basepairs are inferred from the experimental data of the Mutate&Map experiment. Comparison with the set of candidate structures determines the structure most concurrent with the experimental data and thus the most probable secondary structure *in vitro*.

3.2 OBTAINING CANDIDATE STRUCTURES

An RNA's structure space is typically vast, but just few structures are of biological relevance. For our optimization workflow, we want to narrow down the this huge number to few but biologically and thermodynamically relevant structures.

Acquiring a pool of structures is no obstacle due to the availability of secondary structure prediction programs as RNAfold. Therefore, the main difficulty lies in the reliable clustering of structures which could match our *SHAPE* -experiment. Since clustering of data is a well explored field, there exist many different approaches. Not all of them turned out to yield productive solutions (k-means clustering, DIANA-Divisive analysis clustering), so the possibilities were narrowed down to two most promising ways. The first approach, being more intuitive of both, clusters structures based on structural similarity. Approach two is based on RNA folding landscapes.

3.2.1 *Clustering through Abstract RNA Structures*

The program RNAshapes abstracts a secondary structure to a so called shape which represents the nesting pattern of its loops but ignores loop and helix lengths. By this definition, structures are similar to each other if they happen to have the same shape. As each shape can be represented by the structure with the lowest free energy among all structures sharing the same shape ("Shrep"), the choice of an adequate representative structure for a given shape is pretty straightforward.

```
1       $ RNAshapes -i 10000 -t 3 < Sequence.
           fasta
```

RNAshapes is instructed to sample 10000 structures by stochastic sampling, calculate each shape at abstraction level 3 (see Section 2 for detailed information on shape levels) and to output a list of all shapes, their corresponding shreps and shape-probabilities. Out of that list just shape/shrep pairs are kept,

whose probability is above a cutoff probability of 0.01. This cutoff is not arbitrarily set, but based on the observation that less probable structures always are eliminated by subsequent filtering steps as *SHAPE* -pattern simulation. Even if not eliminated here, low-probability structures consistently score very low in the final optimization step. Over many runs, it has never occurred that a low-probability structure was identified as best match for the Mutate&Map data.

After elimination of all improbable structures, the remaining shreps are saved as a list which is used in the following steps.

3.2.2 *Clustering through barrier trees*

A barrier tree visualizes the local minima of an RNA's folding landscape. Depending on the height of the energy barriers to its adjacent minima, a secondary structure will be more or less likely to be populated. As a minimum becomes more likely to be populated, using the conformations of the lowest minima should give a representative sample of which structures can be encountered *in vitro*.

From an RNA sequence, RNAsubopt generates all valid possible structures up to a certain energy above the MFE, hence allows us to explore parts of or even the complete folding space of a sequence. A set of suboptimal structure allows analysis of folding kinetics.

```
1       $ RNAsubopt -s -e E -T 24 -s --noLP <
          Sequence.fasta > subopt.out
```

In the wet lab SHAPE-experiments are commonly performed at room temperature, hence the argument $-T = 24$ sets the temperature to 24 °C. The argument -e sets the height of the energy barrier to the value E. Since structure space grows exponentially with E, it has to be chosen in a way that the resulting set of suboptimal structures still allows calculation of a barrier tree, while being of manageable size. As a first measure, the option -noLP restricts the output to structures where no base pair is without any neighbouring base also being paired. Reason for this restriction is that structural space grows tremendously when lone basepairs are included to our calculations but contribute little structures of interest, as they mostly are energetically unfavourable.

In a further attempt to shrink the amount of output structures to a manageable pool, the energy E is set to a value where not

more than 10 million structures are found. Due to the nature of the RNAsubopt algorithm, it is not possible to know *a priori* how many structures can be found at a given energy. Therefore the energy barrier has to be determined iterativly for each sequence by running RNAsubopt with a small E, counting the structures and re-running with a gradually increasing E, until the number of structures reaches the desired maximum.

The output of RNAsubopt is usually unsorted. Fortunately, the -s option provides in-RAM sorting. However, this is only a viable option if there is enough memory available. Since the output can quickly grow into hundreds of GB or even TB of data, sorting directly inside the RAM is not an option for even well equipped workstations. In case no measures, as the above mentioned iterative procedure, are undertaken, the option -s is be omitted and all output is written to the hard-disk where it can be sorted by Unix' sort-command.

```
1       $ sort -k2 -n subopt.out -T tmp-sorting
            > sorted.out
```

Once sorted by energy, with the most negative at the top, *barriers* calculates the rate matrix and the barrier tree.

```
1       $ barriers -G RNA-noLP --bsize --max
           300 --saddle -M noShift --minh=0.5 <
            subopt.out > barriers.out
```

-G RNA-noLP prohibits lone-Basepairs, since they rapidly expand structure space but occur seldom in nature. The argument -moves=noShift prohibits shifting-movements (a type of move where two paired stretches "slide" along each other), –bsize tells barriers to print out the size of each basin, and –max 300 restricts barriers to the calculation of the 300 lowest local minima. Due to the option –minh=0.5, a barrier separating two basins has to have a height of at least 0.5 kcal/mol, reducing the final number of basins. From the barrier tree, the structures sitting at the lowest point of each local minimum are selected as representatives, and saved in a list.

Since the exploration of a folding landscape is very resource-demanding, it should be noted that this approach is restricted to sequences with at maximum 150 nucleotides.

The central candidate structure filtering step of the whole workflow revolves around evaluating whether a structure is able to produce a SHAPE-pattern which resembles the wet lab experiment. Since experimental data for candidate structures is not available, modeling SHAPE-experiments *in silico* is of vital interest. In addition to bypassing the wet lab, comparison of simulated and measured data also may provide insights into the mechanisms influencing SHAPE-Reactivity.

The challenge in modeling SHAPE-experiments lies in the largely unexplored relationship between computed parameters as secondary structure and Reactivities. Therefore, simulation approaches are not only built on the basis of computational predictions, but also derived from empirical data.

3.3.1 *Simulation from Empirical Data*

Currently, the number of published SHAPE experiments of sequences with known structures is steeply rising, opening the possibility to set up reactivity distributions for recurring structural motives. As the mechanisms behind SHAPE are not yet fully understood, the challenge is to pick an representative set of distributions which approximates the experimental results the most.

Empirical distributions already have been successfully used for the modeling of SHAPE-data [35]. There, three different distributions were set up, depending whether a nucleotide was unpaired, inside or at the end of a stem.

In analogy to previous work, three distinct sets of empirical distributions were investigated. Basis for the construction of the distributions are sequences with known structures and available probing data. The used experimental data was taken from EteRNA, a RNA crowd sourced design project [24]. Only sequences whose probing pattern is consistent with the predicted secondary structure were used.

The following sets of distributions were modeled:

- Paired/Unpaired
  All nucleotides are partitioned into two categories, one for paired, another for unpaired bases.

- Stacked/Helix-End/Unpaired
  Nucleotides are grouped depending on whether they are

unpaired, paired inside a stem, or paired and at the end of a stem.

- Loopbased
  Nucleotides are grouped depending on which loop type they are part of. This discriminates between paired nucleotides or unpaired nucleotides part of Hairpin-, Interior-, Bulge-, Multi-, or exterior loop.

The partitioned data then was used to model a probability density function with the use of Kernel Density Estimates. The Perl module *Statistics::KernelEstimation* offers a convenient framework for this purpose. Out of all three approaches, the ternary model (Stacked/Helix-End/Unpaired) approximated the experiment sufficiently and was subsequently used. The binary model (Paired/Unpaired) tends to oversimplify things, whereas the loop-based approach requires the setup of many distributions but offers no significant increase of simulation quality compared to the ternary model.

Kernel Density Estimates

Kernel Density Estimates (KDE) provide a way to estimate the probability density function of a random variable. From a set of data points $x_1, x_2, .., x_n$ the Kernel density estimator $\hat{f}_h(x)$ approximates the "hidden" probability density function $f(x)$ by the superposition of *Kernel functions*. Formally it is defined as:

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{x - x_i}{h}\Big), \tag{9}$$

where $K(x)$ is a kernel function, and h a smoothening parameter, also called the kernel's bandwidth. The kernel function can be chosen from a variety of model functions. In this case the Gaussian kernel was used,

$$k(t) = \frac{1}{\sqrt{2\pi}} exp(-\frac{1}{2}t^2) \tag{10}$$

which is essentially a normal distribution.

For Gaussian Kernels, the bandwidth h can be estimated by "Silvermans rule of thumb":

$$h = \left(\frac{4\sigma^5}{3n}\right)^{\frac{1}{5}} \approx 1.06\sigma n^{-1/5}, \tag{11}$$

with σ being the samples' standard deviation.

3.3.2 *Simulation from Basepair-probabilities*

Calculation of the partition function of a sequence allows us to access the probabilities of each possible base pair (i, j). For each position i, the sum $\sum_j(i, j)$ represents its overall probability of being involved in a base pair. As a nucleotides' *SHAPE* -reactivity is predominantly governed by base-pairings, a simple model is built which directly correlates reactivity and probability-to-be-paired in a linear equation:

$$R_i = p_i R_{max} \tag{12}$$

where R_i is the calculated reactivity for position i and p_i the computed probability of being paired. The maximum reactivity value R_{max} has to be be chosen according to the normalization technique in use, it is usually set to 2.0 for box-plot normalized data sets. Inspection of probing data (see section 4.1) shows that paired positions frequently have residual reactivity, as well as paired residues often do not exclusively react as weak as assumed. In order to take into account for this deviating behaviour, equation 12 is reshaped:

$$R_i = p_i R_{max} + (1 - p_i) R_{min} \tag{13}$$

With R_{max} and R_{min} are average reactivities of nucleotides with 100% or 0% pairing probability, derived from a set of reference experiments.

3.3.3 *Comparison of SHAPE-patterns*

After simulation of SHAPE experiments, it is necessary to assess the quality of simulated probing patterns. As there is experimental data available, evaluating the similarity between simulation and experiment offers a good measure of quality.

RMSD

The Root Mean Square Deviation (RMSD) is commonly used to measure the differences between values predicted by a model and the actual observed values. It is defined as:

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}. \tag{14}$$

With $\hat{y}_i$ being the experimental observation and y_i the corresponding simulated value. In context with RNA, $i = (1; 2; ..; n)$

with n being the last nucleotide. The RMSD offers the advantage of summing up differences in a single number. Even though easy to understand, when used for SHAPE patterns, little information is given on the nature of the differences between simulated and experimental pattern. On top of that, RMSD values can be strongly biased by outliers, an often encountered issue with SHAPE-data. It is advised that RMSD values are either interpreted with caution or only used after rigorous (manual) inspection of the raw data.

Coarse grained Scoring

In general, SHAPE reactivities above a threshold value are considered to indicate high flexibility and vice versa for values below. Therefore, when comparing two reactivity values, the question whether both lie roughly in the same reactivity range is more interesting than the absolute difference in numbers.

For comparison, all positions are marked according to their simplified reactivity values, which are marked as low if < 0.25, or high if > 0.25. Given two probing patterns s_1 and s_2 with size N, where $s_1[n]$ refers to the pattern's n-th position, a similarity measure based on the comparison of coarse grained reactivity values is defined as following:

$$\text{Score} = \frac{\sum_{n=1}^{N} \Delta_n}{Score_{max}}, \quad \Delta_n \begin{cases} 1 & \text{if } s_1[n] = s_2[n] \\ 0 & \text{if } s_1[n] \neq s_2[n] \end{cases} \tag{15}$$

As coarse grained reactivities are compared, outliers no longer can heavily bias the whole score as their contribution has not more weigh than all other nucleotides.

3.4 DIRECTED MUTAGENESIS

At the core of the optimization procedure, sites of mutations are picked such that the outcome is rich on information, without the need to survey all positions. Based on combinatoric and energetic considerations, we now want to direct mutations in a way that the expected gain of information is maximized. When a base is mutated, it is always exchanged with its complementary base. This reduces the number of possible mutations and increases the expected accuracy, as exchanges to other nucleotides have been shown to score slightly worse [36].

Mutation guidance was implemented in two major ways, which are both explained in this section. Furthermore, the

interpretation of the experimental results also requires close attention, thus is explained separately.

3.4.1 *Combinatorial unambiguous solutions*

As our goal is to reduce the number of necessary mutations, we want to identify the *smallest* combination of mutations that unambiguously discriminates for one structure out of a set of candidate structures. A mutation at position i discriminates between two structures if for both structures, i is paired to different nucleotides.

The problem is solved if every structure can be distinguished from each other in the set. Given our set of secondary structures, we want to find the set of mutations which discriminates unambiguously for exactly one structure.

Table 3 explains the combinatorial method for a simplified set of candidate structures. Following the point-mutation of a single nucleotide, structures which can be distinguished from each other are marked. In this example, for every point mutation there will be two or more structures left which can not be distinguished from each other.

