



universität  
wien

Dissertation

**In silico modelling of RNA-RNA dimer and its application  
for rational siRNA design and ncRNA target search**

angestrebter akademischer Grad  
*Doktor der Naturwissenschaften (Dr.rer.nat)*

Verfasser	Hakim Tafer
Matrikel-Nummer	0547529
Dissertationsgebiet	Physik
Betreuer	Univ.-Prof.Dr. Hofacker

February 14, 2011

## Danksagung

---

**Danke an alle, die zum Entstehen dieser Arbeit beigetragen haben**

Ivo Hofacker, Peter Stadler, Christoph Flamm, Stephan Bernhart, Stefan Ameres, Andreas Gruber, Gregor Obernosterer, Ulrike Mückstein.

## Abstract

---

Non-protein coding region, which constitutes 98.5% of the human genome, were long depreciated as evolutive relict. It is only recently that the biological relevance of the non-coding RNAs associated with these non-coding regions was recognized. The development of experimental and bioinformatical methods aimed at detecting these non-coding RNAs (ncRNAs) lead to the discovery of more than 29,000,000 sequences, grouped into more than 1300 families.

More often than not these ncRNAs function by binding to other RNAs, either protein coding or non-protein coding. Compared to the number of tools to detect and classify ncRNAs, the number of tools to search for putative RNA binding partners is negligible. This leads to the actual situation where the function of the majority of the annotated ncRNAs genes is completely unknown.

The aim of this work is to assess the function of different families of ncRNAs by developing new algorithms and methods to study RNA-RNA interactions. These new methods are extensions of RNA-folding algorithms applied to the problem of RNA-RNA interactions. Depending on the class of ncRNA studied, different methods were developed and tested.

This work shows that the development of RNA-folding algorithms to study RNA-RNA interactions is a promising way to functionally annotate ncRNAs. Still other factors like RNA-proteins interaction, RNA-concentration or RNA-expression, play an important role in the process of RNA hybridization and will have to be taken into account in future works in order to achieve reliable prediction of RNA binding partners.

# Contents

---

<b>1</b>	<b>Introduction</b>	<b>5</b>
1.1	Motivation . . . . .	5
1.2	Structure of this work . . . . .	8
<b>2</b>	<b>Background</b>	<b>9</b>
2.1	RNA secondary structure . . . . .	9
2.1.1	RNA secondary structure: formalism and representation . . .	10
2.2	RNA folding . . . . .	12
2.2.1	Counting structures and maximizing base pairs . . . . .	16
2.2.2	Loop-Based Energy Model . . . . .	17
2.2.3	RNA folding with the loop-energy model . . . . .	19
2.3	Cofolding of two sequences . . . . .	21
2.4	non-coding RNA . . . . .	28
2.4.1	miRNA . . . . .	28
2.4.2	siRNA . . . . .	33
2.4.3	sRNA . . . . .	38
2.4.4	snoRNA . . . . .	39
<b>3</b>	<b>RNAup</b>	<b>43</b>
3.1	Algorithm . . . . .	43
3.2	Free Energy of Interaction . . . . .	48
3.3	Application . . . . .	49
3.3.1	siRNA design . . . . .	49
3.3.2	sRNA targets . . . . .	52
3.4	Conclusion . . . . .	55
<b>4</b>	<b>RNAplfold</b>	<b>61</b>
4.1	From RNAup to RNAplfold . . . . .	61
4.2	Application to RNAi and siRNA design . . . . .	62
4.3	Target site effects in microRNA pathways . . . . .	76

4.4	Conclusion . . . . .	79
<b>5</b>	<b>RNAplex</b>	<b>81</b>
5.1	Methods . . . . .	82
5.1.1	Energy model . . . . .	82
5.1.2	Recursion . . . . .	85
5.1.3	Taking the target accessibility back into RNAplex . . . . .	88
5.1.4	Conserved Interactions . . . . .	92
5.1.5	Model Errors . . . . .	97
5.1.6	Computational efficiency . . . . .	100
5.2	Results . . . . .	106
5.2.1	miRNA targets prediction . . . . .	106
5.2.2	sRNAs targets prediction . . . . .	106
5.2.3	Multiple alignment . . . . .	109
5.3	Conclusion . . . . .	114
<b>6</b>	<b>RNASnoop</b>	<b>119</b>
6.1	Methods . . . . .	120
6.1.1	Single-Sequence RNASnoop . . . . .	120
6.1.2	Machine-Learning Component . . . . .	124
6.1.3	Performance . . . . .	126
6.1.4	A Comparative Version . . . . .	133
6.1.5	SNOPY . . . . .	133
6.2	Results . . . . .	134
6.3	Conclusion . . . . .	138
<b>7</b>	<b>Conclusion</b>	<b>139</b>
	<b>Appendices</b>	<b>145</b>
<b>A</b>	<b>List of Symbols</b>	<b>145</b>
<b>B</b>	<b>Bibliography</b>	<b>147</b>

<i>Contents</i>	3
<b>C List of figures</b>	<b>173</b>
<b>D List of tables</b>	<b>187</b>
<b>E Resume</b>	<b>191</b>



# Introduction

## 1.1 Motivation

---

All living creature contains the necessary information for its structure and function in its genome. Physically the genome is split into one or more chromosomes, which is a long, double-stranded, chain of nucleotides. A nucleotide is a molecule composed of a monophosphate, a pentose and one of the 5 nucleobases, namely adenine, guanine, cytosine, thymine and uracile. These long chains of nucleotide are also called deoxyribonucleic acid (DNA).

Regions in the genome can be either protein-coding or not. Proteins are biological macromolecule composed of a sequence of amino-acids that are involved in almost every functional and structural aspect of the living cell. The production of a protein from the corresponding gene is roughly a three-steps process. First the gene encoding the protein is copied (transcribed) into RNA (ribonucleic acid) called messenger RNA (mRNA). The mRNA is then spliced, i.e it is processed to remove RNA fragments that were transcribed but not protein-coding (introns). Then it is transported into the cytoplasm where it docks onto a ribosome (rRNA) that will translate the genetic code carried by the messenger RNA into a protein. Genomic research originally concentrated exclusively on the protein coding part of the genome, neglecting non protein-coding regions, which make up to 98.5% of the human genome, and their associated non-coding RNA transcripts, as they were long disregarded as evolutionary junk.

50 years ago, Jacob and Monod were the first to raise doubts about the uselessness of non-coding regions [111]. In 1982 Cech showed for the first time that an RNA molecule can have an enzymatic functions [128]. One year later Altman proved that the activity of the RNA cutting enzyme named RNase P was induced by a RNA molecule [88]. In recent years, the plethora of genomic information brought by ncRNA detection programs and high throughput sequencing let the number of known ncRNA-transcripts grew steadily. This is exemplified in [226] and [202], where high throughput techniques allowed the detection of 60 new ncRNAs in *H. pylori* and 1023 new ncRNAs in human cells, respectively. As of 2008, more than 29,000,000 non coding sequences were grouped into 1300 distinct families [75]. Yet despite the abundance and the widespread distribution of ncRNAs, little is known about the biological role they are involved in.

In the few cases where functional annotation of ncRNAs exist, ncRNAs exert their function by binding to other RNAs. For example snoRNAs mediate pseudouridylation and methylation of rRNAs and snRNAs [10] and can influence the splicing of pre-mRNAs [273]. ncRNAs are also involved in sequence editing of other RNAs [15], transcription and translation control (siRNA, miRNA, stRNA) [12, 68, 129] or plasmid replication control [60]. While siRNAs are often fully complementary to their targets, most of the ncRNAs interact in a more intricate manner which does not involve perfect hybridization. For example in *E. Coli*, *OxyS*, which is involved in oxidative stress response, interacts with its target mRNA, *fhlA*, through a two sites kissing complex formation [8].

Systematic target prediction for the plethora of genomic information brought by ncRNA detection programs and high throughput sequencing is a challenging problem and different kinds of tools are currently available to solve it. On one hand, BLAST [3] or FASTA [195] search for long stretches of perfect complementarity between a query and a target sequence. GUUGle [79] can efficiently locate potential complementary regions and, in contrast to BLAST, also allows for G·U pairs. A typical application for these programs is for example siRNA target search. Their main drawback is that they do not give information about the thermodynamics of the interaction between the query and the target RNAs. Moreover their lack of sensitivity is a real issue when looking for more complex interactions found for example

between miRNAs and their targets.

On the other hand, RNA folding algorithm based on the free energy minimization is at present the most accurate and most generally applicable approach for RNA folding [247,275,276]. It is based upon a large number of measurements performed on small RNAs and the assumption that stacking base pairs and loop entropies contribute additively to the free energy of RNA secondary structures [166,168].

A straightforward approach to folding two RNA molecules is to concatenate the two sequences and apply a slightly modified RNA folding algorithm. This approach is taken for example by the `RNAcofold` [18,100] and `pairfold` [246] programs. However, the restriction to pseudo-knot free structures in standard folding algorithms is a more serious issue when dealing with RNA duplexes, as many known RNA-RNA interactions are mediated e.g. by “kissing hairpins” or other structure motifs that appear as pseudo-knots when the sequences are concatenated.

As in the case of single sequences [1] inclusion of pseudo-knots makes the problem of cofolding 2 RNAs NP-complete [2] in the unrestricted case. Polynomial time complexity can be achieved like in Alkan [2,38,107,196], where intramolecular structures of each molecule are pseudoknot free and intermolecular binding pairs are not allowed to cross. While these algorithms can predict complicated interaction motifs, such as the bacterial `OxyS-fhlA` system (see Figure 2.8), they run in  $\mathcal{O}(n^3 \cdot m^3)$  making them prohibitively expensive for most applications. Moreover, these algorithms suffer from a lack of good parameters: Little is known about the energetics of more complicated loop-types, so that predicted optimal structures will often not correspond to reality.

In summary we are currently facing a challenging situation where, on one hand high-throughput sequencing data and bioinformatical approaches unveil a whole new RNA-based world, while on the other hand the sheer scarcity of methods to study RNA-RNA interactions, as well as their limitation, impede us of knowing more about their functions. In this work we will concentrate on extending RNA-folding algorithms to the problem of RNA-RNA interactions and ncRNA target predictions. The extension of the algorithms will not only have to respect accuracies constraints but, due to the large search spaces and huge number of ncRNAs, also runtime constraints. These new approaches will then be used to study known

RNA-RNA interactions and predict ncRNA target RNAs in different organisms.

## 1.2 Structure of this work

---

This thesis is a compilation of the following 9 journal articles [20, 87, 94, 182, 183, 187, 234, 236, 237, 253], one book chapter [102] and of unpublished observations. It is organized as follow: in chapter 2, the necessary background information for understanding the rest of the work is presented. This chapter describes what the concept of RNA secondary structure is as well as how it can be computed for one and two sequences. Further an overview of ncRNAs relevant for this work is presented. The work done in the framework of this dissertation is then presented in the subsequent chapters. In this context chapter 3 contains a description of **RNAup**, a general approach to study RNA-RNA interactions. Chapter 4, presents **RNAplfold** a program developed by [18, 26], which was specifically designed to compute local RNA structures and accessibility. We further show how the information returned by **RNAplfold** can be used to improve siRNA design and miRNA targets search. In 5, we present an approximation to **RNAup** called **RNAplex**, that allows to search for ncRNAs targets with the same accuracy as **RNAup** but with a runtime decreased by three orders of magnitude compared to **RNAup**. Finally in chapter 6, a method is presented to study the complex H/ACA-snoRNA-rRNA interactions. We close this thesis with a discussion in chapter 7.

*Nous sommes comme des nains juchés sur des épaules de géants, de telle sorte que nous puissions voir plus de choses et de plus éloignées que n'en voyaient ces derniers. Et cela, non point parce que notre vue serait puissante ou notre taille avantageuse, mais parce que nous sommes portés et exhaussés par la haute stature des géants.*

Bernard de Chartres (1130-1160)

# 2

## Background

### 2.1 RNA secondary structure

---

RNA is a heteropolymer which consists of nucleotides. A nucleotide is composed of a ribose sugar, a base (adenine, cytosine, guanine, uracil) and a phosphat group. In an RNA chain the phosphat group links the 3' position of the ribose to the 5' position of the next nucleotide.

In contrast to DNA that usually occurs as double strands, RNA molecules are generally single-stranded. RNA structure results from the propensity of the nucleotides to form base pairs with other nucleotides. These base pairs are in general either watson-crick (adenine-uracil, cytosine-guanine) or wobble (guanine-uracil). The intramolecular interactions result in a pattern of double helical regions interspersed with loops. This pattern is termed the RNA secondary structure. These loops and helical structures may further interact to form the tertiary, functional, structure.

In contrast to protein folding programs, where the tertiary structure is predicted, the majority of the currently available RNA folding algorithms concentrate on the secondary structure of the RNAs. The first reason for this difference is a pragmatic one. Current RNA folding algorithms have a polynomial runtime of  $\mathcal{O}(n^3)$  where  $n$  is the sequence length. This is fast enough to allow genome-wide analysis on current off-the-shelf computers. The consideration of the tertiary structure however leads to a superpolynomial-runtime impeding any large-scale application [1]. The second reason is related to the kinetic of RNA folding. Secondary structures form first,

leading to a set of loops and helices, which once formed, interact to yield the tertiary structure. As a consequence, the determination of the tertiary structure depends strongly on the secondary structure [31]. Still the mere knowledge of the secondary structure can be misleading, as two similar tertiary structures can have different secondary structures [125].

### 2.1.1 RNA secondary structure: formalism and representation

A secondary structure  $\mathcal{S}$  on a sequence  $s$  is the set of base pairs  $(s_i, s_j)$ , where  $i < j$  and where  $s_i$  represents the nucleotide at position  $i$  on sequence  $s$ , that has the following properties:

- (i)  $(s_i, s_j) \in \mathcal{S} \implies (s_i, s_j) \in (AU, UA, GC, CG, GU, UG)$
- (ii)  $((s_i, s_j) \wedge (s_k, s_l)) \in \mathcal{S} \wedge (s_i = s_k) \implies j = l$
- (iii)  $((s_i, s_j) \wedge (s_k, s_l)) \in \mathcal{S} \wedge i < k \implies l < j \vee j < k$

In words, constraint **i** means that only watson-crick and wobble base pairs may form. Constraint **ii** states that a nucleotide may be involved in at most 1 base pair. Constraint **iii** implies that all base pairs are nested, i.e. that no pseudoknots are allowed in the secondary structure. While these constraints greatly simplify the folding algorithms, none of the above constraint is biologically relevant. Exotic base pairings, involving more than two nucleotides were reported [76, 77]. Further pseudoknots appear in many important RNAs structures, albeit at a low frequency. For example, in the small ribosomal unit in *E. coli*, from the 447 reported watson-crick and wobble base pairs only 8 are pseudoknots [90].

Any secondary structure generated under these rules can be decomposed into a unique set of loops [167, 243]. The loop is a substructure which consists of a closing base pair  $(s_i, s_j)$  and all nucleotides that are accessible from this base-pair. A nucleotide  $s_p$  is accessible from  $(s_i, s_j)$  if  $i < p < j$  and there exists no other base pair  $(s_k, s_l)$  in  $s$  such that  $i < k < p < l < j$ . Loops can be assigned a degree, i.e. the number of base pairs in the loop, and a size which corresponds to the number of unpaired nucleotide in the loop.

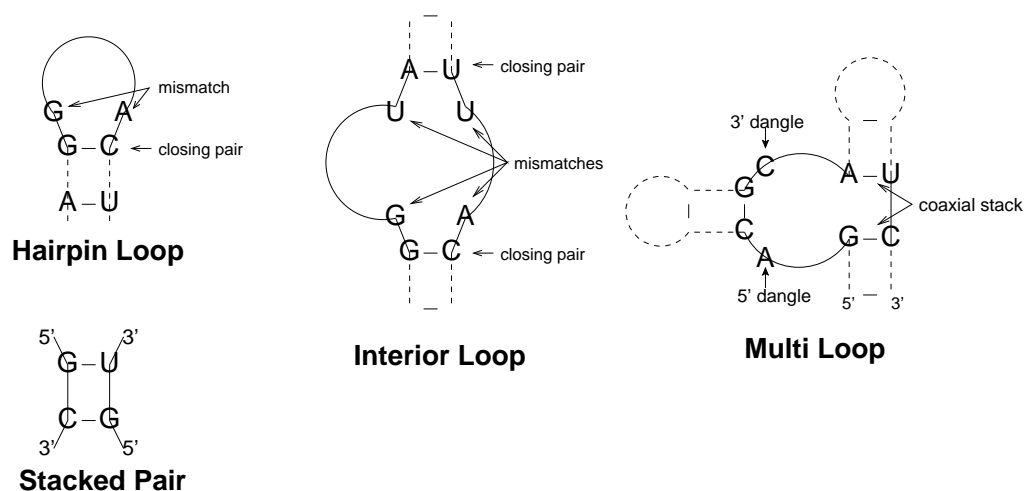


Figure 2.1: The major types of loops in RNA secondary structures. Taken from [101]

There exist different kinds of loop depending on the amount and arrangement of their interior base pairs. Hairpin loops have a degree of 1. Loops of degree 2 are called interior loops. Interior loops of size zero are called stacked pairs. An uninterrupted sequence of stacked pairs represent a stem. Interior loops of size larger than 0, with adjacent interior and exterior base pairs, are called bulge loops. Multiloops are loops of degree greater than 2. Finally exterior loops are the set of nucleotides which are inaccessible by any base pair (see 2.1).

RNA stem-loop structures can be represented in different ways (see Figure 2.3). The dot bracket representation, for example, assigns a “.” to unpaired nucleotides, “(“ and “)” are assigned to nucleotide  $s_i$  and  $s_j$  respectively, if they form a base pair  $(s_i, s_j)$  with  $i < j$ . RNA structure can also be interpreted as a tree [224,225]. For example the *full tree* representation [69] associate base-pairs to internal nodes and unpaired bases to leafs. In a more detailed representation, each interior node is surrounded by a right-most and left-most children which correspond to the 5' and 3' nucleotides of the base pair, respectively. In a Shapiro-Zhang tree, the different loops and stacked regions are represented explicitly with special labels (see Figure 2.2).

The dotplot representation maps the structure to a matrix where a dot at position  $(i, j)$  represents the base pair  $(s_i, s_j)$ . The mountain plot representation maps the



Figure 2.3: Representations of secondary structures. From left to right: Circular representation, NaviView representation, mountain plot, dot plot. Remove the backbone edges from the first two representations leaves the matching  $\Omega$ . Below, the structure is shown in “bracket notation”, where each base pair corresponds to a pair of matching parentheses. The structure shown is the purine riboswitch (Rfam RF00167) taken from [101]

In 1978, Nussinov et al. [186] presented the first algorithm that folded RNA in polynomial time. She considered the maximal matching problem where the best RNA structure is the one having the largest number of base pairs. She proved that the optimal secondary structure can be obtained from the optimal structure of the subsequences. This fact lead her to devise a recursive algorithm based on dynamic programing where the optimal RNA structure can be found in  $\mathcal{O}(n^3)$ , where  $n$  is the RNA sequence length.

(so-called mismatches). Hairpin-loop energies are tabulated for loop size smaller than four. In the other cases approximation based on the mismatches, closing pair and size are used. Finally the multiloop energy model depends linearly on the size, the degree of the multiloop as well as a penalty term for closing the loop. It should be noted that the main reason why the linear multiloop energy model was chosen over more precise models is that it allows to implement RNA-folding algorithm with a runtime of  $\mathcal{O}(n^3)$ .

The main drawback of early approaches employed to predict RNA secondary structures, is that no information about suboptimal structures is returned. This is especially annoying knowing that at physiological conditions base-pair stacking energies and thermal energies are in the same range, allowing the RNA sequences to switch easily between numerous alternative foldings. Moreover, energetically close structures can be radically different. A typical example is the sequence of the 5.8S RNA from *Cryptocodinium cohnii* whose optimal structure does not share any base pair with a suboptimal structure which has an energy within 6% of the global minimum [274] (see Figure 2.4). Furthermore due to the inherent approximation in our model and the errors in the energy parameters, the predicted structure might not even correspond to the real structure. The assessment of suboptimal foldings is therefore crucial.

Different methods were developed to gather information on suboptimal structures. In [274], Zuker et al. designed an algorithm to find the best structures for each admissible base pair in a sequence. For a sequence of length  $n$ , Zuker's approach generates at most  $n \cdot (n - 1)/2$  suboptimals. [267] designed a method that truly retrieves all possible structures situated in an energy band  $\Delta$  above the minimal free energy structure.

Another approach consists in computing the equilibrium partition function for secondary structure [169]. The partition function gives access to the probability of a given structural element in the conformational ensemble. This can be for example the probability to find a given base-pair in the ensemble of structure in thermodynamic equilibrium or for example the probability for a stretch of  $N$  nucleotides to be fully unpaired (accessible) in the ensemble of structures in thermodynamic equilibrium.

An important feature that can be computed is the so-called accessibility of a con-

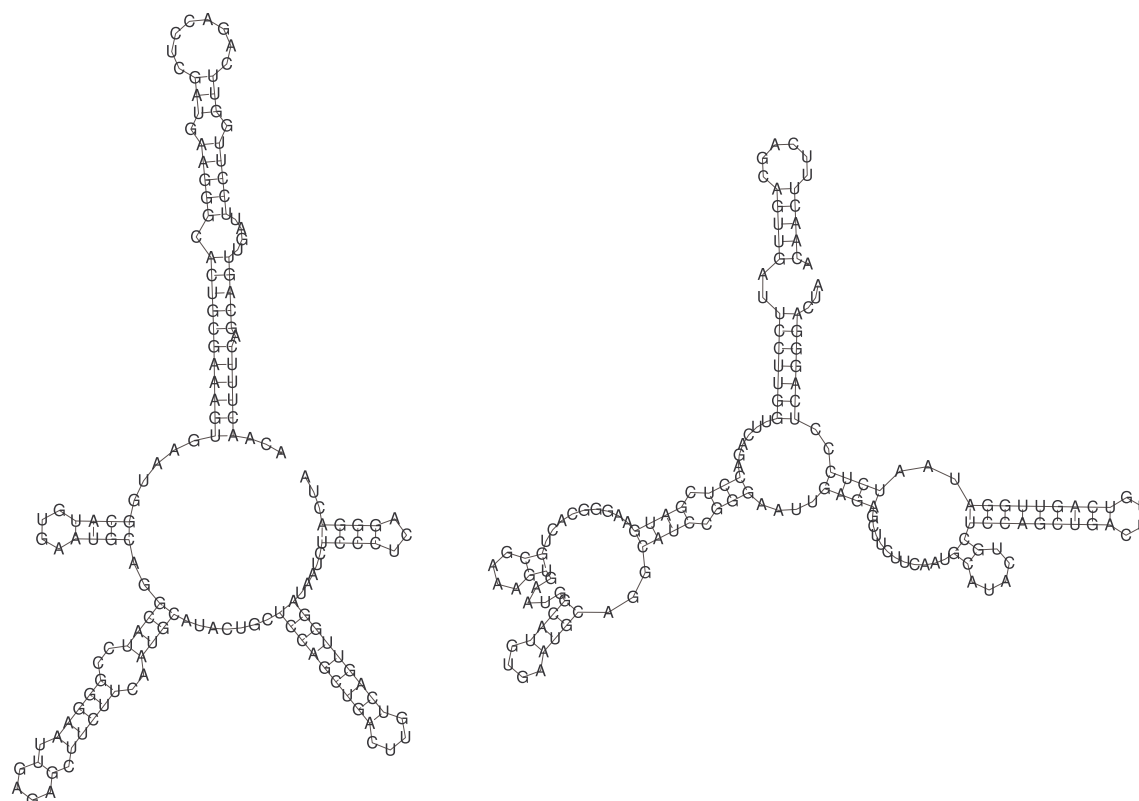


Figure 2.4: **R.h.s** Minimum free energy (MFE) structure of *Crypthecodinium cohnii* 5.8S. Its free energy represents  $-47.10kcal/mol$ . **L.h.s** Suboptimal folding of *Crypthecodinium cohnii* 5.8S sharing no base pair with the MFE structure. This structure has a free energy that differs by only  $2.80kcal/mol$  from the MFE.

tinuous stretch of nucleotides, which corresponds to the probability of this stretch to be completely unpaired. Chapters 3 and 4 are devoted to the computation of methods to compute accessibilities for stretches of sequences of any given length. As it will be shown in this work, accessibility is a key factor for correctly describing the interaction of any two RNAs in both prokaryotes and eukaryotes.

### 2.2.1 Counting structures and maximizing base pairs

Access to the main ideas behind the general approach for RNA folding can be gained by looking at the following combinatorial problem:

**Given an RNA sequence of length  $n$ , enumerate all secondary structures on  $\mathbf{x}$ .**

Let  $s$  be a sequence,  $s_j$  represents the  $j$ -th nucleotides on sequence  $s$ ,  $(s_i, s_j)$  represents the base pair between  $s_i$  and  $s_j$ , and let  $s[i..j]$  stands for the subsequence on  $s$  contained between nucleotides  $s_i$  and  $s_j$ .

Given a subsequence  $s[i..j]$ , the corresponding structure  $\mathcal{S}$  can be derived in exactly two ways from shorter structures. The first nucleotide  $s_i$  is either unpaired and followed by an arbitrary structure on  $s[i+1..j]$  or it binds with an other nucleotide  $s_k$ . Because we do not allow base pairs to cross (see 2.1.1), we have independent structures on the subsequence  $s[i+1..k-1]$  and  $s[k+1..j]$ . This can be graphically represented as:



The number  $N_{ij}$  of structures on  $s[i..j]$  is then given by [263, 264]:

$$N_{ij} = N_{i+1,j} + \sum_{k, (i,k) \text{ pairs}} N_{i+1,k-1} N_{k+1,j} \quad (2.1)$$

where  $N_{ii} = 1$ .

The combinatorial approach is very similar to Nussinov's solution to the folding problem. If we denote by  $E_{ij}$  the maximal number of base pairs on  $s[i..j]$  we see that  $E_{ij}$  is obtained by choosing the optimal substructures among each of the alternatives. The independence of two substructures in the paired cases implies that these substructures can be optimized independently. This yields the Nussinov recursion which can be computed in  $\mathcal{O}(n^3)$  with a dynamic programming approach:

$$E_{ij} = \max \left\{ E_{i+1,j}, \max_{k, (i,k) \text{ pairs}} \{ E_{i+1,k-1} + E_{k+1,j} + 1 \} \right\} \quad (2.2)$$

As already mentioned, at physiological temperatures, RNA molecules switch between different structures rather than being frozen in the single minimum free energy structure (MFE). The probability to find at thermodynamic equilibrium a structure with energy  $\Psi$  is proportional to  $\exp(-E(\Psi)/RT)$ , where  $E(\Psi)$  is the energy of the structure  $\Psi$ . The ensemble of structure is determined by its *partition function*:

$$Z = \sum_{\Psi} \exp(-E(\Psi)/RT), \quad (2.3)$$

Based on the partition function, the equilibrium probability of a structure can be computed as  $p(\Psi) = \exp(-E(\Psi)/RT)/Z$ .  $Z$  can be computed in analogy to 2.2:

$$Z_{ij} = Z_{i+1,j} + \sum_{k, (i,k) \text{ pairs}} Z_{i+1,k-1} Z_{k+1,j} \exp(-\beta_{ik}/RT). \quad (2.4)$$

The equilibrium partition allows not only to compute the probability of a structure in equilibrium but also to list explicitly all possible structures, the number of states with a given energy, to determine structures that optimize certain properties or the probability of a given base pair. The equilibrium base-pair probabilities  $p_{ij}$  for example can be computed by using the outside partition function  $\hat{Z}_{ij}$  of structures outside the subsequence  $s[i..j]$ , yielding:

$$p_{ij} = \hat{Z}_{ij} Z_{i+1,j-1} \exp(-\beta_{ij}/RT)/Z. \quad (2.5)$$

## 2.2.2 Loop-Based Energy Model

The simple energy model used in the maximal matching approach, where base-pairs are assigned a positive scores and loops a negative one, allows only in rare cases to predict correct RNA structures [40]. A more appropriate energy model is the so-called loop-based energy model, where the total free energy of a structure is approximated by the sum of the free energy of its loops.

The main energy contributions are loop entropies, hydrogen bonds and bases stacking. Base stacking energies and hydrogen bond contributions can be theoretically computed with the help of quantum chemistry. Practically however, the energy model considers only energy differences between folded and unfolded states in an

	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	-0.5	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 2.1: Free energies for stacked pairs in kcal/mol. Note that both base-pairs have to be read in 5'-3' direction.

aqueous solution with a high salt concentration. As a result, the corresponding energy parameters are derived from melting experiments.

For small loops, the loop energy is dependent on its sequence composition only [166]. In contrast, the energy of larger loops are dependent on the base composition of the closing and opening base pairs as well as the length and the asymmetry of the loop. Polymer theory predicts that for large loops, the corresponding loop energy grows proportionally to the logarithm of the loop length. To allow efficient dynamic programming algorithms, the free energy of multiloop is modeled in a slightly different way. In this case, energies grows linearly with the loop size and loop degree.

The free energies of stacked base pairs are shown in Table2.1. In Figure 2.5, the interior loop energies for different sizes and asymmetry is shown.

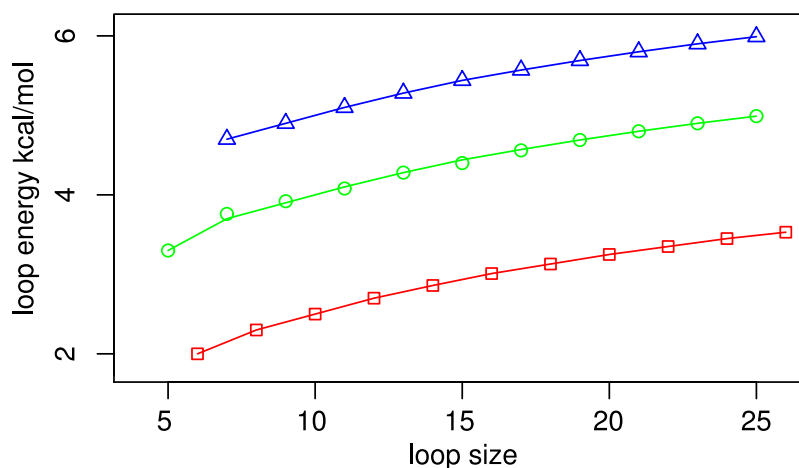


Figure 2.5: Plot of the interior loop free energies against the loop length for different loop asymmetries (red: no asymmetry, green: asymmetry of size 1, blue: asymmetry of size 2)

### 2.2.3 RNA folding with the loop-energy model

The loop based energy model allows to greatly improve RNA structure predictions compared with the pair-matching model. This gain in accuracy comes however at the cost of a slightly more complicated folding algorithm. Still the runtime complexity and memory footprint remains equal at  $\mathcal{O}(n^3)$  and  $\mathcal{O}(n^2)$  respectively. The main difference between both models is that in the case of the loop energy model we have to decompose the set of substructures enclosed by the base pair  $(i, k)$  according to the loop types (see Figure 2.6).

In contrast to hairpin- and interior-loop that decompose into the same kind of loops, multi-loop decomposition needs more attention. Multi-loop energy depends directly on the number of component they have. As a consequence, we need to keep track of the number of components. This is solved by decomposing a multiloop into two parts: a 5' multi-loop component with at least one stem and a 3' part that contains exactly one stem. Both parts can be decomposed into known components: unpaired substructures, shorter multiloops or substructures delimited by a base pair.

The above decomposition enables to easily derive the recursion for computing the minimal energy of a RNA structure. According to figure 2.6 we need the following

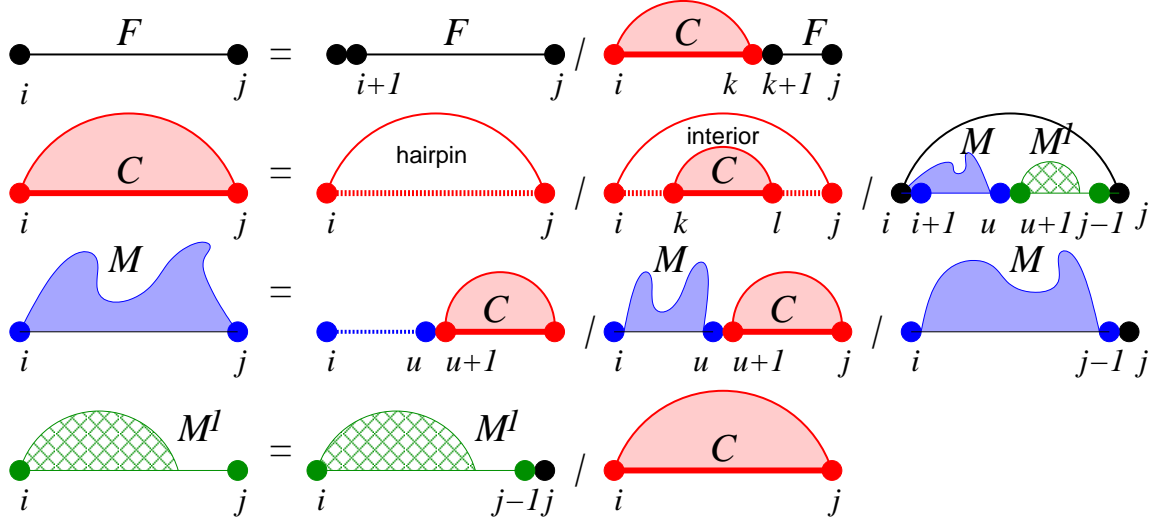


Figure 2.6: Loop decomposition of RNA secondary structure. Hairpin and interior loops are shown in red. Multiloop with more than one component are shown in blue, while multiloop with exactly one component are shown in green. Base Pairs are depicted by arcs. Dotted lines represent unpaired substructures. Taken from [101].

tables during the recursion:

$F_{ij}$  Minimal free energy of the optimal structure on the subsequence  $s[i..j]$ .

$C_{ij}$  Minimal free energy of the optimal structure on the subsequence  $s[i..j]$  given that  $s_i$  and  $s_j$  are paired.

$M_{ij}$  Minimal free energy of the optimal structure on the subsequence  $s[i..j]$  given that there is at least one stem between  $s_i$  and  $s_j$ .

$M_{ij}^1$  Minimal free energy of the optimal structure on the subsequence  $s[i..j]$  given that there is exactly one stem between  $s_i$  and  $s_j$  and  $s_i$  is paired.

The recursion can then be formulated as :

$$\begin{aligned}
F_{ij} &= \min \left\{ F_{i+1,j}, \min_{i < k \leq j} C_{ik} + F_{k+1,j} \right\} \\
C_{ij} &= \min \left\{ \mathcal{H}(i,j), \min_{i < k < l < j} C_{kl} + \mathcal{I}(i,j;k,l), \min_{i < u < j} M_{i+1,u} + M_{u+1,j-1}^1 + a \right\} \\
M_{ij} &= \min \left\{ \min_{i < u < j} (u - i + 1)c + C_{u+1,j} + b, \min_{i < u < j} M_{i,u} + C_{u+1,j} + b, M_{i,j-1} + c \right\} \\
M_{ij}^1 &= \min \{ M_{i,j-1}^1 + c, C_{ij} + b \}
\end{aligned} \tag{2.6}$$

where  $\mathcal{H}(i, j)$  is the energy of hairpin loop enclosed by base pair  $(s_i, s_j)$  and  $\mathcal{I}(i, j; k, l)$  represents the energy of an interior loop delimited by base pair  $(s_i, s_j)$  and  $(s_k, s_l)$ . In this recursion the multiloop energy varies linearly with the loop size and degree and has the form  $E_{\text{ML}} = a + b \cdot \beta + c \cdot l$ , where  $\beta$  is the number of branches and  $l$  is the length of the multiloop (unpaired nucleotide). A careful look at the recursion tells us that the time complexity of this recursion is  $\mathcal{O}(n^4)$ . It can however be reduced to  $\mathcal{O}(n^3)$  by limiting the size of interior loops to an arbitrary constant  $D$ .

There exist alternative implementation of this recursion, however the version shown allows to unambiguously enumerates all possible substructures. Although this is strictly speaking not necessary for retrieving the minimal free energy structure of a given RNA sequence, it does become important when the partition function has to be computed, as each structure has to be counted exactly once.

## 2.3 Cofolding of two sequences

While there exist numerous programs to fold single RNA sequences, only few deals with the problem of folding two or more sequences. The *Hyther* package [197] predicts the hybridization thermodynamics of a given duplex given two strands. It does not produce any secondary structure information nor does it try to minimize the joint free energy.

[229] and others devised a slightly more refined method, where both sequences are linked together and then folded as a pseudo single-sequence with programs like *mfold* or *RNAfold*. As linkers, either short sequences that form very stable structure or

nucleotides that may not interact with other base pair were used. Both type of linkers lead to erroneous predictions of sequence and structure (see Figure 2.7).

The problems resulting from using a linker was first circumvented by Mathews et al. in their program OligoWalk [167]. In their approach the loops containing the linker are considered separately from the other loops. Hofacker [100] was the first to publish a method that could cofold two sequences without linker. In this approach, the sequences are concatenated without a linker and the position where the two sequences are joined is memorized. Loops containing the concatenation point are handled differently from the other loops. A similar approach was published by Hofacker et al. [7] and Bernhart et al. [18].

Use of the modified RNA folding algorithm for the computation of the duplex structure of two or more sequences has the main disadvantage that only regions located in exterior-loops are allowed to undergo intermolecular interactions. This is a direct consequence of the definition of secondary structure from 2.1.1. In other words, duplex-structures containing intermolecular base pairs involving other kind of loops are considered pseudoknotted in the theoretical framework presented in the previous section and cannot be handled by the recursion presented in 2.6 (see figures 2.8 and 2.9).

While pseudoknots are almost absent of single sequence structures, they do play an important role when two RNAs are interacting. Typical examples of RNA-RNA interactions not handled by recursion 2.6 are H/ACA-snoRNA rRNA interactions in eukaryotes or CopA-CopT and OxyS-fhlA interactions in *E. coli* [8, 124] (see Figure 2.9).

For sequences where extensive complementarity to their putative targets is expected, fast string-searching algorithms like BLAST [3], FASTA [195] or Google [79] have been used. Their main drawback is that they do not give information about the thermodynamics of interaction between the query and the target RNA. Their lack of sensitivity is a real issue when looking for interactions where duplex contains large interior loops, as is the case between miRNA and their targets.

Rehmsmeier [204] and Dimitrov [44] developed a cofolding approach where only interior loops are allowed. That means that no information of the native structure of the interacting RNAs is taken into account. These approaches permit to consider

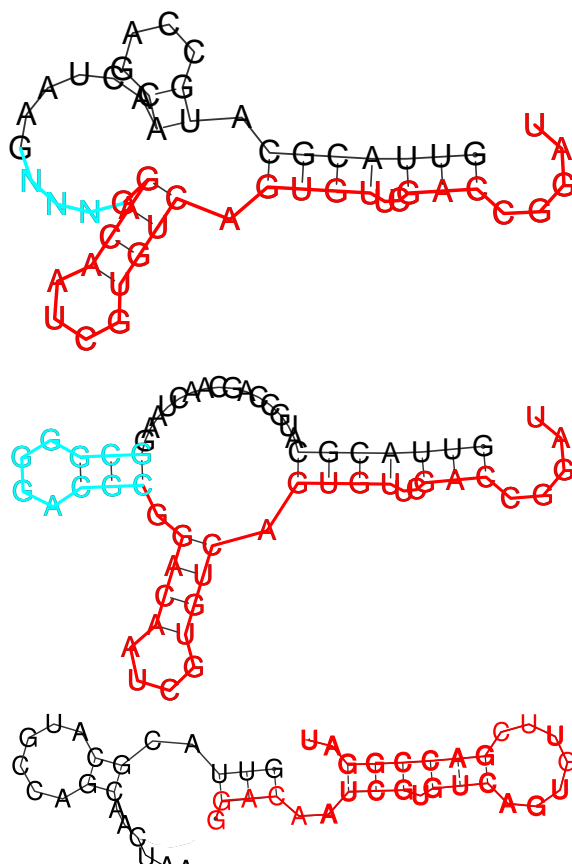


Figure 2.7: Comparison of the minimum free energy of structures of dimers, depending on the kind of linker used to concatenate both sequences. Linkers are drawn in cyan, while the interacting sequences are colored in red and black. **Top:** Structure when using a “poly-N” linker. **Middle:** Structure when using a hair-pin structured linker (from [229]). **Bottom:** Structure from RNAcifold. While the structures are in a narrow energy range (-7.4 to -7.3 kcal/mol), they differ substantially. Taken from [17]

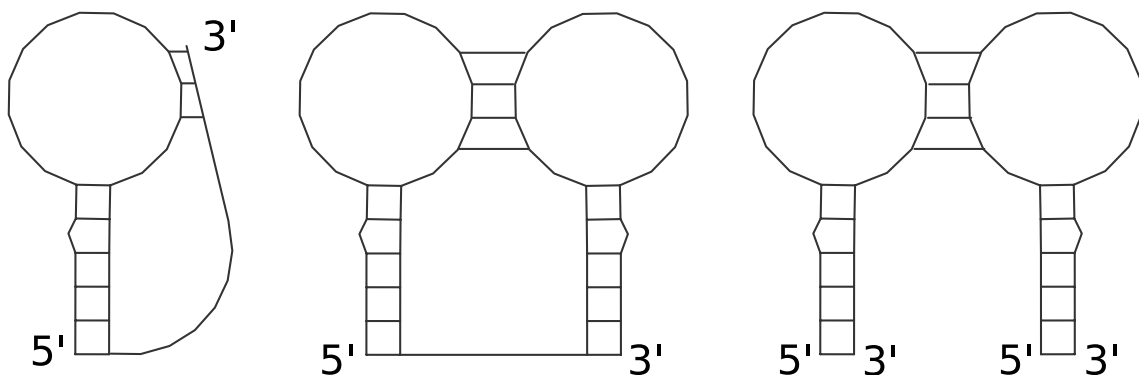


Figure 2.8: Example of pseudoknotted structures. **l.h.s** Typical H-type pseudoknot fold found i.e. in the catalytic core of various ribozymes. **middle** Kissing hairpin pseudoknot found i.e. in the 3' UTR region of the Cocksackie B Virus [252]. **r.h.s** Kissing hairpin-loop interaction between two RNAs. OxyS-fhlA hybrid in *e. Coli* is a typical example of such an interaction. Strictly speaking this is not a pseudoknot, as it involves two distinct sequences. Still this kind of RNA-RNA interactions are not handled correctly by the standard folding algorithm presented in the previous section as it considers it a pseudoknot.

more diverse interactions than sequence based methods alone. Moreover their run-times  $\mathcal{O}(m \cdot n)$ , where  $n$  and  $m$  are the length of the target and ncRNA sequences, respectively, still allow to search genome-wide for putative targets.

A common problem of the sequence based methods as well as the approaches published in [204] and [45] is that no information on the local structures of the interacting RNAs is considered. These approaches consider that all nucleotides are equally able to be involved into intermolecular interactions. While this assumption may hold for short, unstructured RNAs like miRNAs, this is not true for nucleotides involved in very stable intramolecular structures, like bacterial small RNAs.

Neglecting internal structures can lead to important errors in the structure and energy computation of RNA-RNA hybrids. This is exemplified in figure 2.10 where the hybrid involving RybB a ncRNA from *E. Coli* and OmpN, a natural target of RybB, is computed with and without accessibility. In the case where intramolecular structure is not taken into account, the RNA duplex is predicted to extend over the whole length of the ncRNA (bottom of figure 2.10). In contrast, the hybrid found when considering the internal structure is much shorter, involving only 15

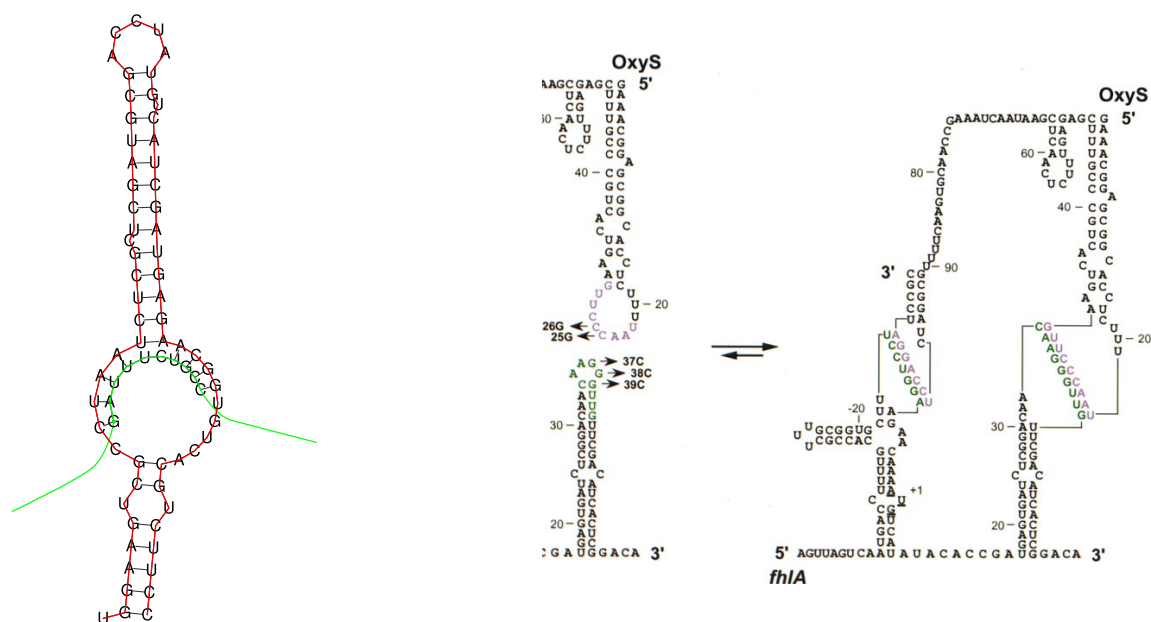


Figure 2.9: Examples of pseudoknotted RNA-RNA interactions. **R.h.s** H/ACA snoRNA (red) interaction with its target (green). **L.h.s** Bound (right) and unbound (left) conformations of OxyS and fhlA.

nucleotides.

Although the interaction region between both RNAs is shorter in the case where accessibility is considered, the stability of the hybrid is higher, with an interaction energy of roughly -16 kcal/mol. In contrast the interaction energy of the first duplex reaches only -1.6 kcal/mol. The difference in energy between both duplexes is due to the high amount of energy needed to **completely** unfold the ncRNA (24.7 kcal/mol) and half of the mRNA (14.9kcal/mol) in order to form the hybrid. The second hybrid is more economical as the cost for opening the mRNA (3.9 kcal/mol) and the ncRNA (1.6 kcal/mol) amounts 5.5 kcal/mol.