In the example case, mutation of positions 3 and 7 are not sufficient for a full identification, when considered alone. However, when taking both experiments into account *together*, which can here be imagined as simply summing both matrices, no structure will be left undistinguished (e.g. marked with "0").

Structures

I	.((..)).
II	..((..))
III	((..))..
IV	((()))..
V	......()

Mutpos = 3

	I	II	III	IV	V
I	x	1	0	1	0
II		x	1	1	1
III			x	1	0
IV				x	1
V					x

Mutpos = 7

	I	II	III	IV	V
I	x	1	1	1	1
II		x	1	1	1
III			x	0	1
IV				x	1
V					x

Table 3: Left: Set of five simplified candidate structures, Middle: Identification status after mutation of position 3, Right: after mutation of position 7. A "1" indicates that the structures can be distinguished from another, otherwise the position is marked 0. As a structure has not to be distinguished from itself, respective cases ar marked with an x.

The proposed solution (Mutated positions: 3&7) is not the only valid solution, there also exist other valid combinations as positions 7&4, 3&8, 2&3, or 3&6. While the solutions here were found manually, this task becomes difficult as RNA lengths and set sizes grow. The major focus of finding a solution therefore lies in the efficient search for solution in the myriads of possible combinations of mutation sites.

Depth-first search

For an RNA of length N, there are

$$\sum_{m=1}^{N} \binom{N}{m} = 2^N - 1$$

possible unique combinations of mutations. As typical sequences of interest are hundreds of nucleotides in length, brute-force enumeration of all possible combinations is not a viable option. Therefore, the combinatoric space is explored by a depth-first search.

Let S be a set containing all N nucleotides $[s_1, s_2, .., s_N]$. We now want to find the *smallest* subset $S' \in S$, which is able to uniquely identfy every structure if all mutations $s' \in S'$ are performed. For that purpose, the resulting combinatorial space is searched in a depth-first manner, which comes with the advantage that an unsolvable problem can be immediately identified once the search reaches maximum depth.

In order to speed up the search, the set of possible mutation sites is filtered. Positions which are unpaired in every structure, as well as positions that form the same basepair in all structures are excluded from our considerations, as they would provide no information. The remaining nucleotides are ranked by how many unique basepairs they participate in all structures. When a layer of the search tree is sampled, combinations containing high ranking positions are preferentially sampled.

3.4.2 *Non-combinatorial Mutation Picking*

The combinatorial approach is based on the assumption that the mutation of a base always reveals its basepairing partner. This assumption is sometimes not completely valid, as there are often unavoidable experimental inaccuracies when inferring basepairs from data (see 19). Therefore, the combinatoric idea

approach is made less tight and mutations are governed by a positions probability to be paired.

In a heuristic scheme, out of all positions just nucleotides with a probability to be paired $pp_i > 0.5$ are considered valid targets. Additionally, we want to place mutations preferentially inside stems, as mutations at the end of stacked regions frequently display fewer positive signals. For all structures in our candidate set S, each position i is assigned a weight w

$$w_i = pp_i \sum_S x_s \begin{cases} 1 & \text{if } s_i \text{ at Helix-End} \\ 2 & \text{if } s_i \text{ Inside Stem} \end{cases}, \tag{16}$$

ranking first residues which are likely to be paired and participate in as many stacked basepairs as possible. Depending on the size of the Mutate&Map experiment, the highest ranking residues are suggested as mutation site .

3.5 SCORING OF MUTATION SETS

After the reduced set of mutations is performed *in vitro*, it is necessary to identify which proposed structures are confirmed by the experimental data and which structures show strong discrepancies. For that purpose, two heuristic procedures are designed.

Z-Scores

For Mutate&Map experiments, SHAPE reactivity data is converted to Z-scores, which serve as the basis for interpretation of any experimental results[36]. For an RNA of the length N, let the observed position reactivities be s_{ij}, with $i = 1, 2..N$ indexing the position numbers and $j = 1, 2..M$ indexing the mutated nucleotides. The Z-scores Z_{ij} are defined as

$$Z_{ij} = \frac{(s_{ij} - \mu_i)}{\sigma_i}, \tag{17}$$

with mean position intensities μ_i and standard deviations σ_i being computed as

$$\begin{aligned} \mu_i &= \frac{1}{N} \sum_{j=1}^{M} s_{ij} \\ \sigma_i &= \left(\frac{1}{M} \sum_{j=1}^{M} (s_{ij} - \mu_i)^2 \right)^{1/2} \end{aligned} \tag{18}$$

Only data with $Z_{ij} \geq 0$ and $\mu_i \leq 0.8$ are kept, as the Mutate&Mapapproach seeks to detect site specific release of nucleotides which are protected in most sequence variants.

3.5.1 *Combinatorial Scoring*

In the combinatorial site directed mutagenesis approach, mutations that maximize information gain are chosen out of purely combinatorial concerns and no thermodynamic considerations are made. The corresponding evaluation procedure closely follows this thought.

Given a list of mutations M and a set $\mathbf{S}$ of candidate structures, for each mutation m it is checked whether it is confirmed by experimental data. A mutatated residue i is considered to confirm a structure if the residue is part of a basepair (i, j) and its Zscore $\mathbf{Z}_{ij} \geq 1.0$. On the opposite, if the site of mutation is predicted to be unpaired, $|\mathbf{Z}_{ii}| \leq 1.0$, since mutation of unpaired residues should not change its flexibility.

If a secondary structure violates just one condition, it is removed from the set of candidates. In theory this should leave us in the end with just one structure. In practice, this evaluation scheme is far too strict, as experimental errors frequently occur in SHAPE experiment. Therefore, this evaluation scheme remains of entire hypothetic nature.

3.5.2 *"Soft"-Scoring*

Since a structure is immediately ruled out in the combinatorial-model at the first discrepancy between prediction and experimental data, it is critically affected by experimental errors. This is problematic, mainly due to frequent events where mutations of a residues do not induce local perturbations but a large-scale rearrangements in the secondary structure. Also, experimental errors resulting in irregular high or low reactivities may hide signals from induced mutations or suggest false positive basepairs.

The effects of a Mutate&Map experiment can be ideally reduced to two possible outcomes. Either the target nucleotide is unpaired and remains that way upon mutation, or a basepair is perturbed, resulting in increased reactivity at both involved nucleotides. Sometimes the signals from the released former pairing-partner are somewhat weak or covered up by local refolding events. Eventhough it is not clear which base this

residue was paired to, the increase in reactivity makes it safe to assume that this residue was involved in any kind of basepair. Therefore, nucleotides which are predicted to be paired and show increased reactivity for the site of mutation but not at the predicted pairing partner, are scored separately from others.

According to these outcomes a scoring scheme is designed, which does not eliminate contradicting structures, but merely "rewards" structures that contain basepairs which are confirmed by experimental evidence. In this heuristic evaluation scheme, a (candidate) structure is rewarded a bonus for each mutation whose effect is consistent with its expected effect on the respective structure.

Given a list of mutations M and a set of candidate structures S, the set $\mathbf{P}_s$ contains all basepairs (i, j) found in S. Since we want to identify significant relative changes in reactivities, all experimental values are compiled into a Z-score matrix $\mathbf{Z}$ (Equation 17), where $\mathbf{Z}_{ij}$ refers to the Z-score of residue i upon point mutation of residue j.

$$\begin{aligned} \text{Score}(s) &= \sum_{i \in M} z_i \\ z_i &= \begin{cases} 1 & \text{if } i \notin \mathbf{P}_s \wedge \mathbf{Z}_{ii} \leq c \\ 3 & \text{if } i \in \mathbf{P}_s \wedge \mathbf{Z}_{ii} \geq c \\ 10 & \text{if } (i,j) \in \mathbf{P}_s \wedge \mathbf{Z}_{ii} \geq c \wedge \mathbf{Z}_{ij} \geq c \end{cases} \end{aligned} \tag{19}$$

The cutoff value c is set to 1.0 and determines how large or small Z-scores have to be before being considered statistically relevant. Each case is weighed differently, with correct basepairs being the most critical as we want to primarily promote structures whose basepairings are the closest to the experimental data.

After evaluation of the whole mutation set for each structure, are ranked with the highest score being the closest to the experimental data. The best structure is presented as solution for for the given mutation-set, eventhough other high-ranking structures mitght also be of interest.

3.5.3 *Visualization of probing efficency*

Closely following the before mentioned evaluation scheme, the quality of a Mutate&Map-experiment is visualized by a secondary structure plot with a special colouring scheme. Each base is coloured according to the outcome of a *SHAPE* -experiment

where it is mutated. Unpaired nucleotides are expected to remain highly reactive upon mutation. Paired nucleotides are expected to show increased reactivity, ideally along with their former pairing partners. On a secondary structure as template, each nucleotide is coloured as follows:

- **Red**:
  - Paired:
    Mutation of this nucleotide led to increased reactivity (Zscore > 1.5).
  - Unpaired:
    Mutation of this nucleotide led to no change in reactivity.
- **Blue**:
  - Paired:
    Mutation of this nucleotide led to increase in reactivity (Zscore > 1.5) at this position *and* at the nucleotide it is paired with.
- **Green**:
  - Paired:
    Mutation led to no significant increase in reactivity.
  - Unpaired:
    Mutation led to a decrease of reactivity (Zscore < -1.5).

The resulting coloured secondary structure (Example: Figure 12) allows the easy and comprehensive assessment of a Mutate&Map-experiment's quality on a given reference structure. Furthermore it is possible to quickly identify stretches of good performance from visual inspectation alone.

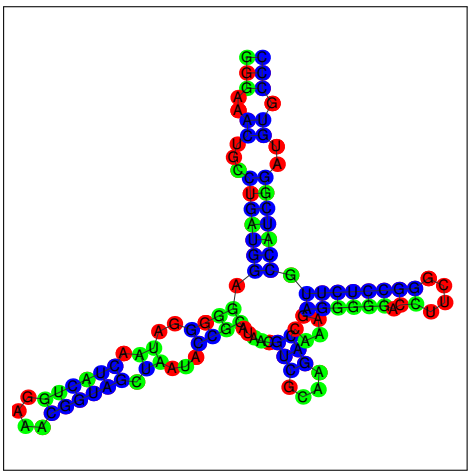


Figure 12: Performance of the Mutate&Map approach visualized on the reference structure of the four-way-junction of *E. coli* 16S rRNA

3.6 SEQUENCE-BLIND BASEPAIR INFERENCE

Kladwang et. al. developed an approach which allows the isolation of statistically probable basepairs from probing data without using any sequence information [36]. Since the Mutate&Map seeks to identify the release of nucleotides upon mutation, not absolute reactivity values, but the change in reactivity for each base in context with all other experiments is of primar interest. Therefore the reactivity values for each residue across all experiments are converted to z-scores. The Z-score measures by how many standard deviations a data point deviates from the average of the whole sample and is defined as:

The Z-score of a residue i of construct j is defined as the difference of each signal from its mean at that residue,divided by the standard deviation of intensity at that residue.

$$Z_i j = \frac{x_j - \mu_j}{\Sigma_j} \tag{20}$$

In order to qualify as part of a basepair, a residue has to fullfill a set of criteria:

1. Mean accessibility
   Residues that show high reactivity already in the majority of experiments are not expected to yield any significant information upon mutation. Here, the cutoff value was set to 0.8.

2. Z-score
   Residues released by a mutation give a significantly higher reactivity than the mean reactivity for this residue. Those events are identified by an above-average Z-score (here: > 1.5).

3. Sequence separation
   Signals that occur less than 3 residues away from the site of mutation are discarded regardless of their Z-score, since Watson-Crick basepairs typically do not occur over such short distances.

4. Punctuate pattern
   Ideally, mutation of a residue only releases its base pairing partner, not its neighbours. Therefore, for a true basepair the residue is required to have at least twice the Z-score of its neighbouring as well as its next-neighbouring residues.

5. Punctuate pattern across constructs
   The mutation of a residue's base pairing partner should affect its chemical accessibility, in contrary to mutations at nearby residues. Therefore, the Z-score of a true signal should be at least twice the Z-score at the same residue induced by the previous and next mutation in the library.

6. Supporting signals Each true basepair should be supported by at least one other signal. One possible confirmation of a basepair (i, j), identified by mutation of i, is derived from mutation of j or constructs where the same residue i was mutated, but to a different base. An alternative kind of supportive mutation comes from the observation that basepairs usually are located inside of stacks, but seldom occur as isolated basepairs. Thus, the requirement is that any signal at (i, j) is supported by another signal at $(i-1, j+1)$ or $(i+1, j-1)$, indicating a stack of two basepairs.

7. Noisy residues
   Residues that show irregular high Reactivities are removed from the set. The reasons for such outliers are caused mostly by wrong data processing or other experimental artifacts as polymerase stopping

4

RESULTS AND DISCUSSION

The following chapter presents the results which were obtained in this work. The results are grouped into two parts. The first part presents analyses and statistics concerning *SHAPE* -experiments, in order to provide a more comprehensive picture of the behaviour of this experimental technique. The second part continues with a focus on the performance of the optimized Mutate&Mapprotocol.