Pervouchine et al. [196], Alkan et al. [2] and more recently [38] and [107] published RNA cofolding methods able to consider complex structures found in the majority of RNA-hybrids. While these methods are able to correctly predict complex interaction structures, their high runtimes, ( $\mathcal{O}(n^3 \cdot m^3)$  for Pervouchine and  $\mathcal{O}((n + m)^6)$  for Alkan, where  $n$  and  $m$  represent the length of the first and second sequences) make them inappropriate for large scale studies.

In light of the available tools devoted to RNA cofolding, it is clear that the functional annotation of the rapidly increasing number of ncRNAs is still in its infancy. On one hand there are tools available that could be used to find ncRNA targets in a reduced amount of time, however with a high trade-off on accuracy. On the other hand, the high runtime of more precise tools able to handle more complex interactions make genome-wide target search for novel ncRNAs impracticable. In chapter 5 a new approach named **RNAplex** that can search for ncRNA targets with the accuracy of methods considering local RNA structures but with a runtime of  $\mathcal{O}(n \cdot m)$  will be presented. In 6 a method will be presented that can treat a very specific RNA-RNA interaction found between H-ACA snoRNAs and their targets. While general RNA-RNA interaction tools like [2, 38, 107, 196] could in theory handle this kind of interactions we will show that the specially tailored algorithm presented in chapter 6 performs very well, with a runtime directly proportional to the length of the target sequence  $\mathcal{O}(n \cdot m^2)$ , suitable for genome-wide target search.



## 2.4 non-coding RNA

---

Non-coding RNAs (ncRNAs) are functional RNA molecules that do not code for proteins. There exist several group of ncRNAs involved in a wide spectrum of cellular process. For example ribosomal RNAs (rRNAs) catalyze the peptide bond formation ([89]) during the translation process. Small nuclear RNAs (snRNAs) are involved in splicing of the mRNAs. Small nucleolar RNAs (snoRNAs) are responsible for the processing of rRNAs, tRNAs and snRNAs and their correct folding [66, 211]. snRNAs (small nuclear RNAs) are critical components of the spliceosome, the large ribonucleoprotein complexes that splice introns out of pre-mRNAs [277]. ncRNAs may also control gene expression. In higher eukaryotes, micro-RNAs (miRNAs) regulate gene expression by binding to targets mRNA [139, 140]. In bacteria, the OxyS ncRNA repress *fhlA* by binding to its ribosome entry site, precluding the translation [8]. In eukaryotes, [68] showed that exogenous small double stranded RNAs are able to regulate protein concentration once transfected in the cells.

Some ncRNAs have more than one functions. SgrS RNA and RNAIII in *E. coli* encode both for a non-coding RNA and a protein [49, 261]. Recently snoRNAs acting as miRNAs were found in human and *Giardi lamblia* [64, 212]. In mammals it has been proved that snoRNAs can also regulate the alternative splicing of mRNA [120]. snoRNA U85 has box a C/D- and a H/ACA-box domain. Accordingly, U85 pseudouridylates and methylates snRNA U5 [121].

In the next subsections we review in more details the classes of non-coding RNAs for which we have been searching for targets. Those are miRNAs, siRNAs, sRNAs and snoRNAs. As already mentioned in previous sections, there exist a lot more ncRNAs than the few families that are going to be presented here.

### 2.4.1 miRNA

miRNAs are single stranded RNA molecules found in eukaryotes, whose main function is to mediate post-transcriptional gene silencing by imperfectly binding to the 3'-UTRs of target mRNAs [130, 136, 139].

The maturation of microRNAs is a multiple-steps process. First a DNA region encoding a miRNA is transcribed, leading to a pri-miRNA transcript whose length

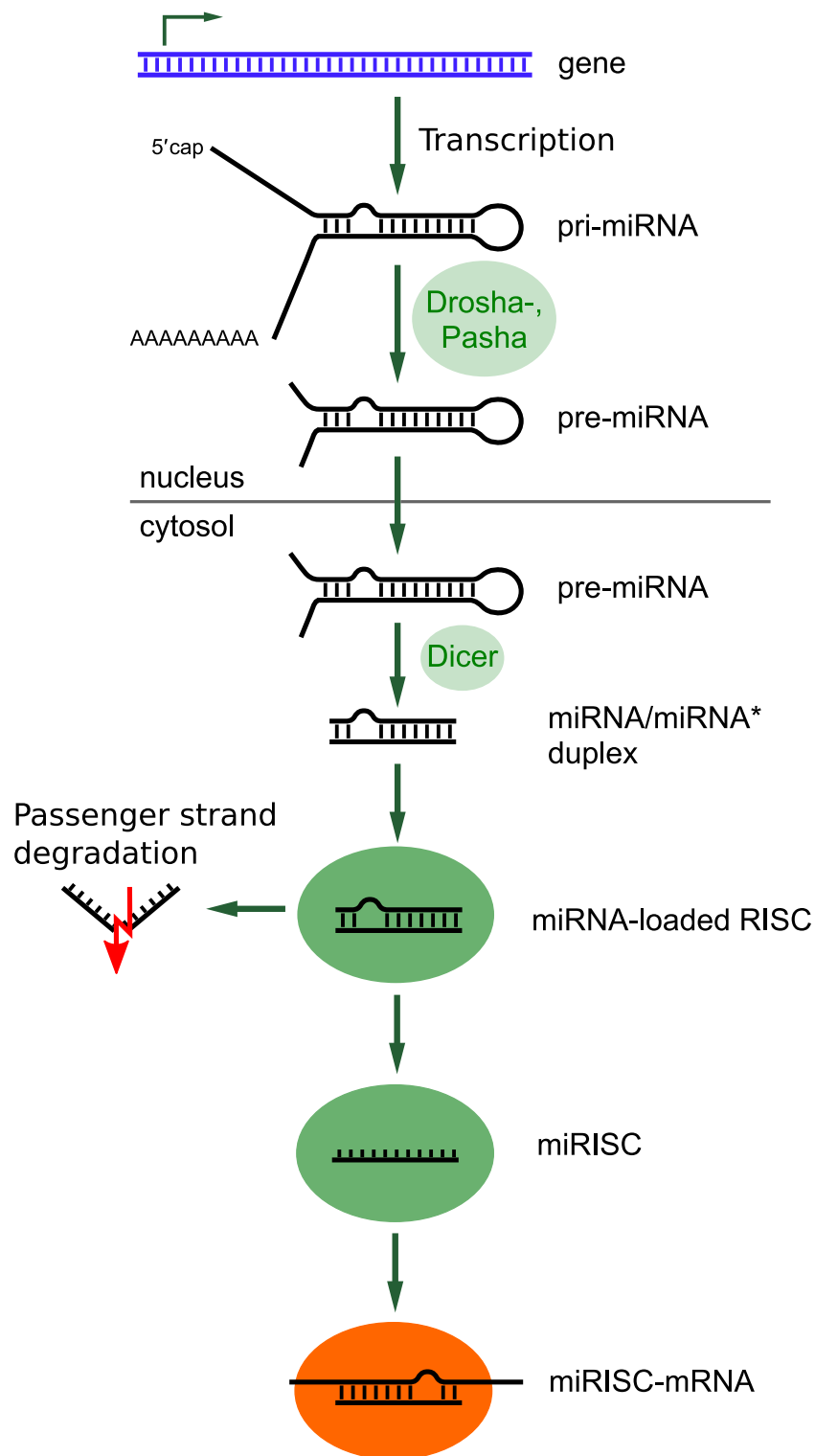


Figure 2.11: Overview of the miRNA maturation process. First miRNAs are transcribed from their loci into pri-miRNAs **top**. pri-miRNAs are then processed by Drosha and Pasha proteins into pre-miRNAs. Dicer processes pre-miRNAs into short double stranded miRNA/miRNA\* duplexes. These duplexes get loaded into RNA-induced silencing complex. Generally the strand with the less stable 5' end is introduced into RISC, while the other strand (passenger strand) is degraded. Once loaded into RISC, miRNAs are ready to recognize their targets through base pairing, leading to the mRNA degradation and/or translation disruption

vary between few hundreds up to tens of kilobases [210]. pri-miRNA are then processed by Drosha and Pasha into 70nts long stem-loops, called pre-miRNA [41]. pre-miRNAs are then processed by Dicer into a 21-23 nts long RNA duplex. Generally the strand that is less stable on its 5' end is subsequently introduced into the RNA-induced silencing complex (RISC). The other strand (the passenger strand) is degraded [84, 218]. Once integrated into the RISC complex, miRNAs hybridize with their cognate mRNAs and can either inhibit translation or cleave their target mRNA, leading to the downregulation of the gene encoded by the mRNA. Cleavage is achieved by argonaute, the catalytic protein which is part of the RISC complex [147] (see Figure 2.11 and Figure 2.12).

Due to the reduced size of miRNAs, the duplex structures that they form with their targets is simple when compared to snoRNA- or sRNA-hybrids (see Figure 2.11 for miRNAs and figure 2.9 for snoRNAs/sRNAs). Still miRNAs targets prediction is a difficult task and actual prediction tools perform poorly. There are different reasons why miRNA target predictions is unsatisfactory. First, miRNAs are very imperfectly bound to their targets [229]. Moreover the search space for putative targets is huge, as miRNAs can potentially bind to any mRNA transcripts [123, 229]. Further the regulation pattern miRNAs is relatively complex as one miRNA can regulate several hundred mRNAs [81, 82], and reversely a mRNA can be targeted by more than one miRNA. Finally until recently the lack of experimental data made it difficult to extract effective prediction rules.

The penury of sufficient data lead to the publication of several miRNA-target prediction rules. However up to now, none of them has been unanimously accepted. The most common rule is the so-called seed-rule. It states that functional miRNA-targets must contain a stretch of 6 contiguous nucleotides complementary to the nucleotides 2-7 of the miRNA [142]. The seed hypothesis derives from the facts that often seed regions are perfect complementary to 3'-UTR functional elements that mediate posttranscriptional downregulation [131]. Furthermore it was shown that the seed region is more conserved than the rest of the miRNA [145]. Many seed targets were validated in-vivo, and several transcriptomics and genomics essays showed that genes containing miRNA seeds were preferentially regulated under miRNA overexpression/inactivation [144, 221]. Still, long before the seed rule was phrased, it had

been shown that *C. elegans* lin-4 and let-7 form non-seeded duplex with their targets [205]. In human, this was demonstrated for miR-10a, which targets ribosomal protein transcripts via non-seed sites [189]. In *D. melanogaster*, it was shown that many miRNAs targets do not have seed matches [58]. Johnston et al. even showed that, at least in *C. elegans*, the presence of seed is a poor target predictor [114].

Another important rule for miRNA target prediction states that miRNA-mRNA interactions in animals happen exclusively in the 3'-UTR region of the target. This hypothesis is mainly a consequence of the first two discovered miRNAs that targeted their cognate mRNAs in the 3'-UTR region [140, 205]. Another element that led to this assumption is that many miRNA target prediction tools use target site conservation as criteria to filter out false positives, i.e. mRNA that are complementary to a miRNA by chance. Because the conservation of miRNA target sites in the open reading frame (ORF) might be a collateral effect of codon conservation, miRNA target prediction programs specialized to the 3'-UTR. Although this rule allowed to predict numerous miRNA targets [222], it also led many scientist to disregard the coding and 5'-UTR regions as potential miRNA targets [86, 144]. Only recently first reports on miRNA targeting regions outside the 3'-UTRs were published [56, 133, 142, 154, 175, 230].

Other factors have drawn the attention of the miRNA community. Recent studies have shown that the target accessibility, i.e. the degree of unstructuredness in and around the target sites, is an important feature for predicting miRNA-mRNA targets [117, 187, 208, 241], similar to what has been shown in other classes of ncRNA-RNA interactions [6, 233–235, 237]. Chapter 5 and 4 will present an overview of our findings on these topics.

Apart from RNA-structures, the presence of proteins on a target site is detrimental to the formation of the miRNA-target hybrid [22, 116]. Contextual sequence features, not directly related to target accessibility, protein binding site or miRNA binding site, have been reported to strongly influence the repression efficiency of miRNAs [43, 254]. Finally the relative in-vivo concentrations of mRNA and miRNA is a further parameter that greatly influence miRNA repression efficiency [53].

Parallel to the publication of new target predictors, new miRNA target predictions methods were published. An non-exhaustive list is presented in Table 2.2.

Name	Method	Availability	Reference
DIANA-microT	Conservation, Hybridization, Accessibility	Flat files	[159]
EIMMo	Conservation,	Flat files, web inter- face	[73]
miRanda	Hybridization	Flat files	[113]
PicTar	Hybridization, Conservation	Flat files	[134]
PITA	Accessibility, Hy- bridization,	Flat files	[117]
RNA22	Hybridization, Pattern	Flat file	[175]
TargetScan	Conservation, others	Flat files, web inter- face	[72]

Table 2.2: Summary of widely used miRNA target prediction tools. The first column contains the name of the tools. The second column indicates the method used by the tools. Conservation means that conservation of the seed/target site is important. Hybridization means that the energy of interaction between the miRNA and its target is relevant. Accessibility means that the structuredness of the target site is taken into account. Besides the conservation of the target site, TargetScan further considers the hybrid structure, the position of the target site on the 3'-UTR as well as the AU content around the target site. The third column lists how target information can be accessed. The last column reports the corresponding literature citation.

Although the prediction accuracies of miRNA target tools grew steadily during the last years, their performances are still unsatisfactory. In a recent review [159], where 10 target prediction tools were confronted, it was shown that the most precise target prediction method only achieved a precision of 58% at a sensitivity level of 4%. More disturbing was the fact that the simple seed rule had a higher precision than 8 of the 10 reviewed tools.

This less than satisfactory situation is mainly due to the quality of the data on which the tools are trained. The majority of the miRNA data published up to now contains information on mRNA concentration variation upon miRNA over-/underexpression ([191]). This kind of data have three important short-comings. First, they do not allow to precisely locate the miRNA binding site on the mRNA. Second, these data allow only to determine mRNA-miRNA interactions leading to the cleavage of the target RNA. We are completely missing interaction leading to the translation inhibition of the target RNA. Finally the quantitative determination of mRNA concentration variation do not allow to precisely model the corresponding protein concentration variation [221].

Still all of the mentioned shortcomings can currently be resolved by using alternative experimental settings. Localization of the interaction site on the mRNA is achieved by site directed mutagenesis. This is however realized in only 10% of the reported targets ([191]). Identification of targets repressed by translation inhibition are identified using high-throughput proteomic methods called stable isotope labeling with amino acids in cell culture (SILAC). In this approach unlabeled ('light' or L) cells are transferred to medium with 'heavy' (H) or 'medium-heavy' (M) amino acids concomitantly with transfection of miRNA. In the subsequent labeling phase the H and M amino acids are incorporated into all newly synthesized proteins. The abundance ratio of H versus M reflects differences in translation of the corresponding proteins under the two conditions [217, 221].

## 2.4.2 **siRNA**

RNA interference (RNAi) describes the post-transcriptional gene silencing process triggered by endogenous or exogenous double stranded RNAs (dsRNAs). After being processed by Dicer, the dsRNAs are transferred to the RNA-Induced Silencing

Complex (RISC), where one of the strands (the guide strand) is introduced while the other strand is degraded (the passenger strand). Target recognition happens through hybridization of the guide RNA with its target gene, which causes the cleavage and the subsequent degradation of the target strand.

The successful utilization of artificial dsRNAs to knockdown specific genes was first reported by Fire et al. [68]. In 2001 Elbashir et al. [61] showed that siRNA-mediated gene knockdown could also be applied in mammalian cells. Initial expectations that there were no need to search for optimal siRNA sequences [231], rapidly proved to be unfounded, as strong variations in silencing efficiency were reported for different siRNAs directed against the same target [104]. Still the potential of RNAi to transiently knockdown genes motivated the scientific community to improve the siRNA design rules (for a review see [193]). Elbashir et al. [63] published the first protocol for designing active siRNAs. They encouraged the use of 21 nucleotides long siRNAs with a G/C content of about 50% and 2 nucleotides 3'overhangs.

In 2003, Khvorova et al. [118] as well as Schwarz et al. [218] proved that even though both strands of the dsRNA could serve as a guide strand [61, 63], the strand with the lower 5' stability was preferentially incorporated into the RISC complex. Subsequent studies concentrated on finding sequence patterns on the guide strand which correlated with the repression efficacy [5, 103, 106, 207, 239, 248]. The majority of those studies confirmed that the relative stability of the siRNA ends was a major determinant of the functionality of siRNAs. Further improvements in the design of siRNA came from the study published by Patzel et al. [194], who showed that the siRNA efficiency directly correlate with the siRNA structuredness.

The small number of siRNAs used in those early studies led to poor agreements on the sequence patterns and to parameter overfitting [209]. The use of heterogeneous data, gathered either from previous work or from siRNA databases (for example siRecords [206]), did not resolve this issue, as the oligonucleotides activity is highly sensitive to biological and experimental parameters (transfection efficiency, cell type, siRNA concentration, target concentration, efficiency measure). To overcome those problems Huesken et al. [108] generated a set of 2431 randomly selected siRNAs targeted against 34 mRNAs, which was used to train an artificial neural network for designing siRNAs. Statistical analysis of this data set confirmed some of the

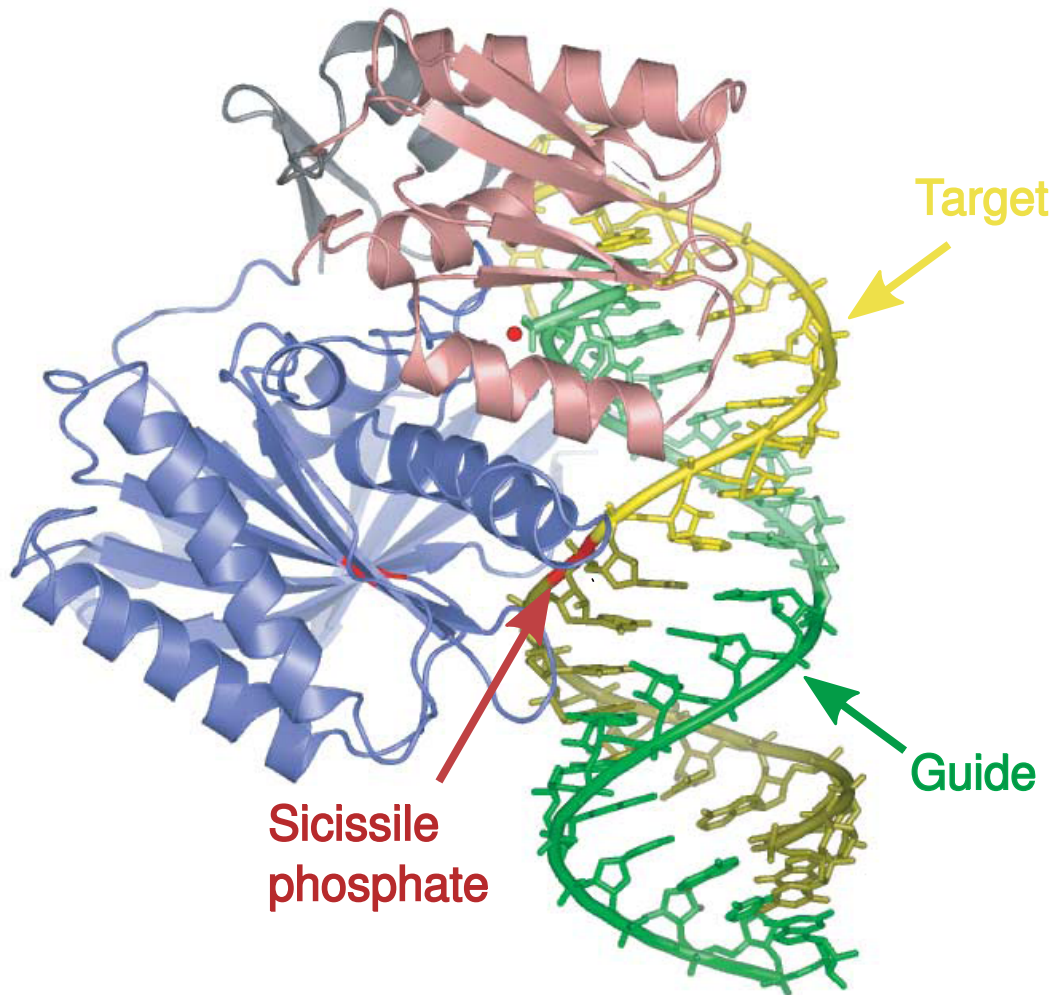


Figure 2.12: Structure of a 19 nucleotides RNA duplex bound to Afpiwi. Afpiwi is an archeal PIWI domain-containing protein which is used to model eukaryotic Argonaute. The guide strand is depicted in green, while the target RNA is in yellow. The region on the mRNA that is cleaved by Argonaute is shown in red. Adapted from [192]

previously published siRNAs features (duplex asymmetry) and revealed new, highly significant sequence motives.

A long debated topic in the field of siRNA design is the influence of the target structure on the siRNA efficiency. While target site structure was recognized as an important feature in the design of antisense oligonucleotides and ribozymes [47,146,173,174,257,272], data arguing for [6,24,32,46,127,151–153,190,216,223,234,256,269,270] and against [23,194,207] the influence of target site accessibility on the siRNA efficiency were reported. As will be shown in chapters 3 and 5, the interaction of two RNAs can be decomposed into two stages. Binding can only occur at positions not already involved in intramolecular base pairs. Thus, base pairs within the target site have to be opened to make the site *accessible*. The energy necessary to do this is termed the disruption or breaking energy. Once the binding site is devoid of structure intermolecular helices can be formed, yielding a stabilizing interaction energy. The total binding energy is then computed as the sum of the hybridization energy and the breaking energy.

In principle such a model could directly predict the fraction of mRNAs that will be bound by siRNAs. This, however, requires knowledge of siRNA and mRNA concentrations which are in general not available. Furthermore, the model implicitly assumes that reactants are free solutes, thus neglecting possible influences of mRNA binding proteins, active translation by the ribosome, and the RISC complex on the siRNA binding. Still, the application of this approach on siRNA data published by Schubert et al. [216] (see chapter 3), where a siRNA was targeted to a gradually less accessible target site, showed that siRNA efficiency is directly correlated to the target site accessibility ([180,181]). Those findings were corroborated by five further studies [151,152], [46,223], [234] (see chapter 4) which looked specifically at the effect of local target secondary structure on RNAi efficiency based on large (100 siRNAs against 3 genes) to very large (3084 siRNAs against 82 genes) homogeneous data sets.

The majority of the siRNA design rules mentioned above can be mapped to key events of the silencing pathways (see Table2.13). The limited length of the siRNA duplex

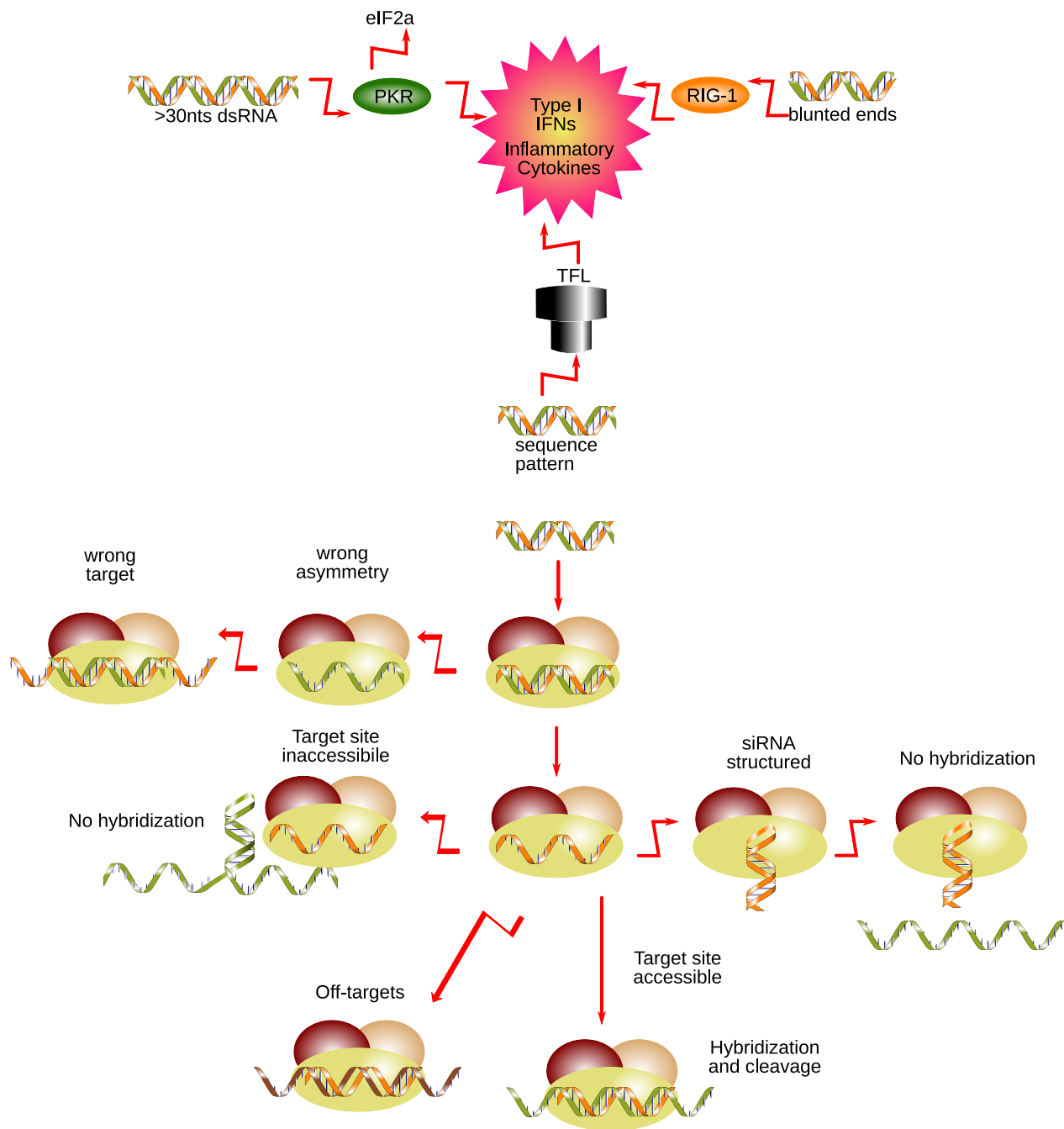


Figure 2.13: Impact of siRNA characteristics along the silencing pathway. The innate immune system may be activated by dsRNAs. dsRNAs with specific sequence patterns or high "U" contents are recognized by Toll Like Receptors (TLRs) inducing inflammatory cytokines and interferon of type I (IFN- $\alpha$ , IFN- $\beta$ ). Large dsRNAs (>30nts) are sensed by PKR (double-stranded RNA-activated protein kinase) which can induce interferon response, expression of inflammatory cytokines and cell death. dsRNAs with 2nts overhangs escape the RIG-1 triggered cytokines and interferon response. Once into RISC, the passenger strand is separated from the guide strand. The strand with the lower 5'-end stability is incorporated into RISC, while the other strand is degraded. A wrong asymmetry results in the selection the bad siRNA strand, leading to no on-target effect. siRNAs that are highly structured are not able to hybridize to their target. Reciprocally siRNAs targeting highly structured region can not bind to their target. Finally sequence specific off-target effects makes it more difficult to gain information from RNAi experiments.

as well as the presence of 3' end dangles allows the siRNA to evade immunorecognition [92, 105, 161]. The rules promoting the sequence/energy asymmetry [118, 218] reflect the ability of Dicer to sense the thermodynamic asymmetry between the two ends of the duplex. The negative effect of structure of the guide strand on the repression efficiency may be explained by a reduced ability of the siRNA to bind to its target and/or hindered interaction of the siRNA with RISC components [194]. Finally the importance of the target site accessibility on the siRNA efficiency derives from a) the ability of RISC to bind to single stranded region only and b) the inability of RISC to unfold structured RNA [6].

### 2.4.3 sRNA

sRNAs are non-coding RNAs found in bacteria. sRNAs are very heterogeneous both in sizes, structures and functions [260]. Most of them act as post-transcriptional regulators by interacting with the 5' untranslated region of mRNA transcripts, modifying their stability and/or their ability to be translated [148].

sRNA-mRNA interaction structures show different levels of complexity. For example *micC*, a siRNA found in enterobacteria, has a stretch of 16 nts fully complementary to its target *ompC*. Other interactions, like *copA-copT* [124], *RNAIII-rotA* [25], and *OxyS-fhlA* [4] rely on more intricate interactions, involving one or more kissing loop complexes (see figure 2.9).

Similar to miRNAs in eukaryotes, sRNAs may target more than one mRNAs and a mRNA may be targeted by more than one sRNA. In some cases sRNAs can work not only as downregulator but also as upregulator [137, 138, 156].

Many approaches have been used to find sRNA targets. BLAST was successfully used to identify the targets of *micC* [37], and *IstR-1* [259]. TargetRNA [244] is a target search tool that computes hybridization score for sRNA-mRNA hybrids and return a ranked list of target RNAs. It is similar to the tools developed by Rehmsmeier [204] and [45] where only interior loops between the sRNA and its target are allowed. The hybridization score is based either on the loop energies model or on the maximum matching model. Mandin et al. [158] followed a similar approach but used an hybrid energy model, where experimental values were used to compute stacking energies, and an empirical approximation of the loop cost.

While these methods proved useful for detecting some sRNA-RNA interactions, they do have some limitations. They are not able to detect upregulating interaction like DsrA-rpoS. Further they neglect the influence of target site accessibility on the interaction, resulting in wrongly predicted target and/or target location (see Figure 2.10). Recently, more complex approaches from [38, 107] described successfully complex interactions like Oxys-fhlA and copA-copT. While their precision make them valuable tools to study known interactions, their high runtime make them unpractical for genome-wide target search.

In the course of this work we will present an algorithm, called **RNAup**(chapter 3), that can predict both the sRNA targets and also the influence of the interaction on the mRNA, i.e. if the interaction upregulate or downregulate its target. We will further develop a program, called **RNAplex** (chapter 5), that have a runtime similar to that of TargetRNA or **RNAhybrid**, but an accuracy similar to that of **RNAup** due to its ability to consider target accessibility.

## 2.4.4 **snoRNA**

snoRNAs are non-coding RNAs that are mainly responsible for two kinds of post-transcriptional nucleotide modifications in rRNA and snRNAs. The first type of modification, called methylation, consists in the attachment or substitution of a methyl group onto either the ribose group or the residue of the target sequence. This modification is conducted by the C/D-Box snoRNAs.

Structurally, those RNAs are characterized by two short conserved motif called C and D box, whose sequence are **UGAUGA** and **CUGA**, respectively. The general shape of these snoRNAs consist of a small stem involving the 5' and 3' ends of the sequence, that closes a large hair-pin loop. Four proteins, that are responsible for the rRNA methylation, associate with the snoRNA [10]. This ribonucleoprotein (RNP) complex recognizes the nucleotide to be methylated with the help of a 10-21 nucleotides region located 10 nucleotides upstream of the D-box that is complementary to the methylation site [122] (see Figure 2.14)

The second type of modification, called pseudouridylation, consists in converting an uridine into a pseudouridine. This modification is conducted by the H/ACA-snoRNAs. Like their C/D-Box counterparts, these snoRNA contain conserved se-

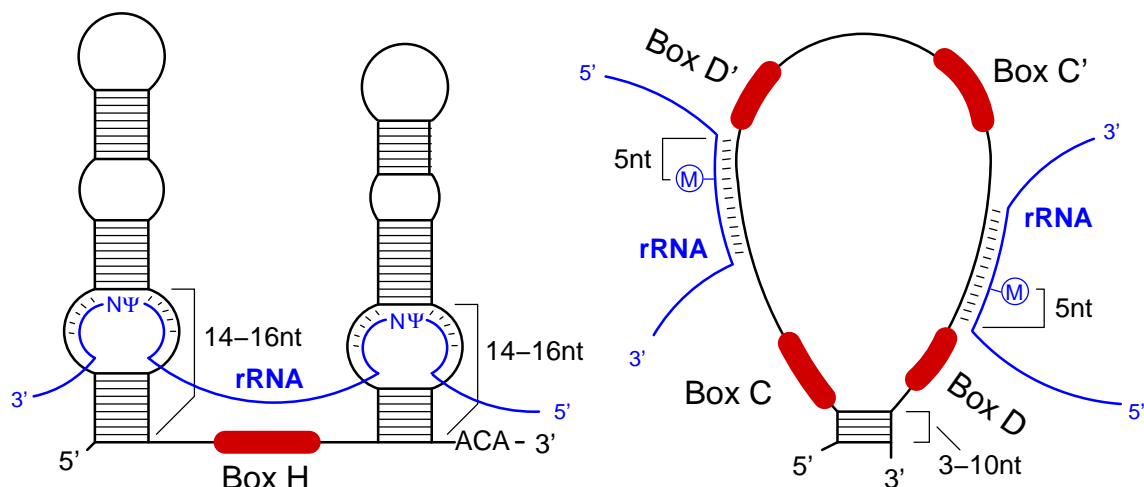


Figure 2.14: Canonical C/D and H/ACA snoRNAs structures. **L.h.s** CD-Box snoRNA, made of a small stem and a large loop. The loop region contains either one or two sets of C/D boxes. The region directly upstream of the D boxes is responsible for the correct target recognition. **R.h.s** HACA-Box snoRNA, made of two target stems separated by an unpaired region containing the H Box. The interior loop in each stem is responsible for the correct target recognition.

quence motifs, called H-Box *ANANNA* and ACA-Box *ACA* [74]. The shape of the H/ACA snoRNAs consists of two hairpins and two single stranded regions [10]. Both hairpin regions contain a bulge, also termed interaction bucket or recognition loops, which are complementary to the pseudouridylation sites. In the framework of *RNAcofold*, the binding pattern is a complex pseudoknot, where both arms of the loop region are involved (see figure 2.8). Similarly to the C/D-Box snoRNAs, the H/ACA-box snoRNAs are associated with four proteins forming a RNP [74] (see Figure 2.14).

Target prediction for C/D box snoRNA is a comparatively easy task, as it only involves the search for complementary regions in the target sequence. *snoTarget* [14], a program specifically designed to find C/D box snoRNA target, uses text based methods to find putative targets. For each targets it then uses *RNAcofold* to gain information about the interaction energy. H/ACA snoRNA target predictions is more complicated as a complex pseudo-knot structure is involved. Currently only the algorithm developed by Pervouchine [196], Alkan [2], Chitsaz [38] and Huang [107] can correctly handle this kind of structure. They are however too slow for any

practical use.

In chapter 6 a new method, named **RNAsnoop**, especially developed to find H/ACA-snoRNA targets will be presented. While it can only be used to detect H/ACA snoRNA-RNA interactions, it does so in  $\mathcal{O}(n \cdot m^2)$ , where  $n$  is the length of the target sequence and  $m$  the length of the snoRNA. This makes **RNAsnoop** suitable to search putative targets not only on rRNAs and snRNA, but also genome-wide for orphan snoRNAs.



In this section we present an alternative approach to the RNA cofolding algorithm reviewed in section 2.3 by taking into account that the oligo can bind also to unpaired sequences in hairpin, interior, or multi-branch loops. These cases could in principle be handled using a generic approach to pseudoknotted RNA structures [50, 51] at the expense of much more costly computations. Instead we conceptually decompose RNA-RNA binding into two stages: (1) we calculate the partition function for secondary structures of the target RNAs subject to the constraint that a certain sequence interval (the binding site) remains unpaired. (2) We then compute the interaction energies given that the binding site is unpaired in the target. The total interaction probability at a possible binding site is then obtained as the sum over all possible types of binding. The advantage is that the memory and CPU requirements are drastically reduced: For a target RNA of length  $n$  and an oligo of length  $m < n$  we need only  $\mathcal{O}(n^2)$  memory and  $\mathcal{O}(n^3 \cdot m)$  time (compared to  $\mathcal{O}(n^2)$  memory and  $\mathcal{O}(n^3)$  time for folding the target alone).

## 3.1 Algorithm

---

Here we present the algorithmic details of folding two RNA sequences based on our two-stages approach. In the following let  $F(\mathcal{S})$  denote the free energy of a secondary structure  $\mathcal{S}$ , and write  $\beta$  for the inverse of the temperature times Boltzmann's constant. The equilibrium partition function is defined as  $Z = \sum_{\mathcal{S}} \exp(-\beta F(\mathcal{S}))$ . Since the frequency of a particular structure  $\mathcal{S}$  in equilibrium is given by  $P(\mathcal{S}) =$

$\exp(-\beta F(\mathcal{S}))/Z$ , partition functions also provide the starting point for computing the frequency of a given structural motif. In particular we are interested in the probability  $P_u[i, j]$  that the sequence interval  $s[i..j]$  is unpaired. Denoting the set of secondary structures in which  $s[i..j]$  remains unpaired by  $\mathcal{S}_{[i,j]}^u$  we have

$$P_u[i, j] = \frac{1}{Z} \sum_{S \in \mathcal{S}_{[i,j]}^u} e^{-\beta F(S)} \quad (3.1)$$

Clearly, the set  $\mathcal{S}_{[i,j]}^u$  will be exponentially large in general. In the special case of an interval of length 1, i.e., a single unpaired base,  $P_u[i, i]$  can be computed by dynamic programming. Indeed,  $P_u[i, i] = 1 - \sum_{j \neq i} P_{ij}$ , where  $P_{ij}$  is the base pairing probability of pair  $(s_i, s_j)$ , which is obtained directly from McCaskill's partition function algorithm [169]. It is natural, therefore, to look for a generalization of the dynamic programming approach to study longer unpaired stretches. Note that we cannot simply use  $\prod_{k=i}^j P_u[k, k]$  since these probabilities are not even approximately independent, as it will be shown in chapter 5. We first observe that the unpaired interval  $s[i..j]$  is either part of the “exterior loop”, (i.e., it is not enclosed by a base pair), or it is enclosed by a base pair  $(s_p, s_q)$  such that  $(s_p, s_q)$  is the closing pair of the loop that contains the unpaired interval  $s[i..j]$ . We can therefore express  $P_u[i, j]$  in terms of restricted partition functions for these two cases:

$$P_u[i, j] = \frac{Z(1, i-1)Z(j+1, n)}{Z(1, n)} + \frac{\sum_{p < i} \sum_{j < q} \hat{Z}(p, q) Z_{pq}[i, j]}{Z(1, n)} \quad (3.2)$$

The first term accounts for the ratio between the partition functions of all substructures on the 5' and 3' side of the interval  $s[i..j]$  and the total partition function. In the second term,  $\hat{Z}(p, q)$  is the partition function outside base pair  $(s_p, s_q)$ , and  $Z_{pq}[i, j]$  the partition function inside a base pair  $(s_p, s_q)$  given that the interval  $s[i..j]$  is unpaired.

The tricky part of the algorithm is the computation of the restricted partition functions  $Z_{pq}[i, j]$ . The recursion is built upon enumerating the possible types of loops that have  $(s_p, s_q)$  as their closing pair and  $s[i..j]$  is unpaired, see figure 3.1. From

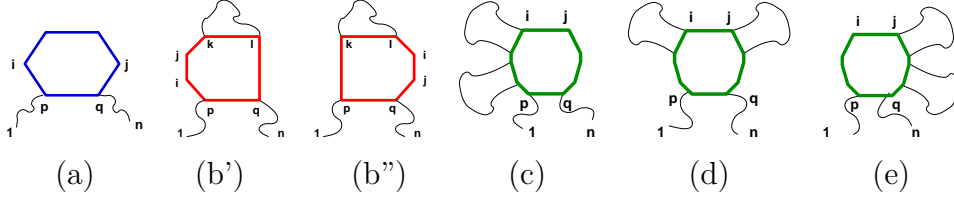


Figure 3.1: A base pair  $(s_p, s_q)$  can close various loop types. According to the loop type different contributions have to be considered. a A hairpin loop is depicted in blue. b In case of an interior loop, which is shown in red, two independent contributions to  $Z_{pq}[i, j]$  are possible: The unstructured region  $s[i..j]$  can be located on either side of the stacked pairs  $(s_p, s_q)$  and  $(s_k, s_l)$ . c If region  $s[i..j]$  is contained within a multiloop we have to account for three different conformations, indicated in the green structures, a more detailed description is given in the text.

this decomposition one derives:

$$\begin{aligned}
 Z_{pq}[i, j] = & \underbrace{\exp(-\beta H(p, q))}_{(a)} \\
 & + \sum_{\substack{p < i \leq j < k \text{ or} \\ l < i \leq j < q}} \underbrace{Z^b[k, l] \exp(-\beta I(p, q; k, l))}_{(b)} \\
 & + \sum_{p < i \leq j < q} \underbrace{Z^{m2}[p+1, i-1] \exp(-\beta c(q-i))}_{(c)} \\
 & + \sum_{p < i \leq j < q} \underbrace{Z^m[p+1, i-1] Z^m[j+1, q-1] \exp(-\beta c(j-i+1))}_{(d)} \\
 & + \sum_{p < i \leq j < q} \underbrace{Z^{m2}[j+1, q-1] \exp(-\beta c(j-p))}_{(e)}
 \end{aligned} \tag{3.3}$$

where  $H(p, q)$  and  $I(p, q; k, l)$  are functions that compute the loop energies of hairpin and interior loops given their enclosing base pairs;  $c$  is an energy parameter for multiloops describing the penalty for increasing the loop size by one. The computation of the multiloop contributions (c-e) requires two additional types of restricted partitions functions:  $Z^m[p, q]$  is the partition function of all conformations on the interval  $s[p..q]$  that are part of a multiloop and contain at least one component, i.e., that contain at least one substructure that is enclosed by a base pair. These quantities are computed and tabulated already in the course of McCaskill's algorithm. There, the computation of  $Z^m$  requires an auxiliary array  $Z^{m1}$  which counts structures in

multiloops that have *exactly* one component, the closing pair of which starts at the first position of the interval. For the one-sided multiloop cases (c) and (e) in figure 3.1 we additionally need the partition functions of multiloop configurations that have *at least* two components. These are readily obtained using

$$Z^{m2}[p, q] = \sum_{p < u < q} Z^m[p, u] Z^{m1}[u + 1, q]. \quad (3.4)$$

It is not hard to verify that this recursion corresponds to a unique decomposition of the “M2” configurations into a 3’ part that contains exactly one component and a 5’ part with at least one component.

It is clear from the above recursions that, in comparison to McCaskill’s partition function algorithm, we need to store only one additional matrix,  $Z^{m2}$ . The CPU requirements increase to  $\mathcal{O}(n^4)$  (assuming the usual restriction of the length of interior loops). In practice, however, the probabilities for very long unpaired intervals are negligible, so that  $P_u[i, j]$  is of interest only for limited interval length  $w$  so that  $|j - i + 1| \leq w$ . Taking this constraint into account shows that the CPU requirements are actually only  $\mathcal{O}(n^3 \cdot w)$ .

We can further divide the runtime by a factor  $w$  by modifying the recursion for  $\hat{Z}(p, q) Z_{pq}[i, j]$ . To achieve this, we start from the observation that  $Z_{pq}[i, j]$  consists of three contributions, of which the summation of all multi-loop energies is the most complex one. This multi-loop part is again split into three parts, depending on whether the unpaired region is to the left or to the right of all components of a multi-loop or in between them (see figures 3.2 and 3.1 and equation 3.3).

$$\begin{aligned} Z^{mult}[i, j] = \sum_{p < i < j < q} \hat{Z}(p, q) \times \\ \left( \underbrace{Z^{m2}[p + 1, i - 1] e^{-\beta c(q-i)}}_c + \underbrace{Z^{m2}[j + 1, q - 1] e^{-\beta c(j-p)}}_e \right. \\ \left. + \underbrace{Z^m[p + 1, i - 1] e^{-\beta c(j-i+1)} Z^m[j + 1, q - 1]}_d \right) \end{aligned} \quad (3.5)$$

The crucial improvement is obtained by replacing the double sum in equation 3.3 by two separate summation steps. For the last, “in-between”, sum term we use the

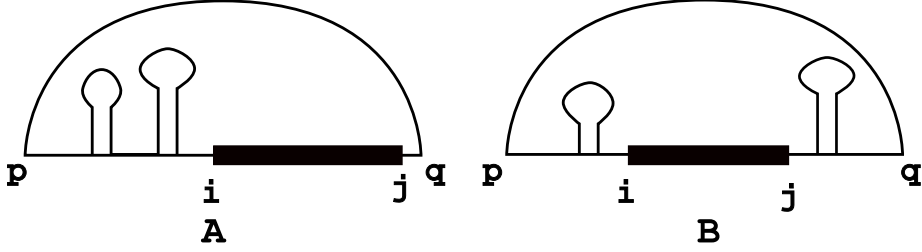


Figure 3.2: Alternative representation of figure 3.1 for multiloops only. Base pair  $(s_p, s_q)$  that includes the unpaired region  $(s_i, s_j)$  is drawn as an arc connecting bases  $s_p$  and  $s_q$ . The unpaired region  $s[i..j]$  is drawn as a bold black line. In the one-sided multiloop case (A) a structured region containing *at least* two structure components is on one side of the unpaired region. In case (B) the unpaired region  $s[i..j]$  is between two structured regions. In case (B) we have to take care to make a unique decomposition of the multiloop into a 3' part that contains exactly one component and a 5' part with at least one component.

auxiliary variables

$$Z^{mm}(q)[i] = \sum_{1 \leq p < i} \hat{Z}(pq) Z^m[p+1, i-1] \quad (3.6)$$

For  $Z_l^m(q)[i]$  where the unpaired region  $s[i..j]$  is to the left of all multi-loop components, we introduce

$$Z_l^m(q)[i] = \sum_{1 \leq p < i} \hat{Z}(p, q) Z^{m2}(p+1, i-1) e^{-\beta c(q-i)} \quad (3.7)$$

and an analogous term is used for the “right” contribution. Computing these values costs  $\mathcal{O}(n^3)$ . By using them, we can compute

$$\begin{aligned} Z^{mult}[i, j] = & \sum_{j < q} Z^m(q)[i] e^{-\beta c(j-i+1)} Z^m[j+1, q-1] \\ & + \sum_{p < i} Z_r^m(p)[j] \\ & + \sum_{j < q} Z^m[j+1, q-1] + Z_l^m(q)[i] \end{aligned} \quad (3.8)$$

in  $\mathcal{O}(n^2 \cdot w)$  time, i.e., the entire algorithm is  $\mathcal{O}(n^3)$ . The computations for hairpin and interior loop contributions are handled in the same way.

In comparison to McCaskill’s partition function algorithm, **RNAup** needs to store five additional matrices ( $Z^{m2}$ ,  $Z^{mm}$ ,  $Z_l$ ,  $Z_r$  and one additional matrix for the interior loop case). Hence we buy the speed-up by  $\mathcal{O}(w)$  by increasing the memory requirements by only about a factor of 2. For interaction lengths of size  $w = 25$ , and sequence lengths below 400, the runtime is decreased by a factor of 20. For sequence lengths between 400 and 2000 nucleotides, the speed up decreases with increasing sequence length, but is always superior to 12. This significant decreases in run-time opens the doors to genome-wide search for ncRNA targets in bacteria.

## 3.2 Free Energy of Interaction

---

The values of  $P_u[i, j]$  as computed above can be of interest in their own right: Hackermüller, Meisner, and collaborators [91, 170] showed that the binding of the HuR protein to its mRNA target depends quantitatively on the probability that the HuR binding site has an unpaired conformation. While not much is known about the energetics of RNA-protein interactions, the case of RNA-RNA interactions can be modelled in more detail:

The free energy of binding  $\Delta G$  consists of the “breaking energies”  $\Delta G_u$  that are necessary to render the binding site on each molecule accessible and a contribution  $\Delta G_h$  that describes the energy gain due to hybridization:

$$\Delta G = \Delta G_u^{s^*} + \Delta G_u^s + \Delta G_h. \quad (3.9)$$

This additivity assumes that the energies of the original loops of the respective RNAs remain unchanged during the hybridization process. For an unpaired binding motif in the interval  $s[i..j]$ , we have  $\Delta G_u^s = (-1/\beta)(\ln Z_u^s[i, j] - \ln Z^s) = (-1/\beta) \ln P_u^s[i, j]$ . Suppose the interaction region covers the intervals  $s^*[i^*..j^*]$  and  $s[i..j]$  in sequence  $s^*$  and  $s$ , respectively. As in **RNAhybrid** and related programs, we allow interior loops and bulges in the interaction region. The partition function over all these binding conformations is obtained by the following recursion:

$$Z^I[i, j, i^*, j^*] = \sum_{\substack{i < k < j \\ i^* > k^* > j^*}} Z^I[i, k, i^*, k^*] e^{-\beta I(k, k^*; j, j^*)}. \quad (3.10)$$

where  $I(k, k^*; j, j^*)$  is the energy contribution for the interior loop delimited by the base pairs  $(k, k^*)$  and  $(j, j^*)$  and  $Z^I[i, j, i^*, j^*]$  stands for the interaction partition function at equilibrium for an interaction region enclosed by base pairs  $(s_j, s_{j^*}^*)$  and  $(s_i, s_{i^*}^*)$

As we want to avoid having to keep track of a four dimensional array, we compute the partition function  $Z^*[i, j]$  over all structures where region  $[i, j]$  in the *longer* molecule is involved in the interaction. While doing this, we keep track of the region where  $Z^I[i, j, i^*, j^*]$  is maximal. The recursion for the calculation of  $Z^*[i, j]$  is shown in equation 3.11.

$$Z^*[i, j] = P_u^s[i, j] \sum_{i^* > j^*} P_u^{s^*}[i^*, j^*] Z^I[i, j, i^*, j^*]. \quad (3.11)$$

From  $Z^*[i, j]$  we can readily compute  $\Delta G[ij]$ , the free energy of binding given the binding site is in region  $[i, j]$  (see equation 3.12). For visual inspection,  $\Delta G[ij]$  can be reduced to the optimal free energy of binding  $\Delta G[i]$  at a given position  $i$ , (see equation 3.12). The memory requirement for these steps is  $\mathcal{O}(n \cdot w^3)$ , the required CPU time scales as  $\mathcal{O}(n \cdot w^5)$ , which, at least for long target RNAs, is dominated by the first step, i.e., the computation of the  $P_u[i, j]$ .

$$\begin{aligned} \Delta G[i, j] &= -RT \ln Z^*[i, j]. \\ \Delta G[i] &= \min_{k \leq i \leq l} \{ \Delta G[k, l] \}. \end{aligned} \quad (3.12)$$

## 3.3 Application

---

### 3.3.1 siRNA design

In order to demonstrate that our algorithm produces biologically reasonable results, we compared predicted binding probabilities with data from RNA interference experiments. Small interfering RNAs (siRNAs) are short (21-23nts) RNA duplexes with symmetric 2-3 nts overhangs [57, 172, 177]. They are used to silence gene expression in a sequence-specific manner in a process known as RNA interference (RNAi) (see chapter 2).

Recently, there has been mounting evidence that the biological activity of siRNAs is influenced by local structural characteristics of the target mRNA [24, 127, 177, 190, 216, 270]: a target sequence must be accessible for hybridization in order to achieve efficient translational repression. An obstacle for effective application of siRNAs is the fact that the extent of gene inactivation by different siRNAs varies considerably. Several groups have proposed basically empirical rules for designing functional siRNAs (see e.g. [62, 207]), but the efficiency of siRNAs generated using these rules is highly variable. Recent contributions [192, 216] suggest two significant parameters: The stability difference between 5' and 3' end of the siRNA, that determines which strand is included into the RISC complex [118, 218] and the local secondary structure of the target site [24, 127, 177, 190, 216, 270].

Schubert et al. [216] systematically analyzed the contribution of mRNA structure to siRNA activity. They designed a series of constructs, all containing the same target site for the same siRNA. These binding sites, however, were sequestered in local secondary structure elements of different stability and extension. They observed a significant obstruction of gene silencing for the same siRNA caused by structural features of the substrate RNA. A clear correlation was found between the number of exposed nucleotides and the efficiency of gene silencing: When all nucleotides were incorporated in a stable hairpin, silencing was reduced drastically, while exposure of 16 nucleotides resulted in efficient inhibition of expression virtually indistinguishable from the wild type.

We applied our methods to study the target sites provided by Schubert et al. [216]. Our predictions, shown in figure 3.3, are in perfect agreement with the experimental results. The target site of the “VR1straight” construct has a high probability of being unstructured, consequently  $\Delta G_i$ , the optimal free energy of binding, is highly favorable and the siRNA will bind almost exclusively to the intended target site. The stepwise reduction of the target accessibility is directly correlated to a weaker optimal free energy of binding and decreasing silencing efficiency. In case of construct VR HP5\_6 the optimal free energy of binding at an alternative binding site at positions 1066 to 1078 nearly equals that at the proposed target site. Since siRNAs can also function as miRNAs [52, 271], the siRNA might act in a miRNA like fashion binding to this alternative target site and contribute to the remaining translational repression

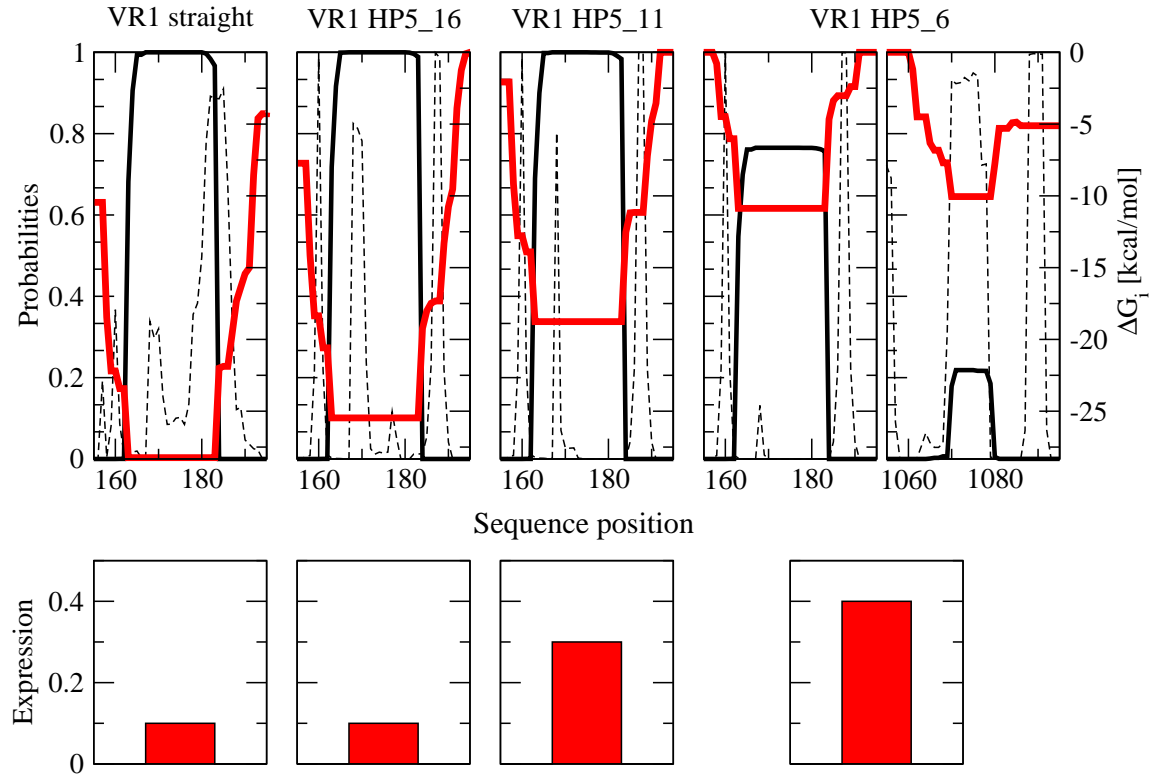


Figure 3.3: Probability of being unpaired  $P_u[i, i]$  (dashed line), probability of binding to siRNA at position  $i$ ,  $P_i^*$ , (thick black line) and  $\Delta G_i$ , the optimal free energy of binding in a region including position  $i$  (thick red line) near the known target site of VsiRNA1. The scale for the probabilities is indicated on the left side, the scale for the minimal free energy of binding on the right side. At the bottom the protein expression levels in experimental data [216] are indicated. The isolated 21mer target sequence, displaying the same activity as the wild type mRNA, and 3 mutants are shown. A decreasing optimal free energy of binding is correlated with increasing expression. In the case of the HP5\_6 mutant an alternative binding site becomes occupied as the optimal free energy of binding due to this alternative interaction nearly equals  $\Delta G_i$  at the proposed target site.

of this construct. The incomplete complementarity of the siRNA to the alternative target site should be no obstacle to functionality, since it was shown that miRNAs can be active even if the longest continuous helix with the target site is as short as 4 - 5 base pairs [29]. Our new accessibility prediction tool can thus be used to identify potential binding sites as well as explain differences in si/miRNA efficiency caused by secondary structure effects.

### 3.3.2 sRNA targets

In the previous example we showed that the target-accessibility of the target RNA is a good descriptor of the siRNA repression efficiency. In this section we show that **RNAup** can predict with great precision the binding location of small bacterial RNAs (sRNAs) on their targets. The main difference between sRNAs and siRNAs, is that the former sRNAs are long enough to be highly structured. Furthermore the binding region usually spans only part of the sRNA. Therefore, the secondary structure of the sRNA will critically influence the exact location of the binding site.

In order to show the importance of the sRNAs structures, we ran **RNAup** with and without considering the sRNA structures on experimentally verified sRNA-mRNA interactions found in [250]. As expected, when omitting the structure within the sRNA the binding energy was markedly higher (mean  $-24.97 \pm 5.97$ ) than when considering it (mean  $-15.54 \pm 1.99$ ).

When comparing binding site location with the location of experimentally verified binding sites, see table 5.5, we found that considering the structure on both the sRNA and the mRNAs predicts binding sites more accurately, i.e. 3 binding sites were predicted with perfect accuracy (the predicted binding site did not deviate by more than one base pair from the binding site reported in literature), and 7 binding sites deviate by at most 17 base pairs, see table 5.5. Neglecting sRNA structure, on the other hand, predicts no binding site with perfect accuracy, 9 binding sites show a deviation between 4 to 45 base pairs, (4, 11, 12, 16, 27, 33, 39, 39, 45), and one binding site prediction was wrong, i.e. did not overlap with the binding site reported in literature.

This comparison emphasizes the importance of the inclusion of secondary structure information of both binding partners when predicting sRNA-mRNA interactions. Neglecting the structure of the sRNA results in an overestimation of the length of the predicted interaction and in most cases hinders the clear localization of the proper target site boundary (see also chapter 2 and 5).

In addition to the location of the binding site, the regulatory effects upon binding of the sRNA to its target mRNA was studied. We used a data set consisting of 9 small regulatory RNAs from *e.coli*, their 9 reported mRNA targets and the fold-change in protein concentration induced by all 81 possible mRNA-ncRNA interactions [250]. Among those interactions, 8 targets were downregulated, 2 were upregulated, and no or only marginal changes were detected for the others (see table 5.5). Downregulation usually occurs when the hybridization of the ncRNA with its cognate mRNA blocks the ribosome entry sites on the target (for a review see [83]). In contrast, upregulation typically takes place when the sRNA-mRNA hybridization disrupts intrinsic inhibitory structures that sequester the ribosome binding site and/or the start codon [156, 157, 199]. In many cases the sRNA-mRNA interactions are assisted by the RNA chaperone protein *Hfq* [251].

Target prediction was performed with the mRNA constructs (117-689 nts) described in [250] and the full length sRNAs (69-220 nts). The mRNA constructs included a long 5'UTR sequence (57-565 nts) and a comparably short fragment of the CDS (35-139 nts). Both the hybridisation energy and the target site position were computed with *RNAup* for all sRNA-mRNA combinations.

For each sRNA we tested which of the mRNA constructs was predicted to bind most strongly. To our satisfaction the most favorable binding energy for each sRNAs was found for its cognate target (see Table 5.5). Since the most common mechanism of translational control is to influence ribosome binding at the Shine-Dalgarno (SD) sequence, we checked the position and structural effects of the predicted interactions. For each of the 8 interactions that resulted in downregulation, we found the binding site to be at or close to the Shine-Dalgarno sequence. This type of inhibition can thus be predicted by comparing *RNAup* predictions with sequence features that are easy to recognize in bacterial genomic sequences.

Our data set contains only two examples of upregulation, namely binding of *DsrA*

Table 3.1: Binding site summary for the 10 functional interactions published by Urban et.al [250]. Column  $\Delta\Delta G$  shows the optimal binding energy calculated with **RNAup**. Column Position gives the binding position relative to the start codon. Column Position lit. gives the binding position found in the literature.

mRNA	sRNA	regulation	$\Delta\Delta G$	Position	Pos.lit.	cite
RyhB	sodB	-	-11.50	-18,+4	-4,+5	[78]
DsrA	hns	-	-14.60	-10,+11	+7,+19	[138]
MicA	ompA	-	-13.60	-21,-6	-21,-6	[201]
MicC	ompC	-	-15.80	-30,-15	-30,-15	[37]
MicF	ompF	-	-17.80	-11,+9	-11,+10	[37]
Spot42	galK	-	-17.00	-18,+30	-19,+21	[178]
SgrS	ptsG	-	-17.33	-28,-10	-28,+4	[115]
GcvB	dppA	-	-17.30	-30,-7	-31,-14	[227]
DsrA	rpoS	+	-14.52	-126,-97	-119,-97	[157]
RprA	rpoS	+	-15.90	-134,-94	-117,-94	[157]

and *RprA* to *rpoS*. In both cases, binding leads to the disruption of a helix which normally sequesters the Shine-Dalgarno sequence as well as the start codon. We remark that this is an example of the modifier RNA mechanism that was proposed in [91, 171].

To assess the ability of **RNAup** to predict upregulating interactions we first compared the accessibility of the region around the start codon of all 9 mRNAs, with the mean accessibility of all 4463 genes in the *E. coli* genome. Mean accessibility was computed for regions of 401 nts, centered at the start codon. For comparability we used the same 401 nts regions of our 9 target genes rather than the constructs used above. The accessibilities and corresponding opening energies were computed with **RNAup** for unpaired regions of length 4. The screen against the *E. coli* genome with all 9 sRNAs took 16 CPU days on one core of an Intel Core2 duo CPU with 2 GB RAM running at 2.40GHz.

In order to compare the accessibility with a random model, we shuffled each sequences and recomputed the accessibility profile. Because the region we are looking at contain

both an untranslated region on the 5' side of the start codon and a protein coding region on the 3' side, we used two different shuffling modes. The untranslated region was dinucleotide shuffled, while the coding region was trinucleotide-shuffled. The start codon was kept untouched. Compared to the random model, the regions located around the start codon show a higher accessibility. This feature is not only specific to *E. coli* but seems conserved in different bacterial families (see figure 3.4). The higher accessibility around the start codon probably facilitates ribosome docking on the mRNA to be translated.

With a local opening energy of 4.51 kcal/mol *rpoS* is the most inaccessible transcript among the 9 transcripts presented here. Genome-wide only 8.8% of the transcripts have a less accessible start codon than *rpoS*. In contrast, the eight downregulated transcripts showed a higher than average (2.23 kcal/mol) accessibility, ranging from 0.30 kcal/mol for *ompA* to a maximum of 1.27 kcal/mol for *ryhB* (see figure 3.5). After binding *DsrA*, the accessibility of the *rpoS* start codon changes dramatically. With only 1.40 kcal/mol, bound *rpoS* is much more accessible than the average transcript and belongs to the 33% most accessible genes, see fig. 3.6. The same effect is seen upon binding with *RprA*, with a local accessibility after binding of 1.90 kcal/mol. Technically, accessibilities after binding can be computed easily by adding the constraint that nucleotides in the binding site remain single stranded.

## 3.4 Conclusion

---

We have demonstrated here that variants of McCaskill's partition function algorithm can be implemented efficiently to compute the probability that a given sequence interval  $s[i..j]$  is unpaired. The computation is rigorous, and can thus be used even for small probabilities, where sampling approaches such as **Sfold** [46, 48] do not work well. As exemplified with the data from [216] and [250], our algorithm is able to capture the most common types of interaction between regulatory RNAs and their targets.

More complicated types of interactions, such as H/ACA snoRNA (see chapter 6) with their target rRNAs or OxyS-fhlA, are neglected. The speed of **RNAup** is clearly sufficient for genome wide searches for sRNA-mRNA interactions in bacteria. In

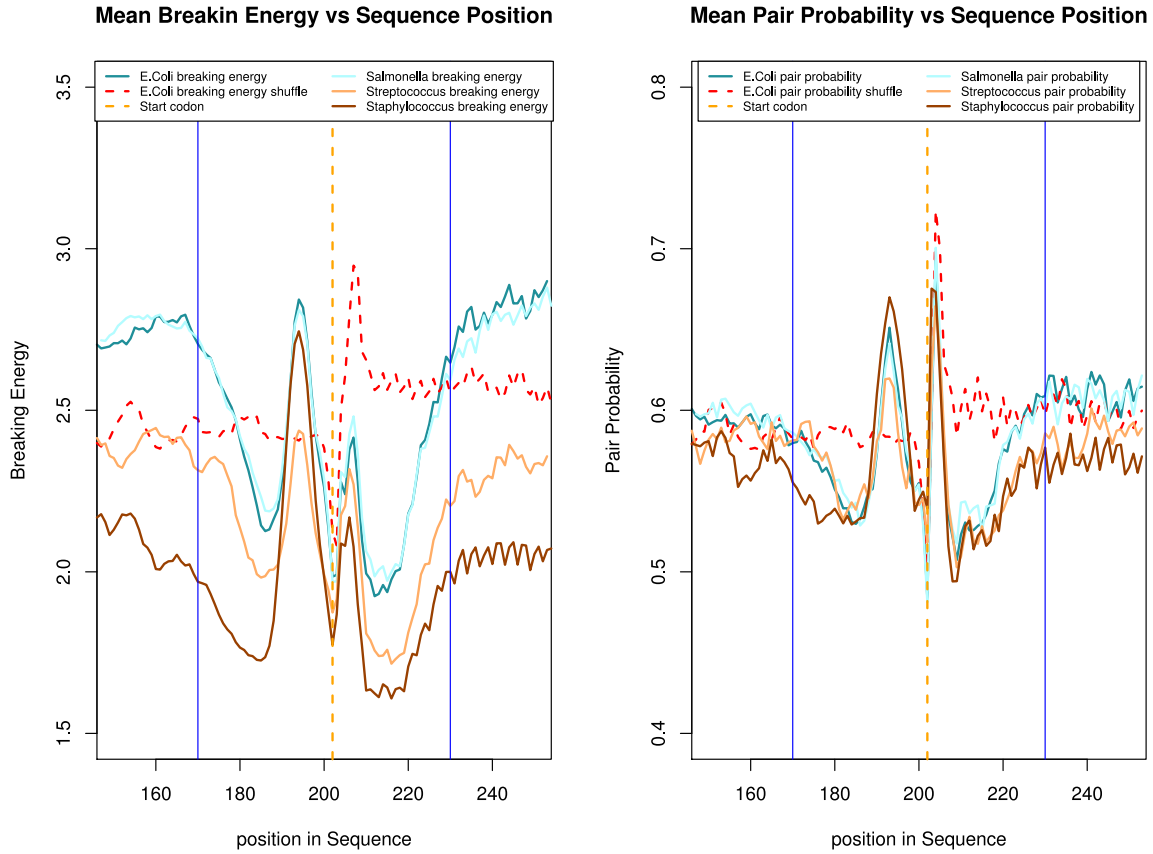


Figure 3.4: Breaking energy profile and pair probability profile around the start codon of all mRNA in four different bacteria species. Boundaries of region with increased accessibility are shown by vertical blue lines. The orange dotted line represents the position of the start codon. The red dotted line represents the mean accessibility measure of the shuffled regions. **R.h.s** Mean breaking energy. **L.h.s** Mean base pair probability.

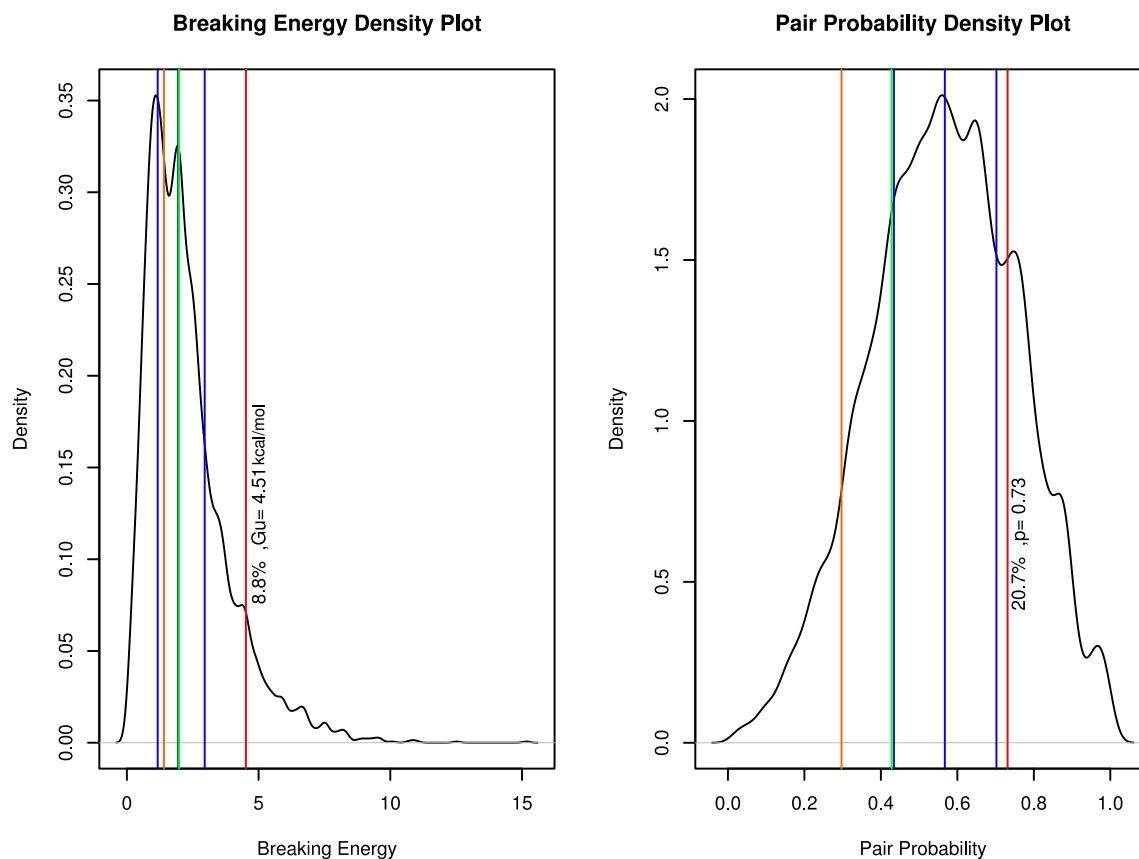


Figure 3.5: Mean Breaking Energy and Pairing Probability distribution around the start codon for all genes in *E. coli*. The black curve represents the density distribution, the red line represents the values for *rpoS* before binding, the blue lines delimits the quartiles of both distributions, the green line represents the values for *rpoS* after hybridization with *RprA*, while the orange line represents the value for *rpoS* after hybridization with *DsrA*. *rpoS* is among the most inaccessible mRNAs before binding, while after binding the local breaking energy belongs to the lower half. For the pairing probability, an even stronger trend is seen, as *rpoS* after binding belongs to the 25% most open targets.

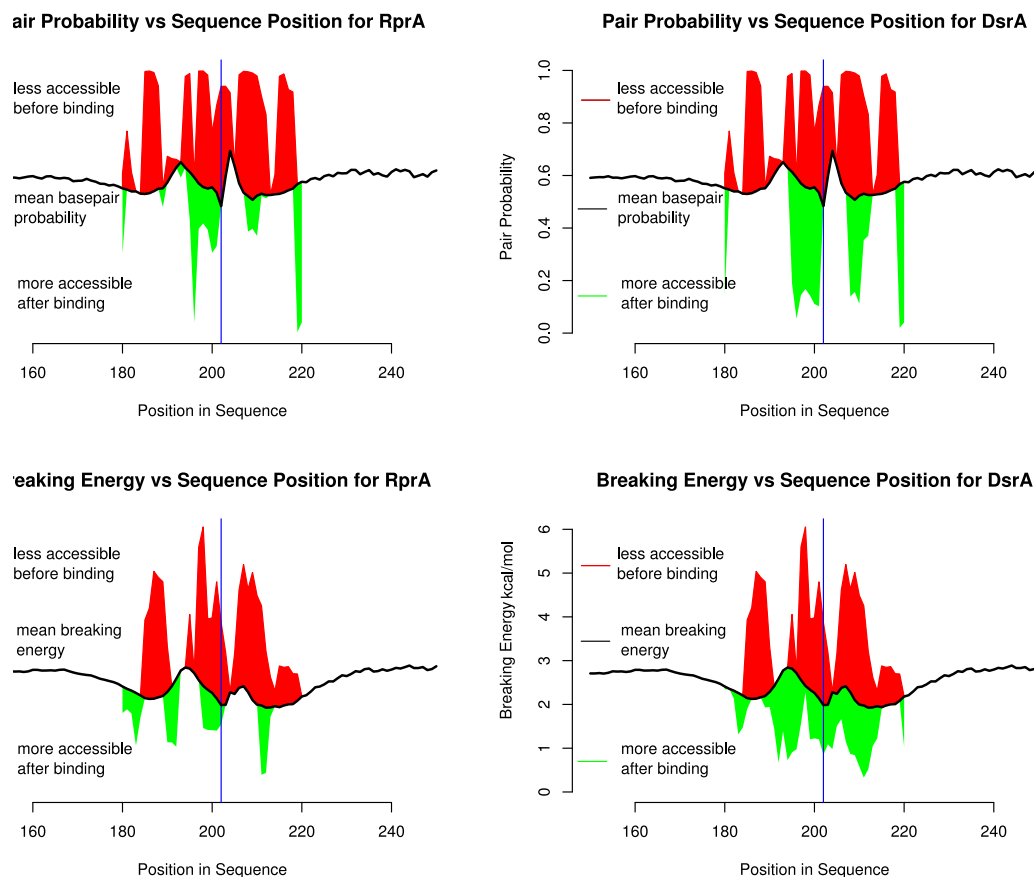


Figure 3.6: Opening energy,  $\Delta G_u$  and single nucleotide base pairing probability plotted around the start codon of *RpoS* versus sequence position for the interaction of *DsrA* and *RprA*. The red area represents regions of higher than average structural stability before sRNA binding on *RpoS*, while the green region represents regions of lesser than average structural stability after sRNA binding to *RpoS*. The blue line represents the position of the start codon.

principle, the approach is equally applicable to interaction search in higher organisms. However, the larger genome size and longer UTR regions pose challenges both in terms of computation time and false positives. In chapter 5 a method is presented that solves the computation time problem, leading to reduction in runtime on average by a factor of 8000.



# 4

## RNAplfold

As shown in the previous sections, target site accessibility as computed by **RNAup**, i.e. the probability  $P_u[i, j]$  that a sequence interval  $s[i..j]$  is devoid of structure, is an important features for correctly predicting RNA-RNA interactions. Accessibility of an RNA sequence is computed by **RNAup** in  $\mathcal{O}(n^3)$ , where  $n$  represents the whole sequence length. While this runtime makes the computation of  $P_u[i, j]$  acceptable for small bacterial genomes, it remains a prohibitive cost for large genome accessibility computation. A reduction in runtime can however be achieved by considering a local approach, where the accessibility is computed for a windows of length  $L$  [18, 27].

This runtime reduction allows to compute the target accessibilities not only for some selected examples like in chapter 3, but also genome-wide. This opens the door to genome-wide ncRNA target predictions, for example for miRNAs, in large genomes. In the next section, the derivation of **RNAplfold** from **RNAup** that was originally realized by Dr. Stefan Bernhart [18, 27] is summarized. Applications to siRNAs and miRNAs are then shown. The importance of **RNAplfold** for RNA-RNA interactions predictions will further be underlined in chapters 5 and 6.

### 4.1 From RNAup to RNAplfold

As shown in chapter 3 the values of  $P_u[i, j]$  can be computed from the equation

$$P_u[i, j] = \frac{Z_{1,i-1}Z_{j+1,n}}{Z_n} + \sum_{h < i, j < l} P_{h,l} \text{Prob} [[i, j] | (h, l)] , \quad (4.1)$$

where  $\text{Prob}[[i, j]|(h, l)]$  is the probability that  $s[i..j]$  is an unpaired region within the loop with closing pair  $(s_h, s_l)$ . This probability depends only on the structures inside the pair  $(s_h, s_l)$ .

We define the average over all folding windows of the probability that  $(s_i, s_j)$  is paired:

$$\pi_{ij}^L = \frac{1}{L - (j - i) + 1} \sum_{u=j-L}^i P_{ij}^{u,L}. \quad (4.2)$$

where  $P_{ij}^{u,L}$  is the probability that  $(s_i, s_j)$  is paired in a window of size  $L$  starting at  $u$ . The average probability of  $P_u[i, j]$  over all windows length  $L$  that contains  $s[i..j]$  can be written as :

$$\begin{aligned} \pi^0[i, j] &= \frac{1}{L - (j - i) + 1} \sum_{u=j-L}^i P_u[i, j] \\ &= \frac{1}{L - (j - i) + 1} \sum_{u=j-L}^i \frac{Z_{1,i-1}^{u,L} Z_{j+1,n}^{u,L}}{Z_{1,n}^{u,L}} \\ &\quad + \sum_{h=j-L}^{i-1} \sum_{l=j+1}^{i+L} \frac{L - (h - l) + 1}{L - (j - i) + 1} \pi_{hl}^L \text{Prob}[[i, j]|(h, l)] \end{aligned} \quad (4.3)$$

Since

$$\text{Prob}[[i, j]|(h, l)] = Z_{hl}[i, j] / Z_{ij} \quad (4.4)$$

is independent of the folding window as long as  $[h, l] \subseteq [u, u + L - 1]$ , and the computation of  $Z_{hl}[i, j]$  requires only partition function entries in the interval  $[h, l]$ , a local version of **RNAup** which can compute the average local accessibility of a sequence can be computed in  $\mathcal{O}(n^2 \cdot L)$ , as shown in [18].

## 4.2 Application to RNAi and siRNA design

RNA interference (RNAi) describes the post-transcriptional gene silencing process triggered by endogenous or exogenous double stranded RNAs (dsRNAs). After being processed by Dicer, the dsRNAs are transferred to the RNA-Induced Silencing Complex (RISC), where one of the strands (the guide strand) is introduced while the other strand is degraded (the passenger strand). Target recognition happens through

hybridization of the guide RNA with its target gene, which causes the cleavage and the subsequent degradation of the target strand (see figure 2.13).

A long debated topic in the field of siRNA design is the influence of the target structure on the target recognition process and subsequently on the siRNA efficiency. While target site structure was recognized as an important feature in the design of antisense oligonucleotides and ribozymes [47, 146, 173, 174, 257, 272], data arguing for [6, 24, 32, 46, 127, 151–153, 190, 216, 223, 234, 256, 269, 270] and against [23, 194, 207] the influence of target site accessibility on the siRNA efficiency were reported.

In order to assess if the target site accessibility, as computed by `RNAplfold`, can be used to discriminate between functional and non-functional siRNAs, and to determine the optimal folding parameters, we used two independent siRNA datasets of measured siRNA efficacies. Dataset 1 was composed of 2433 siRNAs targeting the 3'UTR sequences of 34 different genes, whereas dataset 2 contained 294 siRNAs that were targeted against arbitrary regions of the coding sequences of the human genes `MAP2K1`, `GAPDH`, `PPIB`, and `LMNA` and whose knock-down efficiencies were verified by analyzing mRNA as well as protein levels.

Note, that these datasets are referred to as the “complete” datasets throughout the text. Many of the siRNAs in these datasets showed some but relatively weak repression efficiencies and were therefore not used in the training sets. From the initial datasets we generated reduced subsets of 474 and 99 siRNAs for dataset 1 and 2, respectively by removing siRNAs with intermediate silencing efficiencies, leaving only those that could be clearly assigned as functional or non-functional.

The number of mRNAs targeted in the reduced datasets remained unchanged. Target site accessibilities were then computed for different averaging window sizes  $W$ , maximum base pair spans  $L$ , and lengths of the unpaired region  $u$ . To test for a significant separation of functional and non-functional siRNAs, we performed a Wilcoxon rank sum test comparing the distributions of functional and non-functional siRNAs for each of the two datasets. We found that the silencing efficiency correlated significantly with target site accessibility over a wide range of parameters analyzed, with the most significant separation resulting from 80 nts and 40 nts for  $W$  and  $L$ , respectively (see figure 4.1)

While these values may seem small, it is clear that actively translated coding regions

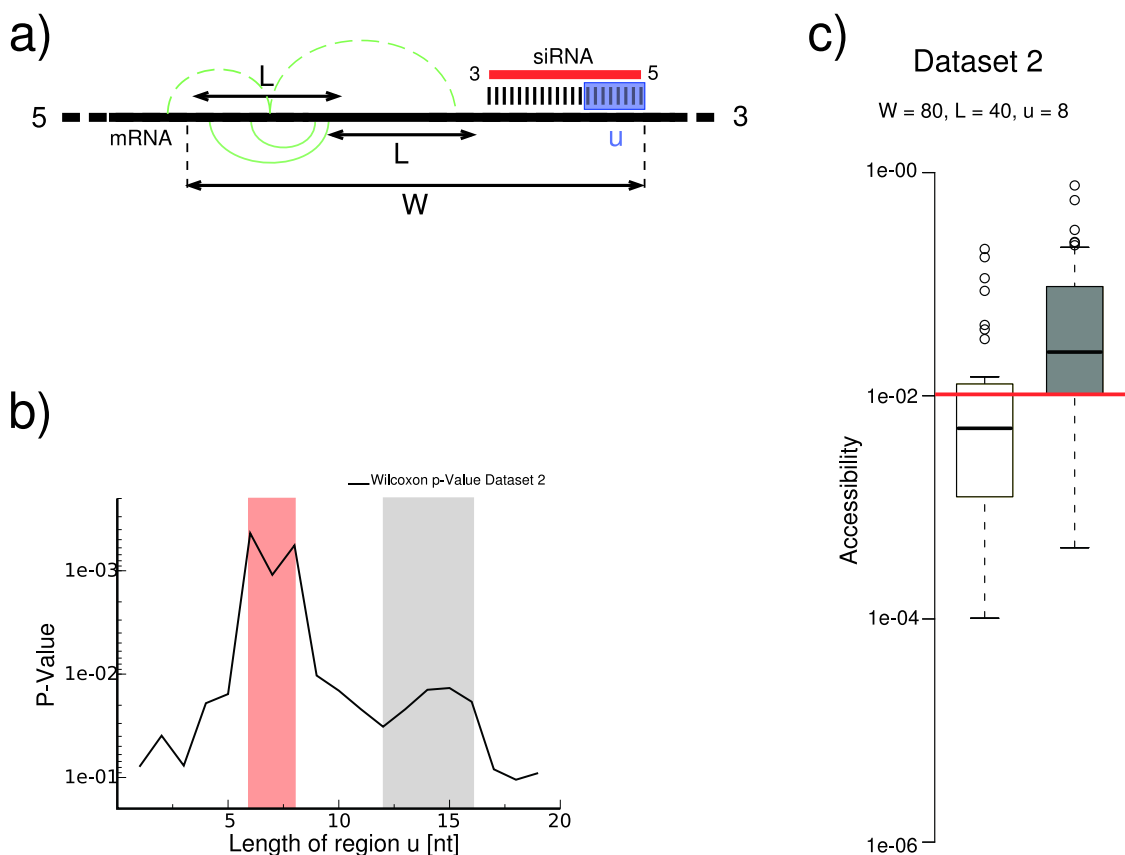


Figure 4.1: Application of **RNAlfold** to separate functional from non-functional siRNAs. (a) The RNA is folded locally in a sliding window approach (window size  $W$ ). Within  $W$ , base pairing is restricted to a maximum distance  $L$ .  $u$  represents the stretch of consecutive nts within a siRNA target site starting at its 3' end for which the accessibility is computed. Green lines represent possible base pairs. Interactions outside the span size of  $L$  or the flanking window  $W$  are not allowed (dotted green lines). (b) Box-plot diagram comparing the accessibility of functional and non-functional siRNAs. The dataset was divided into functional siRNAs (repression efficiency  $> 75\%$ ) and non-functional siRNAs (repression efficiency  $< 25\%$ ); black horizontal lines within the boxes depict medians. The circles represent outliers and dotted lines show the standard deviation. The Wilcoxon p-value is  $5 \cdot 10^{-4}$ . Cutoffs for the accessibility to discriminate functional and non-functional siRNAs was set at 0.01157 (red horizontal line). The parameters  $W$ ,  $L$  and  $u$  are indicated. (c) Accessibility distributions of functional and non-functional siRNAs are best differentiated for a length of 8 and/or 16 nts (according to p-values). p-values, were determined from a Wilcoxon test and are plotted against the length of the analyzed region starting at the 3' end of the target site.

should be devoid of long range structures, since these would be destroyed by the passing ribosome and are slow to reform [240]. Moreover, it is well known that long range structures are much less accurately predicted [54]. Hence, a local structure approach may be more suitable than global mRNA structure prediction programs [108].

When varying the length  $u$  of the unpaired region, we observed two ranges with especially good separation (see figure 4.1). The first range measures the accessibility of the 6-8 nucleotides starting at the 3' end of the target site, and therefore corresponds to the so-called seed region. This is in agreement with previous observations that the 5'-seed region of both siRNAs and microRNAs is the major determinant for RISC-mediated target recognition [6, 28, 53, 110]. Furthermore, a second peak was observed for  $u$  values of 12-16, reminiscent of biochemical data showing that accessibility of the first 16 nts within the target site is required for highly efficient RISC-mediated cleavage [6].

We were not able to detect any further improvements in the separation of functional and non-functional siRNAs by additionally analyzing the energy of siRNA-target RNA duplexes. Presumably, this is because perfect complementarity between siRNAs and their target sites generally implies high duplex energies. However, one cannot exclude such a correlation for siRNA off-target effects or microRNA-mediated gene repression, both of which rely on imperfect base pairing to their target sites [29, 110, 132].

Since it was previously claimed that regions of low G/C content coincide with efficient siRNA silencing [207], and since accessibility correlates with G/C content we separated the datasets by G/C content into five classes and analyzed the impact of accessibility for each class. We noticed that the distinction between functional and non-functional siRNAs remained strong over the whole range of G/C window sizes (see figure 4.2).

Furthermore, we found, that the G/C content is a much poorer predictor of siRNA efficacy than accessibility. For dataset 1, we noted that highly efficient siRNAs targeted regions of higher G/C content (on average 58%) while non-functional siRNAs had an average G/C content of 42%.

We further used the complete dataset 1 (consisting of 2433 siRNAs) to gain insight into the correlation between target site accessibility and siRNA repression efficiency.

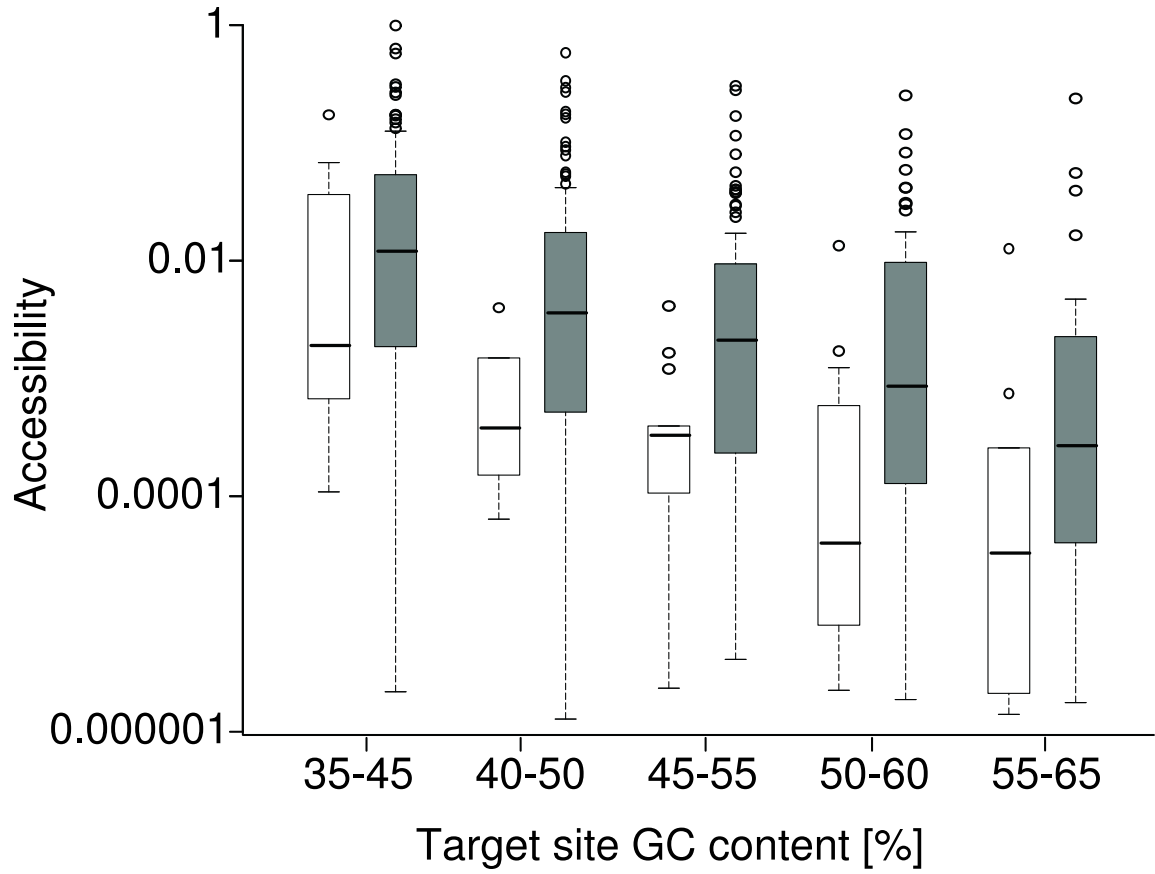


Figure 4.2: Box-plot diagram of functional and non-functional siRNAs for different target site G/C content. Functional and Non-functional siRNAs are partitioned into five groups according to their G/C content. A wilcoxon test was applied showing a significant separation for all G/C windows analyzed.

In order to reduce noise, caused by the fact that the measured mRNA levels have errors of around 30% [108], we binned the data in groups of 36 siRNAs according to their accessibility and plotted for each bin the mean repression score against the mean accessibility. Despite the large variance target accessibility clearly correlates with the repression score over a wide range of accessibilities (from  $10^{-5}$  to  $10^{-1}$ ) (see 4.3).

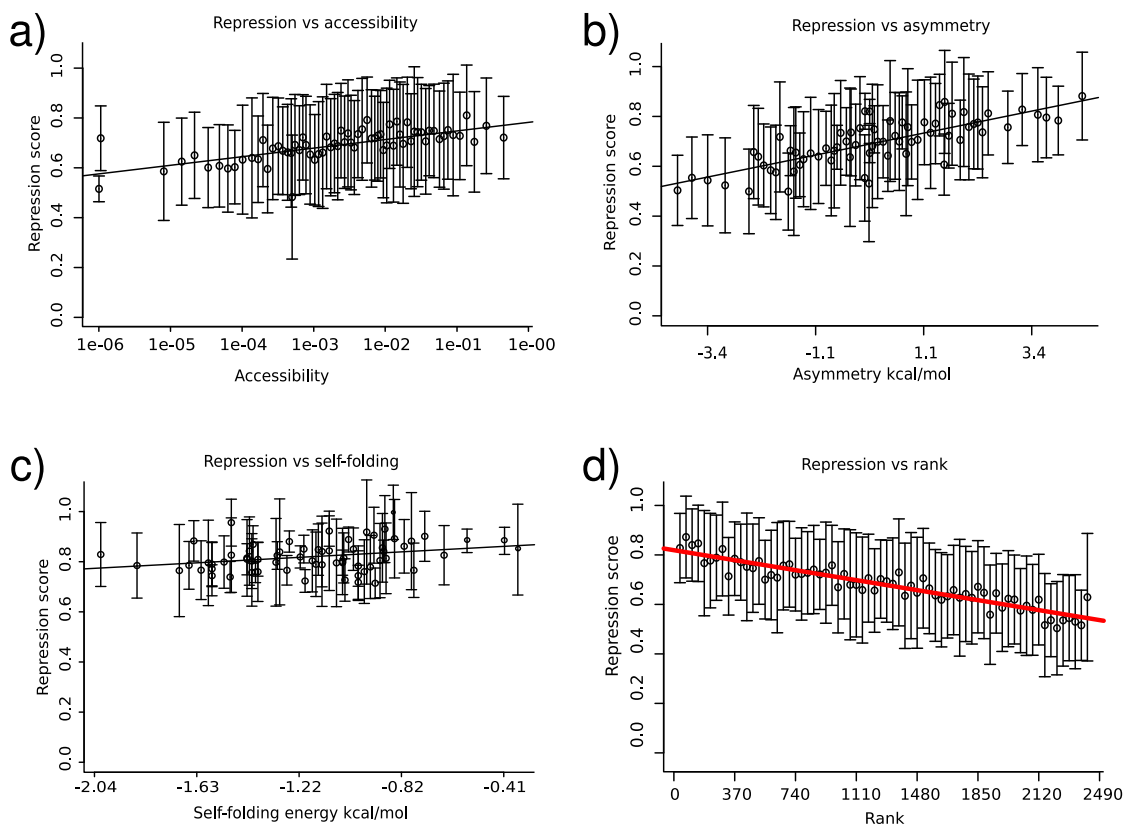


Figure 4.3: Correlation plots for different design criteria for 2433 siRNAs from dataset 1. The siRNAs were grouped into bins, each of them containing 36 siRNAs. The binning was done according to the design criteria. Correlation plots of the novartis repression score against accessibility (a), asymmetry (b) and self-folding (c) are shown. (d) Ranking of siRNAs for the combination of all design criteria including accessibility plotted against the normalized inhibitory activity.