If not explicitly mentioned, the dataset used for all analyses consisted of 500 unique probing experiments from the EteRNA-project [24], rounds 69-77. All used sequences were probed with 1M7 and reached at least 90% of the maximum possible EteRNA-score, indicating that the corresponding predicted structure is very close to the secondary structures found *in vitro*. Not all used sequences reached maximum score, since the number of optimal designs is too low for the calculation of meaningful statistics. Also, the EteRNA-scoring scheme is not perfect, so we wanted to provide a certain frame of tolerance.

Probing experiments are primarily used to discriminate between paired and unpaired bases, sometimes also between helix end and stacked basepairs. As the loop-based energy model is an integral part of every secondary structure prediction approach, this analysis investigates whether *SHAPE* data shows specific behaviour in different looptypes. For each loop type the distribution of reactivities is analyzed and visualized by box-plots.

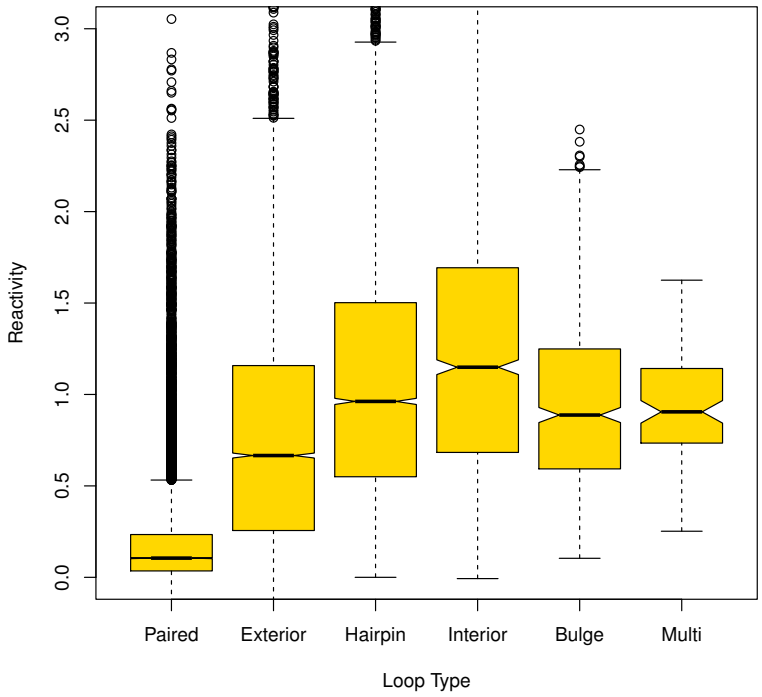


Figure 13: Box-plot graphs showing reactivitiy distributions of nucleotides part of various looptypes; Unpaired refers to nucleotides not part of any loop. Sample sizes: Unpaired/Paired - 10000, Interior loop - 8000, Hairpin loop - 5000, Multi/Bulge loops - 600

Most notable about this overview analysis is the clear difference in general reactivity between Paired and all other unpaired nucleotides, which actually is consistent with the model of how *SHAPE* -reactivity is caused. Interestingly, for paired nucleotides the box-plot shows a very narrow distribution of reactivities, whereas all unpaired nucleotides spread over a much wider range.

A large quantity of outliers can be seen with Paired/Unpaired nucleotides. Irregularly high nucleotides are commonly found in *SHAPE* -data and may occur due to events as polymerase-stopping [25] or highly reactive geometries, whose exact interactions with *SHAPE* -reagents are yet poorly understood [18]. In addition, the box-plot may suggest that outliers are a common feature rather for paired or unpaired (not part of any loop) nucleotides, but here this is mostly due to sample sizes being significantly larger for those groups.

Nevertheless, there is a significant number of datapoints whose reactivities do not fit into their respective group. For all unpaired nucleotides, the lowest quarter of datapoints lie in a reactivity range < 0.5, which is typically more associated with constrained positions. The reasons for this behaviour might be of systematic nature, e.g. artifacts created by data processing steps as electropherogram readout or background subtraction, leading to general lower reactivity values. On top of that, it has to be taken into account that all predicted secondary structures on which this analysis was performed, can be subject to errors. The selection step via EteRNA-scores ensures congruence of structure prediction and experimental data, but as it is not applied too tightly to prevent the case where only data is shown that matches our model, still leaves room for interpretation.

There is also a clear trend in reactivity change among the other looptypes. Combined with the observation that unpaired nucleotides react with a broad variety of reactivities, the nucleotides of each looptype are examined more closely.

Hairpin loops

For all hairpins, reactivity distributions of paired and unpaired nucleotides directly adjacent to the closing basepairs were investigated. In the recent Turner energy parameters, small hairpins of size 3 and 4 are parametrized separately, indicating they are likely to be affected by noncanonical interactions and therefore are analyzed separately.

5′	(	(	.	.	.
	0.10	0.27	0.90	0.51	0.65
	0.14	0.21	0.71	0.78	0.74
3′	(	(	.	.	.

Table 4: Median reactivities of bases at 5′- and 3′-ends of Hairpin loops (N=600)

Notable is that nucleotides of the closing basepair show elevated reactivity, as well as the first unpaired position at the 5′-end, which is more reactive than all other unpaired bases.

5′	(	.	.	.	)	3′	
	0.22	1.52	0.61	0.91	0.19		
5′	(	.	.	.	.	)	3′
	0.27	0.93	0.47	0.58	0.91	0.13	
5′	(	.	.	.	.	.	) 3′
	0.09	0.39	0.84	1.00	1.04	0.60	0.15
5′	(	.	.	.	.	.	.) 3′
	0.15	1.18	1.96	1.25	1.53	1.41	1.61 0.29

Table 5: Median *SHAPE* -Reactivities for all positions in Hairpin loops of Size 3, 4, 5 and 6. (N=50-100)

Even though it is difficult to figure out a clear overall trend, nucleotides are less constrained in longer loops. Hairpin loops of size 3 and 4 show slightly less reactivity, which is mostly due to non-canonical basepairs.

Interior loops

Small symmetric interior loops are often involved in non-canonical basepairs, and are also parametrized separately in the most recent set of energy parameters [37]. Therefore not only the 3′/5′-regions of interor are analyzed, but also symmetric interior loops of size 1,2 and 3.

5′	P	P	.	.	3′
	0.28	0.33	1.05	1.68	

3′	.	.	P	P	5′
	1.6	1.13	0.32	0.22	

Table 6: Median reactivities of 5′- and 3′-regions of interior loops (N=150)

5′	P	.	P	3′
	0.17	0.52	0.16	
	0.11	0.53	0.18	
3′	P	.	P	5′

Table 7: Median reactivities of symmetric interior loops of size 1 (N=9)

5′	P	.	.	P	3′
	0.21	0.49	0.47	0.36	
	0.23	0.56	0.50	0.24	
3′	P	.	.	P	5′

Table 8: Median reactivities of symmetric interior loops of size 2 (N=19)

5′	P	.	.	.	P	3′
	0.17	0.49	1.25	0.82	0.19	
	0.29	0.70	1.33	0.47	0.17	
3′	P	.	.	.	P	5′

Table 9: Median reactivities of symmetric interior loops of size 3 (N=31)

The general analysis of interior loops (table 6) shows no special behaviour. As with other loops, a nucleotide's reactivity increases with the distance to the next paired base e.g. loop size. This can be seen well in the analysis of special loops. Compared to the general analysis, especially (symmetric) interior loops of size 1 or 2 show significant lower reactivity, suggesting that they are structurally constrained even if they do not participate in any (canonical) basepair.

5′	P	.	P	3′
	0.30	1.27	0.27	

5′	P	.	.	P	3′
	0.33	1.64	2.02	0.48	

Table 10: Median reactivities of Bulges of size 1 (N=50) and 2 (N=16)

5′	.	.	.	P	P	3′
	1.05	0.96	0.73	0.14	0.06	

5′	P	P	.	.	.	3′
	0.05	0.18	1.10	1.00	0.75	

Table 11: Median reactivities of 5′/3′ environments in Multiloops (N=269)

Eventhough bulge-loops technically just are special cases of interior loops, analysis reveals drastically different behaviour. There is a steep increase in reactivityin bulge-loops of size 1 and 2, compared with symmetric interior loops of equal sizes (Table 8). This difference is mainly associated with symmetric interior loops, since the overall reactivityof nucleotides in interior loops is considerably higher (figure 13) compared to bulges. A remarkable difference of bulges to all other loop types are the high levels of reactivityfound at nucleotides directly adjacent to the closing basepair(s), since in most other cases reactivityincreases with the distance to the closing basepair. Surprisingly, the closing basepair itself exhibits no striking difference in reactivity, suggesting that local flexibility may be not the only parameter governing the *SHAPE* -Reaction. It can be imagined that respective nulceotides favour conformations which are generally associated with high reactivities [18]. In multiloops no special features can be seen, except for a slightly reduced overall reactivity, as already was seen in the overview picture (Figure 13).

Although the SHAPE-technique has been established for a while, the exact underlying mechanism still has to be characterised conclusively. In previous work it was shown that not only the local flexibility of the backbone, but also the nucleobase itself can catalyse the 2'-acylation [18]. Influences decoupled from the influence of backbone flexibility, such as intrinsically different interaction of purine nucleotides with *SHAPE* -reagents [17], introduce systematic errors to each probing experiment, making it impossible to compare residue reactivities on a normalized scale without further corrections. This emphasizes the need to accurately characterize such influences, and to quantitatively describe them if necessary.

In previous work, analysises were carried out with a relatively small dataset comprising just a few crystal structures. With the availability of large quantities of high quality structure datasets via the EteRNA-project [24] and the introduction of new normalization techniques [38] to SHAPE, this offers a good opportunity to check whether the previous findings also hold true for newer and larger datasets.

For our analysis, the residues were divided into three groups, depending on their structural context: Unpaired nucleotides (U), nucleotides inside of stems (S), and nucleotides at the beginning or end of helices (F). This division makes sense, since it has been shown that the resulting distributions are unique and allow sufficiently accurate modeling of reactivities [35].

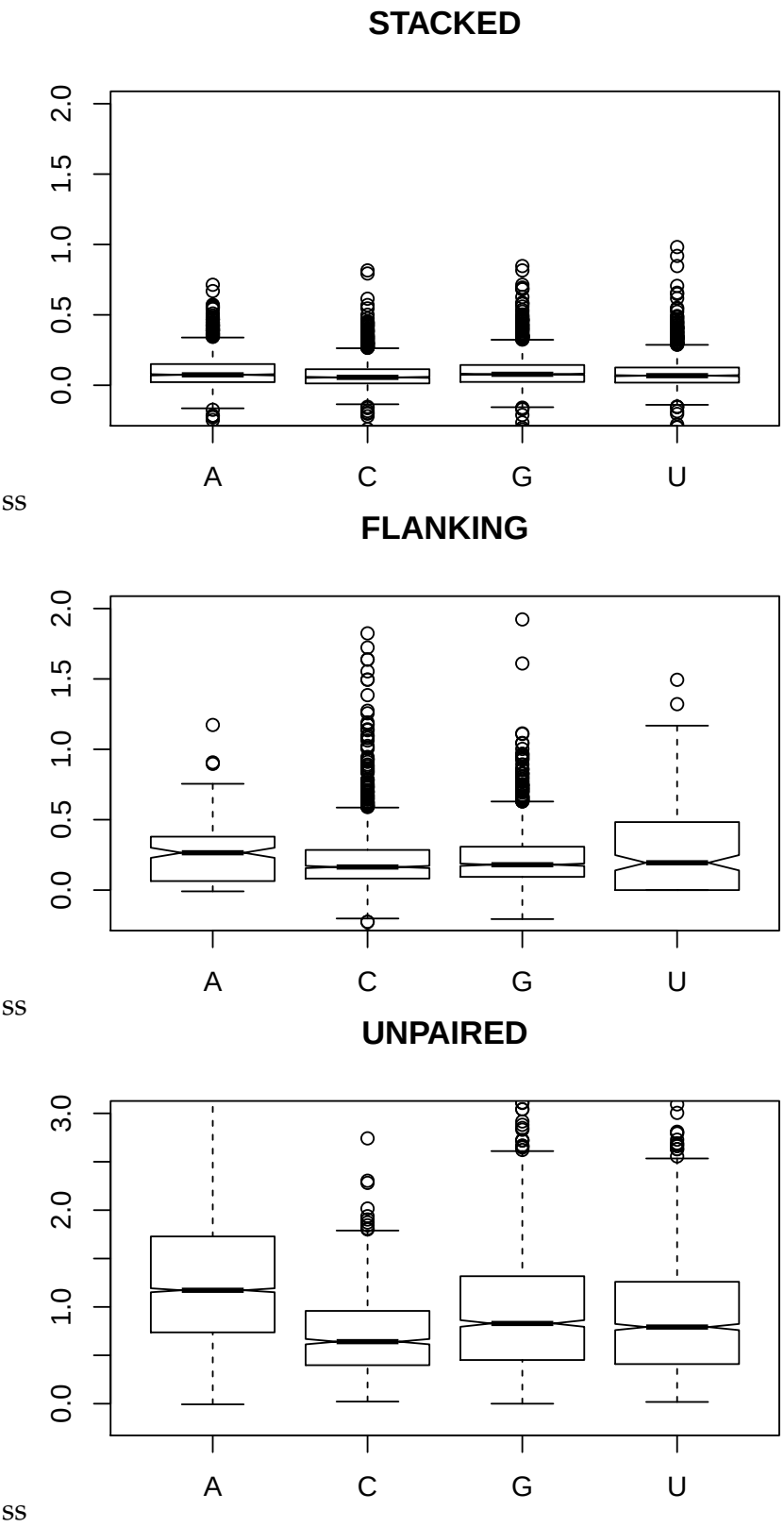


Figure 14: Box-plots showing distributions of *SHAPE* - Reactivities for Stacked, Flanking and Unpaired residues among all four bases.

Stacked residues display no significant difference in median reactivity among all four bases. This also holds true for flanking residues, eventhough the distribution of reactivities is wider, as indicated by size of the box and length of the whiskers. This makes sense, since helix-end residues are more likely to undergo structural change in vitro.

For unpaired residues, the distributions are significantly wider, reflecting the larger degrees of freedom the residues are experiencing. In contrast to the latter two sets, the general median reactivities show significant deviations for adenine and cytosine residues.

In some cases, bases that are predicted to not take part in any Watson-Crick basepairing, are involved in non-canonical basepairs or other interactions which are beyond the scope of current structure prediction algorithms. Nevertheless, in vitro such interactions have significant effects on the local flexibility of the backbone, and therefore would be directly influencing reactivity. Therefore for further analysis, the unpaired residues are more closely dissected into new groups, depending on what type of loop they belong to. There are four loop types: Hairpin loops (H), Interior loops (I), Multiloops (M), Bulges (B). If it holds true that looptypes are prone to errors in secondary structure predictions, there should be notable shift in the resulting median reactivities.

	A	C	G	U
H	1.29	0.70	0.91	0.90
I	1.33	0.85	1.10	0.78
B	1.31	0.33	1.51	0.88
M	0.91	0.38	0.72	0.97
All	1.04	0.41	0.65	0.76

Table 12: Median Reactivities of all four Bases, separately for only Hairpin loops (H), Interior loops (I), Multiloops (M), Bulges (B). Bottom row: Median reactivities for all unpaired nucleotides

Eventhough there are massive differences among single groups, no consistent general trend can be enforced. The data supports the current model, that base Identity plays a minor role in the emergence of reactivity, compared to the influence of local residue flexibility. Moreover, since reactivities are usually displayed on a normalized scale and interpreted in a more

coarse grained way, base identity there is no further need for the implementation of any correcting factors.

4.3 CORRELATION OF PROBABILITY-TO-BE-PAIRED AND REACTIVITY

Generally, *SHAPE* highlights unconstrained nucleotides, e.g. those not involved in any form of basepairing. Thanks to computational efforts, it is also possible to compute base-pair probabilities and subsequently the overall probability to be paired, giving a measure for the nucleotide's overall likelihood to be involved in a (canonical) basepair. Correlating said probabilities with reactivity-values therefore might serve as an interesting benchmark for `RNAfold`'s partition function folding. As in previous examples, the data set consisted only of sequences from the EteRNA-project, whose probing pattern showed maximal congruence (EteRNA-Score=100) with the MFE structures, as predicted by `RNAfold`.

As observed in previous work[33], correlation clearly is present, but still weak. Running calculations at a temperature matching experimental conditions (24C) gave results indistinguishable from `RNAfold`'s default temperature settings, indicating no need to adapt temperature in future calculations. With increasing *SHAPE* -Score comes a steep increase in correlation. While there is a steady trend upwards from scores 60 to 90, the two highest groups (90 and 100) show almost identical correlation coefficients. As *SHAPE* -experiments usually are prone to high error rates, it can be reasoned that sequences at score 90 already are at "peak efficency", and all true basepairs already have been found. The residual difference of 10 *SHAPE* -score are accounted for by (i) experimental errors or (ii) the aberrant behaviour of special looptypes which do not behave as the general model of reactivity would predict. Indeed, Prior analyses (see Section 4.1) have shown that certain loop types in fact are less reactive than it would be expected from unpaired bases. In any case, the fact that higher-scoring designs consistently exhibit better correlations between reactivity and Probability-to-be-paired emphasizes the validity of *SHAPE* -scores as a measure of quality.

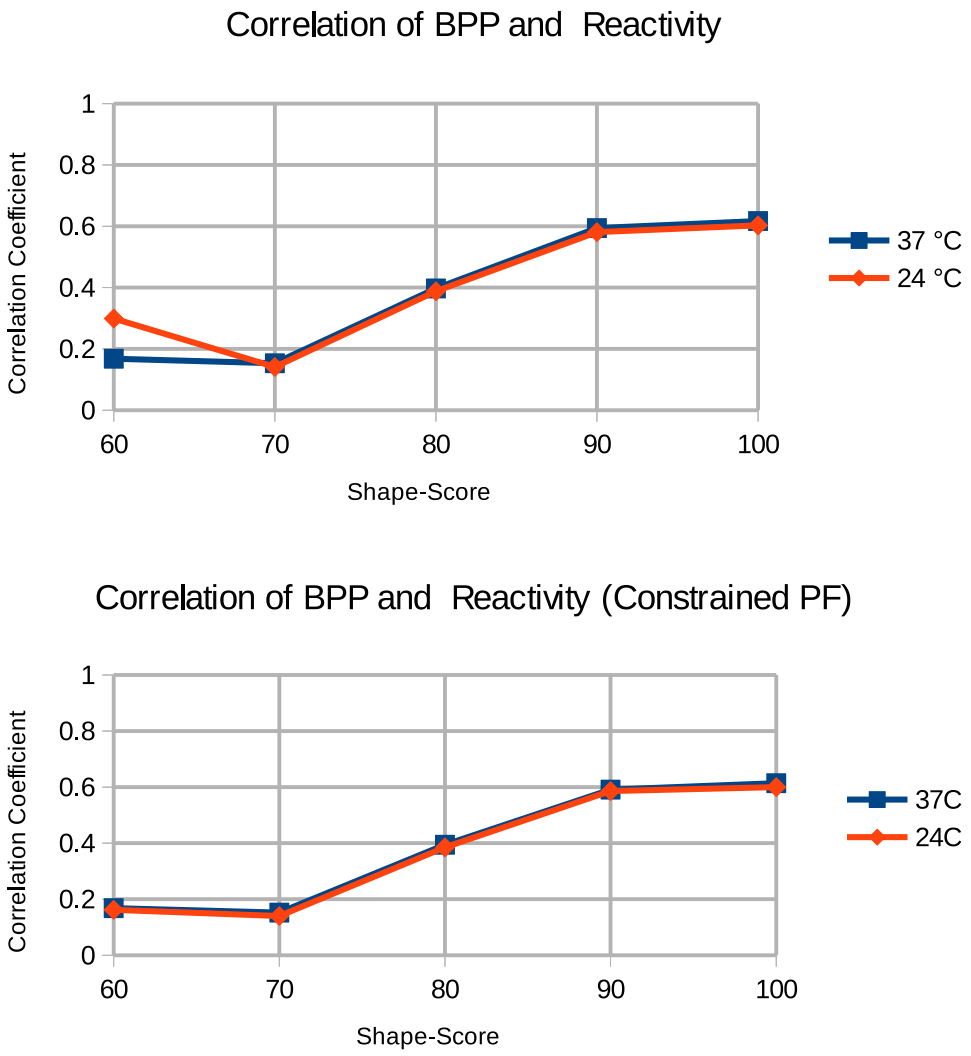


Figure 15: Change of correlation (Pearson) of Probability-to-be-paired and *SHAPE* -reactivity for sequences of different *SHAPE* -scores. Calculations were done at 37C (Default `RNAfold`) and 24C (experimental conditions). Top: Correlations after folding without constraints; Bottom: Same analysis, with *SHAPE* -reactivities used as constraints for `RNAfold`'s partition function calculation.

Surprisingly, inclusion of *SHAPE* -data as constraints into calculation of the partition function has almost no effects on overall correlation. The used version of `RNAfold` incorporates *SHAPE* -reactivities as described in *Deigan et al.* [31], where nucleotides in a stack receive a free energy bonus ΔG_{SHAPE} depending on their reactivity. Sequences selected via EteRNA-

score, are scored on basis of how similar their *SHAPE* -pattern is to the computed MFE structure [24]. If a structure scores high, the computed structure is considered to be very similar to the structure *in vitro* making it plausible that the further addition of energy bonuses will do little more than to confirm the already energetically favourable structure.

However, this does not explain why the correlations between normal and constrained calculations do not change for low scoring sequences, which should actually have room for improvement. Since constraints are only applied to stacks but not unpaired regions, a possible explanation might be that all basepairs in constrained ensembles are already present in the unconstrained ensemble. Further research is needed to explain this discrepancy, if possible also with other methods for constraint inclusion (Washietl et al. [33], Zarringhalam et al. [34]) since the Mathews-method not always yields meaningful results (Luntzer et. al, in preparation).

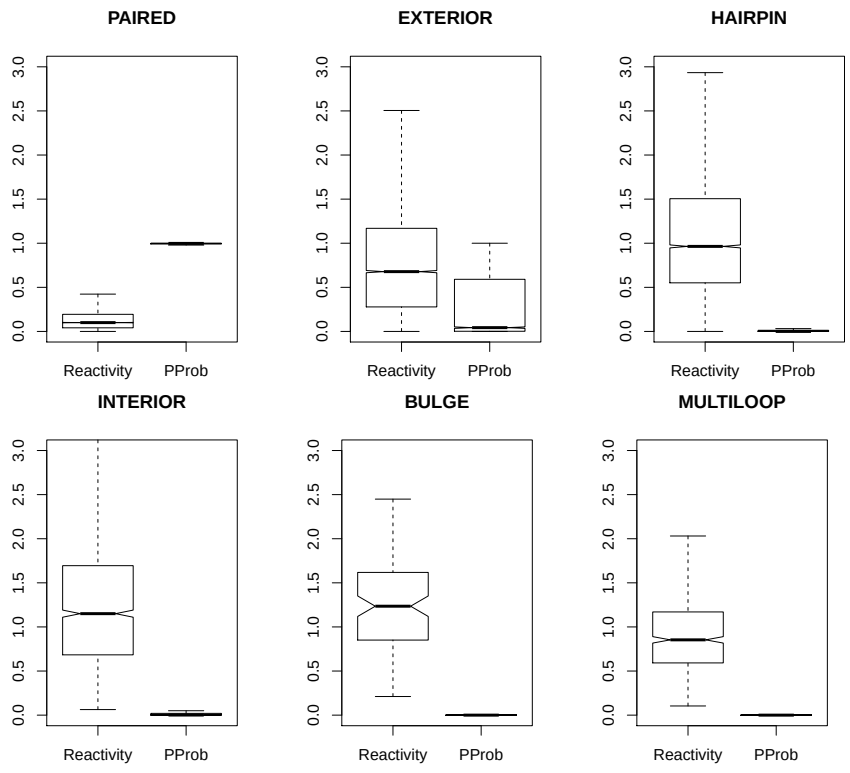


Figure 16: Distribution of reactivities and associated Probabilities-to-be-paired for every loop type. Data: All Sequences with *SHAPE* -score 100.

The in-detail analysis of the distribution of probabilities and reactivities type shows remarkable sharp distributions of pairing-

probabilities, in every loop except for nucleotides in exterior loops. Excluding nucleotides in exterior loops increases the correlation slightly (0.61 -> 0.69). As noted previously (section 4.1) unpaired nucleotides are scattered on a broader range of reactivities.

The EteRNA project[24], source of our sequence and probing data, seeks to design sequences which specifically fold into predefined structures. Despite the designs' constantly increasing quality, shown by better EteRNA-scores, the probing patterns of many sequences still indicate different structures in vitro. Besides flaws in the energy model or unpredictable behavior of *SHAPE* , explanation for those discrepancies could also be rooted in RNA folding kinetics.

In solution, RNA molecules typically do not populate only one state, but constantly undergo a change of conformations until equilibrium is reached. As *SHAPE* -probes continuously react with all RNA molecules present in solution, the outcome of a SHAPE-experiment has to be viewed as the sum of contributions of all structures present *in vitro* throughout the experiment. Therefore it makes sense to see the resulting *SHAPE* -pattern as a superposition of each structures' contribution. In order to avoid cases, where co-transcriptional folding might heavily interfere with the probing experiment, RNAs are usually denatured at high temperatures and refolded under favourable conditions prior to the probing experiment.