To assess the relevance of accessibility for the design of efficient siRNAs, we compared

six commonly used criteria: Two are purely sequence-based (“U at position 10”, “a base other than G at position 13”) [207]; two describe the asymmetry of the siRNA duplex responsible for strand selection, by looking at either the type of base pairs or the interaction energy of the last four base pairs [118, 207, 218]; and the final two features concern the tendency for self-folding of the siRNA (which refers to the level of self complementarity), using either its total folding energy (self-folding) or the number of unpaired nucleotides at the ends of the siRNA (free-end) [194].

We noticed, that the asymmetry resulted in a better correlation with the measured repression than the accessibility criterion (pearson correlation coefficient of 0.48 and 0.23 for asymmetry and accessibility, respectively) (see figure 4.3). The other design parameters free-end or self-folding performed worse (see figure 4.3).

The stronger effect of asymmetry could reflect that siRNA strand selection acts at the level of RISC assembly, and therefore upstream of any target site accessibility effects. We then designed simple filters by defining a threshold for each criterion (other than the two purely sequence-based filters which do not require a threshold). This threshold was chosen conservatively, such that at least 75% of the functional siRNAs were retained from datasets 1 and 2 (see figure 4.4)).

The performance of each filter was assessed on the complete dataset 1 by applying a Wilcoxon test on the distribution of normalized inhibitory activities. We found that the two best single design criteria were accessibility and asymmetry with p-values of  $8.5e^{-8}$  and  $1e^{-16}$ .

In addition, all siRNAs were binned in five functionality classes depending on their inhibitory efficiencies (inhibition smaller than 0.5,  $<F0.5$ , inhibition of at least 0.5  $\geq F0.5$ ,  $0.8 \geq F0.8$ ,  $0.9 \geq F0.9$  or  $1.1 \geq F1.1$ ). Even without any rational design many of the random siRNAs were functional with 82.3% inducing more than 0.5 inhibition ( $\geq F0.5$ ), 31.4% more than 0.8 ( $\geq F0.8$ ), 15.1% more than 0.9 ( $\geq F0.9$ ) and 2% more than 1.1 ( $\geq F1.1$ ). ‘Random siRNAs’ refers to siRNAs that were randomly chosen without any rational design. Free-end and self-folding performed slightly better than random, while the two sequence-based rules did not result in any significant improvement, with “U at position 10” performing even worse than random. The sequence rules were therefore not further considered.

From this analysis it is clear, that accessibility alone, just like any other descriptor

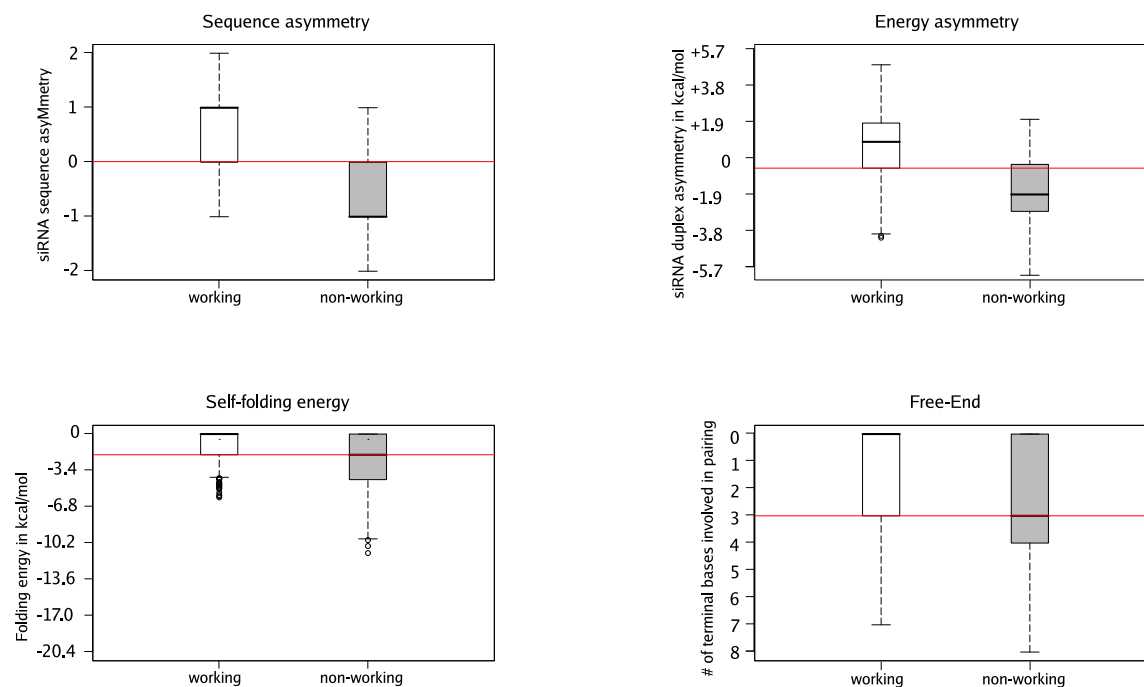


Figure 4.4: Box-plot diagrams comparing asymmetry, self-folding and free-end for functional and non-functional siRNAs. The dataset 1 consisting of 474 siRNAs was used to determine the single criteria thresholds (red line). The dataset was divided into 363 functional siRNAs of ( $>0.900$  repression score) and 109 non-functional siRNAs ( $<0.354$  repression score). The quartiles are represented by the edges of the rectangles, which contain 50% of the data, black horizontal lines within the boxes depict medians. The circles represent outliers and dotted lines show the standard deviation. Thresholds were chosen conservatively, such that at least 75% of the working siRNAs were kept. Note, the same was done for dataset 2 (data not shown).

assessed above, is not sufficient to reliably predict siRNA efficacy. Since most current siRNA design methods neglect the effects of target site accessibility we investigated whether the addition of accessibility to the three most effective conventional design criteria (asymmetry, free-end and self-folding) leads to a superior design of siRNAs. The combinations of asymmetry, self-folding and free-end lead to an increase over random of 16.2%, 9.9% and 5.5% in  $\geq F0.8$ ,  $\geq F0.9$  and  $\geq F1.1$ . The addition of accessibility lead to a further improvement in all functionality classes, specifically the fraction of siRNAs in the  $\geq F0.5$ ,  $\geq F0.8$ ,  $\geq F0.9$  and  $\geq F1.1$  classes increased by 3.4%, 3.9%, 4.2%, and 2.1% respectively. Especially the fraction of siRNAs in the  $\geq F0.9$  (14.2%) and  $\geq F1.1$  (7.6%) classes was doubled compared to random, demonstrating that the accessibility criteria boosts the fraction of very potent siRNAs.

For a typical mRNA about 25% of the sequence positions will pass the combination of all four filter criteria. Thus, the resulting list is usually long enough to choose siRNAs with specific properties from the pool, an important feature e.g. for silencing specific gene splice variants or targeting short exons.

In addition to the filtering we introduced a ranking of the remaining siRNA candidates according to their overall performance in all four criteria. Since different selection criteria recapitulate distinct stages in the RNAi pathway, a poor performance in one descriptor can presumably not be compensated by good values in another. We therefore devised a hierarchical sorting that emphasizes the least favorable criterion for each siRNA, rather than constructing a combined score. More precisely we rank all siRNAs by each of the design criteria separately. The overall sorting is then a hierarchical sort using the worst rank as the primary sorting key.

The distribution of the repression vs. overall rank for dataset 1 can be seen in 4.3 and shows that even among the siRNAs passing the filters the top ranked candidates perform particularly well. The filtering and ranking described above were combined in a user-friendly siRNA design tool, called RNAXs, available as a web service at <http://rna.tbi.univie.ac.at/cgi-bin/RNAXs>. RNAXs returns a ranked list of all siRNA candidates that pass the filters including their performance on each of the design criteria, as well as graphical accessibility plots for the top three candidates. The user can change all parameters and thresholds; in addition RNAXs allows user specific sequence constraints. Subsequently, individual sequences can be submitted to

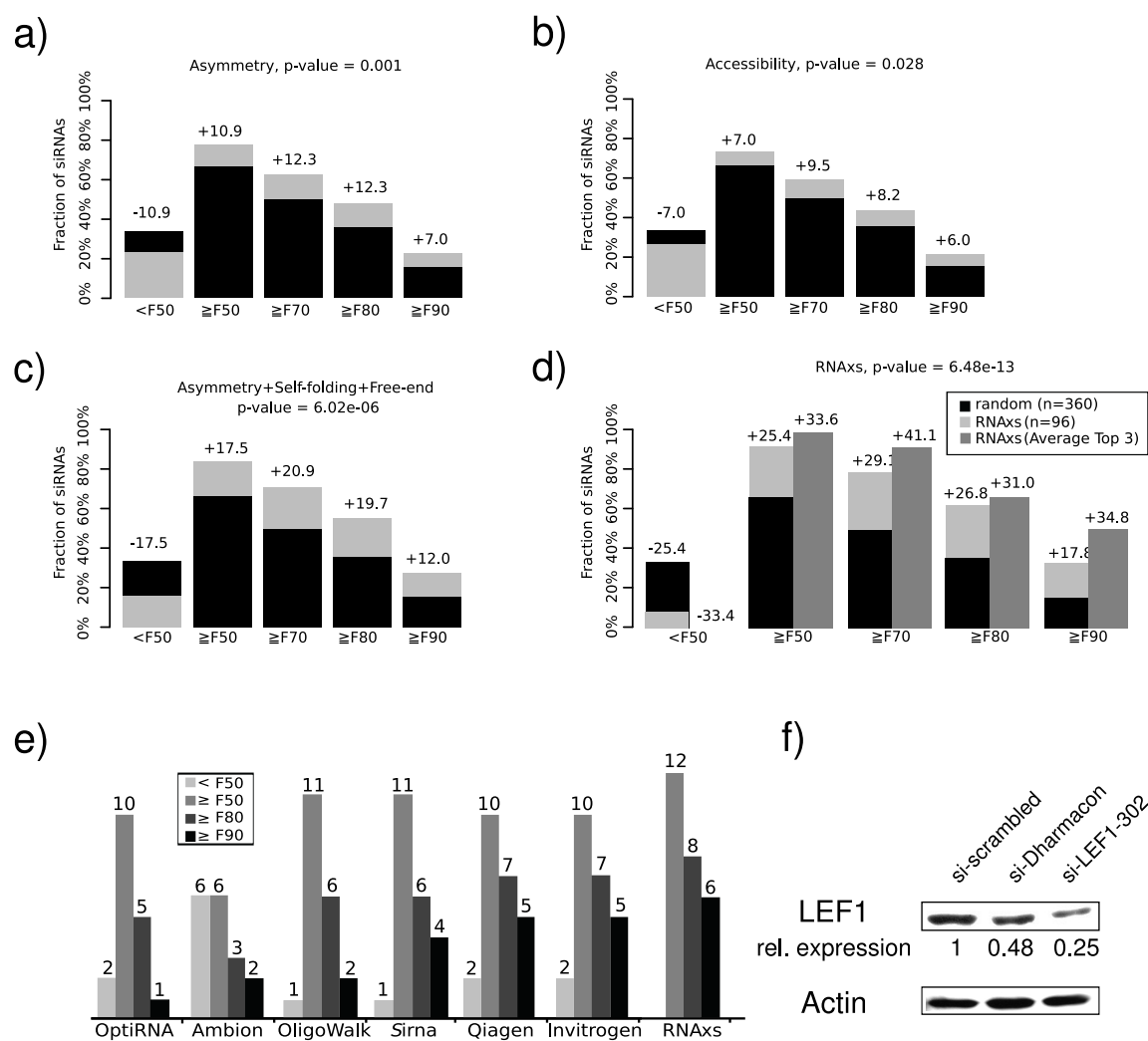


Figure 4.5: Performance of RNAXs on a set of 360 siRNAs targeting the four genes firefly luciferase, human cyclophilin B, ALPPL2 and DBI. SiRNAs were grouped into functionality classes of less than 50% mRNA repression <F50, repression of at least 50% ≥F50, 70% ≥F70, 80% ≥F80 or 90% ≥F90. The random distribution is depicted in black. Functional class enrichments for (a) asymmetry, (b) accessibility, (c) the combination of asymmetry with self-folding plus free-end and (d) all parameters including accessibility (RNAXs) are shown in light gray. The three top ranked siRNAs are all contained in ≥F50 (dark gray). (e) Comparison of RNAXs to other design tools. OptiRNA, Ambion (siRNA Target Finder), Qiagen (siRNA Design Tool), Invitrogen (Block-iT RNAi Designer), oligowalk21 and Sirna (using total score threshold; score > 12) were compared to RNAXs for the four functional classes (<F50, ≥F50, ≥F80, ≥F90). All tools were used with default parameters using the available web servers. For each tool, the repression efficiency of the three best-ranked siRNAs was assessed. RNAXs performed better than the other design tools for all functional classes. (f) Western blot analysis of extracts prepared from Eph4 cells, transiently transfected with scrambled siRNA, Dharmacon mmLEF1 SMARTpool (a combination of four siRNAs) or the single top ranked siRNA designed with RNAXs. Relative LEF1 expression levels are indicated. Actin protein levels show equal loading.

a *BLAST* search in order to detect possible off-target effects (see figure 4.6 and 4.7). To test the performance of our siRNA design tool in comparison to other methods we used a third dataset for validation. This dataset consisted of 360 siRNAs and was independent from the two datasets used to derive optimal design parameters and thresholds. Every second position of an arbitrarily chosen 198 nts stretch within firefly luciferase, human cyclophilin B, human secreted alkaline phosphatase (ALPPL2) and every position of a 108 nts stretch within diazepam binding inhibitor (DBI) was targeted by an individual siRNA.

The performance of the different criteria on this validation set confirmed our previous observations, namely that accessibility and asymmetry are the best single design criteria (see figure 4.5), while free-end and self-folding resulted only in a marginal improvement. The combination of the three traditionally used criteria resulted in a significant enrichment of effective siRNAs (see figure 4.5). The addition of accessibility to the three filters resulted in best performance among all combinations. In contrast, using G/C content in addition to the three traditional criteria resulted only in a slight improvement. Note, that the absolute magnitude of the p-values shown in figure 4.5 is worse than for dataset 1 simply because of the much smaller number of siRNAs.

On average, over 90% of the rationally selected siRNAs were functional and almost every third siRNA reduced gene expression by more than 90%. Furthermore, we looked specifically at the three top ranked siRNAs for firefly luciferase, human cyclophilin B, ALPPL2 and DBI (less amenable for silencing) and found all RNAXs-predicted siRNAs to be functional, half of them reducing gene silencing by more than 90% (see figure 4.5 and figure 4.8).

We compared RNAXs to six existing methods – three commercial siRNA selection tools, a machine learning method – that do not consider target accessibility, **oligowalk** a machine learning method which considers target accessibility and **Sirna** [48], which assesses the target site accessibility as computed by **Sfold** in combination with duplex stability. Since some of these methods return only a few siRNA candidates we limited the comparison to the top three siRNAs predicted by each tool for each of the four genes.

We compared the predicted siRNAs with the measured silencing efficiencies by sort-


RNA XS

1 Data Input

2 View Results

**Welcome to the RNAxs Webserver!** This server will help you to design potent siRNAs to knock down your gene(s) of interest. RNAxs is based on the [RNAfold](#) program to assess the mRNA target site accessibility.

RNAxs was trained on two different datasets, one targeting 3'UTRs and the other one designed to repress coding sequences only.








Simply paste your sequence in FASTA format below. Reasonable default values for siRNA design criteria are pre-chosen, which have shown to give an optimal separation of functional and non-functional siRNAs. To obtain more information about the meaning of the design options hover your mouse over the .

You can test the server using [this sample sequence](#).



Sequence Input

Paste your sequences here: [Clear!](#)

Design Options

8nt (Seed) Accessibility Threshold	0.01157	
16nt Accessibility Threshold	0.001002	
Self Folding Energy	0.9022	
Sequence Asymmetry	0.5	
Energy Asymmetry	0.4655	
Free End	0.625	
Custom Sequence Rules		

Output Option

Maximal Number of siRNAs	3	
E-Mail address (optional):	you@where.org	

REPRESS IT !

Figure 4.6: RNAxs input page. The input page is divided into three areas: a sequence input area, where a FASTA formatted sequence is pasted. A design area where thresholds on different parameters as well as base preferences can be set and the output area which allows to set the number of siRNAs candidates. For each siRNA candidate a plot of the accessibility is generated.

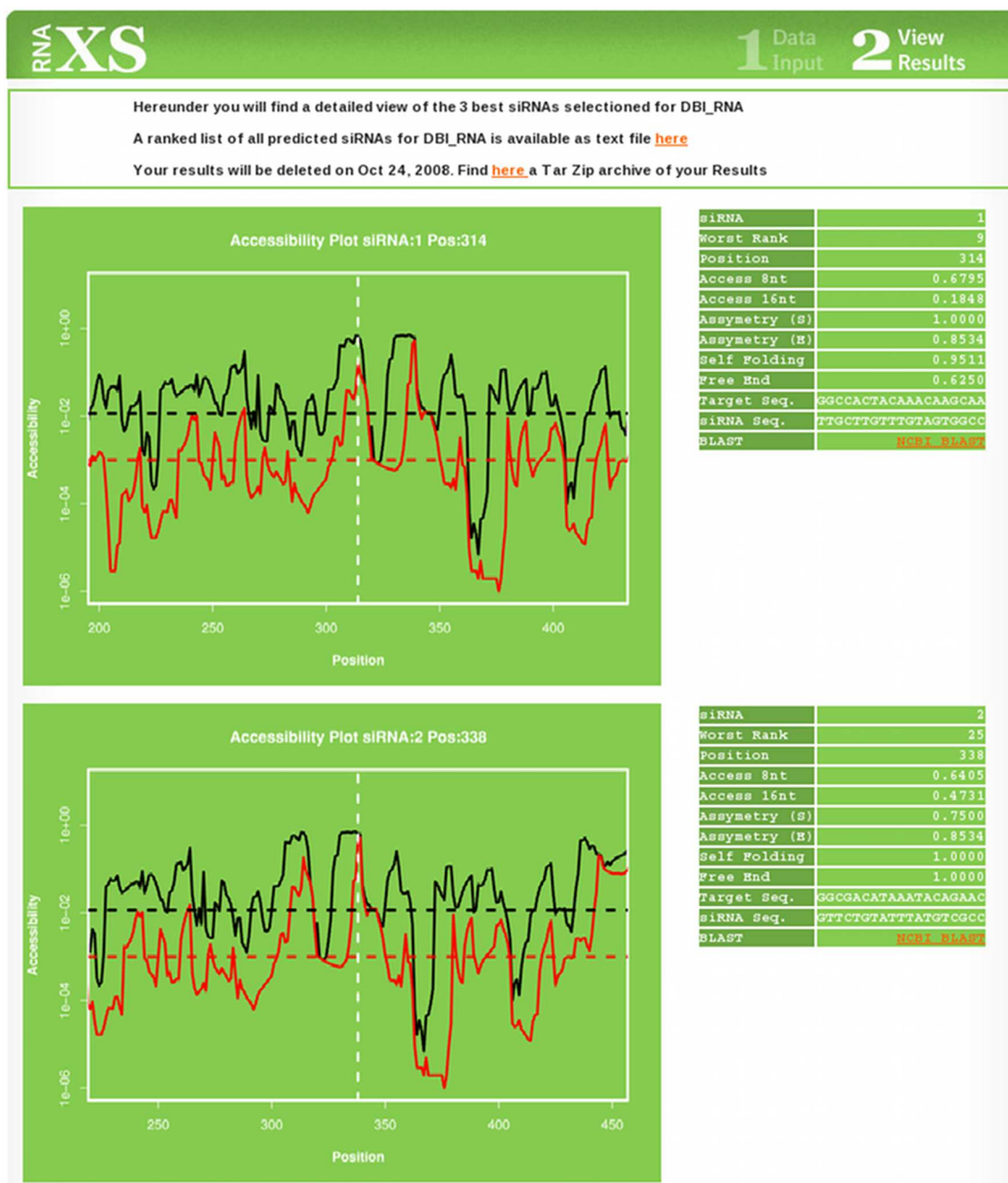


Figure 4.7: Typical output of RNAXS session. A user defined number of siRNA are shown with their features scores as well as a plot of the accessibility profile around the target site. For each siRNA, a link to NCBI blast allows to search for putative off-targets

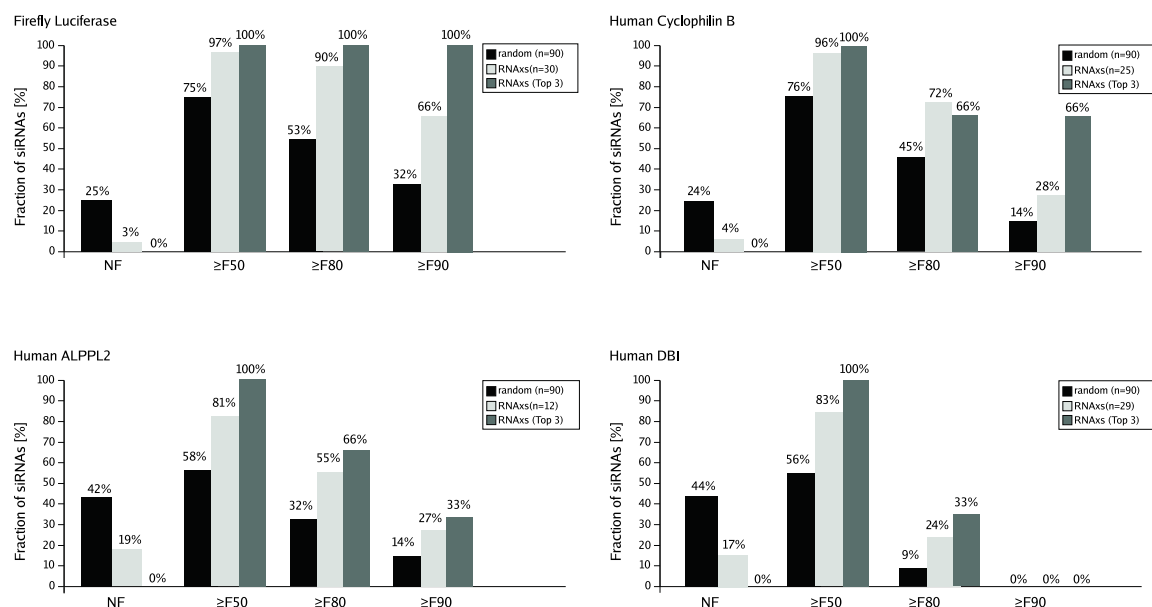


Figure 4.8: Functional siRNA distributions of randomly selected siRNAs (black bars) and rationally designed siRNAs (dark gray bars), as well as the 3 top RNAXs predicted siRNAs (light gray bars) targeting: (A) Firefly luciferase, (B) human cyclophilin B, (C) human ALPPL2, and (D) human DBI

ing them into different functionality classes of less than 50% ( $<F50$ ), more than 50% ( $\geq F50$ ), more than 80% ( $\geq 80$ ), and more than 90% ( $\geq 90$ ) mRNA repression. RNAXs was the only tool where all predicted siRNAs had a measured repression efficiency of  $\geq F50$ ; in fact it outperformed all other programs in each of the functionality classes as shown in figure 4.5.

Furthermore a gene knock-down experiment of the murine Lymphoid Enhancer-Binding Factor 1 (LEF1) protein was performed by Dr. Obernosterer, by using the single top ranked siRNA from RNAXs as well as a commercial siRNA pool and measured the resulting protein levels. The pool, which consisted of a combination of four rationally chosen siRNAs, resulted in 50% knock-down of LEF1, whereas the single siRNA designed with RNAXs resulted in 75% protein knock-down efficiency (figure 4.5). When the respective target sites were analyzed in more details we found that only one of the four siRNAs of the pool would pass the **RNAXs** filters, whereas the other three would be rejected due to accessibility (two out of three) or asymmetry (see figure 4.8).

### 4.3 Target site effects in microRNA pathways

---

miRNAs regulate gene expression in mammals by imperfect base-pairing to the 3-UTR of target mRNAs, thereby mediating either target degradation or translational repression [93]. They control important events during development, differentiation, proliferation, and their deregulation leads to severe diseases such as cancer [28, 39, 176].

Experimental validation of target sites, mostly done in tissue-culture-based reporter gene assays, is difficult since miRNAs form intricate networks targeting more than a hundred 3-UTRs [13, 95]. The development of computational approaches such as PicTar or TargetScan improved the identification of functional target sites [126, 143]. In essence, all of those tools rely on the evolutionary conservation of target sites containing ‘seed regions’. However, there is a constraint in that conservation cannot be applied to non-conserved miRNAs [16]. Some target prediction programs, e.g. RNAhybrid, exclusively compute the free energy between the miRNA and the target RNA [204].

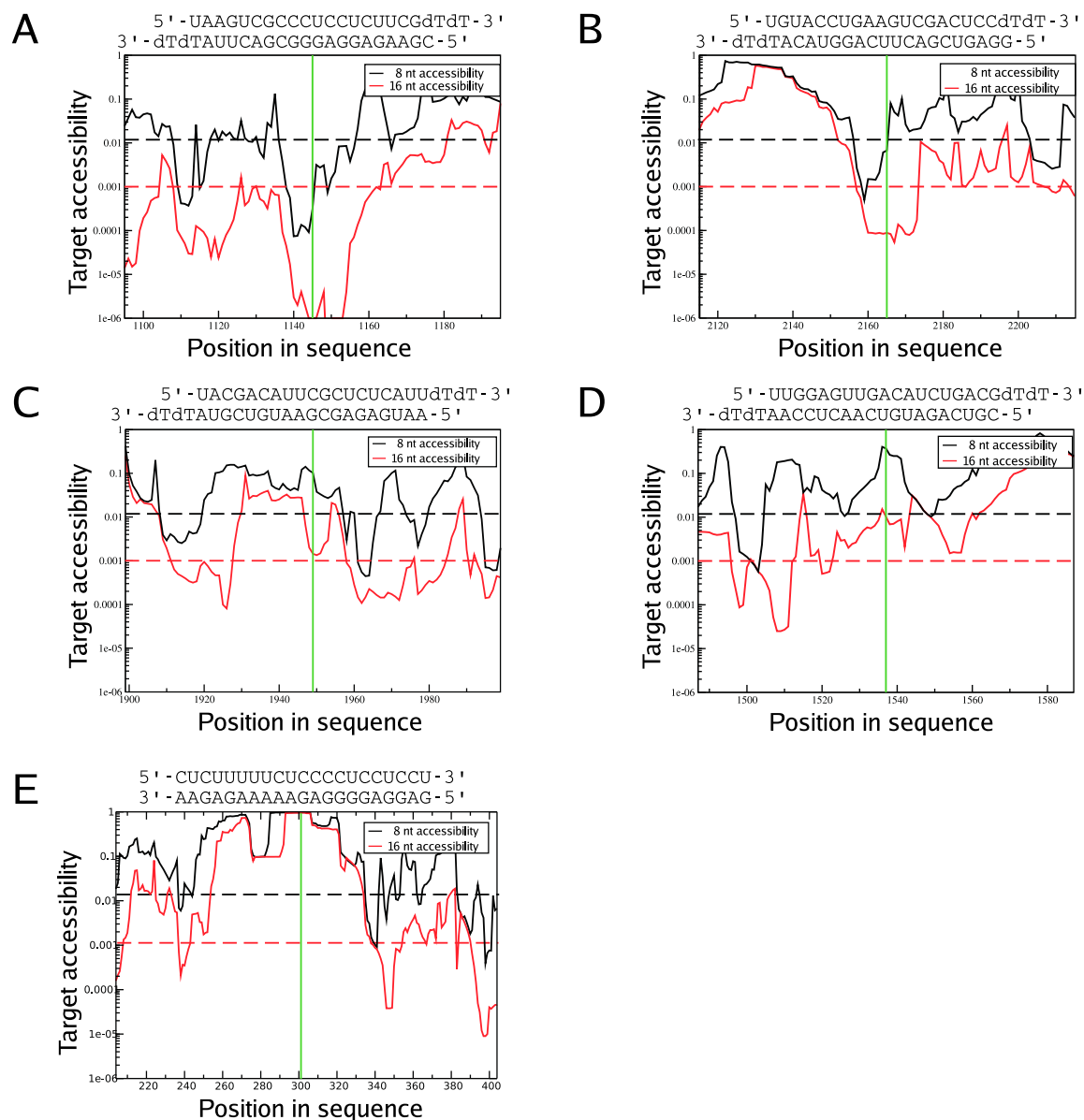


Figure 4.9: Accessibility plots for all four murine LEF1 siRNAs from a commercial siRNA pool (A)-(D) and for the RNAXs designed siRNA (E). The sequence of the respective siRNA duplex is indicated. siRNAs A and B would be rejected by RNAXs because of poor target accessibility. siRNA C would be rejected based on the asymmetry rule.

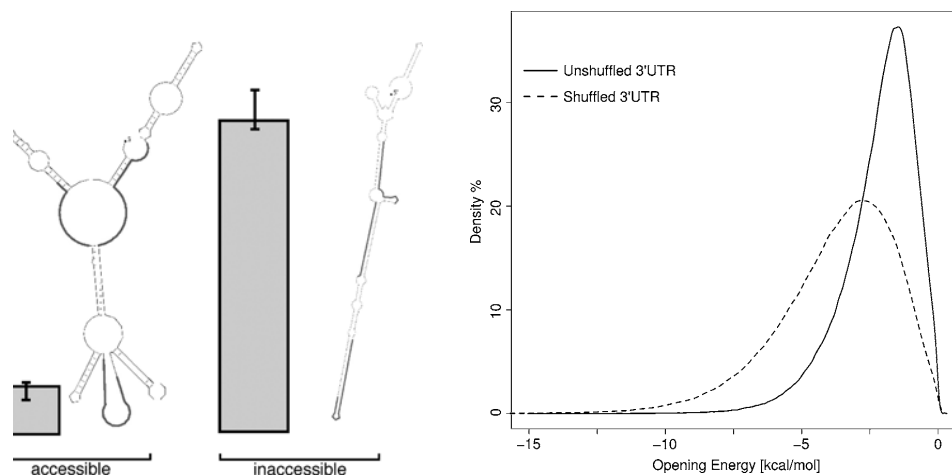


Figure 4.10: **L.h.s** Translational repression of the *Renilla* luciferase (RL), normalized the firefly luciferase (FL), was measured for accessible as well as for non-accessible let-7 reporter constructs. **R.h.s** Distribution of opening energies for human miRNAs. The continuous line represents the density distribution of opening energies corresponding to 3'-UTRs complementary to the seeds for all known human miRNAs. The dotted line shows the density of the shuffled sequences.

Recently, it was shown that the incorporation of context determinants such as A/U (or G/C) content reliably improves the identification of specific sites [86]. Long et al. [149] proposed a structure-based model by combining known features of canonical miRNA target sites such as seed pairing with a two-step hybridization reaction. First, nucleation at the accessible target site takes place followed by miRNA annealing to disrupt closed secondary structures in order to establish a stable miRNA-target duplex. Kertesz et al. [117] have shown that prior to the binding of miRNAs any intramolecular base pairings should be removed. They used conventional RNA folding algorithms to compute the energy cost necessary to remove local secondary structures within the target site. The total miRNA-binding site interaction energy is therefore the sum of the opening energy and the hybridization energy.

We assessed whether RNA sequences complementary to seed regions of known human miRNAs (miRBase release 11.0 [85]) are more accessible than expected. Using RNAplfold, we calculated the opening energy of the reverse complementary seed region and compared this energy with shuffled dinucleotide sequences. Importantly,

the seed region was not shuffled, which ensures (i) that the location of the seed within the 3-UTR remains the same as in the shuffled sequences, avoiding any border effects, (ii) the number of miRNA binding sites stays the same, and (iii) the G/C content is not altered, which might affect accessibility. We found a clear difference in the shape of the distributions for the opening energy between miRNA seeds in real 3-UTRs and seeds where the surrounding 3-UTR sequences were randomized (see figure 4.10). Real miRNA seed sites have a significantly lower opening energy and are therefore more accessible than expected by chance (with a relative enrichment of 17.8%). This result indicates that enhanced accessibility is a feature of miRNA target sites and that it could help separate functional from non-functional target sites. The observation that miRNA targets reside in regions of higher accessibility is in line with other studies [117].

To experimentally test if accessibility affects miRNA-mediated translational repression, three let-7 binding sites behind a luciferase reporter gene [214] were cloned. This original construct was then modified, to contain either highly accessible or non-accessible target sites. In order to alter the accessibility, the surrounding regions around the target sequences were mutated. After transfection of the reporter construct into HeLa cells, we measured the repression efficiency mediated by endogenous let-7. Translational repression was almost completely abolished for the inaccessible construct, but remained unchanged for the accessible one (see figure 4.10). The accessibility and the energy cost to remove intramolecular structures in this small-scale study might explain the difference in the measured repression.

## 4.4 Conclusion

---

In summary, we have shown that the accessibility criterion as computed by the `RNAplfold` strongly influences the RNA hybridization process. Further accessibility combined with biological relevant criteria, like siRNA asymmetry or miRNA target site conservation improve our understanding of how miRNA and siRNA recognize their targets [97].

Chapter 5 and chapter 6 will show that accessibility profiles from `RNAplfold` can be used to search genome-wide for putative ncRNA targets at the same accuracy than

RNAup within a fraction of the time normally required.

...Verein und leite! Besserer Hort.

Johann Wolfgang von Goethe

# 5

## RNAplex

Systematic target prediction for the plethora of genomic information brought by Carena detection programs and high throughput sequencing is a challenging problem [262] and different kinds of tools are available to solve it. Purely sequence based methods like BLAST [3] or FASTA [195] search for long stretches of perfect complementarity between a query and a target sequence. GUUGle [79] can efficiently locate potential complementary regions and, in contrast to BLAST, also allows to consider G·U pairs. A typical application for these programs is for example siRNA target search. Their main drawback is that they do not exploit information about the thermodynamics of the interaction between the query and the target RNA (see chapter 2). Moreover their lack of sensitivity is a real issue even when looking for structurally simple interactions found for example between miRNA and their targets. RNA folding algorithms based on free energy minimization are at present among the most accurate and most generally applicable approaches for RNA folding [247, 275, 276]. Tools like RNAcifold and RNAup (see chapter 3) proved useful for describing precisely RNA–RNA interactions. Still their runtimes  $O(n^3)$  where  $n$  is the length of the longest sequence is prohibitively expensive for most genome-wide applications. A reduction in computational complexity is achieved by omitting the computation of secondary structures within the monomers. This idea was first introduced by RNAhybrid [204] and is also implemented in RNAduplex from the *Vienna RNA* package. It is the simplest and fastest approach with a theoretical time complexity scaling as  $\mathcal{O}(m^2 \cdot n^2)$  which can be reduced to  $\mathcal{O}(m \cdot n \cdot L^2)$  by restricting the maximum loop length to  $L$ . These programs are fast enough e.g. to search for possible targets of a

microRNA. However, for applications where target predictions have to be performed for a large number of small RNAs or when all pairwise comparisons between many RNAs need to be computed, the need for even faster methods still exists. Neglecting the internal structure of the interacting sequences leads to a drastic decrease in specificity, however, see Figure 5.1 and Figure 2.10.

Currently, one therefore has to choose between precise but impractically slow methods or fast but imprecise methods for ncRNA target search, a situation that is quite unsatisfactory. In this chapter the development and application of a new tool called **RNAplex** is presented. **RNAplex** solves the problem of finding *precisely* and *in linear time* putative targets for ncRNAs. Roughly, this is achieved by using a slightly modified energy model together with a scoring system that mimic the effect of the competition between intra- and intermolecular interactions and/or how well the interaction is conserved across different species.

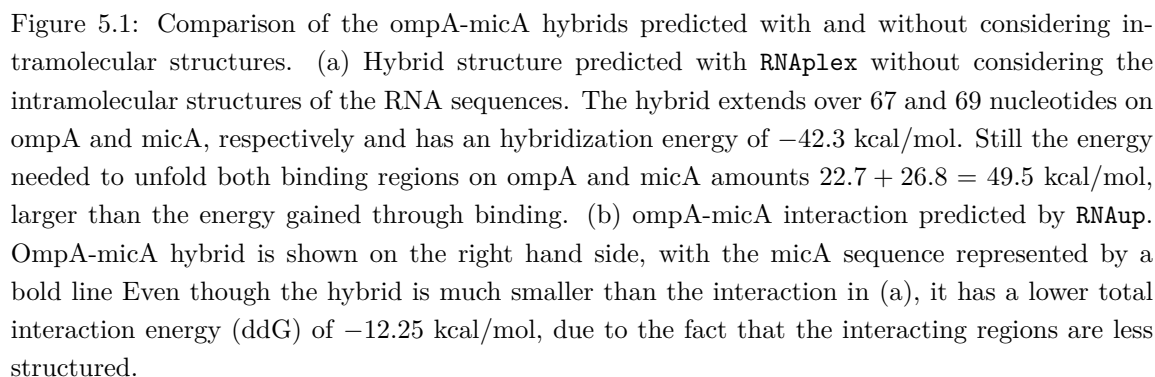
## 5.1 Methods

---

### 5.1.1 Energy model

**RNAduplex**/**RNAhybrid** are essentially equivalent to the classic RNA folding algorithm of Zuker & Stiegler [276] when only interior loops are allowed. As such they have a time complexity of  $\mathcal{O}((n \cdot m)^2)$  in the naive implementation, where  $n$  and  $m$  represent the length of the interacting nucleotide sequences. It is a common practice to speed up these algorithms by restricting the loop size to  $L$  leading to  $\mathcal{O}(n \cdot m \cdot L^2)$ , where  $L = 30$  in the case of **RNAduplex**. Here we use a simplified energy model that allows us to get rid of the constant but fairly large prefactor  $L^2$ .

Since we are neglecting intra-molecular structure here, the only loop types that can appear are stacked pairs, bulge loops, and interior loops. The Turner energy parameters provide look-up tables for the free energies of stacked pairs as well as for small interior loops (1x1, 2x1, and 2x2 loops). These look-up tables are used in **RNAplex** without change. Likewise, bulge loops of length 1 are treated exactly as in the full energy model, namely by adding the stacking energy of the two pairs closing the loop plus a sequence independent penalty. Larger bulge loops are normally assigned



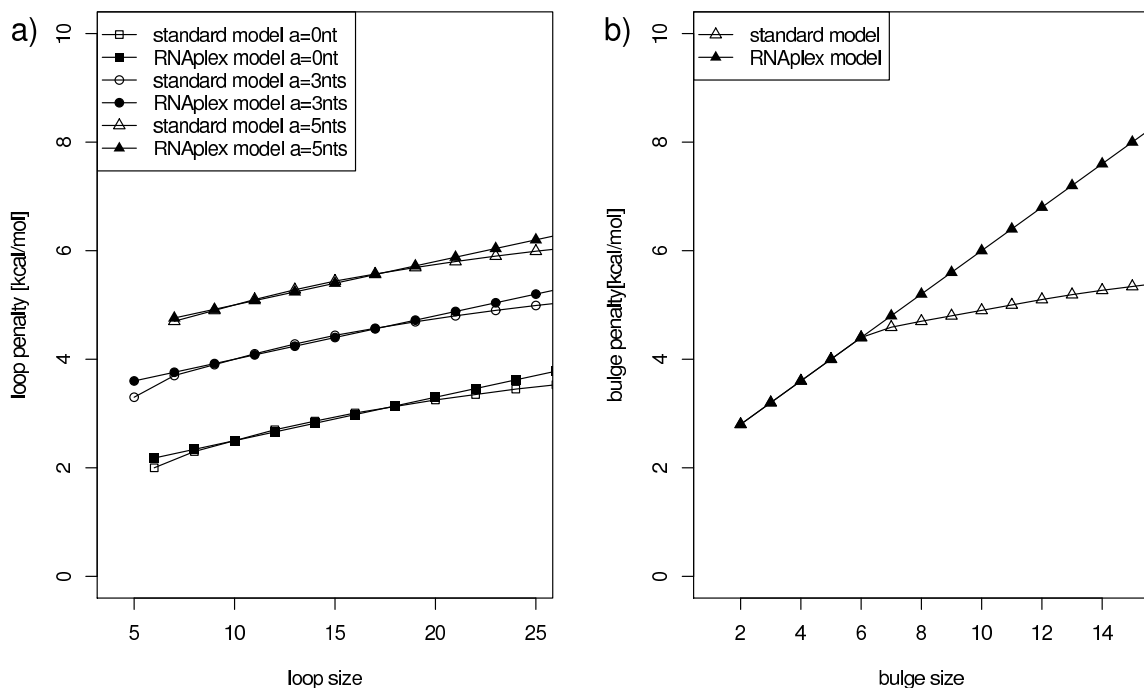


Figure 5.2: Comparison of the **RNAplex** energy model against the Turner energy model for bulges and interior loops *a)* Plot of the interior loop penalty against the total loop size for three different values of asymmetry. The model used in **RNAplex** slightly overestimates the loop energies. *b)* Plot representing the bulge loop penalty against the bulge size. Our model agree exactly with the Thurner model for bulge size up to 6 nts.

a length dependent penalty that grows logarithmically for large loops. In **RNAplex** this bulge energy is approximated by an affine function. Similarly, large interior loops are normally modeled by a size dependent term, an asymmetry penalty, and sequence dependent “terminal mismatches”. Here again, we replace the size dependent loop energy by an affine function. Finally the asymmetry term is approximated by penalizing asymmetrical extension of interior loops (see equation 5.1). The resulting energy model is exact for small loops and slightly overestimates the loop energies of large bulge loops as well as strongly asymmetric loops (see Figure 5.2).

### 5.1.2 Recursion

The structure of RNA duplexes predicted by our model can be decomposed into stacking pairs, interior loops and bulges. Our dynamic programming algorithm therefore employs four tables representing sub-structures that end in a base pair  $C$ , interior loop  $I$  and bulge on the first or second sequence,  $B^x, B^y$ , respectively. The central quantity  $C_{i,j}$  stores the best energy of interaction between sub-sequence  $x[1..i]$  and  $y[j..m]$ . Similarly  $B_{i,j}^{x,y}$  store the best energy of interaction given that residue  $y_j$ , respectively  $x_i$ , is aligned to a bulge. Finally  $I_{i,j}$  stores the best energy of interaction given that  $x_i$  and  $y_j$  are in an interior loop. The asymmetry penalty is modeled by allowing symmetrical extension of the interior loops as well as asymmetrical, penalized, interior loop extension (see equation 5.1). Based on these matrices the recursion relation can be written as:

$$C_{i,j} = \min \left\{ \begin{array}{l} C_{i-1,j+1} + \mathcal{S}(i,j;i-1,j+1) \\ C_{i-1,j+2} + \mathcal{S}(i,j;i-1,j+2) + P_{\text{bulge}} \\ C_{i-2,j+1} + \mathcal{S}(i,j;i-2,j+1) + P_{\text{bulge}} \\ C_{i-2,j+2} + \mathcal{I}(i,j;i-2,j+2) \\ C_{i-3,j+2} + \mathcal{I}(i,j;i-3,j+2) \\ C_{i-2,j+3} + \mathcal{I}(i,j;i-2,j+3) \\ C_{i-3,j+3} + \mathcal{I}(i,j;i-3,j+3) \\ I_{i-1,j+1} + \mathcal{M}(i,j;i-1,j+1) \\ B_{i-1,j+1}^x \\ B_{i-1,j+1}^y \end{array} \right. \quad (5.1)$$

$$I_{i,j} = \min \left\{ \begin{array}{l} C_{i-1,j+1} + \mathcal{M}(i-1,j+1;i,j) + g_{\text{open}}^I + 2g_{\text{ext}}^I \\ I_{i-1,j} + g_{\text{ext}}^I + A \\ I_{i-1,j+1} + 2 * g_{\text{ext}}^I \\ I_{i,j+1} + g_{\text{ext}}^I + A \end{array} \right. \quad (5.2)$$

$$B_{i,j}^x = \min \left\{ \begin{array}{l} C_{i-1,j} + g_{\text{open}}^B + g_{\text{ext}}^B \\ B_{i-1,j}^x + g_{\text{ext}}^B \end{array} \right. \quad (5.3)$$

$$B_{i,j}^y = \min \left\{ C_{i,j+1} + g_{\text{open}}^B + g_{\text{ext}}^B \right. \quad (5.4)$$

where  $\mathcal{S}(i,j,k,l)$  represents the energy gained by stacking the  $(x_i y_j)$  base pair onto

the  $(x_k, y_l)$  base pair. As usual, bulges of length 1 are modeled as the sum of a bulge penalty  $P_{\text{bulge}}$  plus the stacking energy of the adjacent base pairs.  $\mathcal{M}(i, j; i-1, j+1)$  represents the “mismatch” energy of the unpaired nucleotides  $(x_{i-1}, y_{j+1})$  adjacent to the pair  $(x_i, x_j)$ .  $\mathcal{I}$  represents the energy contribution of the small interior loops. Furthermore  $g_{\text{open}}^{B,I}$  and  $g_{\text{ext}}^{B,I}$  represent the parameters of the affine loop energy function that approximates the conventional Turner loop energies. These parameters were gained by linearly fitting the loop energy model. Finally  $A$  represents the asymmetry penalty that approximates the extra destabilizing energy of asymmetrical loops. The above recursion is graphically represented in Figure 5.3.