While eliminating the danger of structures being trapped in suboptimal conformations via co-transcriptional folding, denaturing and refolding does not guarantee that the molecule folds into its MFE conformation. Depending on its folding landscape, secondary structures might get stuck in deep local minima, resulting in a different population of states than thermodynamics alone would predict.

Visual inspection of all sequences' barrier trees shows that among high-scoring structures, folding funnels are commonly found whereas their low-scoring counterparts often form heavily rugged landscapes. One might be tempted to conclude that the structure of the barrier alone is directly related to the quality of a design. However, eventhough rugged landscapes are abundant in low-scoring sequence designs, they are also found in smaller numbers among top sequences, emphasizing the need for the detailed analysis of folding kinetics.

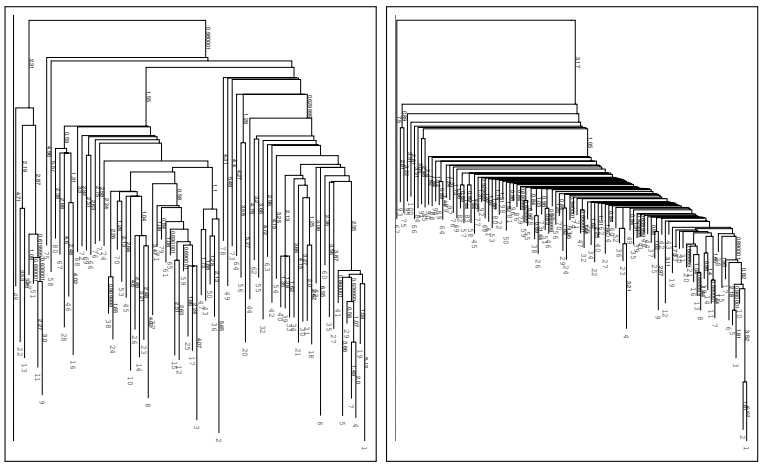


Figure 17: Examples for a rugged landscape (left), and a folding funnel (right).

Fortunately, we do not have to rely on visual analysis alone, but are able to simulate folding kinetics using barrier trees. As the calculation of a barrier tree is involved in some potentially costly steps, some simplifications (Restriction to 10^{10} structures, restriction to lowest 300 basins) were made, respective workflow is detailed described in section 3.2.2. Using barrier trees, folding kinetics was simulated using the program *treekin*. At the start of each simulation, the initial population of states has to be set. Ideally, the open chain would be the starting point of choice, which is not possible here as our barrier trees mostly are restricted to minima with lower free energies. To avoid the case, where the starting point of the simulation sits on a branch not connected to the MFE basin, the state with the highest free energy that is still connected to the global minimum is populated with 100%.

Figure 18 shows the population of the MFE states at the end of the simulation (equilibrium) for sequences of different Shape-Scores. In case the MFE was not the most populated state, Figure 19 depicts how the most populated state scored in comparison to the MFE.

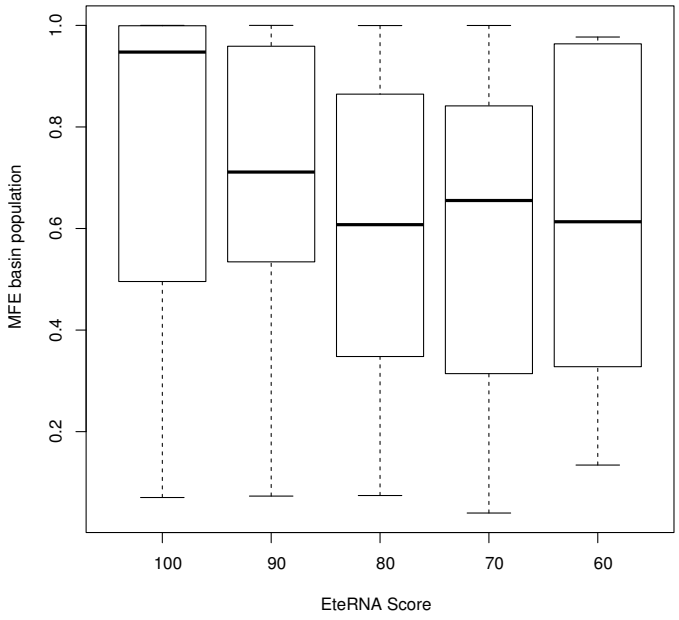


Figure 18: Box-plots showing the distribution of sequences (grouped by EteRNA-score) which fold into the MFE structure at equilibrium.

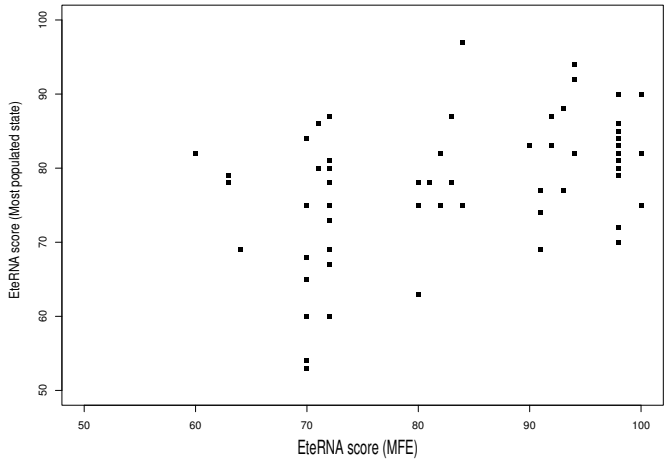


Figure 19: For cases where a non-MFE state is populated the most: EteRNA-score of most populated state versus MFE Shape-score

Discussion

Except for one case, where SHAPE was used to monitor RNA folding kinetics [23], the influence of folding kinetics on probing experiments remains largely unexplored. This is surprising, as boxplot analysis of our simulation data (Figure 18) reveals a clear trend to lower MFE population levels as Shape-scores drop. Boxplots for scores 70 and 60 have to be interpreted with caution, as sample sizes are significantly smaller. Recalculation of barrier trees and simulation of kinetics at temperatures matching the experimental conditions brought no significant changes, as already observed previously (Section 4.3).

Interestingly, MFE structures are not populated in an all-or-nothing manner. Data suggests that as EteRNA-scores drop, the population levels of competing states increase, so the experimental *SHAPE* -pattern has to be interpreted as a superposition of states that were present over the coarse of the experiment *in vitro*. Still, top designs also frequently ($> 50\%$) do not fold into their MFE conformations, but retain a good Shape-score. It has to be noted here, that treekin does not calculate real-time kinectics. Therefore it is uncertain, whether equilibrium as shown here was reached in a time frame corresponding to the duration of the wetlab-experiment. Also, folding mostly was not done from the open chain, as all barrier trees were calculated using simplifications. For simplicity, we assume that in vitro equilibrium was reached prior to the probing experiment.

For sequences, where another state was more prevalent than the MFE structure, EteRNA-scores for respective structures where calculated. Surprisingly, the majority of states show even worse EteRNA-scores, further supporting the hypothesis that the final pattern has to be seen as a superposition of all structures' contributions.

Albeit it is hard to quantify the overall impact of kinetics on *SHAPE* , simulations show significant connections between population levels and the unambiguity of corresponding Probing patterns. There are other plausible sources of errors such as inaccuracies in the energy models which result in biased MFE predictions or the ever-present danger of experimental errors from reading electrophoresis data. Still, RNA folding may be not the major source of errors for EteRNAsequence design, but remains a significant factor affecting the final probing pattern.

Using the optimization workflow outlined in Section 3.1, two RNA sequences were analysed. No wetlab steps were required, as the probing data was already available via the RNA Mapping Database. Both data sets came from full Mutate&Map-experiments and contained probing patterns of the wild-type sequence, as well as of each mutant sequence and can be accessed over the RMDB by the keys 16SFWJ_NMIA_0001 and GLYCFN_SHP_0005_synced.

Test cases are the 16SrRNA-four way junction (16SFWJ), a 110 nt long motif of the E.coli 16S rRNA and a 197 nt long glycine riboswitch from Fusobacterium nucleatum (GLYCFN). Both molecules are characterised by many instances of all loop types. The 16SrRNA-four way junction especially serves as a challenging benchmark for our method, since 4-way junctions structures often are stabilized by higher order interactions [39] and therefore are difficult to predict with common secondary structure prediction methods.

4.5.1 *Application of Mutate&Map data on Reference structures*

For both example cases, figures 20 and 21 show the overall performance of the Mutate&Map approach on their respective reference structures. In both cases, the majority of mutations confirm the respective reference structure's base pairs. Interestingly, most pairs are either identified correctly (blue) or not at all (green). It is especially noteworthy that for many basepairs, just one nucleotide's mutation confirms the basepair (e.g. is coloured blue), whereas the complimentary mutation causes no significant increase in reactivity upon mutation. This suggests that each point-mutation may not inevitably induce local perturbations but can lead to the formation of novel structures where the target is still part of a base pair, but bound to another nucleotide. The relative abundance of such events in the test sets indicates that refolding upon mutation may be a bigger issue than previously assumed and has to be considered more carefully in future experiments.

Beyond that, sequence-blind inference of basepairs from Mutate&Map data, as outlined in Kladwang et al [36] (Section 3.6), was not able to recover any basepairs or even secondary structure in a coherent way. This stresses the need for further assistance by secondary structure prediction algorithms.

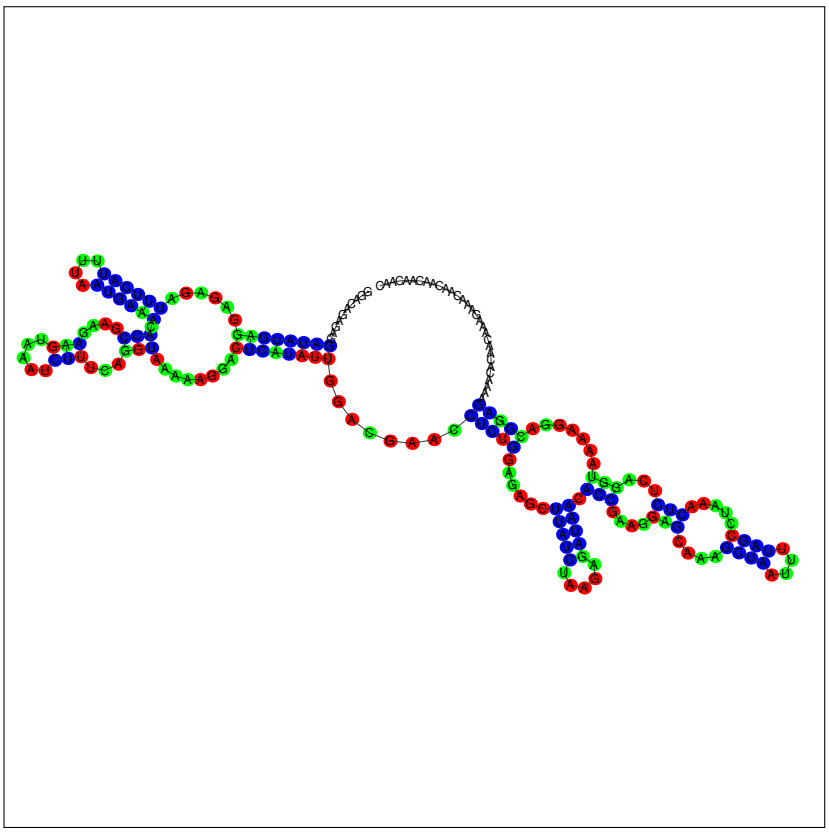


Figure 20: Performance of Mutate&Map visualized on the reference structure of GLYCFN

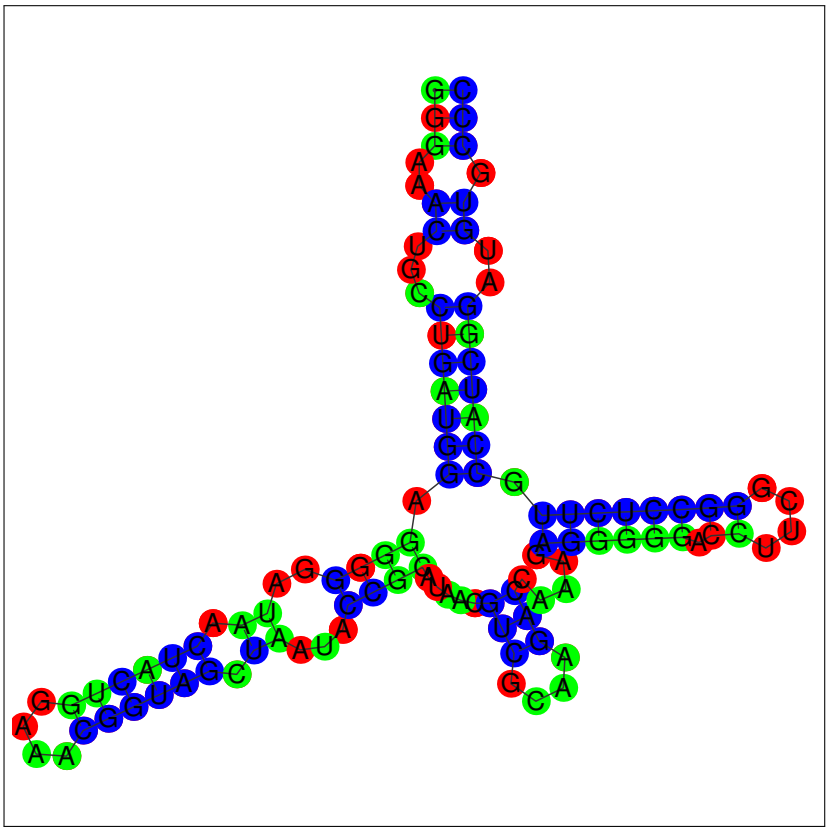


Figure 21: Performance of Mutate&Map visualized on the reference structure of 16SFWJ

4.5.2 *Preparation of Candidate Structure Sets*

Following the procedure outlined in section 3.2, sets of candidate structures were built for each experiment. For the majority of sampled and filtered structures it was observed that at the central step of *SHAPE* -simulation and comparison with the wildtype experiment, the "coarsegrained" scoring metric (Section 3.3.3) performed slightly better than the RMSD comparison method, as it is less influenced by outliers.