In our model a duplex starts with 2 stacked pairs  $(x_i, y_j) \cdot (x_{i-1}, y_{j+1})$ . The initialization of the recursion matrices should ensure that all structural element has to start and end inside the recursion matrices. This means that no interior loops and no bulges on the target sequence may be closed before  $i = 3$ . Moreover no bulge and no interior loop on the query sequence may be closed before  $j = m - 2$ . Finally  $C_{1,0}$  is set to 0. As a consequence the matrices are initialized in the following way

$$\begin{aligned} I_{1,j} &= I_{2,j} &= \infty \quad \forall j \\ B_{1,j}^x &= B_{2,j}^x &= \infty \quad \forall j \\ I_{i,m} &= I_{i,m-1} &= \infty \quad \forall i \\ B_{i,m}^y &= B_{i,m-1}^y &= \infty \quad \forall i \end{aligned}$$

When comparing an RNA of length  $m$  against a large database of length  $n \gg m$  the optimal interaction typically spans the full length of the shorter RNA  $m$ . However, long interactions, extending over many helical turns, are sterically hindered, and moreover have to compete with the tendency to form intra-molecular structure. Therefore, hits consisting of a short but stable duplex should be preferable over interactions that attain a good score only by adding many weak interactions over a long region. To counter this effect, **RNAplex** contains an option that introduces a per nucleotide penalty to the interaction energy. Especially for longer queries, this results in shorter and statistically more significant interactions.

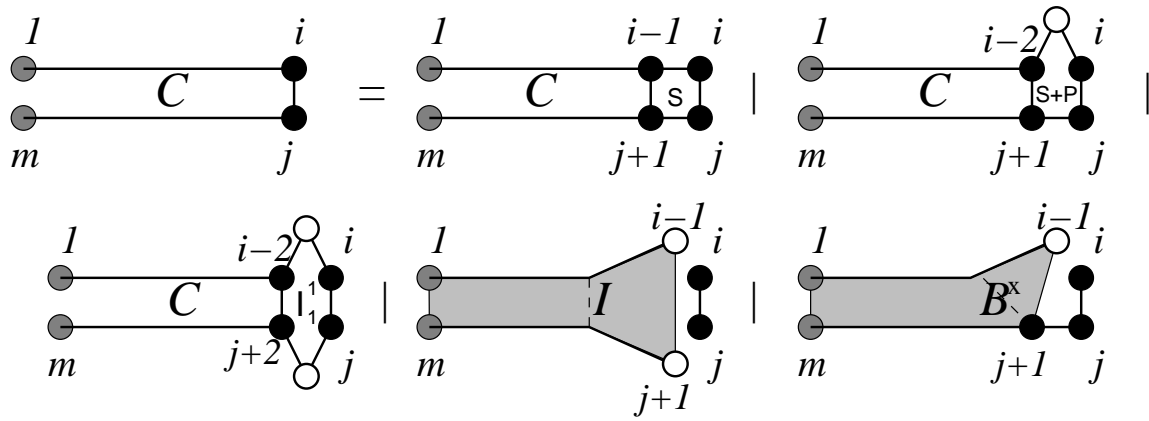


Figure 5.3: Simplified representation of the structure decomposition used in **RNAplex**. For clarity only the decomposition of the closed structure terms (see equation (5.1)) is shown. Black dots represent paired bases. White dots denote unpaired bases. Given that  $x_i$  and  $y_j$  are paired,  $C$  stores the best energy of interaction between  $x_1..x_i$  and  $y_j..y_m$ .  $\mathcal{S}$  is the stacking energy of two pairs of nucleotides.  $P$  is the bulge penalty to add to 1x0 bulges.  $I$  is the matrix holding the best energy of interaction given that  $x_i$  and  $y_j$  are in an interior loop.  $I_1^1$  is the destabilizing energy of a 1x1 interior loop (1x2, 2x1 and 2x2 cases not shown) and  $B^x$  represents the matrix storing the best energy of interaction given that residue  $y_j$  is aligned to a bulge. The cases where  $x_i$  and  $y_j$  do not pair (interior loop and bulge extension and/or creation) are not shown

### 5.1.3 Taking the target accessibility back into RNAp<sub>lex</sub>

RNAp<sub>lex</sub> does have a reduced runtime compared to RNAduplex or RNAhybrid. Still the negligence of target site accessibility make its prediction not as significant as that of approaches that considers target site accessibility (see chapter 2). A straightforward way to increase significance of RNAp<sub>lex</sub> predictions would be to apply RNAup on high confidence RNAp<sub>lex</sub> results and select interactions that are highly-scored by RNAup. While this approach reduces the number of false-positive interactions returned by RNAp<sub>lex</sub>, the number of false-negative remains unaffected.

In the example presented in Table 5.5, RNAp<sub>lex</sub>+RNAup would fail to predict two interactions because RNAp<sub>lex</sub> can not retrieve them. In this section we show how RNAp<sub>lex</sub> can be used in conjunction with accessibility profiles, as produced by RNAp<sub>l</sub>fold and RNAup, to obtain a target prediction accuracy similar to that of RNAup without incrementing the runtime of RNAp<sub>lex</sub>.

RNAp<sub>lex</sub> employs a two steps approach to identify interactions. In the first step, RNAp<sub>lex</sub> identifies positions where putative interactions may end. In this scanning phase RNAp<sub>lex</sub> uses a linear approximation of the size dependence of loop energies used in the standard energy model. For small interior loops (1x1, 2x1, and 2x2) and bulges of size 1, RNAp<sub>lex</sub> employs the look-up tables provided by the Turner Energy Model. The resulting energy model is exact for small loops and slightly overestimates the loop energies of large interior, bulge loops as well as strongly asymmetric loops. A further advantage of the linear model is that RNAp<sub>lex</sub> only needs to record the last 4 columns of the recursion matrix in order to complete the dynamic programming recursion. When all high-scoring interactions were localized along the target sequence, RNAp<sub>lex</sub> uses the standard energy model to recompute the energy and structure of the putative hybrids.

During the scan phase, in order to extend a hybrid by one nucleotide, we need to know the cost of freeing this nucleotide from all the intramolecular interactions it might be involved in. In thermodynamic equilibrium this energy cost can be derived from the probability that the interacting stretch of nucleotides is unpaired. Since it is too expensive to compute this for all intervals, we seek a step-wise procedure. Consider an intermediary hybrid structure  $\mathcal{S}_y^x$  between two sequences  $x$  and  $y$  that starts at base pair  $(x_i, y_j)$  and spans  $w_x$  nucleotides of sequence  $x$  and  $w_y$  nucleotides

of sequence  $y$ . We need to determine the *conditional* probability  ${}^{w_x}P_u^x[i + w_x]$  that nucleotide  $x_{i+w_x}$  is not involved in any intramolecular interaction, *given* that its predecessors  $i + w_x - 1$  is unpaired, and the analogous quantity  ${}^{w_y}P_u^y[j - w_y]$ . The subscript  $u$  emphasizes that the nucleotides  $x$  and  $y$  are supposed to be unpaired. Note that this is not the same as the problem of assessing the probability  $P_u[i + w_x]$  that the individual nucleotides  $x_{i+w_x}$  is unpaired, because base pairing probabilities of adjacent nucleotides are highly correlated [26].

The desired conditional probability can be written as

$${}^{w_x}P_u^x[i + w_x] = P_u^x([i + w_x] | [i, i + w_x - 1]). \quad (5.5)$$

where the notation means that the interval  $[i, i + w_x - 1]$  is unpaired. An analogous expression holds for sequence  $y$ . Using the definition of the conditional probability we can write: we can write:

$${}^{w_x}P_u^x[i + w_x] = \frac{P_u^x([i, i + w_x - 1] \cup [i + w_x])}{P_u^x[i, i + w_x - 1]} \quad (5.6)$$

$$= \frac{P_u^x[i, i + w_x]}{P_u^x[i, i + w_x - 1]} \quad (5.7)$$

Equation 5.6 tells us that the conditional probability  ${}^{w_x}P_u^x[i + w_x]$  depends only on the probabilities  $P_u^x[i, i + w_x]$  and  $P_u^x[i, i + w_x - 1]$ . that the corresponding intervals are unpaired. Conversely, the probability that an intervals is unpaired can be computed from the conditional probabilities and the probabilities that individual nucleotides are unpaired:

$$P_u^x[i, i + w_x] = P_u^x[i] \cdot \prod_{j=1}^{w_x} {}^jP_u^x[i + j] \quad (5.8)$$

A closer look at equation (5.6) shows that the exact start position of the hybrid  $\mathcal{S}_y^x$  has to be known in order to compute the desired conditional probability. Since RNApIex stores only a small number (four) of columns of the dynamic programming matrix, this cannot be done exactly. Instead we employ the approximation

$$\begin{aligned} \frac{P_u^x[i, i + w_x]}{P_u^x[i, i + w_x - 1]} &\approx \frac{P_u^x[i + w_x - \delta + 1, i + w_x]}{P_u^x[i + w_x - \delta + 1, i + w_x - 1]} \\ &= {}^\delta P_u^x[i + w_x] \end{aligned} \quad (5.9)$$

where  $\delta$  represents the number of nucleotides considered in prior to nucleotides  $x_{i+w_x}$  and  ${}^\delta P_u^x[i + w_x]$  represents the conditional probability that  $x_{i+w_x}$  is unpaired for a given  $\delta$ . This approximation is exact for  $\delta = w_x$  and gets worse with decreasing  $\delta/w_x$ . This is a direct consequence of the fact that the state of the nucleotides in the interval  $[i, i + w_x - \delta + 1]$  is not taken into account for the computation of the conditional probability of nucleotide  $x_{i+w_x}$ .

Equation (5.8) can now be rewritten in the form

$$P_u^x[i, i + w_x] \approx {}^\delta P_u^x[i, i + w_x] = P_u^x[i, i + \delta - 1] \cdot \prod_{j=\delta}^{w_x} {}^\delta P_u^x[i + j] \quad (5.10)$$

$$P_u^x[i, i + w_x] \approx {}^\delta P_u^x[i, i + w_x] = P_u^x[i, i + \delta - 1] \cdot \prod_{j=\delta}^{w_x} {}^\delta P_u^x[i + j] \quad (5.11)$$

The probability  $P_u^x[i, i + w_x]$  of being unpaired is related to a corresponding opening energy

$$\Delta G_u^x[i, i + w_x] = -RT \ln P_u^x[i, i + w_x]. \quad (5.12)$$

The energy cost of adding one nucleotide to the hybrid therefore can be written as

$$\begin{aligned} \Delta {}^\delta G_u^x[i + w_x] &= -RT \ln {}^\delta P_u^x[i + w_x] = \\ &\Delta G_u^x[i + w_x - \delta + 1, i + w_x] \\ &- \Delta G_u^x[i + w_x - \delta + 1, i + w_x - 1]. \end{aligned} \quad (5.13)$$

The opening energy of a region of size of  $w$  thus is given by

$$\begin{aligned} \Delta {}^\delta G_u^x[i, i + w_x] &= -RT \ln {}^\delta P_u^x[i, i + w_x] = \\ &\Delta G_u^x[i, i + \delta - 1] + \sum_{j=\delta}^{w_x} \Delta {}^\delta G_u^x[i + j]. \end{aligned} \quad (5.14)$$

Since **RNAplex** only stores the current four columns of the recursion matrix, we set  $\delta = 4$  in practice.

The energy  $\Delta^4 G_u^x[i]$  of freeing nucleotide  $x_i$  from all its intramolecular interactions can now easily be integrated into the dynamic programming recursion of RNAp<sub>lex</sub>. We use the following abbreviations for the opening energies:

$$d_1^x = \Delta^4 G_u^x[i], d_2^x = d_1^x + \Delta^4 G_u^x[i-1], d_3^x = d_2^x + \Delta^4 G_u^x[i-2]; \text{ and } d_1^y = \Delta^4 G_u^y[j], d_2^y = d_1^y + \Delta^4 G_u^y[j+1], d_3^y = d_2^y + \Delta^4 G_u^y[j+2].$$

$$C_{i,j} = \min \left\{ \begin{array}{l} C_{i-1,j+1} + \mathcal{S}(i,j;i-1,j+1) + d_1^x + d_1^y \\ C_{i-1,j+2} + \mathcal{S}(i,j;i-1,j+2) + P_{\text{bulge}} + d_1^x + d_2^y \\ C_{i-2,j+1} + \mathcal{S}(i,j;i-2,j+1) + P_{\text{bulge}} + d_2^x + d_1^y \\ C_{i-2,j+2} + \mathcal{I}(i,j;i-2,j+2) + d_2^x + d_2^y \\ C_{i-3,j+2} + \mathcal{I}(i,j;i-3,j+2) + d_3^x + d_2^y \\ C_{i-2,j+3} + \mathcal{I}(i,j;i-2,j+3) + d_2^x + d_3^y \\ C_{i-3,j+3} + \mathcal{I}(i,j;i-3,j+3) + d_3^x + d_3^y \\ I_{i-1,j+1} + \mathcal{M}(i,j;i-1,j+1) + d_1^x + d_1^y \\ B_{i-1,j+1}^x + d_1^x \\ B_{i-1,j+1}^y + d_1^y \end{array} \right. \quad (5.15)$$

$$I_{i,j} = \min \left\{ \begin{array}{l} C_{i-1,j+1} + \mathcal{M}(i-1,j+1;i,j) + g_{\text{open}}^I + 2g_{\text{ext}}^I + d_1^x + d_1^y \\ I_{i-1,j} + g_{\text{ext}}^I + A + d_1^x \\ I_{i-1,j+1} + 2 * g_{\text{ext}}^I + d_1^x + d_1^y \\ I_{i,j+1} + g_{\text{ext}}^I + A + d_1^y \end{array} \right. \quad (5.16)$$

$$B_{i,j}^x = \min \left\{ \begin{array}{l} C_{i-1,j} + g_{\text{open}}^B + g_{\text{ext}}^B + d_1^x \\ B_{i-1,j}^x + g_{\text{ext}}^B + d_1^x \end{array} \right. \quad (5.17)$$

$$B_{i,j}^y = \min \left\{ \begin{array}{l} C_{i,j+1} + g_{\text{open}}^B + g_{\text{ext}}^B + d_1^y \\ B_{i,j+1}^y + g_{\text{ext}}^B + d_1^y \end{array} \right. \quad (5.18)$$

## Hybrid Structure and Hybrid Energy

The computation of the hybrid structure and interaction energy follows the strategy of RNAup. We assume that the binding region may contain mismatches and bulge loops. Thus the most stable interaction between two segments  $(x_i, y_j)$  and  $(x_k, y_l)$  is

obtained by minimizing over all possible interior loop closed by  $(x_p, y_q)$

$$C(x_i, y_j, x_k, y_l) = \min_{\substack{x_k < x_p < x_i \\ y_l > y_q > y_j}} C(x_i, y_j, x_p, y_q) + I(x_p, y_q, x_k, y_l) + \Delta G_u^x[i, k] + \Delta G_u^y[j, l] \quad (5.19)$$

The overall most stable interaction is then obtained by minimizing over both duplex closing pairs  $(x_i, y_j)$  and  $(x_k, y_l)$ :

$$E_{\min} = \min_{\substack{x_1 < x_k < x_i < x_n \\ y_1 < y_j < y_l < y_m}} C(x_i, y_j, x_k, y_l) \quad (5.20)$$

where  $n$  and  $m$  are the length of sequences  $x$  and  $y$ , respectively. This leads to a theoretical run-time of  $\mathcal{O}(n^3 \cdot m^3)$  and a memory footprint of  $\mathcal{O}(n^2 \cdot m^2)$ .

Here we should note that one end of the hybrid, namely the base-pair  $(x_i, y_j)$ , was already found in the scanning phase of **RNAplex**. As a consequence we only need to minimize over one closing-pair instead of two. Equation 5.20 can thus be rewritten as:

$$E_{\min} = \min_{\substack{x_1 < x_k < x_i \\ y_j < y_l < y_m}} C(x_i, y_j, x_k, y_l) \quad (5.21)$$

Equations 5.20 and 5.21 show that the knowledge of base-pair  $(x_i, y_j)$  allows to reduce memory and run-time by a factor  $n \cdot m$ . Furthermore, the size of the interaction regions as well as the size of interior loops can be limited to arbitrary lengths  $\omega$  and  $L$ , respectively, leading to a run-time of  $\mathcal{O}(\omega^2 \cdot L^2)$  and a memory usage of  $\mathcal{O}(\omega^2)$ , that is, the same complexity as **RNA duplex** or **RNA hybrid**.

#### 5.1.4 Conserved Interactions

The absence of conserved target-site in closely related species may indicate that the proposed interaction does not occur in nature. The presence of compensatory mutations between the sRNA and the target site, on the other hand, can lend further credibility to single-sequence target predictions [35]. Alignments thus can improve the specificity of target search by focusing on evolutionary conserved interactions.

We therefore extended **RNAplex** to alignments. The approach follows the same idea as **RNAalifold** [19, 98], where a thermodynamic energy minimization folding algorithm

is coupled with a simple scoring model to assess structural evolutionary conservation. Base pairs are therefore restricted to pairs of positions in the alignments in which most or all sequences can form canonical pairs.

In case of conserved interactions, **RNAplex** takes two multiple sequences alignments as input and computes the interactions based on these alignments. Let  $\mathbb{X}$  and  $\mathbb{Y}$  be the target and query alignments, respectively, each with the same number of sequences  $N$ .  $\mathbb{X}_i$  represents the  $i^{th}$  column of alignment  $\mathbb{X}$ , and  $\mathbb{X}^\alpha$  represents the  $\alpha^{th}$ -sequence in the alignment.

Similar to the single case,  $C_{i,j}$  represents the smallest sum over all sequences  $\alpha \in \mathbb{X}, \mathbb{Y}$  of the interaction energy between the subsequences  $\mathbb{X}_1^\alpha \dots \mathbb{X}_i^\alpha$  and  $\mathbb{Y}_j^\alpha \dots \mathbb{Y}_m^\alpha$ , where  $m$  is the length of the query alignment. Similarly,  $B_{i,j}^{\mathbb{X},\mathbb{Y}}$  stores the optimal interactions energy given that residues  $\mathbb{X}_i^\alpha$  or residue  $\mathbb{Y}_j^\alpha$ ,  $\forall \alpha \in \mathbb{X}, \mathbb{Y}$  are part of a bulge;  $I_{i,j}$  stores the optimal interaction energy given that all residues  $\mathbb{X}_i^\alpha$  and  $\mathbb{Y}_j^\alpha$ ,  $\forall \alpha \in \mathbb{X}, \mathbb{Y}$  are in an interior loop.

The asymmetry penalty  $A$  models asymmetric extension of interior loops.  $\mathcal{S}(i, j, i - 1, j + 1)$  represents the sum  $\sum_{\alpha=1}^N \mathcal{S}(\mathbb{X}_i^\alpha, \mathbb{Y}_j^\alpha, \mathbb{X}_{i-1}^\alpha, \mathbb{Y}_{j+1}^\alpha)$  over all sequences  $\alpha \in \mathbb{X}, \mathbb{Y}$  of the energies gained by stacking base-pairs  $(\mathbb{X}_i^\alpha, \mathbb{Y}_j^\alpha)$  onto  $(\mathbb{X}_{i-1}^\alpha, \mathbb{Y}_{j+1}^\alpha)$ .

$\mathcal{M}(i, j, i - 1, j + 1)$  represents the sum  $\sum_{\alpha=1}^N \mathcal{M}(\mathbb{X}_i^\alpha, \mathbb{Y}_j^\alpha, \mathbb{X}_{i-1}^\alpha, \mathbb{Y}_{j+1}^\alpha)$  over all sequences  $\alpha$  of the mismatch energy for the unpaired nucleotides  $(\mathbb{X}_{i-1}^\alpha, \mathbb{Y}_{j+1}^\alpha)$  adjacent to the pair  $(\mathbb{X}_i^\alpha, \mathbb{Y}_j^\alpha)$ . The energy contribution of the small interior loops is represented by  $\mathcal{I}$ . Furthermore, we use the following abbreviations for the opening energies:

$$\begin{aligned} d_1^{\mathbb{X}} &= \sum_{\alpha=1}^N d_1^{\mathbb{X}^\alpha} = \sum_{\alpha=1}^N \Delta^4 G_u^{\mathbb{X}^\alpha}[i], \quad d_1^{\mathbb{Y}} = \sum_{\alpha=1}^N d_1^{\mathbb{Y}^\alpha} = \sum_{\alpha=1}^N \Delta^4 G_u^{\mathbb{Y}^\alpha}[j], \\ d_2^{\mathbb{X}} &= d_1^{\mathbb{X}} + \sum_{\alpha=1}^N d_2^{\mathbb{X}^\alpha} = d_1^{\mathbb{X}} + \sum_{\alpha=1}^N \Delta^4 G_u^{\mathbb{X}^\alpha}[i - 1], \quad d_2^{\mathbb{Y}} = d_1^{\mathbb{Y}} + \sum_{\alpha=1}^N d_2^{\mathbb{Y}^\alpha} = d_1^{\mathbb{Y}} + \sum_{\alpha=1}^N \Delta^4 G_u^{\mathbb{Y}^\alpha}[j - 1], \\ d_3^{\mathbb{X}} &= d_2^{\mathbb{X}} + \sum_{\alpha=1}^N d_3^{\mathbb{X}^\alpha} = d_2^{\mathbb{X}} + \sum_{\alpha=1}^N \Delta^4 G_u^{\mathbb{X}^\alpha}[i - 2], \quad d_3^{\mathbb{Y}} = d_2^{\mathbb{Y}} + \sum_{\alpha=1}^N d_3^{\mathbb{Y}^\alpha} = d_2^{\mathbb{Y}} + \sum_{\alpha=1}^N \Delta^4 G_u^{\mathbb{Y}^\alpha}[j + 2]. \end{aligned}$$

The recursion of **RNAplex** can then be rewritten as:

$$C_{i,j} = \min \left\{ \begin{array}{l} C_{i-1,j+1} + \sum_{\alpha=1}^N (\mathcal{S}(\mathbb{X}_i^\alpha, \mathbb{Y}_j^\alpha, \mathbb{X}_{i-1}^\alpha, \mathbb{Y}_{j+1}^\alpha) + d_1^{\mathbb{X}^\alpha} + d_1^{\mathbb{Y}^\alpha}) \\ C_{i-1,j+2} + \sum_{\alpha=1}^N (\mathcal{S}(\mathbb{X}_i^\alpha, \mathbb{Y}_j^\alpha, \mathbb{X}_{i-1}^\alpha, \mathbb{Y}_{j+2}^\alpha) + P_{\text{bulge}} + d_1^{\mathbb{X}^\alpha} + d_2^{\mathbb{Y}^\alpha}) \\ C_{i-2,j+1} + \sum_{\alpha=1}^N (\mathcal{S}(\mathbb{X}_i^\alpha, \mathbb{Y}_j^\alpha, \mathbb{X}_{i-2}^\alpha, \mathbb{Y}_{j+1}^\alpha) + P_{\text{bulge}} + d_2^{\mathbb{X}^\alpha} + d_1^{\mathbb{Y}^\alpha}) \\ C_{i-2,j+2} + \sum_{\alpha=1}^N (\mathcal{I}(\mathbb{X}_i^\alpha, \mathbb{Y}_j^\alpha, \mathbb{X}_{i-2}^\alpha, \mathbb{Y}_{j+2}^\alpha) + d_2^{\mathbb{X}^\alpha} + d_2^{\mathbb{Y}^\alpha}) \\ C_{i-3,j+2} + \sum_{\alpha=1}^N (\mathcal{I}(\mathbb{X}_i^\alpha, \mathbb{Y}_j^\alpha, \mathbb{X}_{i-3}^\alpha, \mathbb{Y}_{j+2}^\alpha) + d_3^{\mathbb{X}^\alpha} + d_2^{\mathbb{Y}^\alpha}) \\ C_{i-2,j+3} + \sum_{\alpha=1}^N (\mathcal{I}(\mathbb{X}_i^\alpha, \mathbb{Y}_j^\alpha, \mathbb{X}_{i-2}^\alpha, \mathbb{Y}_{j+3}^\alpha) + d_2^{\mathbb{X}^\alpha} + d_3^{\mathbb{Y}^\alpha}) \\ C_{i-3,j+3} + \sum_{\alpha=1}^N (\mathcal{I}(\mathbb{X}_i^\alpha, \mathbb{Y}_j^\alpha, \mathbb{X}_{i-3}^\alpha, \mathbb{Y}_{j+3}^\alpha) + d_3^{\mathbb{X}^\alpha} + d_3^{\mathbb{Y}^\alpha}) \\ I_{i-1,j+1} + \sum_{\alpha=1}^N (\mathcal{M}(\mathbb{X}_i^\alpha, \mathbb{Y}_j^\alpha, \mathbb{X}_{i-1}^\alpha, \mathbb{Y}_{j+1}^\alpha) + d_1^{\mathbb{X}^\alpha} + d_1^{\mathbb{Y}^\alpha}) \\ B_{i-1,j+1}^{\mathbb{X}} + \sum_{\alpha=1}^N (d_1^{\mathbb{X}^\alpha}) \\ B_{i-1,j+1}^{\mathbb{Y}} + \sum_{\alpha=1}^N (d_1^{\mathbb{Y}^\alpha}) \end{array} \right. \quad (5.22)$$

$$I_{i,j} = \min \left\{ \begin{array}{l} C_{i-1,j+1} + \sum_{\alpha=1}^N (\mathcal{M}(\mathbb{X}_{i-1}^\alpha, \mathbb{Y}_{j+1}^\alpha, \mathbb{X}_i^\alpha, \mathbb{Y}_j^\alpha) + g_{\text{open}}^I + 2g_{\text{ext}}^I + d_1^{\mathbb{X}^\alpha} + d_1^{\mathbb{Y}^\alpha}) \\ I_{i-1,j} + \sum_{\alpha=1}^N (g_{\text{ext}}^I + A + d^{\mathbb{X}^\alpha}) \\ I_{i-1,j+1} + \sum_{\alpha=1}^N (2g_{\text{ext}}^I + d_1^{\mathbb{X}^\alpha} + d_1^{\mathbb{Y}^\alpha}) \\ I_{i,j+1} + \sum_{\alpha=1}^N (g_{\text{ext}}^I + A + d_1^{\mathbb{Y}^\alpha}) \end{array} \right. \quad (5.23)$$

$$B_{i,j}^{\mathbb{X}} = \min \left\{ \begin{array}{l} C_{i-1,j} + \sum_{\alpha=1}^N (g_{\text{open}}^B + g_{\text{ext}}^B + d_1^{\mathbb{X}^\alpha}) \\ B_{i-1,j}^{\mathbb{X}} + \sum_{\alpha=1}^N (g_{\text{ext}}^B + d_1^{\mathbb{X}^\alpha}) \end{array} \right. \quad (5.24)$$

$$B_{i,j}^{\mathbb{Y}} = \min \left\{ \begin{array}{l} C_{i,j+1} + \sum_{\alpha=1}^N (g_{\text{open}}^B + g_{\text{ext}}^B + d_1^{\mathbb{Y}^\alpha}) \\ B_{i,j+1}^{\mathbb{Y}} + \sum_{\alpha=1}^N (g_{\text{ext}}^B + d_1^{\mathbb{Y}^\alpha}) \end{array} \right. \quad (5.25)$$

The interaction energy between two alignments must be combined with a score quantifying the sequence variation of the base-pairs between columns  $\mathbb{X}_i$  and  $\mathbb{Y}_j$ . This quantification should take into account both the allowed and inconsistent base-pairs, i.e. base-pairs that involved a nucleotide and a gap or unallowed base-pairs.

Let us define the abbreviation:

$$d_{ij}^{\alpha,\beta} = 2 - \delta(\mathbb{X}_i^\alpha, \mathbb{X}_i^\beta) - \delta(\mathbb{Y}_j^\alpha, \mathbb{Y}_j^\beta) \quad (5.26)$$

where  $\delta(a', a'') = 1$  if  $a' = a''$  and 0 otherwise.  $d_{ij}^{\alpha\beta} = 0$  if the nucleotides  $\mathbb{X}_i^\alpha$  and  $\mathbb{Y}_j^\alpha$  coincide with nucleotides  $\mathbb{X}_i^\beta$  and  $\mathbb{Y}_j^\beta$ , respectively,  $d_{ij}^{\alpha\beta} = 1$  if they differ in one position, and  $d_{ij}^{\alpha\beta} = 2$  if they differ in both positions. A straightforward covariation measure for conserved base-pairs can then be written as:

$$c_{i,j} = \frac{1}{\binom{N}{2}} \sum_{\alpha < \beta} d_{ij}^{\alpha,\beta} \Pi_{ij}^\alpha \Pi_{ij}^\beta \quad (5.27)$$

where  $\Pi_{ij}^\alpha = 1$  if nucleotide  $\mathbb{X}_i^\alpha$  and  $\mathbb{Y}_j^\alpha$  form an allowed base pairs, in our case Watson-Crick and wobble base-pairs, and 0 else.

In case of inconsistent base-pairs, a simple score would count the combinations of nucleotide and a gap as well as the unallowed base-pairs and ignore gap-gap and allowed base-pairs. This can be written as:

$$q_{i,j} = 1 - \frac{1}{N} \sum_{\alpha=1}^N \{ \Pi_{ij}^\alpha + \delta(\mathbb{X}_i^\alpha, gap) + \delta(\mathbb{Y}_j^\alpha, gap) \}. \quad (5.28)$$

Now both scores can be linearly combined into one scoring scheme for consistent and inconsistent base-pairs:

$$b_{i,j} = c_{i,j} - \phi_1 q_{i,j} \quad (5.29)$$

It should be noted that alignments with a large number of sequences, sequencing and alignment errors must be expected. Thus two columns  $\mathbb{X}_i$  and  $\mathbb{Y}_j$  should not be marked as unpaired if a single base-pair is inconsistent. Thus a threshold value  $b^*$  for the combined score  $b_{i,j}$  can be defined. This allows to set the pairing matrix  $\Pi_{ij}^{\mathbb{X},\mathbb{Y}}$  for the interaction of two alignments  $\mathbb{X}$  and  $\mathbb{Y}$  as

$$\Pi_{ij}^{\mathbb{X},\mathbb{Y}} = \begin{cases} 0 & \text{if } b_{ij} < b^* \\ 1 & \text{if } b_{ij} \geq b^* \end{cases} \quad (5.30)$$

Parameter		Default
Threshold for pairing	$b^*$	-1.00
Relative weight of inconsistent sequences	$\phi_1$	1.00
Weight of sequence covariation	$\phi_2$	1.00 kcal/mol

Table 5.1: Additional “energy” parameters for alignment folding

The full energy model used in **RNAplex** is obtained as a linear combination of the average pairing energy and the combined covariation score

$$C'_{i,j} = \frac{1}{N}C_{i,j} - \phi_2 b_{i,j} \quad (5.31)$$

In addition to the standard energy model for RNA folding we only need three additional variables: the threshold value  $b^*$  and the two scaling factors  $\phi_1$  and  $\phi_2$ . In **RNAplex** we use their default values as listed in Table 5.1. In addition, non-standard base pairs can occur in the alignment folding for which no measured energy parameters are available. We substitute the default stacking energy of 0.0kcal/mol in this case.

The evolutionary model used in **RNAplex**, while straightforward, performs well in predicting consensus secondary structure. Its simplicity allows it to be integrated into **RNAplex** without runtime overhead.

A potential weakness is the **RNAalifold** scoring model, which was trained and optimized for intramolecular interaction, instead for the intermolecular interactions to which it is applied here. More complex scoring schemes such as the one used in **PETfold** and **PETcofold**, where a maximum expected scoring approach combines the evolutionary probabilities of a consensus structure given an alignment with the thermodynamic probabilities of the associated structures in each sequence [219, 220], perform slightly better than the **RNAalifold** scoring scheme. However, they can be incorporated only at the cost of a greatly increased runtime, and thus are incompatible with the purpose of **RNAplex**.

## 5.1.5 Model Errors

### Simplified Energy Model

**RNAplex** uses an approximation both for the computation of the interaction energy as well as for the determination of the opening energy. The accuracy of the interaction and opening energy models can be tested separately. To test whether the simplified interaction model affects the sensitivity of **RNAplex**, we assessed how well **RNAplex**, **RNAhybrid** and **RNA duplex** recovered experimentally confirmed miRNA–mRNA interactions.

A set of 27 interactions taken from TarBase [222] involving 25 mRNAs and 22 miRNAs was used. For each of the reported interactions, the hybridisation energy of the reported target site with its cognate miRNA was computed with **RNAplex**, **RNA duplex** and **RNAhybrid**. Moreover for each miRNA-mRNA pairs, the 10 best binding sites were identified using **RNAplex**, **RNA duplex** and **RNAhybrid**. For **RNAhybrid** we constrained the hybridisation to target sites which were fully complementary to the miRNA seed region, since this gave the highest sensitivity in the test. The experimentally confirmed binding site was then reported as recovered if it overlapped with any of the 10 best hits (see Table 5.2). All three programs performed similarly well with **RNA duplex** retrieving 22 out of 27 interactions, while **RNAplex** and **RNAhybrid** each recovered 20 interactions.

### Opening Energy

In order to investigate the accuracy of the accessibility profiles, we used a set of 11460 randomly generated sequences of length 400nts for which the accessibility profiles was computed with **RNAup**. For each sequence, we then determined the difference of the **RNAup** opening energy and the **RNAplex** opening energy for the region located between nucleotides 181 and 200. Figure 5.4 shows the relative energy differences between both models as bar plots for different values of  $\delta$ . The largest variations are seen for  $\delta = 1$  with differences larger than 100%.  $R^2$  (triangle) and the Pearson correlation coefficient (square) reach their minimum there (0.09 and 0.37, respectively). Both coefficients then steadily improve with  $\delta$  and reach their theoretical maximum of 1 for  $\delta = w$ . For  $\delta < w$ , our approximation slightly overestimates the opening energy.

mRNA	miRNA	$\Delta G_{RNA duplex}^*$	$\Delta G_{RNAplex}^*$	$\Delta G_{RNAhybrid}^*$	mRNA	miRNA	$\Delta G_{RNA duplex}^*$	$\Delta G_{RNAplex}^*$	$\Delta G_{RNA hybrid}^*$
AGTR1	miR-155	<b>-11.50(NF)</b>	<b>-11.50(NF)</b>	<b>-17.2(NF)</b>	HOXA1	miR-10a	-15.93(3)	-15.93(5)	-22.7(1)
BCL2	miR-16	-18.90(2)	-18.90(1)	-24.1(1)	KIT	miR-221	-17.70(3)	<b>-17.70(NF)</b>	-23.4(2)
CAT-1	miR-122	-23.80(1)	-23.80(1)	-29.0(1)	KIT	miR-222	<b>-14.70(NF)</b>	<b>-14.70(NF)</b>	-19.8(3)
CGI-38	miR-16	-20.80(2)	-20.80(2)	<b>-26.0(NF)</b>	KRAS	let-7a	<b>-14.10(NF)</b>	-14.10(7)	-18.9(6)
Clock	miR-141	-16.40(1)	-16.40(1)	-22.1(1)	Lin28	let-7b	-27.40(1)	-27.40(1)	-33.5(1)
CXCL12	miR-23a	<b>-8.90 (NF)</b>	<b>-8.90 (NF)</b>	-14.0(5)	MAPK14	miR-24	-27.10(1)	-27.10(1)	-32.2(1)
CYP1B1	miR-27b	-28.20(1)	-28.20(1)	-33.6(1)	MYCN	miR-101	<b>-13.80(NF)</b>	<b>-13.80(NF)</b>	-20.7(1)
E2F3	miR-34a	-19.10(2)	<b>-19.10(NF)</b>	-25.1(1)	NRAS	let-7a	-16.10(2)	-16.10(3)	<b>-21.1(NF)</b>
Enx-1	miR-101	-16.90(1)	-16.90(1)	<b>-22.4(NF)</b>	PTEN	miR-19a	-17.70(1)	-17.70(1)	-23.2(1)
FLJ2130	miR-145	-21.80(1)	-21.80(1)	<b>-27.4(NF)</b>	R1CS	miR-132	-18.80(1)	-18.80(1)	-25.1(1)
Fstl1	miR-206	-18.40(6)	<b>-18.40(NF)</b>	-23.2(2)	SMC1L1	let-7e	-22.20(1)	-22.20(1)	-27.5(1)
GJA1	miR-1	-14.30(3)	-14.30(4)	-20.6(2)	TMSB4X	miR-1	-16.90(1)	-16.90(1)	
GJA1	miR-206	-14.53(10)	-14.53(10)	-20.5(4)	TPM1	miR-21	-15.60(1)	-15.60(1)	<b>-21.9(NF)</b>
Hand2	miR-1	-12.20(1)	-12.20(1)	-18.1(2)					<b>-19.6(NF)</b>

Table 5.2: Binding site summary for 27 functional miRNA–mRNA interactions in Human, taken from TarBase [222]. Columns 1 and 2 contain the name of the mRNA and miRNA, respectively. The column 3 to 5 contain the interaction energy for the reported miRNA mRNA interactions as computed by **RNAplex**, **RNAplex** and **RNAhybrid**, respectively. The number in parenthesis represent the rank of the experimental target site where 1 stands for the most stable interaction and 10 for the 10th best interaction. NF means that the reported target site was not found among the 10 best interaction sites and are shown in red.

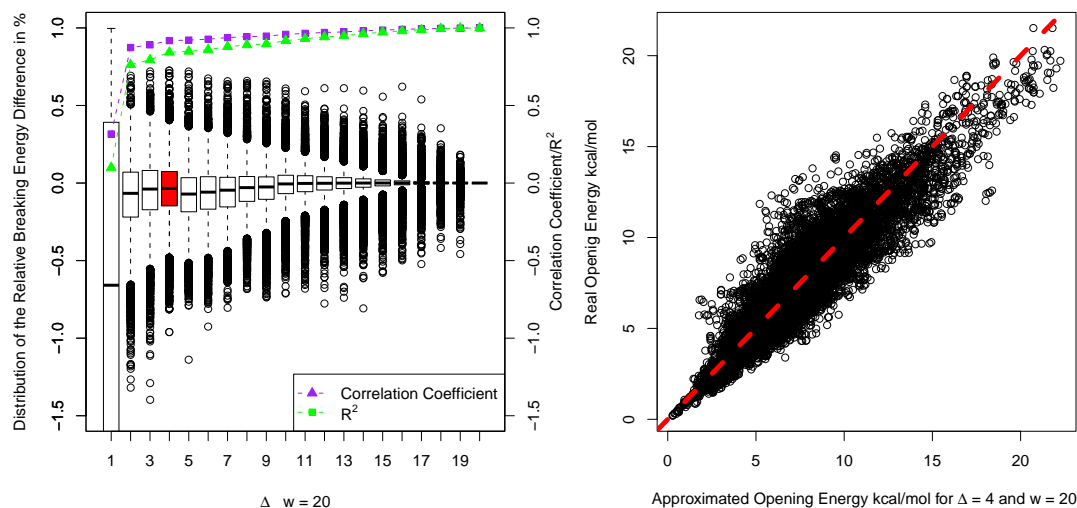


Figure 5.4: Error representation of our accessibility model. **l.h.s** Boxplot representation of the distribution of the relative breaking energy between our approximated model and the standard energy model for different  $\Delta$  size and a fixed target size of 20 nts. The larger  $\Delta$  the smaller the error in our approximation. **RNAplex** uses  $\Delta = 4$ . At this level of approximation, the pearson correlation coefficient between the approximated model and the real model reaches 0.92

This can be seen for  $\delta = 4$ , the value used in **RNAplex** in the scatterplot in the middle of Figure 5.4. Half of the relative deviation are contained between +7% and -14%.

## Whole Approximation

The influence of the different approximations made in **RNAplex** on the quality of the predictions were evaluated by surveying how well the boundaries of known duplexes was recovered. The knowledge of the exact localization of RNA-RNA interactions is important, because ncRNAs may regulate their targets in different ways depending on the location of the binding sites.

The accuracy of the energy model (interaction and opening energy) used in **RNAplex** was compared to that of **RNAup**, **biRNA** [38], and the old version of **RNAplex** (**RNAplex-c**) on a dataset of 17 known bacterial small RNA-mRNA interactions [38] (see

Table 5.3). In this dataset both the opening energy of the interacting sequences and the hybridization energy affects the prediction.

**RNAplex -c** (old version) missed four interactions, while all **RNAplex -a** (with accessibility information) predictions overlapped with the corresponding experimentally determined interactions, as did the predictions of **RNAup** and **biRNA**. These results emphasize the importance of accessibility for the correct prediction of RNA-RNA interactions. Furthermore, it confirms that the approximations used in **RNAplex** are sufficient to reach a level of accuracy similar to that of **RNAup** and **biRNA** (see Table 5.3).

The location of the predicted closing pairs was compared to the confirmed locations. For each prediction tool, the average over all 17 interactions of the sum of the magnitude of the deviation between the predicted and confirmed locations of the four closing nucleotides was computed (see Table 5.3). All three accessibility based methods performed similarly with an average deviation of 16.76 for **RNAup**, 19.88 for **biRNA** and 20.60 for **RNAplex -a**, much smaller than the average deviation of **RNAplex -c** (59.76 nts).

It should be noted that **RNAup** and **RNAplex**, in contrast to **biRNA**, cannot handle interactions involving two or more interacting regions, such as the two kissing-hairpin complexes found in *OxyS-fhlA*. Still, in contrast to **RNAup**, **RNAplex** can return sub-optimal predictions, without runtime overhead, that can be used to identify disjoint interaction regions. For *OxyS-fhlA*, the confirmed binding regions are located at positions [22, 30] and [98, 104] on *OxyS* and [87, 95] and [39, 45] on *fhlA*, in accord with the two best suboptimals returned by **RNAplex** which are located on [23, 28] and [96, 100] on *OxyS* and [87, 92] and [41, 45] on *fhlA*.

### 5.1.6 Computational efficiency

When comparing search speed to **RNAhybrid**, we found that the speedup varied with sequence length and program options, but was at least 10. **RNAhybrid** performed best for miRNA target search when limiting the search to targets with a perfect seed matches, i.e. Watson-Crick pairs only at microRNA positions 2 to 7. Without this constraint **RNAplex** speed-up increased from 10 to 20-fold. Furthermore the speedup increased slightly for longer query sequences, reaching 27 for query sequences of

mRNA	sRNA	Position lit.		Position biRNA		Position RNAup		Position RNAplex -c		Position RNAplex -a	
gltI	GcvB	66,77	31,44	64,81	26,44	65,76	32,43	<b>137,154</b>	<b>34,53</b>	65,76	32,43
argT	GcvB	75,91	89,104	71,90	90,108	70,91	89,109	69,88	91,110	72,90	90,107
dppA	GcvB	65,90	133,150	62,81	135,153	57,81	135,157	56,74	141,158	57,76	138,157
livJ	GcvB	63,87	59,82	66,84	54,73	63,87	59,82	64,83	62,81	63,81	65,82
livK	GcvB	68,77	165,177	67,86	156,175	65,88	155,177	<b>117,126</b>	<b>240,249</b>	65,82	161,177
oppA	GcvB	65,90	155,179	67,86	158,176	65,89	155,178	64,76	167,179	65,75	168,178
STM4351	GcvB	70,79	44,52	69,77	44,52	62,86	34,58	<b>35,53</b>	<b>106,124</b>	62,77	44,58
lamB	MicA	8,36	122,148	8,26	131,148	8,32	126,148	7,26	130,149	8,22	132,148
ompA	MicA	8,24	113,128	8,24	113,128	8,24	113,128	7,25	112,129	8,24	113,128
rpoS	DsrA	8,36	10,38	21,40	7,25	21,40	7,25	9,28	18,37	12,30	16,34
rpoS	RprA	33,62	16,39	40,51	22,32	40,62	16,32	32,46	26,40	40,51	22,32
tisA	IstR	65,87	57,79	66,85	59,78	65,87	57,79	64,83	60,80	65,83	61,79
ompC	MicC	1,30	93,139	1,16	104,119	1,16	104,119	<b>40,59</b>	<b>77,93</b>	1,16	104,119
ompF	MicF	1,33	100,125	14,30	99,118	5,28	105,122	1,14	113,126	1,13	114,125
sdhD	RyhB	9,50	89,128	22,41	98,116	22,41	98,116	8,26	112,129	22,40	99,116
sodB	RyhB	38,46	52,60	38,46	48,64	38,49	50,60	37,50	49,61	38,49	50,60
ptsG	SgrS	157,187	76,107	174,187	76,89	168,187	76,95	167,186	76,96	168,187	76,95
average deviation				19.88		16.76		59.76		20.60	

Table 5.3: Binding site summary for a set of 17 functional interactions from [38]. The first and second columns contain mRNAs- and sRNAs-ID, respectively. We compared biRNA, RNAup and to RNAplex. biRNA and RNAup were run using the default parameters, while RNAplex was run with either an extension penalty of 0.3 [kcal/mol] (RNAplex -c) or the accessibility files produced by RNAup (RNAplex -a). All predictions made by RNAup, biRNA and RNAplex -a overlapped with the experimentally reported interactions, while RNAplex -c missed four interactions. The last row reports the average deviation between the experimentally found locations and the predicted ones

length 320. In the tests above, we searched only for the single most stable interaction site. While **RNAplex** can return suboptimal interaction sites without a speed penalty, **RNAhybrid** needs to repeat the whole dynamic programming procedure for each desired suboptimal, making it accordingly expensive.

Compared to **IntaRNA** and **RNAup**, **RNAplex** has also a much lower time-cost. This is exemplified in Figure 5.5. Here we compared how fast RNA-RNA interactions tools were able to search for targets for 19 bacterial sRNA in a set of 100 sequences of length 1200nts. **RNAplex** achieved this task in 35.7 seconds, **RNAup** in 86487[s] and **IntaRNA** in 34150[s]. Note that all tools were compiled with the same optimization and were run on the same machine.

We further compared the runtime and the memory consumption of **RNAup** and **IntaRNA** against that of the new **RNAplex**, by generating a set of random target sequences of size 400, 800, 1600, 3200 and 6400 nts and query sequences of size 100, 200, 400 and 800 nts and searching for targets with all three tools. On this dataset the new **RNAplex** is between 575 and 1600 times faster than **IntaRNA** and between 1500 and 65400 times faster than **RNAup**. The memory consumption is also drastically reduced. **RNAplex** needs at least 17 and at most 1330 times less memory than **IntaRNA**, and 15 to 626 times less memory than **RNAup** (see Table 5.4). Compared to the old version without accessibilities, the new **RNAplex** needs only four times more memory.