In the following step of filtering by pattern simulation and comparison to the experiment, the lower similarity cutoff was set to 0.6, as further variation led to no improvement in optimization quality. The final composition of each candidate structure set can be found in Tables and .

	N_{Sample}	N_{Shapes}	$N_{p>0.001}$	N_{Filter}
GLYCFN	10^4	121	32	19
16SFWJ	10^4	1334	97	20

Table 13: Yield of the filtering workflow for both experiments. All N_{Sample} structures obtained by stochastic backtracking can be reduced to N_{Shreps} coarsegrained RNA-Shapes, from which $N_{p>0.001}$ have a probability > 0.001. Simulation of *SHAPE* -patterns using each Shape's minimum energy structure (Shrep) and comparison with the experimentally obtained pattern, gives a similarity score. Structures with a score < 0.6 were discarded, resulting in N_{Filter} final structures.

4.5.3 *Optimization I: Combinatoric picking of mutation sites*

From the set of candidate structures, the smallest set of mutations which still unambiguously identifies one structure, was picked for both examples. For evaluation with experimental data, the "soft" method (section 3.5.2) was used. The also proposed "hard-scoring" method, which follows the premise that the correct mutation set has to be backed entirely by experimental evidence, failed for every candidate structure, even when including the reference structure into the set of candidate structures. However, the method's failure is not surprising in view of the strong fluctuations intrinsic to *SHAPE* -replica experiments. It in fact underlines the need for flexible and error-tolerant scoring procedures.

Using the soft-scoring for GLYCFN, the proposed solution structure is very similar to its reference. However, the significance of this result has to be interpreted with caution, as (i) only 1 of 3 mutations perturbing a paired base also correctly inferred the same (e.g. is coloured blue) and (ii) the scores of all other candidate structures are just slightly worse. This strongly suggests that such a small set of mutations (4 nucleotides) is by far too small to carry enough information as it is needed for any meaningful distinction between possible structures through Mutate&Map. For 16SFWJ, the proposed solution structure is even worse and again most other structures score nearly equally well.

Picking a larger, also unambiguous set (10 nucleotides), did not bring any improvements. Here, the main problem is that due to the nature of the picking procedure, mutations are placed on purely combinatorial concerns, without any further consideration of additional helpful parameters as pairing-probabilities. Also note that, using a larger set basically means abandoning the whole idea of a minimal mutation set.

Furthermore it should be noted that with a minimal set of mutations, the small amount of experimental data hinders the calculation of statistically significant Z-scores, which are an integral part of the used scoring metric.

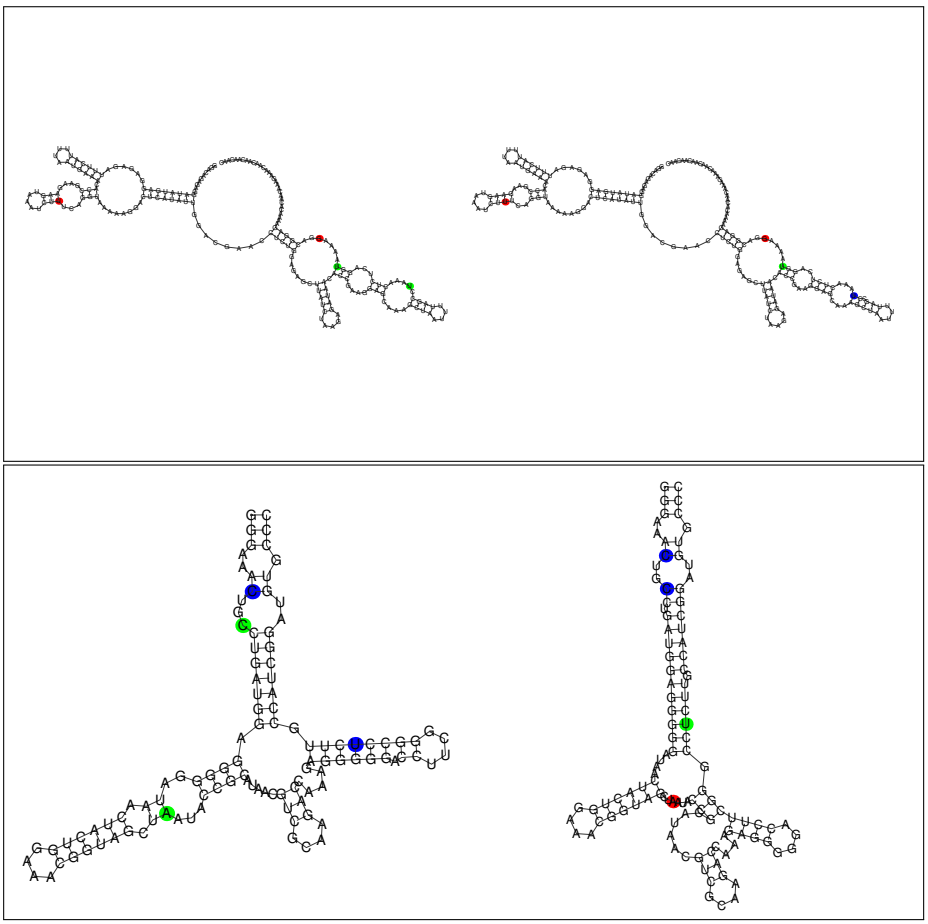


Figure 22: Solution of smallest possible mutation sets (4 in both cases). Picked bases are coloured according to Section 3.5.3; Left: reference structure, Right: candidate structure which scored best; Top: GLYCFN, Bottom: 16SFWJ

Summary

Drastic reduction of a Mutate&Map experiment to a few combinatorial relevant mutations resulted in no visible improvement of structure prediction accuracy. The low success of this method is mainly due to (i) the low number of mutation experiments and (ii) the structurally arbitrary picking of mutation sites. Moreover, as fewer experiments are performed, Z-scores become less significant, making interpretation difficult.

4.5.4 *Optimization II: Weighed picking of mutation sites*

Starting from the same set of candidate structures, the reduced pool of mutations is picked according to probability-to-be-paired and a nucleotide's structural conservation over all candidate structures (see Section 3.4.2). All sets were evaluated with experimental data, using the "soft" scoring method (section 3.5.2).

Optimal set size

Prior analyses suggest that the number of mutations critically influences the outcome of the optimization procedure. In contrary to the combinatoric approach, the "weighed" mutation picking procedure is not subject to the constraint of having the absolute minimum number of mutations. As we now are now free from the constraint of picking a set of minimal size, all sample sizes are tested systematically (Figure 23).

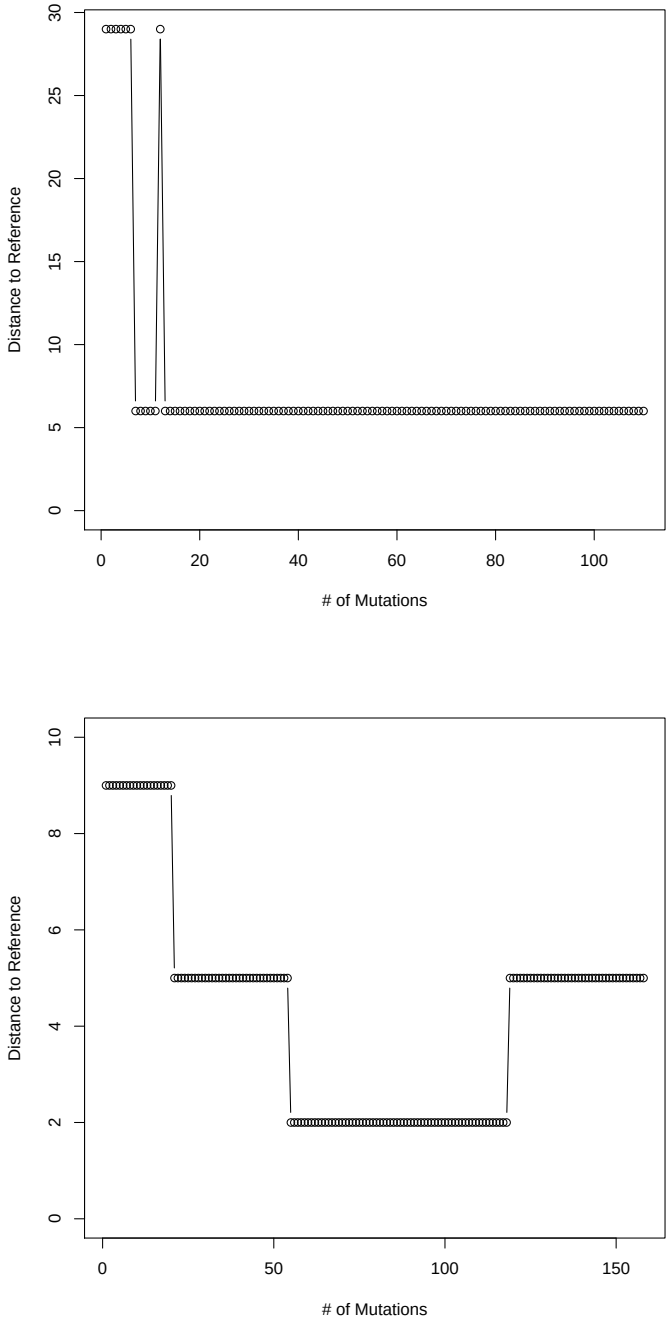


Figure 23: Basepair-distance of the highest scoring structure to the reference, as a function of the number of mutations. Top: 16SFWJ, Bottom: GLYCFN

As already speculated, the solution's quality (similarity to the reference structure) steeply increases with the number of performed mutations. Nevertheless, there is some variety on how *quickly* the best solution is found.

In the case of 16SFWJ, after a few (7) mutations the candidate structure which is closest to the reference structure, is found as solution. Except for one case at $n = 12$, all following mutation sets also identify the same candidate structure as the solution. For GLYCFN, the analysis shows a gradual improvement of quality, with a local optimum at mutation sizes 55-118. Larger mutation sets again favour slightly worse candidate structures.

Both test cases demonstrate that eventhough mutations are now picked by a weighed measure, set size still influences the experiments accuracy.

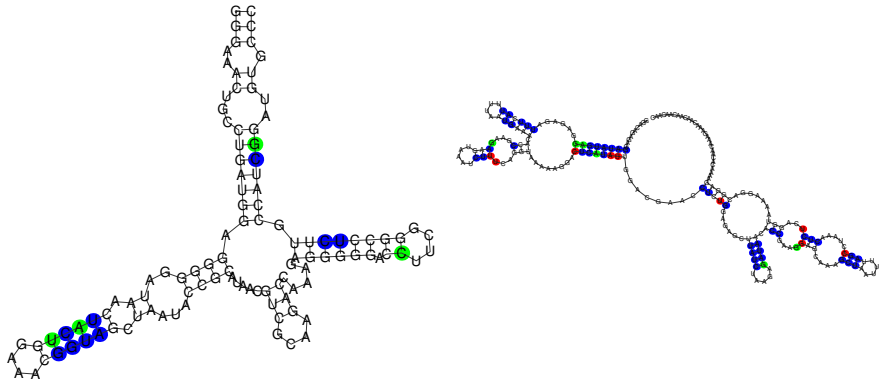


Figure 24: Solution structures for 16SFWJ (left) and GLYCFN (right). Mutated positions are coloured according to their predictive power as outlined in section 3.5.3

Ranking of mutations

The picking procedure favours positions which are likely to be paired and are located in stems in most candidate structures. Figures 25 and 26 progressively show which mutations have been picked. As expected, nucleotides in stems are chosen before any unpaired positions are considered. Also, in the high-ranking segments the majority of mutations result in the inference of a basepair (e.g. are coloured blue), a remarkable improvement to the earlier proposed "minimal"-picking procedure.