Furthermore we also benchmarked the runtime of the alignment version of **RNAplex** with accessibility on a dataset containing 9 sRNAs (*dsrA*, *gcvB*, *micA*, *micC*, *micF*, *rprA*, *ryhB*, *sgrS*, *spot42*) and 100 mRNAs multiple sequence alignments of the homologs of the 100 genes from *e.coli* K12 used to benchmark the single sequence version of **RNAplex** (see the Dataset section).

In Figure 5.6, we show the runtime of **RNAplex** with alignment and accessibility for those 900 interactions against the number of sequences in the alignments in a log-log plot. Compared to the single sequence version, **RNAplex** with alignment is twice slower. The runtime dependency on the number of sequences is proportional to  $\sqrt{N}$ , where  $N$  is the number of sequences in the alignments.

length target	length query	Speedup vs IntaRNA	Memory vs IntaRNA	Speedup vs RNAup	Memory vs RNAup
400	100	NA	1.74e+01	NA	1.57e+01
400	200	NA	2.70e+01	NA	2.69e+01
400	400	5.74e+02	4.60e+01	1.49e+03	4.87e+01
400	800	1.01e+03	8.91e+01	1.67e+03	5.54e+01
800	100	NA	3.13e+01	NA	2.37e+01
800	200	5.93e+02	4.98e+01	2.10e+03	3.45e+01
800	400	6.26e+02	8.65e+01	1.75e+03	5.60e+01
800	800	6.75e+02	1.63e+02	1.32e+03	9.82e+01
1600	100	8.56e+02	6.10e+01	6.30e+03	5.35e+01
1600	200	7.28e+02	9.71e+01	3.86e+03	6.39e+01
1600	400	5.92e+02	1.70e+02	2.09e+03	8.54e+01
1600	800	6.87e+02	3.16e+02	1.61e+03	1.27e+02
3200	100	1.20e+03	1.42e+02	1.98e+04	1.90e+02
3200	200	8.45e+02	2.13e+02	8.48e+03	1.88e+02
3200	400	8.14e+02	3.55e+02	4.75e+03	1.92e+02
3200	800	8.91e+02	6.33e+02	2.91e+03	2.30e+02
6400	100	1.58e+03	3.82e+02	6.54e+04	6.72e+02
6400	200	1.42e+03	5.20e+02	3.35e+04	6.59e+02
6400	400	1.35e+03	7.95e+02	1.71e+04	6.41e+02
6400	800	1.44e+03	1.33e+03	8.93e+03	6.26e+02

Table 5.4: Speedup and memory improvement of the accessibility based **RNAplex** against **IntaRNA** and **RNAup** for different random query and target sequences as measured by the *time* application. The first two columns show the target and query length, respectively. The third and fifth column show the runtime improvement of **RNAplex** against **IntaRNA** and **RNAup**, respectively. The difference between **RNAplex** and the two other tools slightly grow with increasing target length, but diminish with increasing query length. On this dataset, **RNAplex** is between 600 and 1600 times faster than **IntaRNA** and from 1500 up to 65400 times faster than **RNAup**. Note that for very short sequences (less than 400 nts), **RNAplex** needs less than 1/100th second to compute the hybrid, too fast to be precisely measured by time, hence the NA for the first entries. The memory consumption of **RNAplex** is 17.4 to 1330 times smaller than that of **IntaRNA** and 15 to 626 times smaller than that of **RNAup**. Note that larger sequences could not be used because the memory need of both **RNAup** and **IntaRNA** exceeded the available RAM (4G)

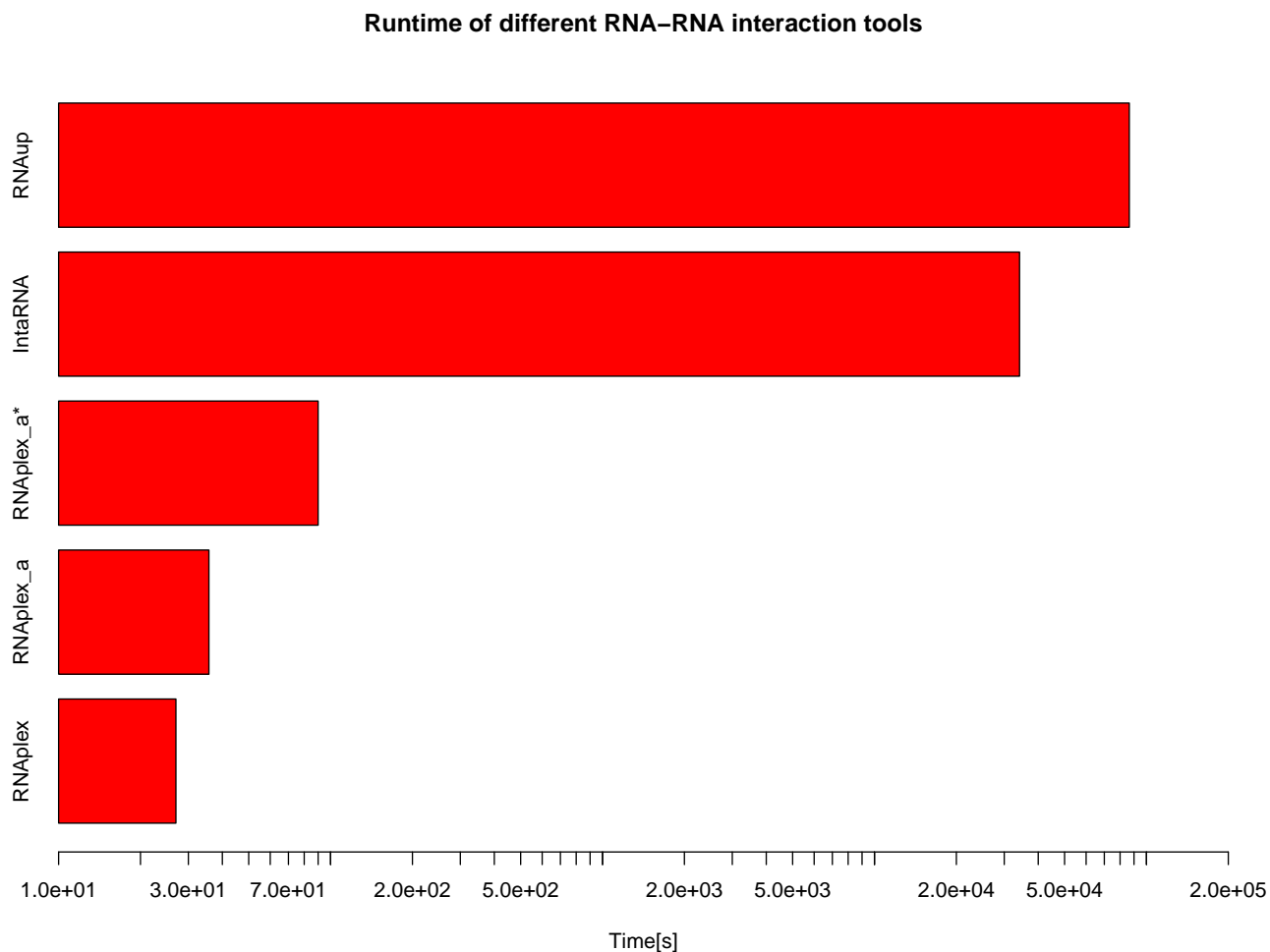


Figure 5.5: Bar plots representing the time necessary to complete the target search for 19 bacterial sRNAs in 100 random sequences of length 1200 nts for different RNA-RNA interaction tools. **RNAPlex -c** is the fastest application with a completion time of 27[s]. **RNAPlex -a** needs 36[s] to achieve the same task. This grows to 90[s] if one considers the time necessary to compute the accessibility profile. **RNAPlex -a** is 1000 times faster than **IntaRNA** and 2422 times faster than **RNAup**.

### Prefiltering based on Google

We tested whether the computation time of **RNAPlex** was reduced further by identifying stretches of complementarity before attempting the more time consuming

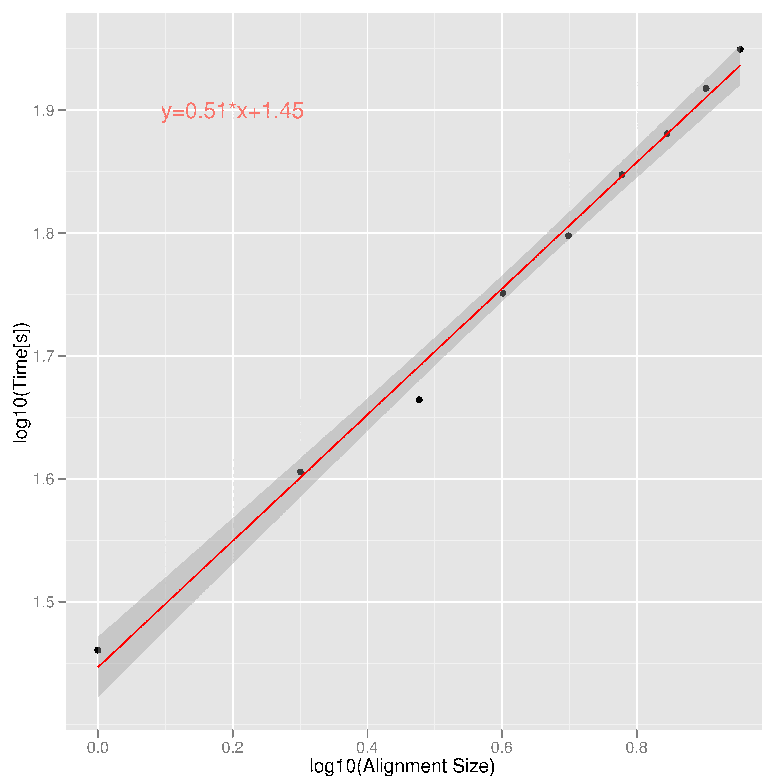


Figure 5.6: Runtime of **RNAplex** with alignment and accessibility against the number of sequences in alignments for a set of 9 query and 100 target sequences. The runtime of **RNAplex** increases proportionally to  $\sqrt{N}$ , where  $N$  is the number of sequences in the alignments.

dynamic programming procedure. **GUUGle**, which locates potential helical regions under RNA base pairing rules with the help of suffix arrays to find these highly complementary regions, was used as a prefilter mechanism. The trade-off between speed and sensitivity is controlled by the *ktup* parameter, which specifies the size of complementarity to search for (word size). We compared the CPU time and sensitivity of **RNAplex** and **GUUGle+RNAplex** when searching for experimentally verified miRNA targets. Up to a word size of 7 **RNAplex** is faster than **GUUGle+RNAplex**, while the sensitivities of both programs are the same. For larger word size, **GUUGle+RNAplex** performs better than **RNAplex** however at the cost of a reduced sensitivity. As such **RNAplex+GUUGle** may prove to be useful for searching of gapped interactions with

complementary regions longer than 7 nts.

## 5.2 Results

---

### 5.2.1 miRNA targets prediction

As a first application example, target sites of mouse miR-134, an miRNA involved in regulating dendritic development and in the differentiation of mouse embryonic stem cells [119,255], were searched with **RNAplex** in all mouse 3'UTRs. These results were compared with the target predicted by a two steps approach where first **RNAplex** selects putative targets that are then further filtered with **RNAup**. The specificity of both methods was assessed by recording the number of sequences that had a better interaction energy than the experimentally confirmed miR-134/Limk1 hybrid [215]. For each 3'UTRs sequences the minimal free energy of interaction (MFE) was computed. All sequences that had an MFE smaller than -15 kcal/mol were stored for subsequent inspection with **RNAup** (7503 sequences). Instead of using the whole 3'UTR sequence into **RNAup**, a 200 nts regions centered around the binding sites reported by **RNAplex** was selected. Then each reported interactions was ranked based either on its **RNAup** or **RNAplex** interaction energy.

In case of the two steps method, where first putative targets are rapidly identified with **RNAplex** and further inspected with **RNAup**, Limk1 had a **RNAup** binding energy of -19.97 kcal/mol and was ranked among the 74 best targets (0.9%). In contrast, the same interaction was ranked 1057 when looking at the **RNAplex** energy (14.10%) (see Figure 5.7). Similarly using **RNAhybrid** instead of **RNAplex** would have resulted in 1445 hits. Those results are inline with the one presented in chapter 4, where it was shown that highly structured regions cannot be targeted by miRNAs.

The 73 target mRNAs scoring higher than Limk1 are likely to contain additional true targets. 8 of the 73 targets were actually contained in a recent study of [175]. For all of them miR-134 reduced the respective protein concentration by at least 45%.

### 5.2.2 sRNAs targets prediction

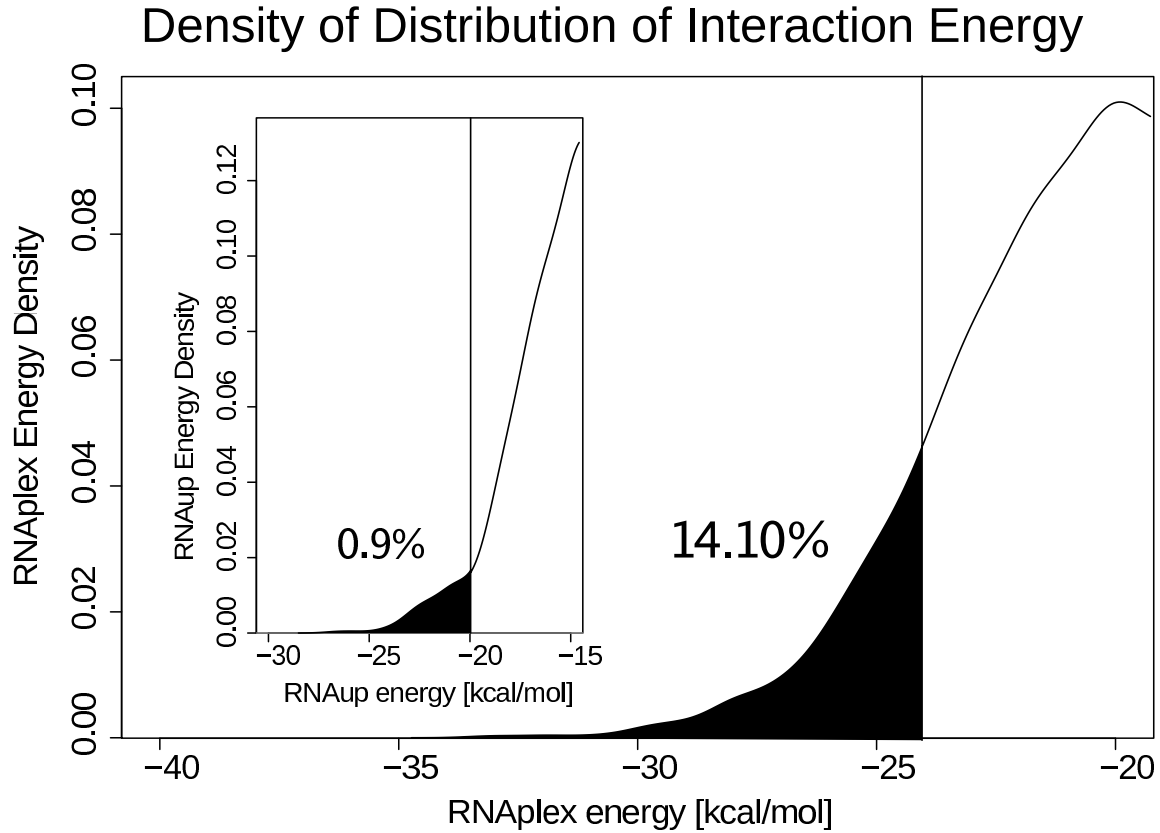


Figure 5.7: Density distribution of interaction energy as computed by **RNAplex** for miR-134 against all mouse 3'UTRs. The vertical line represents the energy of the experimentally confirmed miR-134/Limk1 interaction as computed by **RNAplex**. The black area represents the proportion of 3'UTRs having a higher energy of interaction than the experimentally predicted one. The number in parenthesis represents the percentage of the entire distribution falling below the threshold defined binding energy, as computed by **RNAplex**, of the experimentally verified interaction. The inset shows the density of distribution of interaction energy as computed by **RNAup** for miR-134 against the mouse 3'UTRs. The vertical line represents the energy of binding as computed by **RNAup** for the experimentally confirmed miR-134/Limk1 interaction. The number in parenthesis represents the percentage of the entire distribution falling below the threshold defined by the binding energy, as computed by **RNAup**, of the experimentally verified interaction.

As an application example, we consider the genome-wide prediction of sRNA targets in *e.coli*. As a reference set, we use the experimentally confirmed interactions published by [250]. We expect that, for a given sRNA, the number of predicted

interactions with other (false positive) targets should decrease when accessibility of the target mRNA is included. Ideally, it should reach the low levels observed for **RNAup** [180].

We used the following methodology: for a given sRNA and the corresponding confirmed sRNA-mRNA interactions, we looked genome-wide at how many mRNAs binds with a lower energy to the sRNA than the reported sRNA-mRNA interaction, i.e. the number of false positive. For each 4463 *e.coli* genes, a mRNA of length 1200 nts, including 200nts upstream and 1000nts downstream of the start codon were defined. Accessibility profiles were computed with **RNAplfold**, with a folding windows (option **-W**) of 240 nt and a maximal base-pair distance of 160 (option **-L**). An interaction was reported if the corresponding sRNA-mRNA interaction energy is smaller than the experimentally confirmed interaction, and if it occurs in region encompassing 80 nts, 50 nts upstream and 30 nts downstream of the start codon.

With these settings **RNAplex -c** was able to precisely locate 7 out of 9 interactions, with a maximal difference of 30 nts (see Table 5.5). In two cases **RNAplex -c** failed to predict the correct target sites. **RNAhybrid** maximized the length of hybridization, leading to substantially longer target sites. In 6 out of 9 cases, experimental and predicted target sites overlapped. However the size of the predicted interactions did not allow a clear localization of the proper target boundaries.

The inclusion of the accessibility profiles in the new version of **RNAplex** leads to a substantial improvement as can be seen from Table 5.6. All native interaction sites are among the predictions, and the detailed target site localization is improved. Most importantly, the number of predictions with better interaction energies, i.e., the false positives, is reduced to a level similar to that of **RNAup**.

The average number of better binding interactions for **RNAhybrid** was 997, ten times higher than for **RNAplex -c**, showing that even a constant extension penalty is better than no accessibility correction. **RNAup**, with an average of 17 better binding interactions performed significantly better than **RNAplex -c**, however at the cost of a much higher runtime. This problem was however solved with **RNAplex -a** which performed on par with **RNAup** but with a runtime similar to that of **RNAplex -c**.

In order to better assess the number of false positives, the same method was applied on the dinucleotide shuffled sRNAs and mRNAs. To this end, we compared the

interaction energy of the non-shuffled, experimentally confirmed interactions, to the energy distribution of the shuffled sequences. Interestingly, in 7 out of 9 cases, the number of false positives is smaller (see Table 5.7) in the shuffled case than in the non-shuffled one. This can be explained by the fact that in various bacteria, the region around the ribosomal entry site, which is also the preferred region of sRNA binding, is more accessible than the rest of the mRNA (see Figure 3.4). This in turn implies that compared to shuffled sequences, sRNAs have a greater chance to bind to the region around the start codon in non-shuffled mRNAs. Depending on the ncRNAs, one can expect between  $7.5 \times 10^{-7}$  false positives per nucleotide for *micC* and  $1.5 \times 10^{-4}$  false positives for *gcvB* (see Table 5.7).

It should be noted that the *RhyB-sodB* interactions were badly predicted by most of the interaction tools. The main reason is that the *Hfq*-protein promotes this interaction. The mechanism how this happens is currently not well understood. *Hfq* could work as a chaperone and unfold sRNAs, facilitating the interactions [179]. An other explanation could be that through its RNA-binding ability, *Hfq* could artificially increase the local concentration of sRNAs, leading to an increased rate of reaction [30].

### 5.2.3 Multiple alignment

While *RNAplex* recovers all interactions, some of them like *RyhB-sodB* or *GcvB-oppA* are ranked lowly. A comparative version of *RNAplex* was designed (see *methods*) to reduce the number of false positives. Similar to consensus RNA folding, the quality of the input alignments is crucial to obtain meaningful results [19].

The comparison of the performance of the single sequence with the comparative version of *RNAplex* was achieved by generating multiple sequences alignments *clustalw* [135] for the 8 sRNAs from Table 5.6 and with *MUSCLE* [59] for the 4463 *e.coli* mRNAs. The bacteria genome used were:

- *Yersinia pestis*
- *Yersinia pestis* 92
- *Yersinia pseudotuberculosis* IP 31758

- *Sodalis glossinidus*
- *Salmonella enterica* serovar Typhi Ty2
- *Salmonella typhimurium* LT2
- *Salmonella enterica* Paratyphi
- *Shigella flexneri* 2a
- *Enterobacter* sp. 638
- *Escherichia coli* K12-MG1655
- *Escherichia coli* O157:H7 EDL933
- *Escherichia coli* 536
- *Escherichia coli* APEC O1
- *Photobacterium luminescens* TTO1
- *Pectobacterium atrosepticum*

In many cases **MUSCLE** and **clustalw** were not able to satisfactorily align the sequences. This was caused e.g. by misannotations of the start codon as for the **ompA** gene in *Escherichia coli* APEC O1, which was incorrectly annotated 70 nts upstream of the true start codon. In order to better handle these cases, we devised a method to produce multiple alignments of highly similar and strongly binding target sites.

Given a reference gene in *eColi\_K12*, the corresponding sequences in the 14 remaining species are retrieved and clustered based on their sequence similarity. We let **RNAplex** run on each of these sequences separately and store for each genes the  $n$  (in our case 3) best targets (see Figure 5.8a and 5.8b). This gives a total of  $3 \cdot 15 = 45$  high scoring sequences.

Once all high-scoring target-sequences are found, we need to know which set of 15 sequences represents the set of most conserved, best pairing target sequences. To

this aim we developed a dynamic programming approach which minimizes a scoring function based on sequence conservation and hybridization energy. Let us set the number  $m$  of sequences in the alignment to 15 and for each target sequences the number of best targets  $n$  to 3. Further let us define by  $\sigma_{i,j}$  the  $j^{th}$  best target sequences in the  $i^{th}$  species ( $1 \leq j \leq 3, 1 \leq i \leq 15$ ). Further let us define the interaction energy between the  $i^{th}$  sRNA sequence and the  $j^{th}$  best sequence in the  $i^{th}$  species,  $\sigma_{i,j}$ , as  $\Delta\Delta G_{i,j}$ . Next we define with  $\mathcal{P}(\sigma_{i,j}; \sigma_{i-1,k})$  the sequence identity between  $\sigma_{i,j}$  and  $\sigma_{i-1,k}$ . The sequence identity is computed with the myers bit-vector algorithm [184]. Next we define  $\Gamma_{i,j}$  as the best set of target sequences from species 1 till species  $i$  and containing the target sequence  $\sigma_{i,j}$ . We define the recursion relation as:

$$\Gamma_{i,j} = \min_{1 \leq k \leq n} \Gamma_{i-1,k} + \mathcal{P}(\sigma_{i,j}; \sigma_{i-1,k}) \cdot \Delta\Delta G_{i,j} \quad (5.32)$$

(see Figure 5.8c). In our approach  $\Gamma_{1,j}$  is set to  $\Delta\Delta G_{1,j}$ . The score of the best set of target sequences is defined as  $\Gamma_{min}$  is equal to  $\min_{1 \leq i \leq n} \min_{1 \leq j \leq m} \Gamma_{i,j}$ .

Starting the backtracking procedure from the target sequences  $\sigma_{i,j}$  for which  $\Gamma_{i,j} = \Gamma_{min}$ , we can retrieve the ensemble of sequences that shows high binding affinity to the sRNAs sequences and that are well conserved (see Figure 5.8d and e).

Because highly conserved interactions are more credible than non-conserved interactions, ranking of interactions based on multiple sequences alignments should not only take the interaction energy into account, but also the number of organisms (in which a predicted interactions is detectable). This can be achieved by using  $Z$ -scores as alternative ranking criterion. The  $Z$ -scores can be computed for all interactions having the same number of sequences in the alignments. This is important as highly conserved interactions tend to have a higher consensus interaction energy than interactions that are conserved in only few organisms (see Figure 5.9).

In this way, extremely stable interactions can be compared without having to worry about the number of sequences in the alignments. The main drawback of this method is that highly conserved interactions with more than 10 sequences are rare, making the  $Z$ -score analysis unreliable. This is the case for example for the *micA-ompA* pair, which has the highest interaction energy among the interactions involving 14 species. In this case the rank of MicA drops from 2 for the single sequence approach

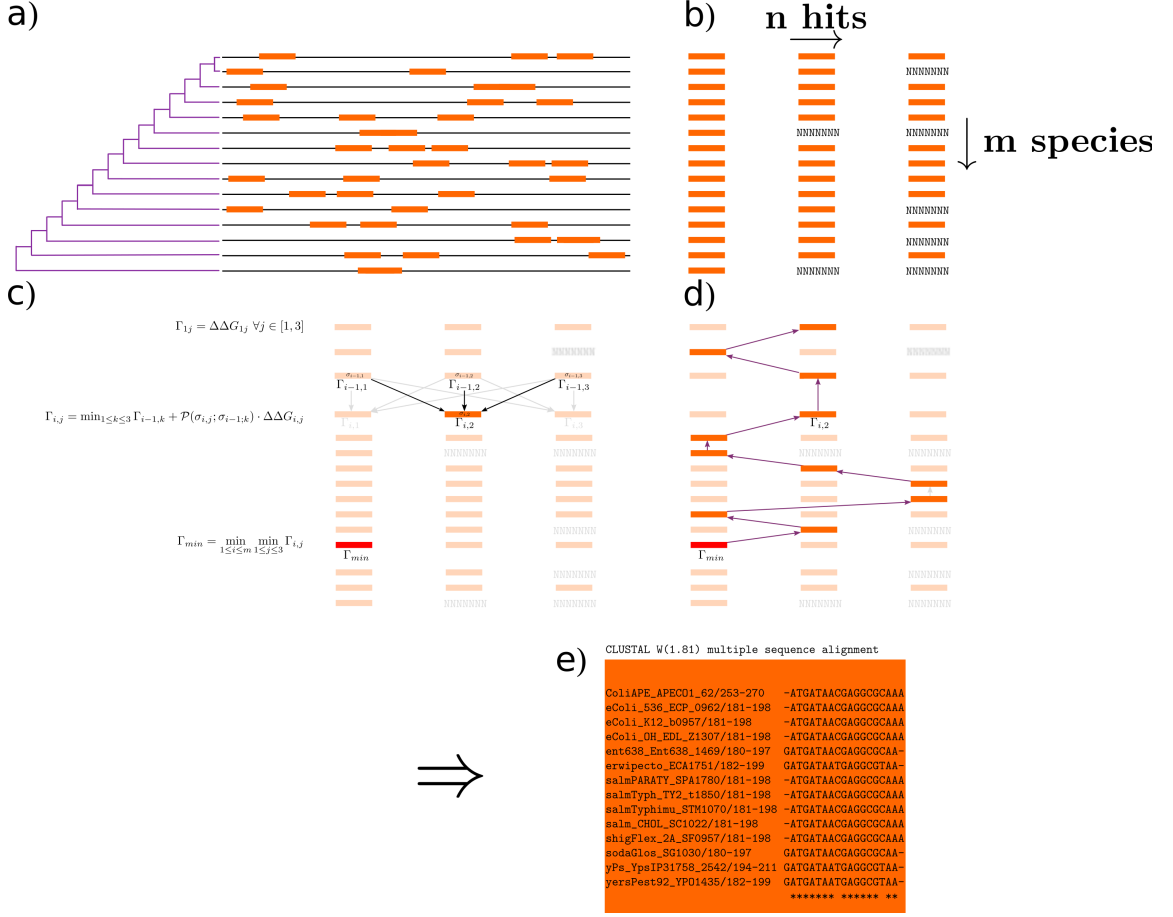


Figure 5.8: Procedure used to select high-binding, highly similar target binding site in multiple sequence alignments. a) Sequences are sorted based on their sequence similarities with `clustalw`. b) RNAplex is ran on each sequences in order to select the  $n$ -best hits for each sequences. In this study  $n = 3$  was used. c) A recursive approach based on the sequence similarities and the strength of interaction of target sites is used to find the best set of target sites among the  $m$ -species. d) Starting from the target site with the minimum score, the best set of target sites is retrieved through backtracking. e) The set of target sites is realigned. It is used to compute the multiple-alignment interaction between the sRNAs and the selected target sites. Accessibility information are retrieved thanks to the coordinates found in the multiple alignments, e.g. 253-270 for gene ColiAPE\_APEC01\_62 and 181-198 for eColi\_K12\_b0957.

to 11 for the alignment approach.

Table 5.6 shows that the rank based on the interaction energy or the  $Z$ -score is

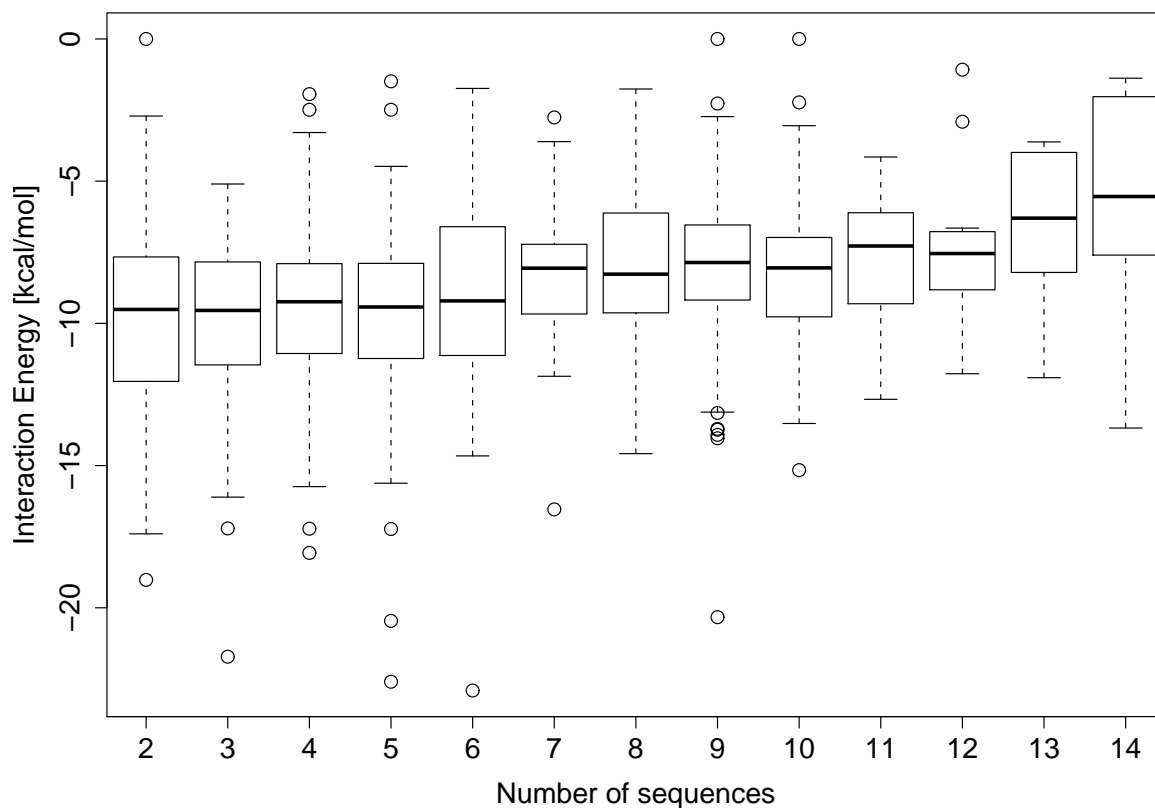


Figure 5.9: Boxplots showing the interaction energy distribution as a function of the number of sequences in the alignments for sRNA *GcvB*. Well conserved interactions have in average a higher interaction energy than interactions involving less sequences.

similar to that of the single sequence energy ranking. However, when considering only interactions having a greater or equal number of sequences and a higher  $Z$ -score, the number of interactions that score better than the native one decreases significantly, with the greatest reduction being seen for *ryhB*.

Similar to the single sequence case, the use of accessibility information in the case of multiple sequences alignments allows to improve the rank of the known interactions. This can be seen in the last column of Table 5.6.

It should be noted that some false positives turned out to be real interactions: For example, *iscS* and *acnB* score better than *sodB* as targets for *ryhB* and are true targets [42, 164].

Similar trends can be seen if the  $Z$ -score threshold is set to 0 and the number of sequences in the multiple alignment remains unchanged. If we look at the gene ontology of these targets in the case of *ryhB* (43 targets), we see that 35 are involved in catalytic activities ( $p = 0.006$ ), 9 are involved in iron-sulfur cluster binding ( $p = 0.007$ ), 39 are involved in binding ( $p = 0.01$ ). *ryhB* targets are also significantly overrepresented in the CO<sub>2</sub> fixation ( $p = 0.0001$ ) as well as citrate cycle cellular pathways ( $p = 0.0002$ ), in line with the gene ontology analysis.

For *micA* (18 targets), 3 targets are located in the outer membrane ( $p = 0.005$ ), 2 in the pore complex ( $p = 0.027$ ), and 6 are located in the plasma membrane ( $p = 0.044$ ). Similarly for *gcvB*-sRNA *ilvJ*, *dppA*, and *cycA* [200, 227, 250] are in the list of the interactions scoring better than the *gcvB*-*oppA* interaction listed in the benchmark set.

A further interesting example are targets of *gcvB* (86 targets), for which 8 targets are implicated in cellular respiration ( $p = 0.0008$ ), in line with recently published results [34]. *micF* is significantly involved in the valine, leucine and isoleucine biosynthesis pathways ( $p = 0.003$ ), in line with results from [65], where *micF* was shown to be regulated by the leucine repression protein (lrp).

## 5.3 Conclusion

---

This chapter introduced **RNAplex**, a RNA-RNA interaction method derived from **RNAduplex**. In contrast to **RNAduplex**, **RNAplex** can consider accessibilities to predict RNA-RNA interactions. Although the approximation used in **RNAplex** can lead to large errors in the computation of strongly asymmetric loop, large bulges as well as accessibilities, all interactions contained in our test samples (see Table 5.5, Table 5.3 and Table 5.6) were predicted by **RNAplex** with a precision similar to that of **RNAup**. The ability of **RNAplex** to perform comparative target search allows to discard poorly conserved interaction and to lend further credibility to interactions showing compensatory mutations. Based on a dataset of experimentally confirmed interactions, we

show that **RNAplex** in its present form is an useful tool to predict new sRNA targets. We further show that suboptimal predictions from **RNAplex** may actually be real targets. Application of the comparative version of **RNAplex** on larger genomes and other ncRNAs, e.g. miRNAs, is straightforward.

sRNA	mRNA	$\Delta G_{RNAhybrid}$	Position RNAhybrid	$\Delta G_{RNAplex-c}$	Position RNAplex -c	$\Delta G_{RNAup}$	Position RNAup	$\Delta G_{RNAplex}$	Position RNAplex -a	Pos.lit.
RyhB	sodB	-58.4(1247)	<b>-239, -100</b>	-25.20(87)	<b>+183, +162</b>	-10.50(60)	-18, +4	-8.57(100)	-7, +5	-4, +5
DsrA	hns	-49.0(1296)	<b>-170, -61</b>	-21.90(128)	+1, +20	-10.90(17)	-10, +11	-13.50(0)	+5, +22	+7, +19
MicA	ompA	-54.2(58)	-87, +30	-23.90(67)	-22, -5	-13.46(0)	-21, -6	-13.82(5)	-20, -5	-21, -6
MicC	ompC	-71.1(120)	-86, +36	-22.00(97)	-31, -14	-15.85(1)	-30, -15	-18.48(0)	-30, -15	-30, -15
MicF	ompF	-47.5(1010)	<b>-150, -61</b>	-26.80(34)	-27, +10	-17.00(3)	-11, +9	-14.95(1)	-12, +8	-16, +10
Spot42	galK	-79.4(28)	-112, +40	-29.30(38)	+4, +37	-18.92(0)	-18, +30	-13.02(3)	-19, +14	-19, +21
SgrS	ptsG	-139.0(1938)	-68, +200	-23.30(170)	<b>+150, +171</b>	-17.17(1)	-28, -10	-20.10(0)	-28, -8	-28, +4
GcvB	dppA	-125.2(1436)	-154, +110	-29.40(80)	-31, -6	-16.90(16)	-30, -7	-21.17(10)	-31, -6	-31, -14
GcvB	oppA	-122.8(1837)	-156, +189	-25.10(263)	-3, 45	-11.94(58)	-2, 14	-16.69(22)	-4, 21	-8, +16

Table 5.5: Binding site summary for the 9 functional interactions from [250]. The number in parenthesis represents the quantity of predicted interactions involving the same ncRNA, overlapping with a 401 nts long region centered around the start codon and having a higher interaction energy than the functional hybrid. Positions in red indicate target sites that were misspredicted by the respective tool. For RNAplex -c a per nucleotide penalty of 0.3 kcal/mol was used.

sRNA	mRNA	Pos.lit.	Pos <sub>RNAplex</sub>	$\Delta G$ RNAup	$\Delta G$ RNAplex	$\Delta G$ RNAplex -A	$\Delta G$	Z-score	Z-score N°seq
RyhB	sodB	-7, +5	-4, +5	-10.50(60)	-11.08(50/87)	-9.31(12)	65	57	2 (7)
DsrA	hns	+6,+21	+7, +19	-10.90(17)	-12.74(2/128)	-11.25(10)	1	12	0 (0)
MicA	ompA	-21, -6	-21, -6	-13.46(0)	-14.35(1/67)	-14.04(14)	0	11	0 (0)
MicC	ompC	-30, -15	-30, -15	-15.85(1)	-16.24(2/97)	-17.50(9)	0	0	0 (0)
MicF	ompF	-8, +10	-16, +10	-17.00(3)	-13.65(8/34)	-18.28(6)	0	0	0 (2)
Spot42	galK	-19,+14	-19, +21	-18.92(0)	-13.02(25/38)	-7.31(9)	25	28	5 (12)
SgrS	ptsG	-28, -8	-28, +4	-17.17(1)	-17.53(0/170)	-11.17(10)	5	4	0 (1)
GcvB	dppA	-31, -10	-31, -14	-16.90(16)	-17.11(8/80)	-13.15(9)	14	14	7 (19)
GcvB	oppA	-4, 21	-8, 16	-11.64(58)	-12.00(36/263)	-14.43(5)	27	26	14 (19)

Table 5.6: Summary of the predicted binding sites for the 9 functional interactions reported by [250]. The first and second columns show the name of interaction partners. Column 3 and 4 give the predicted and experimentally reported binding regions, respectively. Column 5 and 6 report the binding  $\Delta G$  computed by **RNAup** and **RNAplex**, respectively. The numbers in parenthesis in the sixth column represent the number of interactions, located within a window of 80 nts centered around the start codon, with a lower interaction energy than the experimentally reported interaction for the predictions made by **RNAplex** with and without considering the opening energy, respectively. Column 7 gives the interaction energy for the multiple sequences interactions. The numbers in parenthesis in column 7 represent the number of sequences in the final alignments. Column 8 shows the rank of the interaction when looking only at the interaction energy. Column 9 shows the rank of the interactions based on the Z-score corrected for the number of sequences in the alignment. Finally column 10 shows the rank of the interaction based on the Z-score, given that only interactions with a greater or equal number of sequences in the alignment are taken into account. The number in parenthesis in the last column represent the number of better scoring elements in the case of alignment when no accessibility information are taken into account.

sRNA	mRNA	Rank <b>RNAplex</b>	False positive rate 1/nt
RyhB	sodB	20	$1.6e^{-4}$
DsrA	hns	9	$7.2e^{-5}$
MicA	ompA	0	$3.7e^{-6}$
MicC	ompC	0	$7.5e^{-7}$
MicF	ompF	17	$8.7e^{-4}$
Spot42	galK	0	$7.3e^{-6}$
SgrS	ptsG	1	$3.9e^{-6}$
GcvB	dppA	21	$1.4e^{-4}$
GcvB	oppA	1	$4.1e^{-6}$

Table 5.7: Summary of the number of false positives under different condition. In the third column, for each confirmed interaction, the number of better scoring interactions involving the corresponding dinucleotide shuffled sRNA and any dinucleotide shuffled *e.coli* mRNAs is reported. It should be noted that the interaction should take place in the region located 50 nts upstream and 30 nts downstream of the start codon. In the last column, the number of expected hits per nucleotide is reported. In this case there is no location restriction.



## RNAsnoop

Box H/ACA snoRNA facilitates the conversion of Uracil to pseudouracil ( $\Psi$ ) in a specific sequence context [10]. The specificity for a particular target site is the consequence of the hybridization of snoRNA and target RNA, in most cases a ribosomal RNA. The target U is positioned by two specific interactions of the flanking target RNA sequence with the complementary sequence of the recognition loop of the snoRNA [185](see figure 6.1). The “correct” secondary structures of snoRNAs are typically hard to predict. Thus, the exact structure of the interior loop, and hence the sequence motifs complementary to the binding site, are unknown.

The prediction of putative snoRNA target sites is an integral part of two programs (**snoGPS** [213] and **Fisher** [71]) that attempt to detect H/ACA snoRNAs in genomic DNA. Both programs search for sequence complementarities between a list of possible target sites and the binding region of the snoRNA candidate. In these models, mismatches between the target and the snoRNA are not allowed. Furthermore, neither program provides information on the energetics of the interaction or the stability of the stems, two factors that were recently shown to be important for correctly predicting snoRNA-target interactions [268].

The idea of Thermodynamic Matchers [109] is employed in this chapter to determine the energetically optimal structure of an H/ACA snoRNA that is bound to a given putative target sequence. The implementation of Thermodynamic Matchers [203] is not directly applicable, however, since the snoRNA-target interaction corresponds to a complex pseudoknot that is beyond the scope of existing RNA folding software. We present here a dynamic programming algorithm, **RNAsnoop**, that specifically cap-

tures the structure of the snoRNA-target interaction and is optimized for scanning speed. The thermodynamic considerations are combined with a Machine Learning component to increase the specificity of target predictions, which can be improved even further by including comparative information.

## 6.1 Methods

---

### 6.1.1 Single-Sequence RNAsnoop

RNAsnoop implements a specialized co-folding algorithm that takes into account that stringent structural constraints must be satisfied for a functional interaction of a box H/ACA snoRNA stem-loop and its target. As input, RNAsnoop takes one of the typical two stem-loop components of a known or predicted H/ACA snoRNA. The closing stem,  $T$  is assumed to be known from the *a priori* prediction of the snoRNA structure. The part of the snoRNA sequence enclosed by  $T$  is allowed to interact with the target structure. Figure 6.1 outlines the general principle.

The interaction structure can be decomposed into the unbranched stem-loop “above” the pseudouridylation site, and the left and right “arms” of the binding site itself. The total energy of these components will be optimized by dynamic programming. In addition, the snoRNA-target interaction is influenced by the short closing stem of the interaction loop.

The upper stem-loop structure of the snoRNA (with sequence  $y$ ) is simply modeled as an unbranched fold. The energies of its optimal substructures satisfy the recursion

$$M_{p,q} = \min \begin{cases} \mathcal{H}(y[p, q]) \\ \min_{k,l} M_{p-k, q+l} + \mathcal{I}(y[p-k, p], y[q, q+l]) \end{cases} \quad (6.1)$$

where  $\mathcal{H}(y[p, q])$  denotes the energy parameters [150, 167] for a hairpin loop formed by the sub-sequence  $y[p, q] = y_p y_{p+1} \dots y_q$  including the closing pair  $(y_p, y_q)$ . Analogously,  $\mathcal{I}(y[u, p], y[q, v])$  is the energy of an interior loop composed of the sequences  $y[u, p]$  and  $y[q, v]$ , again including the delimiting base pairs  $(y_p, y_q)$  and  $(y_u, y_v)$ .

Inspection of known snoRNA-rRNA interactions revealed that the interaction region can contain only single and tandem mismatches but no bulges. Therefore we allow

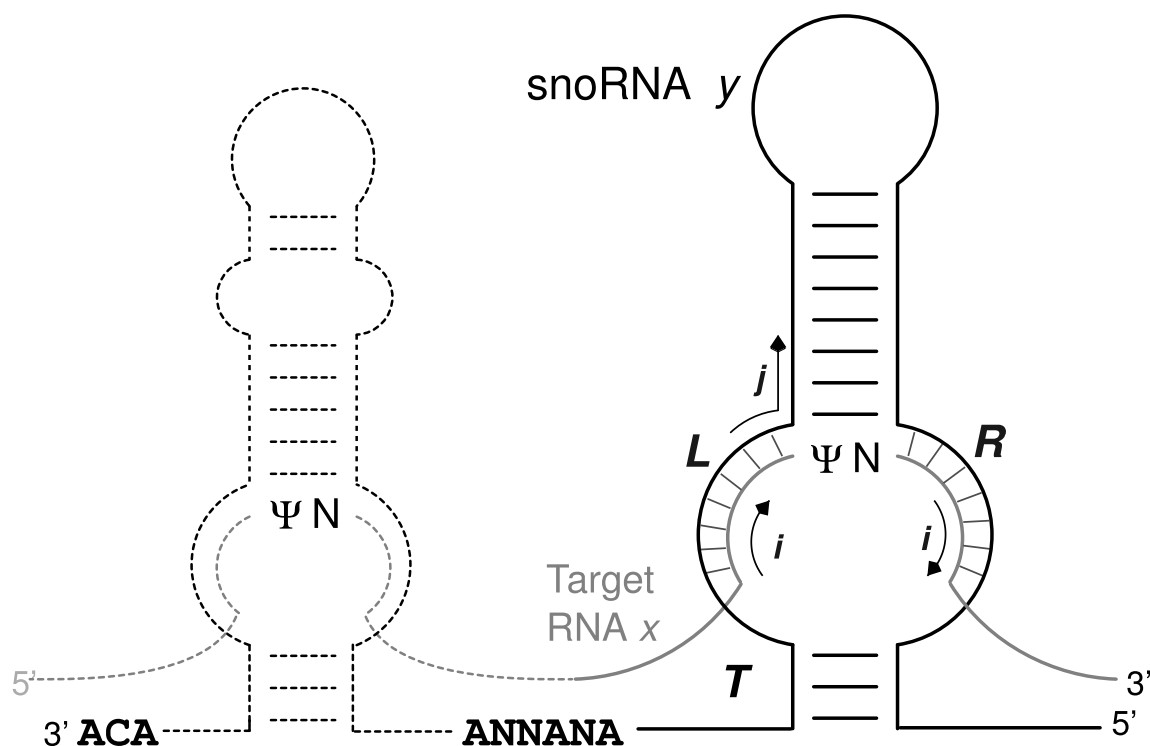


Figure 6.1: Box H/ACA snoRNAs typically interact with both stem-loop structures with regions of a target RNA flanking the Uracil residue that is to be pseudouridylated. Computation of the interaction structure is performed separately for the two stems-loop components of a H/ACA snoRNA. The closing stem  $T$  at the root of each branch is assumed to be given from the structure prediction. The region inside of  $T$  is decomposed into the upper stem-loop structure with an energy contribution  $M$ , l.h.s. and r.h.s. interaction structures with their energy contribution  $L$  and  $R$ , respectively. Since *RNA*snoop scans the target RNA in  $5' - 3'$  direction, the snoRNA is read in  $3' - 5'$  direction.

only stacked base pairs and symmetrical loops of length 2 and 4. Thus the left part satisfies the recursion

$$L_{i,j} = \min_{k=1,2,3} L_{i-k,j+k} + \mathcal{I}(x[i-k,i], y[j,j+k]) \quad (6.2)$$

The index  $i$  runs along the target RNA  $x$ , while  $j$  refers to the position on the snoRNA  $y$ . To ensure that all interactions start inside the recursion matrix we set  $L_{i,j} = 0$

The r.h.s. array  $R$  contains the optimal folding energies of the interaction structure up to positions  $i$  on the target and  $j$  on the snoRNA consisting of the l.h.s. binding region  $L$ , the snoRNA stem-loop  $M$ , and the partial r.h.s. binding region  $R_{i,j}$ . It thus extends a r.h.s. binding region or refers to its first base pair. In the latter case, nucleotide  $x_{i-2}$  is the uracil that is pseudouridylated. The corresponding recursion reads

$$R_{i,j} = \min \begin{cases} \min_{k,l \leq 2} R_{i-k,j+l} + \mathcal{I}(x[i-k,i], y[j,j+l]) \\ \min_{l \in [3, |y|-j]} L_{i-3,j+l+1} + M_{j+1,j+l} \\ \text{if } x[i-2] = 'U' \end{cases} \quad (6.3)$$

For each  $i$ , the best binding energy at target position  $i$  is  $\max_j R_{i,j}$ .

Space and time requirements for the  $M$ -matrix are limited by the size  $|y|$  of the snoRNA stem-loop structure, which is a user specified constant, typically 120 nts. Formally, the space and time complexity is  $\mathcal{O}(|y|^2)$  and  $\mathcal{O}(|y|^4)$ , respectively. Similarly to **RNAplex** (see chapter 5), our implementation, the space requirements for the  $L$  and  $R$  arrays are limited to  $5 \times |y|$  independent of the target  $|x|$  of the target RNA. This is possible because the length of interior loops in the recursions is restricted to not more than 4 and the transition from  $L$  to  $R$  recursion only looks back to  $i - 4$ . The time complexity for  $L$  is  $\mathcal{O}(|x| \cdot |y|)$ , while for  $R$  we need  $\mathcal{O}(|x| \cdot |y|^2)$  operations. The total run time is thus  $\mathcal{O}(|x| |y|^2 + |y|^4)$ , i.e., we have a linear “scanning algorithm” for long target RNAs.

Due to the difference in accessibility between sites with pseudouridine and uridine residues in both human and yeast (see figure 6.2), we extended **RNA snoop** so that accessibility information are considered in the folding step. Accessibility profiles as computed by **RNAup** or **RNAplfold** describe the energy necessary to open the

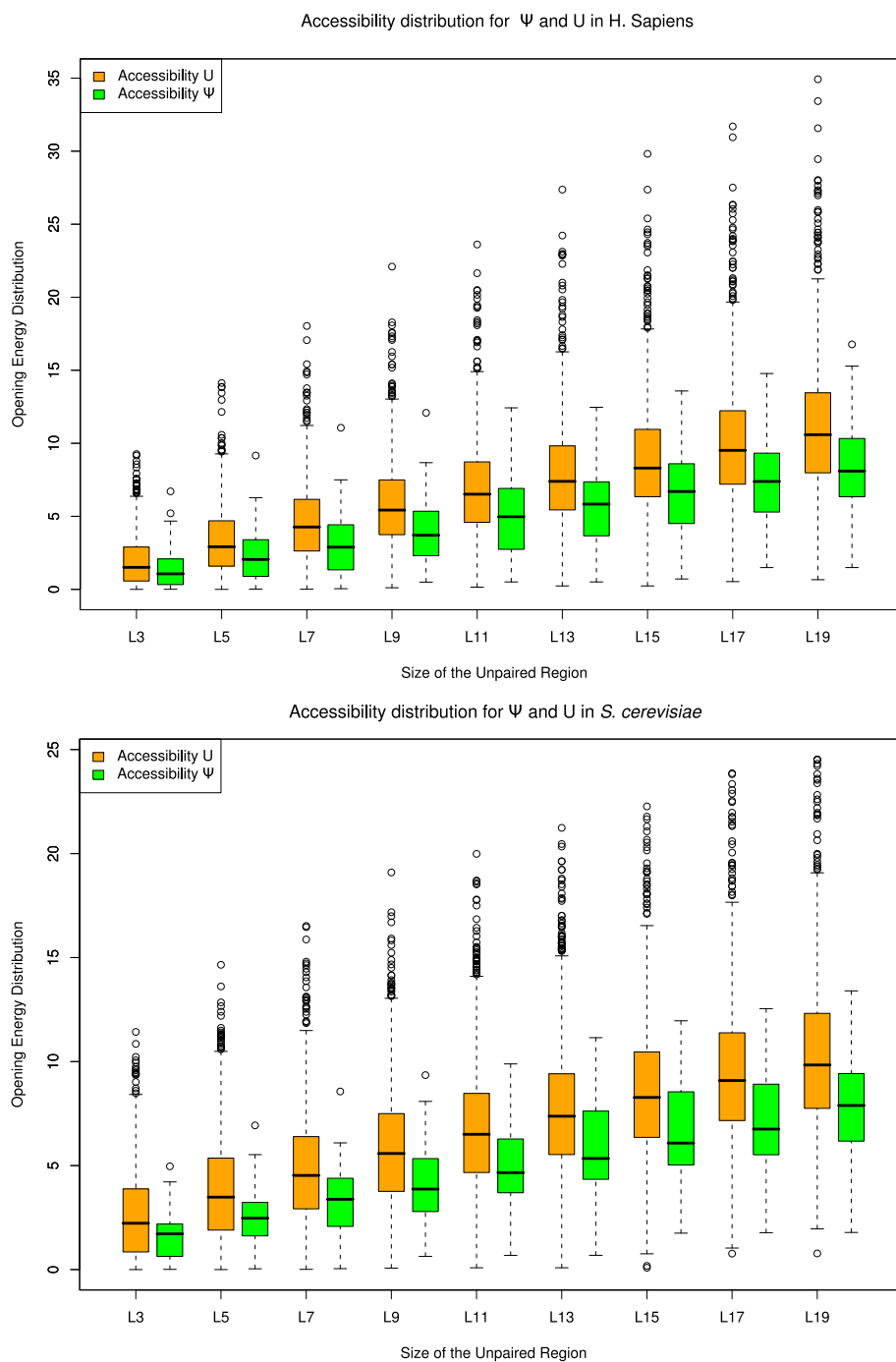


Figure 6.2: Boxplots showing the accessibility distribution for all known uridines in human (top) and yeast (bottom) 28S and 18S rRNAs. The target accessibility was computed by using *RNAup* on the whole length sequences of 28S and 18S rRNAs. The target size was varied between 3 and 19 nts in steps of 2 nts and was centered around the (pseudo)uridine site.

secondary structure on an interval of the target sequence. The full implementation of RNA-RNA interactions is too expensive in terms of computational resources for a target search program. We therefore borrow the approach from **RNAplex** (see chapter 5), which uses an affine approximation to speed up the computation of RNA-RNA interaction energies. We further use pre-computed accessibility profiles in a way similar to that of **RNAplex** in order to improve the prediction accuracy.

### 6.1.2 Machine-Learning Component

[268] showed that the interaction energy is necessary but not sufficient to distinguish functional from non-functional snoRNA-rRNA interactions. Stability of the stems enclosing the pseudouridylation pocket as well as structural features relative to the stems and the interaction regions are equally relevant. In order to take those parameters into account, a machine-learning method (SVM) was used to analyze the output of **RNAsnoop**. Two models were developed depending on whether or not **RNAsnoop** considers the target site accessibility. The SVM was trained on verified interactions from yeast [213] and human [268], respectively. Because the training data set did not contain experimentally confirmed non-functional interactions we augmented it by adding artificial ones. For each snoRNA-stem involved in a verified interaction, **RNAsnoop** was run against yeast 28S and 18S sequences. All hits that had an interaction energy smaller than the one of the experimentally validated interaction and that do not target a known pseudouridylation site were considered non-functional. The final training data set contained 43 positive and 103 negative interactions.

For both models we derived a set of 29 features to pass to the SVM, and then selected a subset following the approach described by [36]. Features that were ultimately selected are described in some details in figure 6.3. We used different feature set depending on whether accessibility is taken into account or not.

For the case where the target accessibility was neglected, only five features are used, four of which describe the geometry of the interaction itself (**t\_i\_gap**, **U\_gap**, **i\_t\_gap**, and **gap\_right**) and the length of the intervening stem **stem\_length**.

For the model with accessibility, 11 features are used. In addition to features describing the geometry of the interaction (**t\_i\_gap**, **U\_gap**, **i\_b\_gap**, **i\_t\_gap**, **gap\_right**) and of the upper stem (**stem\_length**, **stem\_asymmetry**), four energy-related values

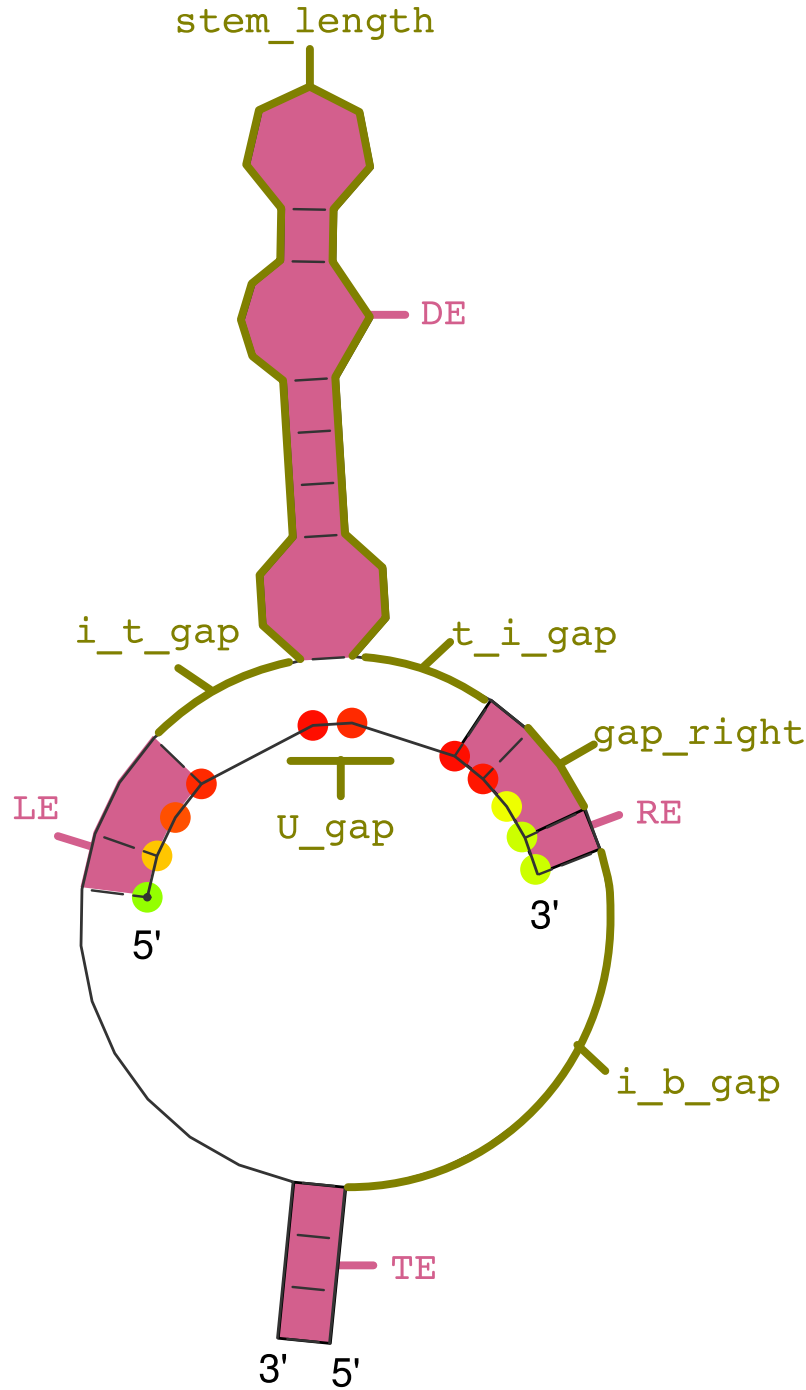


Figure 6.3: Features considered in the SVM model. Structural (black bold lines) and energy features (shaded regions). TE: lower stem energy, LE: 5' interaction energy, DE: upper stem energy, RE: 3' interaction energy, For each nucleotide in the target, its local opening energy is represented by a gray circle, where light gray represents low local opening energy and dark gray high local opening energy. The target total opening energy (OE) is the sum of all local opening energies, YE:  $YE = LE + RE + TE + DE$ , XE:  $XE = LE + RE + DE$ , dYE:  $dYE = YE + OE$ , t\_i\_gap: number of nucleotides between the 5' end of the upper stem and the 3' end of the 5' interaction on the snoRNA, U\_gap: number of nucleotides between the 3' end of the 5' interaction and the 5' end of the 3' interaction on the mRNA, i\_b\_gap: number of nucleotides between the end of the lower stem and the 3' end of the 5' interaction on the snoRNA, i\_t\_gap: number of nucleotides between the 5' end of the 5' interaction and the 5' end of the snoRNA stem, stem\_length: length of the upper stem of the snoRNA.

YE, DE, XE, and dYE were selected. (see figure. 6.3).

### 6.1.3 Performance

The prediction accuracy of **RNASnoop**, **snoGPS** and **fisher** was assessed on the human [268] and yeast [213] datasets of experimentally confirmed/rejected snoRNA-rRNA interactions. For a given snoRNA involved in a confirmed interaction, we determined how many target sites were predicted to bind with a better score/energy than the experimentally reported one. Table 6.1 summarizes these rank values for the confirmed interactions in yeast. We clearly see that **fisher** is less sensitive, detecting only 16 of the 44 interactions in yeast. Still, these 16 interactions were all ranked first, indicating that **fisher** has a high specificity. In comparison, **RNASnoop** and **snoGPS** detect 43 and 41 of the 44 verified interactions in yeast, and 11 and 10, resp., in human. We remark that **RNASnoop** did not identify the interaction of snR82 with LSU-U2349, because **RNASnoop** predicts the adjacent position LSU-U2351 as preferred target. On average, **RNASnoop** ranks the confirmed interactions higher in the list than **snoGPS**. This trend is also seen in the ROC curve in figure 6.4, where **RNASnoop** shows a higher prediction accuracy than **snoGPS**.

In human, **RNASnoop** performs better than **snoGPS**. In particular, the SVM version successfully rejects the four non-functional snoRNA-rRNA interactions and successfully ranks 11 out of the 12 confirmed interactions first (see Table 6.2). Still, one of the confirmed interaction was rejected by the SVM.

The run time of **RNASnoop** was compared to that of **snoGPS** and **RNAhybrid**. **fisher** was modified to turn it into a target finder; the resulting run time, however, was so high that we decided to not evaluate it further. **RNAhybrid** uses a dynamic programming algorithm to find putative miRNA-targets and has a run time of  $\mathcal{O}(|x| \cdot |y|)$ . Because the run time of **RNASnoop** is linear in the target size but quadratic in the snoRNA size, we varied the length of both sequences. Due to the important variance of H/ACA snoRNA stems length [11, 245], we incremented the snoRNA stem size in steps of 30 nucleotides from 60 up to 420 nucleotides, keeping the target RNA length fixed to 5000 nucleotides. Conversely, the target length was varied between 1000 and 256000 nucleotides with a snoRNA stem length set to 200. We set the threshold for each program so that they returned at most one hit. Independently of the snoRNA

snoRNA	Target	Position	snoGPS	fisher	RNASn. A	snoRNA	Target	Position	snoGPS	fisher	RNASn. A	RNASn. A
snR11	25S	2416	3	—	12	snR10	25S	2923	2	1	28	26
snR161	18S	632	6	1	8	snR46	25S	2865	1	1	1	1
snR161	18S	766	1	—	11	snR49	18S	120	3	1	1	1
snR189	18S	466	2	1	1	snR49	18S	211	2	—	5	5
snR189	25S	2735	1	—	1	snR49	18S	302	1	—	5	4
snR191	25S	2258	1	—	5	snR49	25S	990	4	—	—	1
snR191	25S	2260	99	—	8	snR5	25S	1004	3	1	1	1
snR3	25S	2129	4	—	1	snR5	25S	1124	1	—	8	1
snR3	25S	2133	1	—	1	snR8	25S	960	68	—	3	5
snR3	25S	2264	2	—	3	snR8	25S	986	55	1	2	3
snR31	18S	999	1	1	1	snR80	18S	759	—	—	2	2
snR32	25S	2191	1	1	1	snR80	25S	776	—	—	2	2
snR33	25S	1042	1	1	1	snR81	25S	1052	57	1	2	1
snR34	25S	2826	2	—	1	snR82	25S	2349	1	1	—	—
snR34	25S	2880	1	—	1	snR82	25S	2351	1	—	1	2
snR35	18S	1191	1	—	1	snR82	25S	1110	—	—	2	4
snR36	18S	1187	12	1	7	snR83	18S	1290	1	—	58	7
snR37	25S	2944	1	—	2	snR83	18S	1415	4	—	1	1
snR42	25S	2975	1	1	4	snR84	25S	2266	1	—	2	2
snR43	25S	966	1	—	1	snR85	18S	1181	1	1	1	1
snR44	18S	106	1	—	2	snR86	25S	2314	13	—	3	1
snR44	25S	1056	2	1	1	snR9	25S	2340	33	—	18	19

Table 6.1: Prediction comparison of **RNASnoop** (abbreviated **RNASn.**), **snoGPS** and **fisher** for the known snoRNA-rRNA interactions in yeast. **RNASn. A** stands for the accessibility version of **RNASnoop**.

snoRNA	Target	Position	Type	snoGPS	RNASn.	RNASn. A	SVM
ACA19_1	28S	3709	+	1	1	1	1
ACA19_2	28S	3618	+	25	2	1	1
ACA19_1	18S	863	-	10	1	4	—
ACA19_1	18S	866	-	10	—	—	—
ACA24_1	18S	863	+	—	1	1	1
ACA24_2	18S	612	-	86	3	6	—
ACA28_1	18S	815	+	1	4	1	1
ACA28_2	18S	866	+	—	2	4	1
ACA42_1	18S	572	-	3	4	19	—
ACA42_2	18S	109	+	1	1	1	1
ACA50_1	18S	34	+	1	1	1	—
ACA50_2	18S	105	+	2	1	1	1
ACA62_1	18S	34	+	3	24	1	1
ACA62_2	18S	105	+	2	1	1	1
ACA67_1	18S	572	+	2	2	1	1
ACA67_2	18S	109	+	1	1	1	1

Table 6.2: Prediction performance in human for **snoGPS**, **RNASnoop** (**RNASn.**), **RNASnoop** with accessibility (**RNASn. A**) and the **SVM** in human. The numbers represent the rank of the interaction for the corresponding snoRNA stem. In column Type, +, − represent experimentally confirmed or rejected interactions, respectively. When using the human interactions for testing, we trained the SVM exclusively on the yeast dataset.

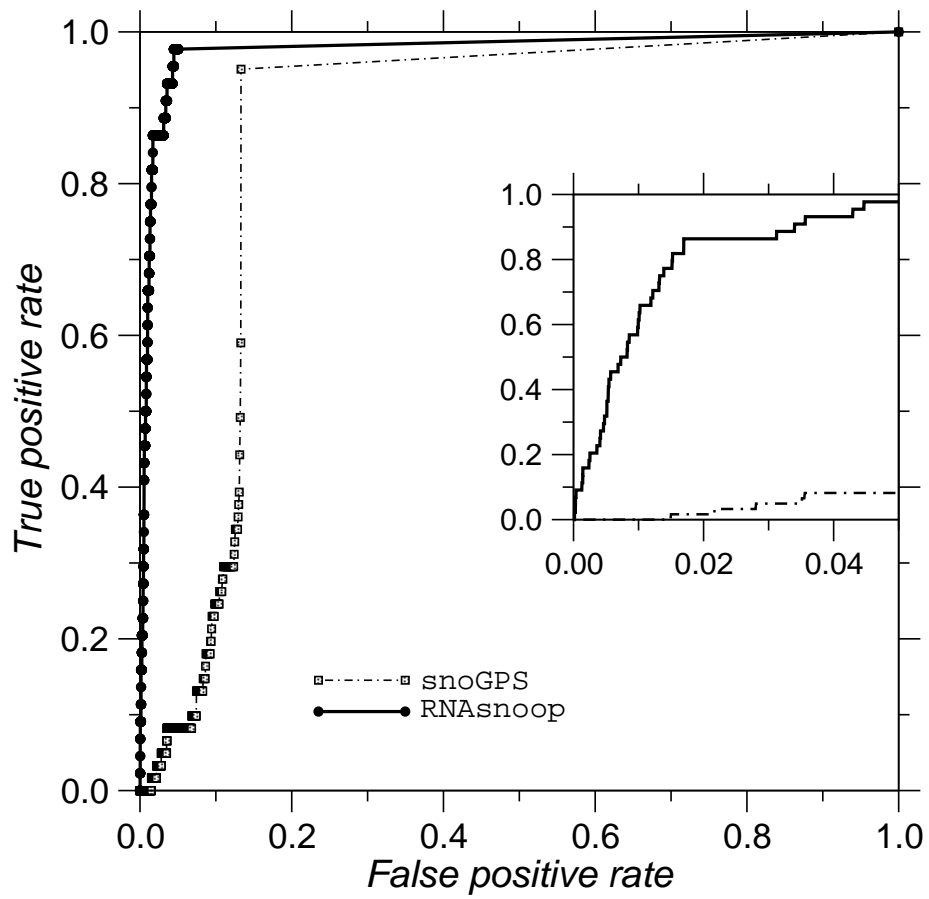


Figure 6.4: ROC curve for RNASnoop and snoGPS on the yeast data set [213]. RNASnoop was used without the SVM functionality.

or target sequence size, **snoGPS** and **RNAsnnoop** have a similar run time. They are around 15 times faster than **RNAhybrid** (see figure 6.5 and 6.6).

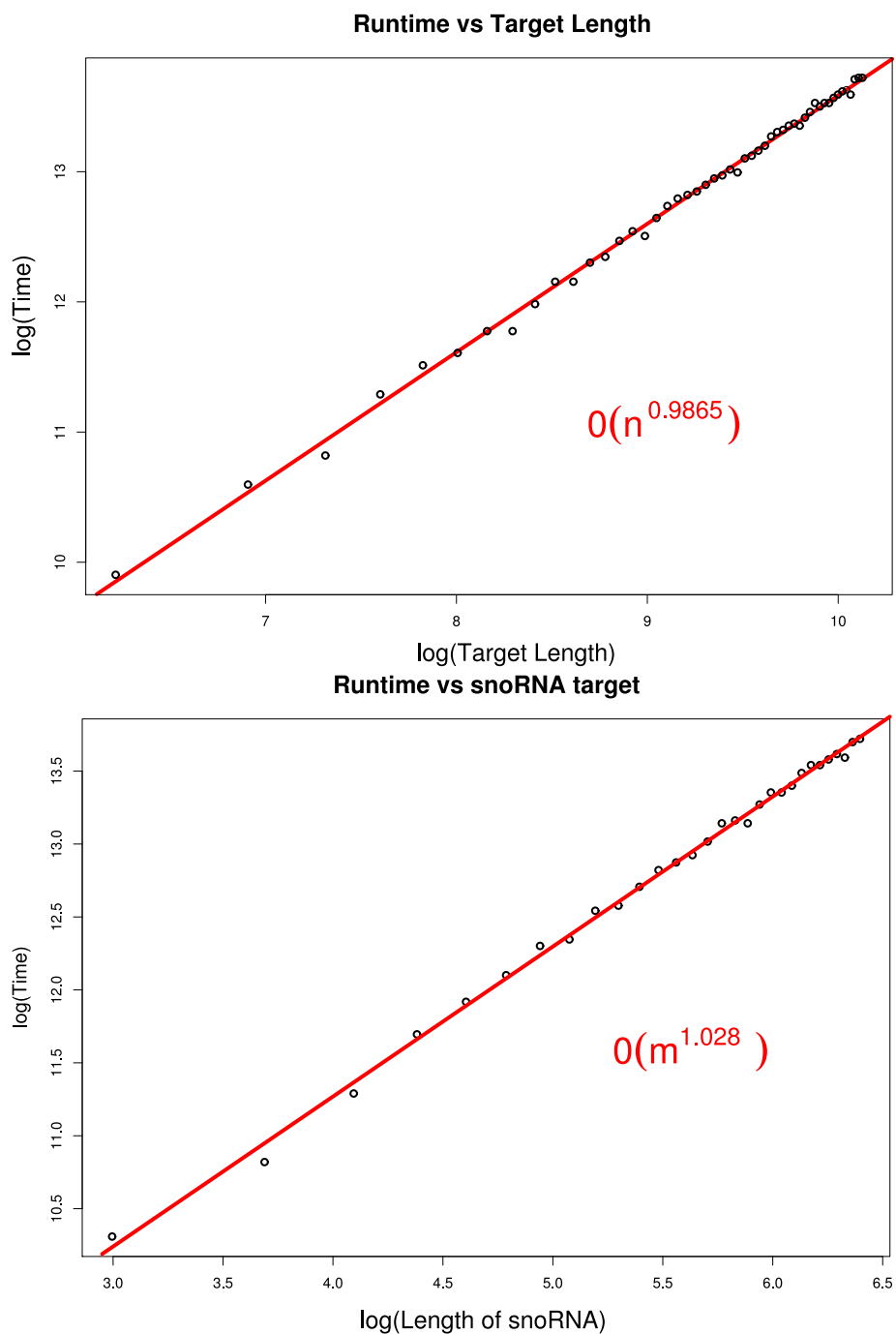


Figure 6.5: Time dependency of RNAsnoop on the target size (top) and snoRNA size (bottom). The target size was varied between 500 and 25000 nts while the snoRNA sizes were varied between 20 and 500 nts. The runtime of RNAsnoop grows linearly with the target size and grows more rapidly than linear with the snoRNA size.

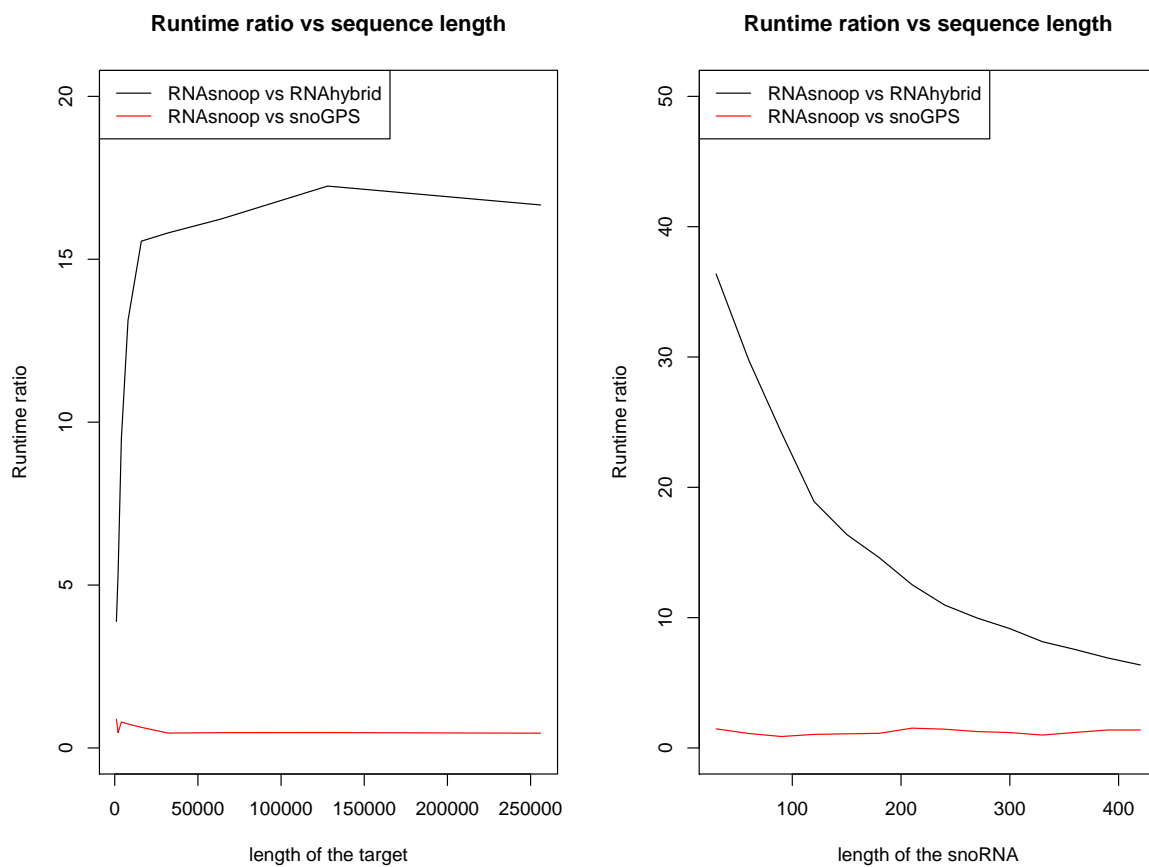


Figure 6.6: Ratio of the time dependency of RNAsnoop against RNAhybrid and snoGPS. (left) Dependence of the ratio on the target size. (right) Dependence of the ratio on the snoRNA size. All three programs were run so that only the best interaction was returned. Under these conditions RNAsnoop has a runtime similar to that of snoGPS (red curve), while RNAhybrid is about 15 times slower than RNAsnoop (black curve). Due to the higher than linear runtime dependency this difference becomes smaller for larger snoRNA (right, black curve).

### 6.1.4 A Comparative Version

The use of alignments in the target search can further help to find real snoRNA-RNA interactions. On one hand, the absence of conserved target-site in closely related species may indicate that the proposed interaction does not occur in nature. The presence of compensatory mutations between the snoRNA binding bucket and the target site, on the other hand, can lend further credibility to single-sequence target predictions [35].

The alignment extension of **RNASnoop** is based on the same approach used in **RNAalifold** [19, 99], where a thermodynamic energy minimization folding algorithm is coupled with a simple scoring model to assess evolutionary conservation. As in the single sequence algorithm, the upper-stem is modelled as an unbranched fold by a slightly modified **RNAalifold** algorithm. The interaction part uses the same approach as **RNAalifold**, with the sole difference that only interior loops are allowed between the snoRNA and its target.

### 6.1.5 SNOOPY

For an efficient analysis of data we provide and recommend the `perl` script **SNOOPY**. It uses both the SVM as well as the homology information to predict putative target-interactions. **SNOOPY** takes as input a snoRNA alignment and a target alignment. In a first step **SNOOPY** uses **mLocARNA** to obtain sequence/structure alignments of the snoRNAs [266]. If the sum of scores of **mLocARNA** pairwise alignments for a sequence is lower than  $< 2500$ , then the sequence is discarded. Duplicates and sequences belonging to species that are present in only one of the two alignments are also removed. **SNOOPY** pre-selects possible targets in a user defined reference organism by means of the single-sequence version of **RNASnoop** and one of the two SVM-models. For each reported targets, **SNOOPY** extracts the corresponding slice from the alignments and then realigns the corresponding subsequences with **Clustalw** [242]. Target sequences for which the pairwise-alignment score is below a threshold, or which do not exhibit a U residue at the previously predicted site, are removed together with the snoRNA sequences from the same organisms. Whenever the number of retained sequences is above a user-defined threshold, the alignment version of **RNASnoop** is applied. Fi-

nally **SNOOPY** reports for each snoRNA alignment a user-specified number of putative interactions. These interactions can be ranked either by their SVM-score or by the single sequence interaction energy for the reference organism.

## 6.2 Results

---

In order to test the usability of **RNASnoop** we consider the problems of finding snoRNAs associated with “orphan” pseudouridylation sites in human rRNAs. Although the role of snoRNAs in locating target uridine residues was discovered more than a decade ago, there are still a few pseudouridylation sites in human rRNAs [155, 188] for which the responsible snoRNAs have not yet been determined. We used the single sequence version of **RNASnoop** to predict the possible snoRNAs that may pseudouridylate these orphan sites. For this we used all the known human H/ACA sequences reported in **snoRNA-LBME-db** [141] and tested them against the 11 reported orphan sites in the human LSU and SSU. Based on the currently available snoRNA data, 8 orphan sites can be mapped to existing snoRNA stems. Interestingly, 2 orphan snoRNAs (ACA38B, ACA51), and 2 stems, for which no function was reported, were among the predictions. Additionally, 4 stems with known targets were predicted to target four of the orphan sites. The predicted interactions are listed in table 6.3 and figure 6.7.

We used **SNOOPY** to assign putative targets to the 5 orphan snoRNAs found in *Drosophila* (Or-aca1, Or-aca2, Or-aca3, Or-aca4, Or-aca5). For each orphan snoRNA reported in **Flybase** [9], we searched for homologous sequences in the 11 other *Drosophila* species by using **blast** [3]. For each species the sequence with the highest homology with *Drosophila melanogaster* was selected. The sequences were then aligned with **mLocARNA**, a variant of the Sankoff algorithm. For each snoRNA, the full length alignment was then divided into a 5' stem and 3' stem alignments.

The rRNA alignments were retrieved from the **arb-silva** database [198]. In order to get the best possible alignments, we realigned them with **Clustalw**, **Muscle** [59], and **RNASalsa** [232]. The quality of the alignments was assessed by determining how well the conserved pseudouridylation sites in *Drosophila melanogaster* and *Homo sapiens* were aligned in the twelve drosophilid rRNA sequences. Based on this quality

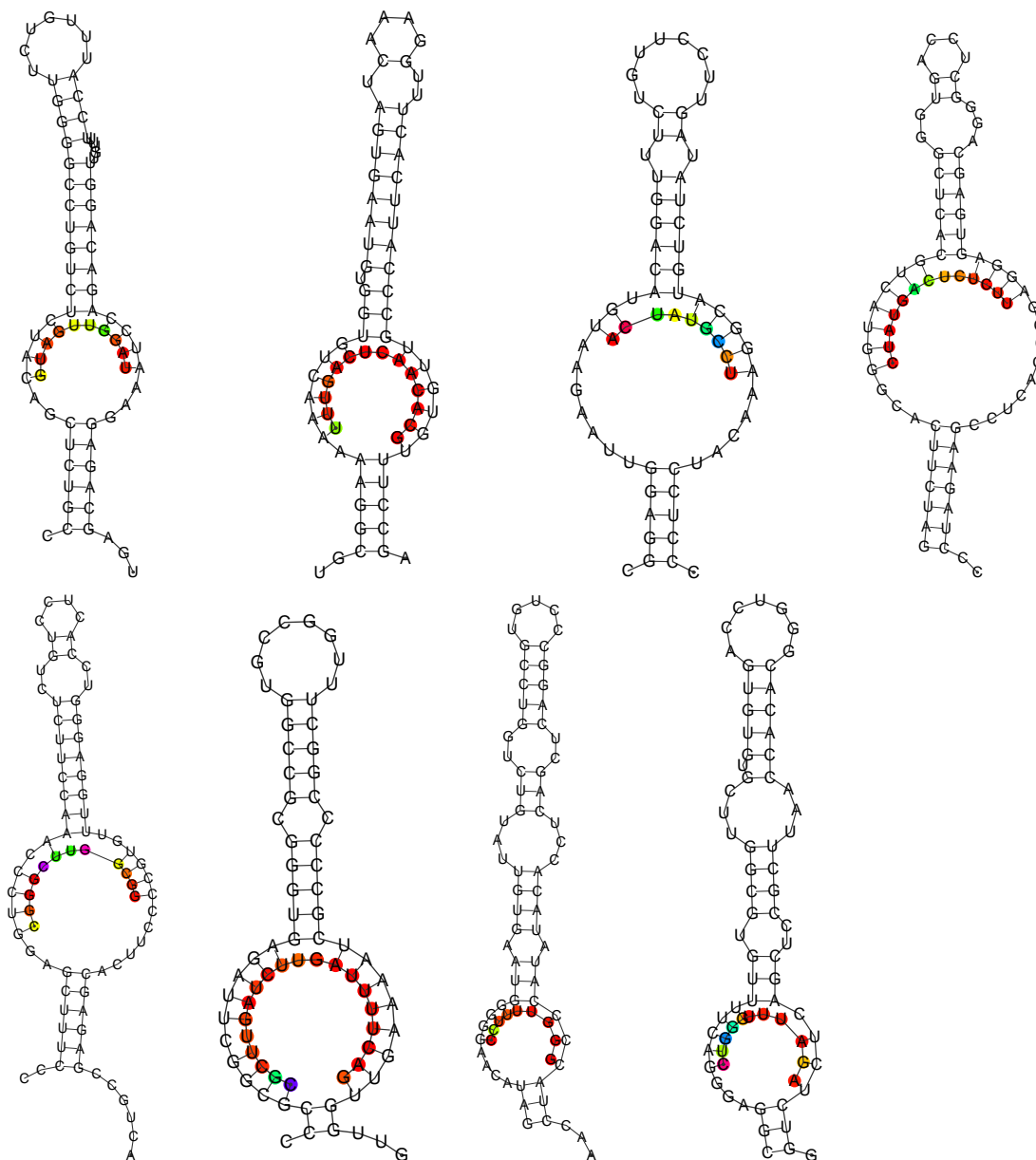


Figure 6.7: Structure of the interactions between the human orphan  $\Psi$  sites and their predicted snoRNAs as returned by *RNAsnoop*. From left to right: ACA55-2:18S-681, ACA13-1:18S-1248, SNORA38B-1:28S-1523, ACA52-2:28S-3747, U71c-2:28S-3863, ACA64-1:28S-4266, ACA51-2:28S-4323, ACA10-1:28S-4501, where i.e. ACA51-2:28S-4323, means that the second stem of ACA51 binds to position 4323 on rRNA 28S. All structures were generated by *RNAsnoop*. The accessibility for each nucleotide is color-coded, with a red representing accessible and green inaccessible nucleotides.

rRNA	Position	snoRNA	stem	function	SVM-score	Energy
18S	681	ACA55	2	18S-36	0.76	-34.32
18S	918	ACA13	1	18S-1248	0.81	-35.90
28S	1523	SNORA38B*	1	—	0.66	-18.08
28S	1849	—	—	—	—	—
28S	3674	—	—	—	—	—
28S	3747	ACA52	2	—	0.87	-28.94
28S	3749	—	—	—	—	—
28S	3863	U71c	2	18S-406	0.53	-19.14
28S	4266	ACA64	1	—	0.75	-32.00
28S	4323	ACA51*	2	—	0.63	-20.39
28S	4501	ACA10	1	28S-4491	0.54	-15.00

Table 6.3: Predicted snoRNAs targeting the orphan pseudouridines in human ribosomal RNAs. No snoRNAs were found for position 1849, 3674 and 3749 on rRNA 28S. ACA51 and SNORA38B are orphan snoRNAs while ACA52-2 and ACA64-1 are orphan stems

measure, **RNAalsa** was found to perform best. Alignments of snRNAs were taken from [162].

Of the 5 orphan snoRNAs, only Oaca-4 was reported to have a target. We predict that the first stem modifies U2499 on the 28S rRNA (see figure 6.8). This target site is interesting since it was reported to be pseudouridylated [80], but no corresponding snoRNA is known. Moreover, in human and yeast this position, which correspond to U3674 in human and U2191 in yeast, is conserved and pseudouridylated [141]. U3674, finally, remains an orphan site in human.

Interestingly, both the target and binding buckets are completely conserved from *Drosophila melanogaster* to *Drosophila willistoni*, see figure 6.8. On the other hand, 6 out of the 12 base pairs found in the upper stem exhibit compensatory mutations.

The fact that no credible targets have been predicted for the remaining four orphan snoRNAs is not unexpected. First, snoRNAs have also been implicated in modifying “non-canonical targets” such as mRNAs [14, 120, 249], some cause cleavage of pre-rRNAs [66], and [238] recently showed that Or-aca5 is processed by *Dicer*, suggesting a function in the RNA interference pathway.

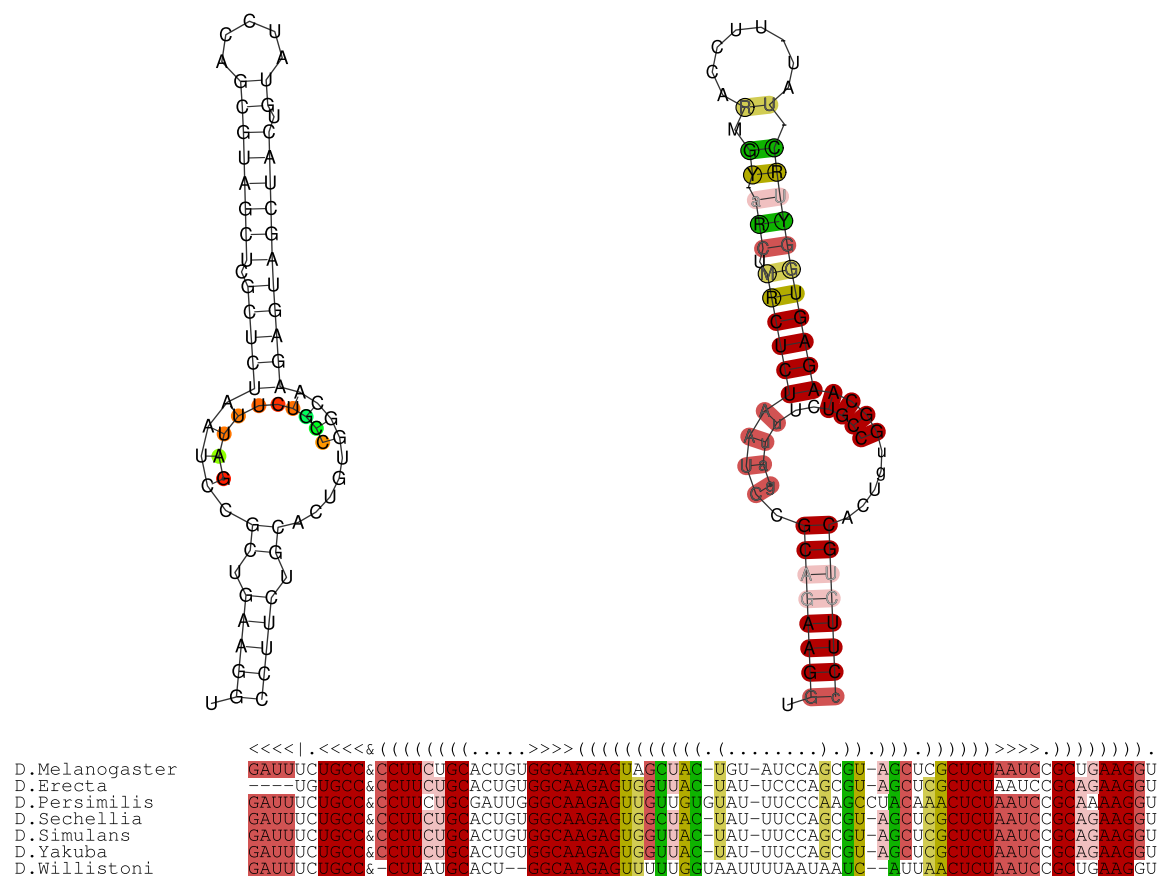


Figure 6.8: . Structure of the interactions between Or-aca4 and its putative target. **L.h.s.:** Single sequence structure. **R.h.s.:** Multiple sequence structure. **Below:** Alignment of the target (up to the & column) and the snoRNA. For the multiple sequence and alignment figures, the color in the order red, ochre, green indicate 1 through 3 different type of base pairs. The consensus structure is represented in dot bracket format on top of the alignment. The angle brackets represent intermolecular base pairs and the braces represent intramolecular base pairs.

## 6.3 Conclusion

---

We presented here **RNAsnoop**, a tool specifically designed to predict complex pseudoknotted H/ACA snoRNA-RNA interaction. In contrast to previous tools, it uses a dynamic programming approach coupled with a nearest-neighbor energy model to identify putative targets. This allows **RNAsnoop** to capture structural and energetic features essential for correctly predicting snoRNA-target interactions [268]. Coupled with a SVM-Classification **SNOOPY** achieves good performance, ranking first 11 out of 12 confirmed snoRNA-mRNA interactions in human and excluding all experimentally rejected interactions. These good results should however not be overestimated as both the training and test datasets are small and were extracted from only two species.

The run time of **RNAsnoop** is comparable to that of **snoGPS**, and scales linearly with the length of the target sequence. Together with the improved accuracy, this makes **RNAsnoop** not only suitable for target search in rRNA and snRNA sequences or in specific putative mRNA candidates, but also for large-scale genome-wide surveys.

## Conclusion

Thanks to bioinformatics and experimental methods the number of known ncRNAs sequences has risen over 29,000,000. Still, besides the fact that a majority of ncRNAs exert their function through binding to other RNAs, only little is known about their functions. In this work, we extended RNA-folding algorithms to the problem of RNA-RNA interactions to accurately and rapidly predict ncRNA targets and, as a consequence, obtain functional annotation of ncRNAs.

The RNA-RNA interaction tools that were developed in this work can be divided into general and specialized approaches. General approaches, like **RNAup** and **RNAplex**, only assume that the interactions involve a continuous stretch of nucleotides on both interacting sequences. These algorithms capture the most common types of interaction between regulatory RNAs and their targets, even though more complicated types of interactions, such as H/ACA snoRNA with their target rRNAs or OxyS-fhlA, are neglected.