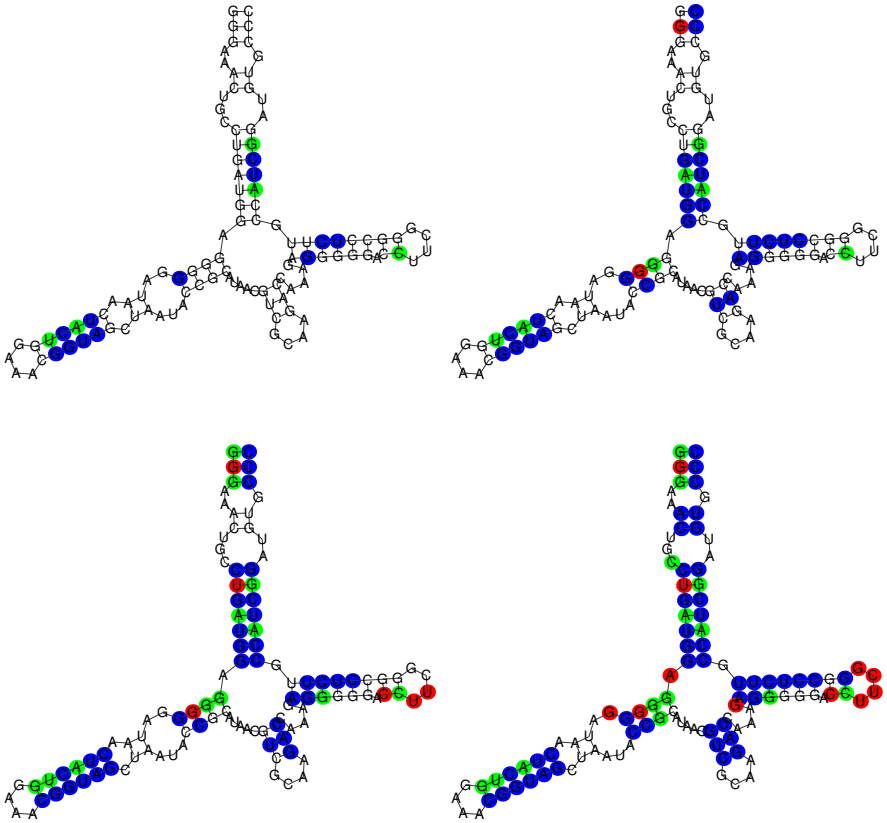


Figure 25: 16SFWJ: Mutations mapped on reference structure - Top left: best 15% of mutations, Top right: best 30%, Bottom left: best 45%, Bottom right best 60%

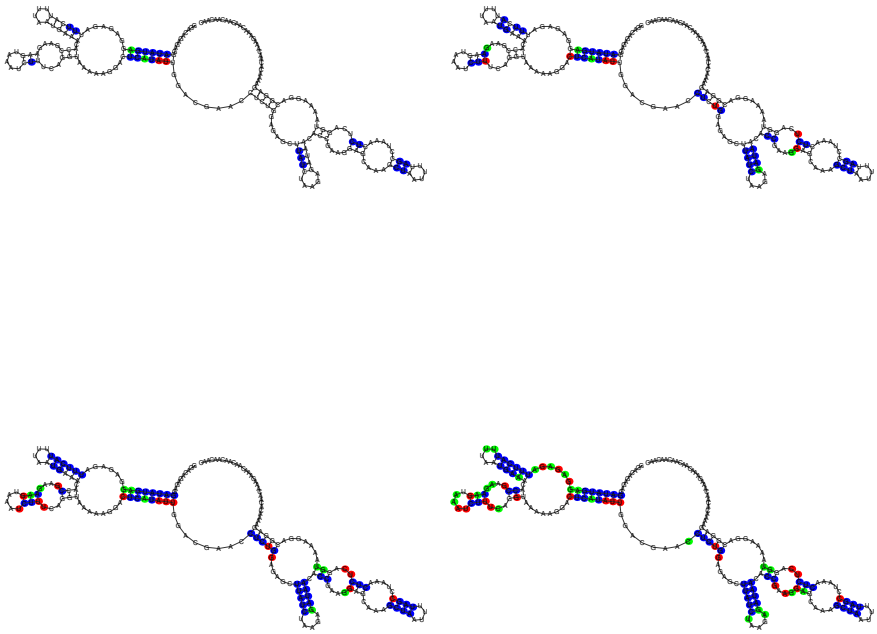


Figure 26: GLYCFN: Mutations mapped on reference structure - Top left: best 15% of mutations, Top right: best 30%, Bottom left: best 45%, Bottom right best 60%

5

SUMMARY

A *SHAPE* -experiment provides information about the conformational constraints a nucleotide experiences *in vitro*. Surprisingly, all hitherto existing interpretations of *SHAPE* -data, fit this information into a simplified framework where the experiment just indicates whether a nucleotide is paired or not paired. This view dominates not only the qualitative analysis of *SHAPE* -data, but is also common for computational procedures which include *SHAPE* -data as folding constraints.

This model is not without reason. In this thesis, it is shown that there is a significant correlation of a nucleotide's probability to be involved in a basepair, and its corresponding *SHAPE* -reactivity. However, detailed analysis of *SHAPE* -data shows that among unpaired nuclotides, there is a broad variability in the corresponding reactivity-values. Most important, these intrinsic differences in interaction with the *SHAPE* -reagent can be understood not only as a function of loop type, of which unpaired nucleotide is part, but are also heavily influenced by its structural context. Moreover, it was shown that unique structural motives, such as small hairpins or interior loops, are associated with unique reactivity patterns which can not be understood in the current framework.

This detailed analysis provides a highly relevant insight into the nature of *SHAPE* -experiments. With help of the newly acquired knowledge, it will be possible to establish a more sophisticated model for structure prediction algorithm to include *SHAPE* -data as folding constraints, leading to a significantly improved performance.

The knowledge gained by the extensive analysis of *SHAPE* -data was also applied with the optimization of Mutate&Map-experiments. The main thought behind Mutate&Map-experiments is, that an unkonwn structure's base pairs can be inferred from the induced perturbations alone. This approach turned out to be inaccurate and prone to errors if performed in a sequence-blind manner. However, with the help of structure prediction algorithms, the Mutate&Map-approach serves as a robust guide for the building of accurate secondary structure models.

A Mutate&Map-experiment typically consists of many sub-experiments, where not all of them are strictly relevant for the outcome. Using the knowledge obtained in the previous mentioned analyses, a simple heuristic was designed, which narrows down the number of necessary mutations to a small, but considered most informative set. When the interpretation of a Mutate&Map-experiment's muations is performed with the assistance of folding algorithms -i.e. *not* in the already mentioned "sequence-blind" fashion - the number of necessary mutations can be greatly reduced by a factor 2-3. However, even when optimal mutations are selected, the number of mutations needed to obtain high quality predictions can usually not be lower than 25% of the number of possible mutations. This is mostly due to the high level of noise observed in *SHAPE* reactivity data.

APPENDIX A

Shrep	Score	D_{bp}	p
((....))..(((((((..((((....((((((....))))))......(((......(((....)))....((((....)))))))..))))..)))))))........	0.736	32	0.003
((....))..(((((((((((((....((((((....))))))......(((......(((....)))....((((....)))))))..))))))..)))))))........	0.691	34	0.039
(((..((..((.((((((((((((....((((((....))))))......(((......(((....)))....((((....)))))))..))))))..)))))))..)).)))	0.682	31	0.031
(((.......((((((((((((.((.((((((....))))))..))...(((......(((....)))....((((....)))))))..))))))))))).....)))	0.682	27	0.044
(((.......((((((((((((....((((((....))))))......(((......(((....)))....((((....)))))))..))))))))))).....)))	0.673	29	0.090
(((..((...(((((((..((((....((((((....))))))......(((......(((....)))....((((....)))))))..))))..)))))))..)).)))	0.655	25	0.025
(((......((.(((((..((((.((.((((((....))))))..))...(((......(((....)))....((((....)))))))..))))..)))))))).....)))	0.645	29	0.004
(((......((.((((((((((((....((((((....))))))......(((......(((....)))....((((....)))))))..))))))))))).....)))	0.645	33	0.097
(((..((...(((((((..((((....((((((....))))))......(((......(((....)))....((((....)))))))..))))..)))))))..)).)))	0.645	25	0.004
((....))..((((((((((((.((.((((((....))))))..))...(((......(((....)))....((((....)))))))..))))))))))).....)).	0.645	32	0.018
(((..((..((.(((((..((((....((((((....))))))......(((......(((....)))....((((....)))))))..))))..)))))))..)).)))	0.645	29	0.002
(((......((.(((((..((((....((((((....))))))......(((......(((....)))....((((....)))))))..))))..)))))))).....)))	0.645	31	0.009
(((.....((((((((((((((....((((((....))))))......(((......(((....)))....((((....)))))))..))))))))))))..)).)))	0.636	31	0.011
((....))((((((((((((((.((.((((((....))))))..))...(((......(((....)))....((((....)))))))..))))))))))))....))..	0.636	34	0.002
((....))((((((((((((((....((((((....))))))......(((......(((....)))....((((....)))))))..))))))))))))....))..	0.636	36	0.004
(((.......((((((..((((....((((((....))))))......(((......(((....)))....((((....)))))))..))))..)))))))).....)))	0.636	27	0.006
(((..((..((.((((((((((((.((.((((((....))))))..))...(((......(((....)))....((((....)))))))..))))))..)))))))..)).)))	0.636	29	0.015
(((.......((((((..((((.((.((((((....))))))..))...(((......(((....)))....((((....)))))))..))))..)))))))).....)))	0.627	25	0.003
(((..((...((((((((((((....((((((....))))))......(((......(((....)))....((((....)))))))..))))))..)))))))..)).)))	0.627	27	0.306
(((..((...(((((((..((...((.((((((....))))))..))...))......(((....)))....(((((.((....)))))))..)))))))..)).)))	0.618	6	0.006

Reference Structure/Sequence
(((..((...(((((((.((((..((.((((((....))))))..))...))))).....(((....))).....((((((.((....)))))))..)))))))..)).)))
GGGAAACUGCCUGAUGGAGGGGGAUAACUACUGGAAACGGUAGCUAAUACCGCAUAACGUCGCAAGACCAAAGAGGGGGACCUUCGGGCCUCUUGCCAUCGGAUGUGCCC

Table 14: 16SrRNA Four-way junction (16SFWJ): Candidate structures; Green/Yellow: Structure which scored best (Weighed/-Combinatoric picking), D_{bp}: Basepair-distance to the reference structure, p: Shape-probability.

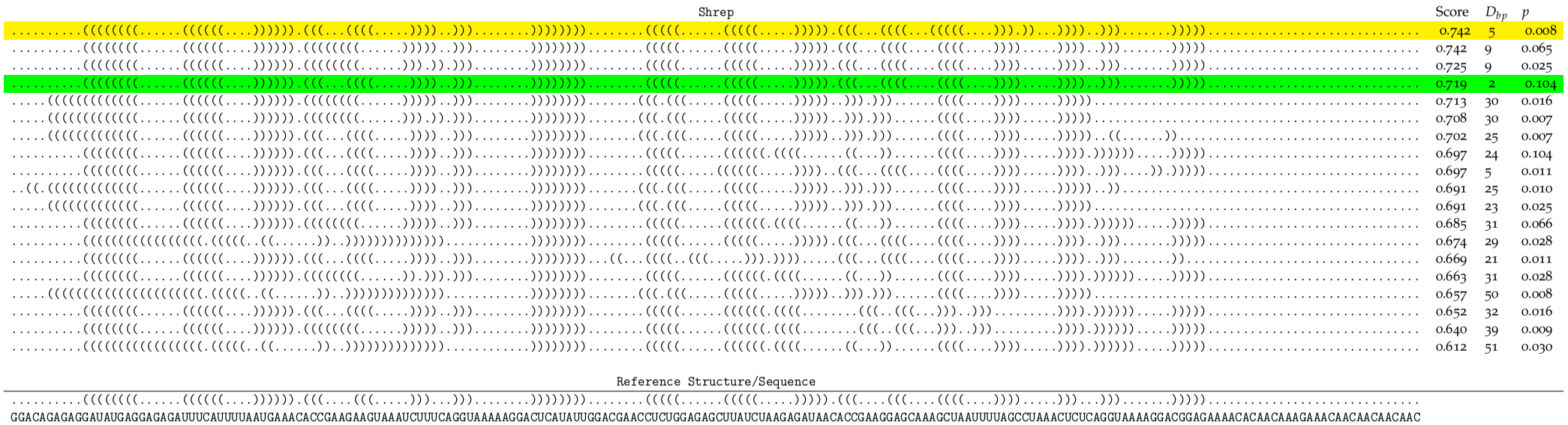


Table 15: Glycine Riboswitch (GLYCFN): Candidate structures; Green/Yellow: Structure which scored best (Weighed/Combinatoric picking), D_{bp}: Basepair-distance to the reference structure, p: Shape-probability.

BIBLIOGRAPHY

1. Weeks, K. M. Advances in RNA Secondary and Tertiary Structure Analysis by Chemical Probing. *Current opinion in structural biology* **20,** 295–304. ISSN: 0959-440X (June 2010).
2. Low, J. T. & Weeks, K. M. SHAPE-directed RNA secondary structure prediction. *Methods. RNA: From Sequence to Structure and Dynamics* **52,** 150–158. ISSN: 1046-2023 (Oct. 2010).
3. Kladwang, W. & Das, R. A Mutate-and-Map Strategy for Inferring Base Pairs in Structured Nucleic Acids: Proof of Concept on a DNA/RNA Helix. *Biochemistry* **49,** 7414–7416. ISSN: 0006-2960 (Sept. 7, 2010).
4. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35,** 849–857. ISSN: 0092-8674 (Dec. 1983).
5. Cech, T. R. A model for the RNA-catalyzed replication of RNA. *Proceedings of the National Academy of Sciences of the United States of America* **83,** 4360–4363. ISSN: 0027-8424 (June 1986).
6. Robertson, M. P. & Joyce, G. F. The Origins of the RNA World. *Cold Spring Harbor Perspectives in Biology* **4,** a003608. ISSN: , 1943-0264 (May 1, 2012).
7. Watson, J. D. & Crick, F. H. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171,** 737–738. ISSN: 0028-0836 (Apr. 25, 1953).
8. Rich, A. & Watson, J. D. SOME RELATIONS BETWEEN DNA AND RNA*. *Proceedings of the National Academy of Sciences of the United States of America* **40,** 759–764. ISSN: 0027-8424 (Aug. 1954).
9. Varani, G. & McClain, W. H. The G?U wobble base pair. *EMBO Reports* **1,** 18–23. ISSN: 1469-221X (July 17, 2000).
10. Yakovchuk, P., Protozanova, E. & Frank-Kamenetskii, M. D. Base-stacking and base-pairing contributions into thermal stability of the DNA double helix. *Nucleic Acids Research* **34,** 564–574. ISSN: 0305-1048, 1362-4962 (Jan. 1, 2006).

11. Giegerich, R., Voß, B. & Rehmsmeier, M. Abstract shapes of RNA. *Nucleic Acids Research* **32,** 4843–4851. ISSN: 0305-1048, 1362-4962 (Jan. 1, 2004).

12. Steffen, P., Voß, B., Rehmsmeier, M., Reeder, J. & Giegerich, R. RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics* **22,** 500–503. ISSN: 1367-4803, 1460-2059 (Feb. 15, 2006).

13. Peattie, D. A. & Gilbert, W. Chemical probes for higher-order structure in RNA. *Proceedings of the National Academy of Sciences* **77,** 4679–4682. ISSN: 0027-8424, 1091-6490 (Aug. 1, 1980).

14. Gopinath, S. C. B. Mapping of RNA–protein interactions. *Analytica Chimica Acta* **636,** 117–128. ISSN: 0003-2670 (Mar. 23, 2009).

15. Lindell, M., Romby, P. & Wagner, E. G. H. Lead(II) as a probe for investigating RNA structure in vivo. *RNA* **8,** 534–541. ISSN: 1355-8382 (Apr. 2002).

16. Merino, E. J., Wilkinson, K. A., Coughlan, J. L. & Weeks, K. M. RNA Structure Analysis at Single Nucleotide Resolution by Selective 2'-Hydroxyl Acylation and Primer Extension (SHAPE). *Journal of the American Chemical Society* **127,** 4223–4231. ISSN: 0002-7863 (Mar. 1, 2005).

17. Wilkinson, K. A. *et al.* Influence of nucleotide identity on ribose 2?-hydroxyl reactivity in RNA. *RNA* **15,** 1314–1321. ISSN: 1355-8382 (July 2009).

18. McGinnis, J. L., Dunkle, J. A., Cate, J. H. D. & Weeks, K. M. The Mechanisms of RNA SHAPE Chemistry. *Journal of the American Chemical Society* **134,** 6617–6624. ISSN: 0002-7863 (Apr. 18, 2012).

19. Gherghe, C. M., Shajani, Z., Wilkinson, K. A., Varani, G. & Weeks, K. M. Strong Correlation Between SHAPE Chemistry and the Generalized NMR Order Parameter (S2) in RNA. *Journal of the American Chemical Society* **130,** 12244–12245. ISSN: 0002-7863 (Sept. 17, 2008).

20. Mitra, S., Shcherbakova, I. V., Altman, R. B., Brenowitz, M. & Laederach, A. High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. *Nucleic Acids Research* **36,** e63. ISSN: 0305-1048 (June 2008).

21. Yoon, S. *et al.* HiTRACE: high-throughput robust analysis for capillary electrophoresis. *Bioinformatics* **27,** 1798–1805. ISSN: 1367-4803, 1460-2059 (July 1, 2011).

22. Wang, Q., Barr, I., Guo, F. & Lee, C. Evidence of a novel RNA secondary structurein the coding region of HIV-1 pol gene. *RNA* **14,** 2478–2488. ISSN: 1355-8382 (Dec. 2008).

23. Mortimer, S. A. & Weeks, K. M. Time-resolved RNA SHAPE chemistry: quantitative RNA structure analysis in one-second snapshots and at single-nucleotide resolution. *Nature Protocols* **4,** 1413–1421. ISSN: 1754-2189 (Sept. 2009).

24. Lee, J. *et al.* RNA design rules from a massive open laboratory. *Proceedings of the National Academy of Sciences of the United States of America* **111,** 2122–2127. ISSN: 0027-8424 (Feb. 11, 2014).

25. Kladwang, W., VanLang, C. C., Cordero, P. & Das, R. Understanding the errors of SHAPE-directed RNA structure modeling. *Biochemistry* **50,** 8049–8056. ISSN: 0006-2960 (Sept. 20, 2011).

26. Kladwang, W., VanLang, C. C., Cordero, P. & Das, R. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nature Chemistry* **3,** 954–962. ISSN: 1755-4330 (Dec. 2011).

27. Tian, S., Cordero, P., Kladwang, W. & Das, R. High-throughput mutate-map-rescue evaluates SHAPE-directed RNA structure and uncovers excited states. *RNA.* ISSN: 1355-8382, 1469-9001. doi:10.1261/rna.044321.114. <http://rnajournal.cshlp.org/content/early/2014/09/01/rna.044321.114> (visited on 09/22/2014) (Sept. 2, 2014).

28. Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms for Molecular Biology : AMB* **6,** 26. ISSN: 1748-7188 (Nov. 24, 2011).

29. Nussinov, R., Pieczenik, G., Griggs, J. R. & Kleitman, D. J. Algorithms for Loop Matchings. *SIAM Journal on Applied Mathematics* **35,** 68–82. ISSN: 0036-1399 (July 1, 1978).

30. Zuker, M & Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research* **9,** 133–148. ISSN: 0305-1048 (Jan. 10, 1981).

31. Deigan, K. E., Li, T. W., Mathews, D. H. & Weeks, K. M. Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy of Sciences* **106,** 97–102. ISSN: 0027-8424, 1091-6490 (Jan. 6, 2009).

32. Wilkinson, K. A. *et al.* High-Throughput SHAPE Analysis Reveals Structures in HIV-1 Genomic RNA Strongly Conserved across Distinct Biological States. *PLoS Biology* **6.** ISSN: 1544-9173. doi:10.1371/journal.pbio.0060096. <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2689691/> (visited on 05/20/2014) (Apr. 2008).

33. Washietl, S., Hofacker, I. L., Stadler, P. F. & Kellis, M. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Research* **40,** 4261–4272. ISSN: 0305-1048, 1362-4962 (May 1, 2012).

34. Zarringhalam, K., Meyer, M. M., Dotu, I., Chuang, J. H. & Clote, P. Integrating Chemical Footprinting Data into RNA Secondary Structure Prediction. *PLoS ONE* **7,** e45160 (Oct. 16, 2012).

35. Sukosd, Z., Swenson, M. S., Kjems, J. & Heitsch, C. E. Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *Nucleic Acids Research* **41,** 2807–2816. ISSN: 0305-1048 (Mar. 2013).

36. Kladwang, W., Cordero, P. & Das, R. A mutate-and-map strategy accurately infers the base pairs of a 35-nucleotide model RNA. *RNA* **17,** 522–534. ISSN: 1355-8382 (Mar. 2011).

37. Turner, D. H. & Mathews, D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Research* **38,** D280–D282. ISSN: 0305-1048 (Database issue Jan. 2010).

38. Kladwang, W. *et al.* Standardization of RNA Chemical Mapping Experiments. *Biochemistry* **53,** 3063–3065. ISSN: 0006-2960 (May 20, 2014).

39. Laing, C. & Schlick, T. Analysis of four-way junctions in RNA structures. *Journal of molecular biology* **390,** 547–559. ISSN: 0022-2836 (July 17, 2009).

ABSTRACT

Ribonucleic acid (RNA) plays a fundamental role in life's biochemical processes. Since a molecules function cannot be studied without knowledge of its structure, the development of structure elucidation methods was always subject of research. Today, there exist numerous biochemical and computational tools for this purpose.

Recently, the whole field has gained new momentum, based on the development of the probing method **SHAPE** and **Mutate and Map** - SHAPE. By overcoming flaws, which had been troubling probing methods for decades, SHAPE allows the investigation of RNA structure at an unprecedented resolution - but not without a steep increase of cost and effort.

The popularity of *SHAPE* was also boosted by the development of new high-throughput methods and adequate analysis software. Today it is possible to probe enormous amounts of RNA sequences in short time.

Eventhough there is a large amount of data availbale, the mechanism of the *SHAPE* reaction remains poorly understood. The goal of this masters thesis was, to understand the behaviour of the *SHAPE* -reaction by analysis of available probing data and subsequently investigate the possibility of optimization of *SHAPE* by computational means.

ZUSAMMENFASSUNG

Ribonukleinsäuren (RNS) sind von fundamentaler Bedeutung für viele essentielle biochemische Prozesse des Lebens. Nachdem die Funktion eines Moleküls nur im Kontext seiner Struktur sinnvoll erfasst werden kann, war die Entwicklung von Strukturaufklärungsmethoden seit jeher Gegenstand von Untersuchungen. Als Resultat dieser jahrzehntelangen Anstrengungen, kann heute auf ein breites Repertoire aus biochemischen und computergestützten Werkzeugen zur Strukturaufklärung zurückgegriffen werden.

In jüngster Zeit erlebt das Forschungsgebiet eine neue Dynamik, getragen durch die Entwiklung der biochemischen Probingmethode **SHAPE**, und davon weiterführend **Mutate and Map** - SHAPE. Durch die Beseitigung einiger limitierender Faktoren bisheriger Probingmethoden, verspricht SHAPE die Aufklärung von RNA-Struktur auf einer bisher mit biochemischen Methoden unerreichten Auflösung - wenn auch mit einem deutlich höheren materiellen Aufwand, verglichen mit anderen Probingmethoden.

Der Erfolg von *SHAPE* wurde in weiterer Folge durch die Entwicklung neuer Hochdurchsatzverfahren und den dazugehörigen Analysetools begünstigt. Dadurch ist es heute möglich, in kurzen Zeiträumen enorme Mengen an RNA-Sequenzen mittels probings zu untersuchen.

Trotz der dadurch mittlerweile beträchtlichen Menge an verfügbaren Daten, ist die genaue Funktionsweise von *SHAPE* immer noch unzureichend untersucht. Das Ziel dieser Masterarbeit war es, durch die Analyse der verfügbaren Probingdaten genauer das Verhalten von *SHAPE* zu verstehen, um darauf aufbauend die Mutate&Map-Technik durch Zurhilfenahme von computergestützten Methoden optimieren und in ihrer Aufwändigkeit zu reduzieren.

Curriculum Vitae

Personal data

Last name:	Ochsenreiter
First name:	Roman Wilhelm
Address:	Institute of Theoretical Chemistry Währinger Str. 17, 1090 Vienna, Austria
Nationality:	Austrian

Academic career

4/2013 12/2014	-	Master thesis in the group of Prof. Dr. Ivo Hofacker Title: *Computational Refinement of* SHAPE - *RNA probing experiments*
10/2012- 12/2014		Master Biological Chemistry, **University of Vienna**
10/2008- 07/2009		Studies paused for Civil Service
2007-2012		Bachelor Molecular Biology, **University of Vienna**
06/2007		Matura, **Bundesrealgymnasium Rainergasse**

Further qualifications

Language skills	German (mother tongue) English (fluent) Italian (fluent) Latin (advanced) French (intermediate) Russian (basic)