In contrast, approaches like **RNAxs** and **RNAsnoop** are especially designed to study a given type of RNA-RNA interactions. **RNAxs** utilizes besides accessibility a series of biologically comprehensible design criteria describing distinct stages in the RNAi pathway. **RNAsnoop** is specifically designed to predict complex pseudoknotted H/ACA snoRNA-RNA interaction. It uses a dynamic programming approach with a nearest-neighbor energy model to identify putative targets.

The performance of the presented algorithm can be regarded as good. **RNAup** and **RNAplex** retrieved successfully all sRNA-mRNA interactions studied and ranked known interactions high. **RNAxs** performs on par with the best siRNA design tools

available. **RNASnoop** outperforms previously published methods and successfully classifies functional from non-functional snoRNA-RNA interactions.

A central feature influencing all classes of interactions reviewed in this work, was accessibility. Its consideration allows to dramatically improve the performances of **RNAplex**, **RNASnoop** as well as siRNA design.

Due to the large amount of ncRNAs data, the approaches presented here are not only optimized for predictions accuracy but also for scanning speed. Compared to the method developed by [2] that can theoretically handle the complex snoRNA-rRNA interactions, **RNASnoop** is many order of magnitude faster, with a runtime of  $\mathcal{O}(|x||y|^2 + |y|^4)$  compared to  $\mathcal{O}((|x| + |y|)^6)$  for [2], where  $|x|$  and  $|y|$  are the target and snoRNA lengths, respectively. Even though **RNAplex** and **RNAup** have similar prediction performances, **RNAplex** returns its predictions about 2000 times faster than **RNAup**. As such **RNAplex** and **RNASnoop** tools are able to cope with the task of searching targets for the rapidly growing number of ncRNAs.

While the requirement of developing fast and accurate methods to find putative ncRNA targets could be met satisfactorily, it should be stressed that the ncRNA target search field is far from being closed. A general limitation is the lack of knowledge concerning the energetics of RNA-RNA interactions within loops: the binding of the oligo to a loop will of course alter the energy contribution of the loop itself. In this work we implicitly assume that this energy change is a constant. Additional measurement along the lines of the investigation of kissing-interactions are required to improve the energy parameters for interacting RNAs. A further concern is whether the underlying assumption of thermodynamically controlled binding is correct; it is possible that in particular when RNA binding is associated with large structural changes, kinetic effects of structure formation might be important.

Correct predictions of ncRNA targets is not only a function of thermodynamics of RNA-RNA interaction but also depends on protein factors. Proteins may sit on the predicted target region of a miRNA and impede its function. **Hfq** can promote thermodynamically unfavorable interactions between a sRNA and its target.

Concentration dependence of the solutes is a further factor to be taken into account. Event hough the theoretical framework of the solution dependence of finite length RNA hybridization has already been studied [45] and implemented [21, 38, 160], only

few experimental data are currently available [38].

Successful predictions of ncRNA targets should further consider the metabolic pathways and cellular processes in which the sRNA and the targets are found. ncRNAs should not only be regarded as mere protein regulators, but rather as cellular process regulators. As such, true ncRNA targets should share similar gene ontologies and be involved in related metabolic pathways. Classical examples thereof are targets of microRNA miR-134 that are heavily involved in neuron development [67, 215, 258, 265] as well as targets of RyhB in bacteria that are involved in iron regulation [112, 163–165, 199, 228].



# Appendices





# List of Symbols

Table A.1: List of Symbols

symbol	meaning
$s$	sequence $s$
$s_i$	$i$ th nucleotide of sequence $s$
$(s_i, s_j)$	base-pair between nucleotide $s_i$ and $s_j$
$s[i..j]$	subsequence on $s$ contained between nucleotides $s_i$ and $s_j$
$\mathcal{S}$	secondary structure
$\mathcal{S}_y^x$	hybrid structure of sequence $x$ with sequence $y$
$Z$	equilibrium partition function
$\beta$	inverse of the temperature times Boltzmann's constant
$P_u[i, j]$	probability that the sequence interval $s[i..j]$ is unpaired
$\mathcal{S}_{[i, j]}^u$	set of secondary structures in which $s[i..j]$ remains unpaired
$P_{ij}$	base pairing probability of pair $(s_i, s_j)$
$\hat{Z}(p, q)$	partition function outside base pair $(s_p, s_q)$
$Z_{pq}[i, j]$	partition function inside a base pair $(s_p, s_q)$ given that the given that the interval $s[i..j]$ is unpaired
$H(p, q)$	loop energies of hairpin loops given their enclosing base pairs $(s_p, s_q)$
$I(p, q; k, l)$	loop energies of interior loops given their enclosing base pairs $(s_p, s_q)$ and $(s_k, s_l)$ ;

$Z^m[p, q]$	partition function of all conformations on the interval $s[p..q]$ that are part of a multiloop and contain at least one component
$Z^{m1}[p, q]$	partition function of all multiloops on the interval $s[p..q]$ that have <i>exactly</i> one component
$Z^{m2}[p, q]$	functions of multiloop configurations that have <i>at least</i> two components
$w_x \mathbf{P}_u^x[i]$	probability that nucleotide $x_i$ is unpaired given that subsequence $x[i - w_x..i - 1]$ is unpaired
$\Delta^{w_x} \mathbf{G}_u^x[i]$	energy necessary to remove $x_i$ from all intramolecular interaction, given that subsequence $x[i - w_x..i - 1]$ is unpaired
$C_{i,j}$	best energy of interaction between sub-sequence $x[1..i]$ and $y[j..m]$
$B_{i,j}^{x,y}$	best energy of interaction given that residue $y_j$ , respectively $x_i$ , is aligned to a bulge
$I_{i,j}$	best energy of interaction given that $x_i$ and $y_j$ are in an interior loop
$\mathcal{S}(i, j, k, l)$	the energy gained by stacking the $(x_i y_j)$ base pair onto the $(x_k, y_l)$ base pair
$\mathcal{M}(i, j; i - 1, j + 1)$	the “mismatch” energy of the unpaired nucleotides $(x_{i-1}, y_{j+1})$ adjacent to the pair $(x_i, y_j)$
$\mathcal{I}$	energy contribution of the small interior loops
$g_{\text{open}}^{B,I}$	gap-open penalty for interior and bulge loops, respectively
$g_{\text{ext}}^{B,I}$	gap-extension penalty for interior and bulge loops, respectively
$d_{1,2,3}^x$	energy contribution to free 1,2 or 3 consecutive nucleotides on sequence $x$ in RNAPlex
$\mathbb{X}$	multiple sequences alignment
$\mathbb{X}_i$	$i^{th}$ column of alignment $\mathbb{X}$
$\mathbb{X}^\alpha$	$\alpha^{th}$ -sequence in the alignment $\mathbb{X}$
$d_1^{\mathbb{X}}$	$\sum_{\alpha=1}^N d_1^{\mathbb{X}^\alpha} = \sum_{\alpha=1}^N \Delta^4 \mathbf{G}_u^{\mathbb{X}^\alpha}[i]$

# B

## Bibliography

### Bibliography

---

- [1] T Akutsu. Dynamic programming algorithms for RNA secondary structure with pseudoknots. *Discrete Applied Mathematics*, 104:45–62, 2000.
- [2] C Alkan, E Karakoc, J Nadeau, S Sahinalp, and K Zhang. RNA-RNA interaction prediction and antisense RNA target search. *J Comput Biol*, 13(2):267–282, Mar 2006.
- [3] S Altschul, W Gish, W Miller, E Myers, and D Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.
- [4] S Altuvia, A Zhang, L Argaman, A Tiwari, and G Storz. The escherichia coli oxys regulatory RNA represses fhla translation by blocking ribosome binding. *EMBO J*, 17(20):6069–6075, Oct 1998.
- [5] M Amarzguioui and H Prydz. An algorithm for selection of functional siRNA sequences. *Biochem Biophys Res Commun*, 316(4):1050–1058, Apr 2004.
- [6] S Ameres, J Martinez, and R Schroeder. Molecular basis for target RNA recognition and cleavage by human RISC. *Cell*, 130(1):101–112, 2007.
- [7] M Andronescu, R Aguirre-Hernandez, A Condon, and H Hoos. RNAsoft: A suite of RNA secondary structure prediction and design software tools. *Nucleic Acids Res.*, 31(13):3416–3422, 2003.

- [8] L Argaman and S Altuvia. fhla repression by oxys RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. *J Mol Biol*, 300(5):1101–1112, Jul 2000.
- [9] M Ashburner and R Drysdale. Flybase—the drosophila genetic database. *Development*, 120(7):2077–2079, Jul 1994.
- [10] J Bachellerie, J Cavaillé, and A Hüttenhofer. The expanding snoRNA world. *Biochimie*, 84:775–790, 2002.
- [11] M Bally, J Hughes, and G Cesareni. Snr30: a new, essential small nuclear RNA from *saccharomyces cerevisiae*. *Nucleic Acids Res*, 16(12):5291–5303, Jun 1988.
- [12] D Banerjee and F Slack. Control of developmental timing by small temporal RNAs: a paradigm for RNA-mediated regulation of gene expression. *Bioessays*, 24(2):119–129, Feb 2002.
- [13] D Bartel and C Chen. Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. *Nat Rev Genet*, 5(5):396–400, May 2004.
- [14] P Bazeley, V Shepelev, Z Talebizadeh, M Butler, L Fedorova, V Filatov, and A Fedorov. snoTARGET shows that human orphan snoRNA targets locate close to alternative splice junctions. *Gene*, 408:172–179, 2008.
- [15] R Benne. RNA editing in trypanosomes. the us(e) of guide RNAs. *Mol Biol Rep*, 16(4):217–227, Sep 1992.
- [16] I Bentwich. Prediction and validation of microRNAs and their targets. *FEBS Lett*, Sep 2005.
- [17] S Bernhart. *Variations of RNA folding - Locally stable structures and RNA hybridization*. PhD thesis, Vienna University, 2007.
- [18] S Bernhart, I Hofacker, and P Stadler. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–615, Mar 2006.
- [19] S Bernhart, I Hofacker, S Will, A Gruber, and P Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:474, 2008.

- [20] S Bernhart, H Tafer, U Mückstein, C Flamm, P Stadler, and I Hofacker. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol*, 1(1):3–3, 2006.
- [21] S Bernhart, H Tafer, U Mückstein, C Flamm, P Stadler, and I Hofacker. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol*, 1(1):3–3, 2006.
- [22] S Bhattacharyya, R Habermacher, U Martine, E Closs, and W Filipowicz. Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell*, 125(6):1111–1124, Jun 2006.
- [23] Q Boese, D Leake, A Reynolds, S Read, S Scaringe, W Marshall, and A Khvorova. Mechanistic insights aid computational short interfering RNA design. *Methods Enzymol*, 392:73–96, 2005.
- [24] E Bohula, A Salisbury, M Sohail, M Playford, J Riedemann, E Southern, and V Macaulay. The efficacy of small interfering RNAs targeted to the type 1 insulin-like growth factor receptor (IGF1R) is influenced by secondary structure in the IGF1R transcript. *J. Biol. Chem.*, 278(18):15991–15997, 2003.
- [25] S Boisset, T Geissmann, E Huntzinger, P Fechter, N Bendridi, M Possedko, C Chevalier, A Helfer, Y Benito, A Jacquier, C Gaspin, F Vandenesch, and P Romby. Staphylococcus aureus RNaiii coordinately represses the synthesis of virulence factors and the transcription regulator rot by an antisense mechanism. *Genes Dev*, 21(11):1353–1366, Jun 2007.
- [26] A Bompfünewerer, R Backofen, S Bernhart, J Hertel, I Hofacker, P Stadler, and S Will. Variations on RNA folding and alignment: lessons from benasque. *J Math Biol*, 56(1-2):129–144, Jan 2008.
- [27] A F Bompfünewerer, R Backofen, S H Bernhart, C Flamm, C Fried, G Fritzsche, J Hackermüller, J Hertel, I L Hofacker, K Missal, A Mosig, S J Prohaska, D Rose, P F Stadler, A Tanzer, S Washietl, and S Will. Rnas everywhere: genome-wide annotation of structured rnas. *J Exp Zoolog B Mol Dev Evol*, 308(1):1–25, Jan 2007.
- [28] J Brennecke and S Cohen. Towards a complete description of the microRNA complement of animal genomes. *Genome Biol*, 4(9):228–228, 2003.

- [29] J Brennecke, A Stark, R Russell, and M Cohen. Principles of microRNA-target recognition. *PLoS Biol*, 3(3):e85, Mar 2005.
- [30] C Brescia, P Mikulecky, A Feig, and D Sledjeski. Identification of the hfq-binding site on dsra RNA: Hfq binds without altering dsra secondary structure. *RNA*, 9(1):33–43, Jan 2003.
- [31] P Brion and E Westhof. Hierarchy and dynamics of RNA folding. *Annu Rev Biophys Biomol Struct*, 26:113–137, 1997.
- [32] J Brown and P Sanseau. A computational view of microRNAs and their targets. *Drug Discov Today*, 10(8):595–601, Apr 2005.
- [33] R Brucoleri and G Heinrich. An improved algorithm for nucleic acid secondary structure display. *Comput Appl Biosci*, 4(1):167–173, Mar 1988.
- [34] M Brynildsen and J Liao. An integrated network approach identifies the isobutanol response network of escherichia coli. *Mol Syst Biol*, 5:277–277, 2009.
- [35] C Chen, R Perasso, L Qu, and L Amar. Exploration of pairing constraints identifies a 9 base-pair core within box c/d snoRNA-rRNA duplexes. *J Mol Biol*, 369(3):771–783, Jun 2007.
- [36] H Chen, B Fan, C Zhao, L Xie, C Zhao, T Zhou, K Lee, and G Allaway. Computational studies and drug design for hiv-1 reverse transcriptase inhibitors of 3',4'-di-o-(s)-camphanoyl-(+)-cis-khellactone (dck) analogs. *J Comput Aided Mol Des*, 19(4):243–258, Apr 2005.
- [37] S Chen, A Zhang, L Blyn, and G Storz. MicC, a second small-RNA regulator of Omp protein expression in Escherichia coli. *J Bacteriol*, 186(20):6689–6697, 2004.
- [38] H Chitsaz, R Backofen, and S Sahinalp. *Algorithms in Bioinformatics*, volume 5724 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.
- [39] C Croce and G Calin. miRNAs, cancer, and stem cell division. *Cell*, 122(1):6–7, Jul 2005.

- [40] D Crothers, P Cole, C Hilbers, and R Shulman. The molecular mechanism of thermal unfolding of escherichia coli formylmethionine transfer RNA. *J Mol Biol*, 87(1):63–88, Jul 1974.
- [41] A Denli, B Tops, R Plasterk, R Ketting, and G Hannon. Processing of primary microRNAs by the microprocessor complex. *Nature*, 432(7014):231–235, Nov 2004.
- [42] G Desnoyers, A Morissette, K Prévost, and E Massé. Small rna-induced differential degradation of the polycistronic mrna iscrsua. *EMBO J*, 28(11):1551–1561, Jun 2009.
- [43] D Didiano and O Hobert. Molecular architecture of a miRNA-regulated 3' utr. *RNA*, 14(7):1297–1317, Jul 2008.
- [44] R Dimitrov and M Zuker. Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys. J.*, 87:215–226, 2004.
- [45] R Dimitrov and M Zuker. Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys. J.*, 87:215–226, 2004.
- [46] Y Ding, C Chan, and C Lawrence. Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Research*, 32(Web Server issue):W135–141, 2004.
- [47] Y Ding and C Lawrence. Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. *Nucl. Acids Res.*, 29:1034–1046, 2001.
- [48] Y Ding and C Lawrence. A statistical sampling algorithm for RNA secondary structure prediction. *Nucleic Acids Res.*, 31:7280–7301, 2003.
- [49] M Dinger, K Pang, T Mercer, and J Mattick. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol*, 4(11), Nov 2008.
- [50] R Dirks and N Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *J. Comput. Chem.*, 24(13):1664–1677, 2003.
- [51] R Dirks and N Pierce. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. *J. Comput. Chem.*, 25(10):1295–1304, 2004.

- [52] J Doench, C Petersen, and P Sharp. siRNAs can function as miRNAs. *Genes Dev*, 17(4):438–442, Feb 2003.
- [53] J Doench and P Sharp. Specificity of microRNA target selection in translational repression. *Genes Dev*, 18(5):504–511, Mar 2004.
- [54] K Doshi, J Cannone, C Cobaugh, and R Gutell. Evaluation of the suitability of free-energy minimization using nearest-neighbor energy parameters for RNA secondary structure prediction. *BMC Bioinformatics*, 5:105–105, Aug 2004.
- [55] P Doty, J Marmur, J Eigner, and C Schildkraut. Strand separation and specific recombination in deoxyribonucleic acids: Physical chemical studies. *Proc Natl Acad Sci U S A*, 46(4):461–476, Apr 1960.
- [56] A Duursma, M Kedde, M Schrier, C leSage, and R Agami. mir-148 targets human dnmt3b protein coding region. *RNA*, 14(5):872–877, May 2008.
- [57] D Dykxhoorn, C Novina, and P Sharp. Killing the messenger: short RNAs that silence gene expression. *Nat. Rev. Mol. Cell Biol.*, 4(6):457–467, 2003.
- [58] G Easow, A Teleman, and S Cohen. Isolation of microRNA targets by mirnp immunopurification. *RNA*, 13(8):1198–1204, Aug 2007.
- [59] R Edgar. Muscle: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113–113, Aug 2004.
- [60] Y Eguchi and J Tomizawa. Complex formed by complementary RNA stem-loops and its stabilization by a protein: function of coe1 rom protein. *Cell*, 60(2):199–209, Jan 1990.
- [61] S Elbashir, J Harborth, W Lendeckel, A Yalcin, K Weber, and T Tuschl. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, 411(6836):494–498, May 2001.
- [62] S Elbashir, H J, K Weber, and T Tuschl. Analysis of gene function in somatic mammalian cells using small interfering RNAs. *Methods*, 26(2):199–213, 2002.
- [63] S Elbashir, J Martinez, A Patkaniowska, W Lendeckel, and T Tuschl. Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO J*, 20(23):6877–6888, Dec 2001.

- [64] C Ender, A Krek, M Friedländer, M Beitzinger, L Weinmann, W Chen, S Pfeffer, N Rajewsky, and G Meister. A human snoRNA with microRNA-like functions. *Mol Cell*, 32(4):519–528, Nov 2008.
- [65] B R Ernsting, M R Atkinson, A J Ninfa, and R G Matthews. Characterization of the regulon controlled by the leucine-responsive regulatory protein in escherichia coli. *J Bacteriol*, 174(4):1109–1118, Feb 1992.
- [66] E Fayet-Lebaron, V Atzorn, Y Henry, and T Kiss. 18s rRNA processing requires base pairings of snr30 h/aca snoRNA to eukaryote-specific 18s sequences. *EMBO J*, 28(9):1260–1270, May 2009.
- [67] R Fiore, S Khudayberdiev, M Christensen, G Siegel, S Flavell, T Kim, M Greenberg, and G Schratt. Mef2-mediated transcription of the mir379-410 cluster regulates activity-dependent dendritogenesis by fine-tuning pumilio2 protein levels. *EMBO J*, 28(6):697–710, Mar 2009.
- [68] A Fire, S Xu, M Montgomery, S Kostas, S Driver, and C Mello. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature*, 391(6669):806–811, Feb 1998.
- [69] W Fontana, D Konings, P Stadler, and P Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33(9):1389–1404, Sep 1993.
- [70] W Fontana, D Konings, P Stadler, and P Schuster. Statistics of RNA secondary structures. *Biopolymers*, 33(9):1389–1404, Sep 1993.
- [71] E Freyhult, S Edvardsson, I Tamas, V Moulton, and A Poole. Fisher: a program for the detection of h/aca snoRNAs using mfe secondary structure prediction and comparative genomics - assessment and update. *BMC Res Notes*, 1:49, 2008.
- [72] R Friedman, K Farh, C Burge, and D Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19(1):92–105, Jan 2009.
- [73] D Gaidatzis, E vanNimwegen, J Hausser, and M Zavolan. Inference of miRNA targets using evolutionary conservation and pathway analysis. *BMC Bioinformatics*, 8:69–69, 2007.
- [74] P Ganot, M Bortolin, and T Kiss. Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. *Cell*, 89(5):799–809, May 1997.

- [75] P Gardner, J Daub, J Tate, E Nawrocki, D Kolbe, S Lindgreen, A Wilkinson, R Finn, S Griffiths-Jones, S Eddy, and A Bateman. Rfam: updates to the RNA families database. *Nucleic Acids Res*, 37(Database issue):136–140, Jan 2009.
- [76] D Gautheret, S Damberger, and R Gutell. Identification of base-triples in RNA using comparative sequence analysis. *J Mol Biol*, 248(1):27–43, Apr 1995.
- [77] D Gautheret, D Konings, and R Gutell. A major family of motifs involving g.a mismatches in ribosomal RNA. *J Mol Biol*, 242(1):1–8, Sep 1994.
- [78] T Geissmann and D Touati. Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J*, 23(2):396–405, 2004.
- [79] W Gerlach and R Giegerich. Guugle: a utility for fast exact matching under RNA complementary rules including g-u base pairing. *Bioinformatics*, 22(6):762–764, Mar 2006.
- [80] E Giordano, I Peluso, S Senger, and M Furia. minify, a drosophila gene required for ribosome biogenesis. *J Cell Biol*, 144(6):1123–1133, Mar 1999.
- [81] A Giraldez, R Cinalli, M Glasner, A Enright, J Thomson, S Baskerville, S Hammond, D Bartel, and A Schier. MicroRNAs regulate brain morphogenesis in zebrafish. *Science*, 308(5723):833–838, May 2005.
- [82] A Giraldez, Y Mishima, J Rihel, R Grocock, S VanDongen, K Inoue, A Enright, and A Schier. Zebrafish mir-430 promotes deadenylation and clearance of maternal mRNAs. *Science*, 312(5770):75–79, Apr 2006.
- [83] S Gottesman. Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet*, 21(7):399–404, 2005.
- [84] R Gregory, T Chendrimada, N Cooch, and R Shiekhattar. Human risc couples microRNA biogenesis and posttranscriptional gene silencing. *Cell*, 123(4):631–640, Nov 2005.
- [85] S Griffiths-Jones. The microRNA registry. *Nucleic Acids Res*, 32(Database issue):109–111, Jan 2004.

- [86] A Grimson, K Farh, W Johnston, P Garrett-Engele, L Lim, and D Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 27(1):91–105, Jul 2007.
- [87] A R Gruber, D Koper-Emde, M Marz, H Tafer, S Bernhart, G Obernosterer, A Mosig, I L Hofacker, P F Stadler, and B J Benecke. Invertebrate 7sk snrnas. *J Mol Evol*, 66(2):107–115, Feb 2008.
- [88] C Guerrier-Takada, K Gardiner, T Marsh, N Pace, and S Altman. The RNA moiety of ribonuclease p is the catalytic subunit of the enzyme. *Cell*, 35(3 Pt 2):849–857, Dec 1983.
- [89] R Gutell, J Cannone, Z Shang, Y Du, and M Serra. A story: unpaired adenosine bases in ribosomal RNAs. *J Mol Biol*, 304(3):335–354, Dec 2000.
- [90] R Gutell, M Gray, and M Schnare. A compilation of large subunit (23s and 23s-like) ribosomal RNA structures: 1993. *Nucleic Acids Res*, 21(13):3055–3074, Jul 1993.
- [91] J Hackermüller, N Meisner, M Auer, M Jaritz, and P Stadler. The effect of RNA secondary structures on RNA-ligand binding and the modifier RNA mechanism: A quantitative model. *Gene*, 345:3–12, 2005.
- [92] C Haro and J Santoyo. The eIF-2 $\alpha$  kinases and the control of protein synthesis. *FASEB J*, 10(12):1378–1387, Oct 1996.
- [93] L He and G Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet*, 5(7):522–531, Jul 2004.
- [94] J Hertel, D deJong, M Marz, D Rose, H Tafer, A Tanzer, B Schierwater, and P Stadler. Non-coding RNA annotation of the genome of trichoplax adhaerens. *Nucleic Acids Res*, 37(5):1602–1615, Apr 2009.
- [95] O Hobert. Architecture of a microRNA-controlled gene regulatory network that diversifies neuronal cell fates. *Cold Spring Harb Symp Quant Biol*, 71:181–188, 2006.
- [96] M Höchsmann, T Töller, R Giegerich, and S Kurtz. Local similarity in RNA secondary structures. *Proc IEEE Comput Soc Bioinform Conf*, 2:159–168, 2003.
- [97] I Hofacker. How microRNAs choose their targets. *Nat Genet*, 39(10):1191–1192, Oct 2007.

- [98] I Hofacker, M Fekete, and P Stadler. Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, 319(5):1059–1066, Jun 2002.
- [99] I Hofacker, M Fekete, and P Stadler. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.*, 319:1059–1066, 2002.
- [100] I Hofacker, W Fontana, P Stadler, S Bonhoeffer, M Tacker, and P Schuster. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.*, 125:167–188, 1994.
- [101] I L Hofacker and P F Stadler. RNA secondary structures. *unpublished*, 2004.
- [102] I L Hofacker and H Tafer. Designing optimal sirna based on target site accessibility. *Methods Mol Biol*, 623:137–154, 2010.
- [103] H Hohjoh. Enhancement of RNAi activity by improved siRNA duplexes. *FEBS Lett*, 557(1-3):193–198, Jan 2004.
- [104] T Holen, M Amarzguioui, M Wiiger, E Babaie, and H Prydz. Positional effects of short interfering RNAs targeting the human coagulation trigger Tissue Factor. *Nucleic Acids Res*, 30(8):1757–1766, Apr 2002.
- [105] V Hornung, M Guenthner-Biller, C Bourquin, A Ablasser, M Schlee, S Uematsu, A Noronha, M Manoharan, S Akira, A d Fougerolles, S Endres, and G Hartmann. Sequence-specific potent induction of IFN- $\alpha$  by short interfering RNA in plasmacytoid dendritic cells through TLR7. *Nat Med*, 11(3):263–270, Mar 2005.
- [106] A Hsieh, R Bo, J Manola, F Vazquez, O Bare, A Khvorova, S Scaringe, and W Sellers. A library of siRNA duplexes targeting the phosphoinositide 3-kinase pathway: determinants of gene silencing for use in cell-based screens. *Nucleic Acids Res*, 32(3):893–901, 2004.
- [107] F Huang, J Qin, C Reidys, and P Stadler. Target prediction and a statistical sampling algorithm for RNA-RNA interaction. *Bioinformatics (Oxford, England)*, 26(2):175–81, January 2010.
- [108] D Huesken, J Lange, C Mickanin, J Weiler, F Asselbergs, J Warner, B Meloon, S Engel, A Rosenberg, D Cohen, M Labow, M Reinhardt, F Natt, and J Hall. Design of a genome-wide siRNA library using an artificial neural network. *Nat Biotechnol*, 23(8):995–1001, Aug 2005.

- [109] T H chsmann, M H chsmann, and R Giegerich. Thermodynamic matchers: strengthening the significance of RNA folding energies. *Comput Syst Bioinformatics Conf*, pages 111–121, 2006.
- [110] A Jackson, S Bartz, J Schelter, S Kobayashi, J Burchard, M Mao, B Li, G Cavet, and P Linsley. Expression profiling reveals off-target gene regulation by RNAi. *Nat Biotechnol*, 21(6):635–637, Jun 2003.
- [111] F Jacob and J Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3:318–356, Jun 1961.
- [112] J Jacques, S Jang, K Prévost, G Desnoyers, M Desmarais, J Imlay, and E Massé. Ryhb small RNA modulates the free intracellular iron pool and is essential for normal growth during iron limitation in escherichia coli. *Mol Microbiol*, 62(4):1181–1190, Nov 2006.
- [113] B John, A Enright, A Aravin, T Tuschl, C Sander, and D Marks. Human microRNA targets. *PLoS Biol*, 2(11), Nov 2004.
- [114] R Johnston and O Hobert. A microRNA controlling left/right neuronal asymmetry in caenorhabditis elegans. *Nature*, 426(6968):845–849, Dec 2003.
- [115] H Kawamoto, Y Koide, T Morita, and H Aiba. Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq. *Mol Microbiol*, 61(4):1013–1022, 2006.
- [116] M Kedde and R Agami. Interplay between microRNAs and RNA-binding proteins determines developmental processes. *Cell Cycle*, 7(7):899–903, Apr 2008.
- [117] M Kertesz, N Iovino, U Unnerstall, U Gaul, and E Segal. The role of site accessibility in microRNA target recognition. *Nat Genet*, 39(10):1278–1284, 2007.
- [118] A Khvorova, A Reynolds, and S Jayasena. Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115(2):209–216, Oct 2003.
- [119] D H Kim, L M Villeneuve, K V Morris, and J J Rossi. Argonaute-1 directs siRNA-mediated transcriptional gene silencing in human cells. *Nat Struct Mol Biol*, 13(9):793–797, Sep 2006.

- [120] S Kishore and S Stamm. Regulation of alternative splicing by snoRNAs. *Cold Spring Harb Symp Quant Biol*, 71:329–334, 2006.
- [121] T Kiss. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J*, 20(14):3617–3622, Jul 2001.
- [122] Z Kiss-László, Y Henry, and T Kiss. Sequence and structural elements of methylation guide snoRNAs essential for site-specific ribose methylation of pre-rRNA. *EMBO J*, 17(3):797–807, Feb 1998.
- [123] W Kloosterman, E Wienholds, R Ketting, and R Plasterk. Substrate requirements for let-7 function in the developing zebrafish embryo. *Nucleic Acids Res*, 32(21):6284–6291, 2004.
- [124] F Kolb, E Westhof, C Ehresmann, B Ehresmann, E Wagner, and P Romby. Bulged residues promote the progression of a loop-loop interaction to a stable and inhibitory antisense-target RNA complex. *Nucleic Acids Res*, 29(15):3145–3153, Aug 2001.
- [125] A Krasilnikov, Y Xiao, T Pan, and A Mondragón. Basis for structural diversity in homologous RNAs. *Science*, 306(5693):104–107, Oct 2004.
- [126] A Krek, D Grün, M Poy, R Wolf, L Rosenberg, E Epstein, P MacMenamin, I daPiedade, K Gunsalus, M Stoffel, and N Rajewsky. Combinatorial microRNA target predictions. *Nat Genet*, 37(5):495–500, May 2005.
- [127] R Kretschmer-KazemiFar and G Sczakiel. The activity of siRNA in mammalian cells is related to structural target accessibility: a comparison with antisense oligonucleotides. *Nucleic Acids Res*, 31(15):4417–4424, Aug 2003.
- [128] K Kruger, P Grabowski, A Zaug, J Sands, D Gottschling, and T Cech. Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. *Cell*, 31(1):147–157, Nov 1982.
- [129] J Kugel and J Goodrich. An RNA transcriptional regulator templates its own regulatory RNA. *Nat Chem Biol*, 3(2):89–90, Feb 2007.
- [130] M Lagos-Quintana, R Rauhut, W Lendeckel, and T Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 294(5543):853–858, Oct 2001.

- [131] E Lai. Micro RNAs are complementary to 3' utr sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet*, 30(4):363–364, Apr 2002.
- [132] E Lai, P Tomancak, R Williams, and G Rubin. Computational identification of Drosophila microRNA genes. *Genome Biol*, 4(7), 2003.
- [133] A Lal, H Kim, K Abdelmohsen, Y Kuwano, R Pullmann, S Srikantan, R Subrahmanyam, J Martindale, X Yang, F Ahmed, F Navarro, D Dykxhoorn, J Lieberman, and M Gorospe. p16(ink4a) translation suppressed by mir-24. *PLoS One*, 3(3), 2008.
- [134] S Lall, D Grün, A Krek, K Chen, Y Wang, C Dewey, P Sood, T Colombo, N Bray, P Macmenamin, H Kao, K Gunsalus, L Pachter, F Piano, and N Rajewsky. A genome-wide map of conserved microRNA targets in c. elegans. *Curr Biol*, 16(5):460–471, Mar 2006.
- [135] M Larkin, G Blackshields, N Brown, R Chenna, P McGettigan, H McWilliam, F Valentin, I Wallace, A Wilm, R Lopez, J Thompson, T Gibson, and D Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23:2947–2948, Nov 2007.
- [136] N Lau, L Lim, E Weinstein, and D Bartel. An abundant class of tiny RNAs with probable regulatory roles in caenorhabditis elegans. *Science*, 294(5543):858–862, Oct 2001.
- [137] R Lease and M Belfort. A trans-acting RNA as a control switch in escherichia coli: DsrA modulates function by forming alternative structures. *Proc Natl Acad Sci U S A*, 97(18):9919–9924, Aug 2000.
- [138] R A Lease, M E Cusick, and M Belfort. Riboregulation in Escherichia coli: DsrA RNA acts by RNA:RNA interactions at multiple loci. *Proc Natl Acad Sci U S A*, 95(21):12456–12461, 1998.
- [139] R Lee and V Ambros. An extensive class of small RNAs in caenorhabditis elegans. *Science*, 294(5543):862–864, Oct 2001.
- [140] R Lee, R Feinbaum, and V Ambros. The c. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–854, Dec 1993.
- [141] L Lestrade and M Weber. snoRNA-lbme-db, a comprehensive database of human h/aca and c/d box snoRNAs. *Nucleic Acids Res*, 34(Database issue):158–162, Jan 2006.

- [142] B Lewis, C Burge, and D Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, Jan 2005.
- [143] B Lewis, I Shih, M Jones-Rhoades, D Bartel, and C Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–798, Dec 2003.
- [144] L Lim, N Lau, P Garrett-Engele, A Grimson, J Schelter, J Castle, D Bartel, P Linsley, and J Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027):769–773, Feb 2005.
- [145] L Lim, N Lau, E Weinstein, A Abdelhakim, S Yekta, M Rhoades, C Burge, and D Bartel. The microRNAs of *Caenorhabditis elegans*. *Genes Dev*, 17(8):991–1008, Apr 2003.
- [146] W F Lima, B P Monia, D J Ecker, and S M Freier. Implication of RNA structure on antisense oligonucleotide hybridization kinetics. *Biochemistry*, 31(48):12055–12061, Dec 1992.
- [147] J Liu, M Carmell, F Rivas, C Marsden, J Thomson, J Song, S Hammond, L Joshua-Tor, and G Hannon. Argonaute2 is the catalytic engine of mammalian RNAi. *Science*, 305(5689):1437–1441, Sep 2004.
- [148] J Livny, A Brencic, S Lory, and M Waldor. Identification of 17 *Pseudomonas aeruginosa* sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2. *Nucleic Acids Res*, 34(12):3484–3493, 2006.
- [149] D Long, C Y Chan, and Y Ding. Analysis of microRNA-target interactions by a target structure based hybridization model. *Pac Symp Biocomput*, pages 64–74, 2008.
- [150] Z Lu, D Turner, and D Mathews. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucleic Acids Res.*, 34:4912–4924, 2006.
- [151] Z J Lu and D H Mathews. Efficient siRNA selection using hybridization thermodynamics. *Nucleic Acids Res*, 36(2):640–647, 2008.
- [152] Z J Lu and D H Mathews. Efficient siRNA selection using hybridization thermodynamics. *Nucleic Acids Res*, 36(2):640–647, Feb 2008.

- [153] K Luo and D Chang. The gene-silencing efficiency of siRNA is strongly dependent on the local structure of mRNA at the targeted region. *Biochem Biophys Res Commun*, 318(1):303–310, May 2004.
- [154] J Lytle, T Yario, and J Steitz. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5' utr as in the 3' utr. *Proc Natl Acad Sci U S A*, 104(23):9667–9672, Jun 2007.
- [155] E Maden and J Wakeman. Pseudouridine distribution in mammalian 18 s ribosomal RNA. a major cluster in the central region of the molecule. *Biochem J*, 249(2):459–464, Jan 1988.
- [156] N Majdalani, C Cunning, D Sledjeski, T Elliott, and S Gottesman. DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription. *Proc Natl Acad Sci U S A*, 95(21):12462–12467, 1998.
- [157] N Majdalani, D Hernandez, and S Gottesman. Regulation and mode of action of the second small RNA activator of RpoS translation, RprA. *Mol Microbiol*, 46(3):813–826, 2002.
- [158] P Mandin, F Repoila, M Vergassola, T Geissmann, and P Cossart. Identification of new noncoding RNAs in listeria monocytogenes and prediction of mRNA targets. *Nucleic Acids Res*, 35(3):962–974, 2007.
- [159] M Maragkakis, M Reczko, V Simossis, P Alexiou, G Papadopoulos, T Dalamagas, G Giannopoulos, G Goumas, E Koukis, K Kourtis, T Vergoulis, N Koziris, T Sellis, P Tsanakas, and A Hatzigeorgiou. Diana-microt web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res*, 37(Web Server issue):273–276, Jul 2009.
- [160] N Markham and M Zuker. Unafold: software for nucleic acid folding and hybridization. *Methods Mol Biol*, 453:3–31, 2008.
- [161] J T Marques and B R G Williams. Activation of the mammalian immune system by siRNAs. *Nat Biotechnol*, 23(11):1399–1405, Nov 2005.
- [162] M Marz, T Kirsten, and P Stadler. Evolution of spliceosomal snRNA genes in metazoan animals. *J. Mol. Evol.*, 2008. doi: 10.1007/s00239-008-9149-6.

- [163] E Massé, F Escorcia, and S Gottesman. Coupled degradation of a small regulatory RNA and its mRNA targets in escherichia coli. *Genes Dev*, 17(19):2374–2383, Oct 2003.
- [164] E Massé and S Gottesman. A small RNA regulates the expression of genes involved in iron metabolism in escherichia coli. *Proc Natl Acad Sci U S A*, 99(7):4620–4625, Apr 2002.
- [165] E Massé, C Vanderpool, and S Gottesman. Effect of ryhb small RNA on global iron use in escherichia coli. *J Bacteriol*, 187(20):6962–6971, Oct 2005.
- [166] D Mathews. Using an RNA secondary structure partition function to determine confidence in base pairs predicted by free energy minimization. *RNA*, 10(8):1178–1190, Aug 2004.
- [167] D Mathews, M Burkard, S Freier, J Wyatt, and D Turner. Predicting oligonucleotide affinity to nucleic acid targets. *RNA*, 5(11):1458–1469, Nov 1999.
- [168] D H Mathews, M E Burkard, S M Freier, J R Wyatt, and D H Turner. Predicting oligonucleotide affinity to nucleic acid targets. *RNA*, 5(11):1458–1469, 1999.
- [169] J McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29:1105–1119, 1990.
- [170] N Meisner, J Hackermüller, V Uhl, A Aszódi, M Jaritz, and M Auer. mRNA openers and closers: A methodology to modulate AU-rich element controlled mRNA stability by a molecular switch in mRNA conformation. *Chembiochem.*, 5:1432–1447, 2004.
- [171] N Meisner, J Hackermüller, V Uhl, A Aszódi, M Jaritz, and M Auer. mRNA openers and closers: A methodology to modulate AU-rich element controlled mRNA stability by a molecular switch in mRNA conformation. *Chembiochem.*, 5:1432–1447, 2004.
- [172] G Meister and T Tuschl. Mechanisms of gene silencing by double-stranded RNA. *Nature*, 431(7006):343–349, 2004.
- [173] N Milner, K U Mir, and E M Southern. Selecting effective antisense reagents on combinatorial oligonucleotide arrays. *Nat Biotechnol*, 15(6):537–541, Jun 1997.
- [174] K U Mir and E M Southern. Determining the influence of structure on hybridization using oligonucleotide arrays. *Nat Biotechnol*, 17(8):788–792, Aug 1999.

- [175] K Miranda, T Huynh, Y Tay, Y Ang, W Tam, A Thomson, B Lim, and I Rigoutsos. A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. *Cell*, 126(6):1203–1217, Sep 2006.
- [176] E Miska. MicroRNAs—keeping cells in formation. *Nat Cell Biol*, 10(5):501–502, May 2008.
- [177] V Mittal. Improving the efficiency of RNA interference in mammals. *Nat. Rev. Genet.*, 5(5):355–365, 2004.
- [178] T Moeller, T Franch, C Udesen, K Gerdes, and P Valentin-Hansen. Spot 42 RNA mediates discoordinate expression of the *E. coli* galactose operon. *Genes Dev*, 16(13):1696–1706, 2002.
- [179] I Moll, D Leitsch, T Steinhauser, and U Bläsi. RNA chaperone activity of the sm-like hfq protein. *EMBO Rep*, 4(3):284–289, Mar 2003.
- [180] U Muckstein, H Tafer, S Bernhart, M Hernandez-Rosales, J Vogel, P Stadler, and I Hofacker. Translational control by RNA-RNA interaction. *Submitted to BIRD 2008*, 2008.
- [181] U Mückstein, H Tafer, J Hackermüller, S Bernhart, P Stadler, and I Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10):1177–1182, May 2006.
- [182] U Mueckstein, H Tafer, S H Bernhart, M Hernandez-Rosales, J Vogel, P F Stadler, and I L Hofacker. Translational control by RNA-RNA interaction: Improved computation of RNA-RNA binding thermodynamics. *Communications in Computer and Information Science*, 13:114–127, 2008.
- [183] U Mueckstein, H Tafer, J Hackermueller, S H Bernhart, P F Stadler, and I L Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10):1177–1182, 2006.
- [184] G Myers. A fast bit-vector algorithm for approximate string matching based on dynamic programming. In *CPM '98: Proceedings of the 9th Annual Symposium on Combinatorial Pattern Matching*, pages 1–13, London, UK, 1998. Springer-Verlag.
- [185] J Ni, A Tien, and M Fournier. Small nucleolar RNAs direct site-specific synthesis of pseudouridine in ribosomal RNA. *Cell*, 89(4):565–573, May 1997.

- [186] R Nussinov, G Piecznik, J R Griggs, and D J Kleitman. Algorithms for loop matching. *SIAM J. Appl. Math.*, 35(1):68–82, 1978.
- [187] G Obernosterer, H Tafer, and J Martinez. Target site effects in the RNA interference and microRNA pathways. *Biochem Soc Trans*, 36(Pt 6):1216–1219, Dec 2008.
- [188] J Ofengand and A Bakin. Mapping to nucleotide resolution of pseudouridine residues in large subunit ribosomal RNAs from representative eukaryotes, prokaryotes, archaeobacteria, mitochondria and chloroplasts. *J Mol Biol*, 266(2):246–268, Feb 1997.
- [189] U Ørom, F Nielsen, and A Lund. MicroRNA-10a binds the 5’utr of ribosomal protein mRNAs and enhances their translation. *Mol Cell*, 30(4):460–471, May 2008.
- [190] M Overhoff, M Alken, R Far, M Lemaitre, B Lebleu, G Sczakiel, and I Robbins. Local RNA target structure influences siRNA efficacy: a systematic global analysis. *J Mol Biol*, 348(4):871–881, May 2005.
- [191] G Papadopoulos, M Reczko, V Simossis, P Sethupathy, and A Hatzigeorgiou. The database of experimentally supported targets: a functional update of tarbase. *Nucleic Acids Res*, 37(Database issue):155–158, Jan 2009.
- [192] J Parker, S Roe, and D Barford. Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. *Nature*, 434(7033):663–666, Mar 2005.
- [193] V Patzel. In silico selection of active siRNA. *Drug Discov Today*, 12(3-4):139–148, Feb 2007.
- [194] V Patzel, S Rutz, I Dietrich, C Köberle, A Scheffold, and S H E Kaufmann. Design of siRNAs producing unstructured guide-RNAs results in improved RNA interference efficiency. *Nat Biotechnol*, 23(11):1440–1444, Nov 2005.
- [195] W Pearson and D Lipman. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8):2444–2448, Apr 1988.
- [196] D Pervouchine. IRIS: Intermolecular RNA interaction search. *Proc. Genome Informatics*, 15:92–101, 2004.
- [197] N Peyret, P Seneviratne, H Allawi, and J SantaLucia. Nearest-neighbor thermodynamics and nmr of dna sequences with internal a.a, c.c, g.g, and t.t mismatches. *Biochemistry*, 38(12):3468–3477, Mar 1999.

- [198] E Pruesse, C Quast, K Knittel, B Fuchs, W Ludwig, J Peplies, and F Glöckner. Silva: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with arb. *Nucleic Acids Res*, 35(21):7188–7196, 2007.
- [199] K Pr vost, H Salvail, G Desnoyers, J Jacques, E Phaneuf, and E Mass . The small RNA ryhb activates the translation of shia mRNA encoding a permease of shikimate, a compound involved in siderophore synthesis. *Mol Microbiol*, 64(5):1260–1273, Jun 2007.
- [200] S C Pulvermacher, L T Stauffer, and G V Stauffer. Role of the srna gcvb in regulation of cyca in escherichia coli. *Microbiology*, 155(Pt 1):106–114, Jan 2009.
- [201] A A Rasmussen, M Eriksen, K Gilany, C Udesen, T Franch, C Petersen, and P Valentin-Hansen. Regulation of ompA mRNA stability: the role of a small regulatory RNA in growth phase-dependent control. *Mol Microbiol*, 58(5):1421–1429, 2005.
- [202] M Rederstorff, S Bernhart, A Tanzer, M Zywicki, K Perfler, M Lukasser, I Hofacker, and A Hüttenhofer. Rnpomics: Defining the ncRNA transcriptome by cdna library generation from ribonucleo-protein particles. *Nucleic Acids Res*, Feb 2010.
- [203] J Reeder, J Reeder, and R Giegerich. Locomotif: from graphical motif description to RNA motif search. *Bioinformatics*, 23(13):i392–i400, Jul 2007.
- [204] M Rehmsmeier, P Steffen, M Hochsmann, and R Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA.*, 10(10):1507–17, 2004.
- [205] B Reinhart, F Slack, M Basson, A Pasquinelli, J Bettinger, A Rougvie, H Horvitz, and G Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in caenorhabditis elegans. *Nature*, 403(6772):901–906, Feb 2000.
- [206] Y Ren, W Gong, Q Xu, X Zheng, D Lin, Y Wang, and T Li. siRecords: an extensive database of mammalian siRNAs with efficacy ratings. *Bioinformatics*, 22(8):1027–1028, Apr 2006.
- [207] A Reynolds, D Leake, Q Boese, S Scaringe, W Marshall, and A Khvorova. Rational siRNA design for RNA interference. *Nat. Biotechnol.*, 22(3):326–30, 2004.

- [208] H Robins, Y Li, and R Padgett. Incorporating structure to predict microRNA targets. *Proc Natl Acad Sci U S A*, 102(11):4006–4009, Mar 2005.
- [209] P Saetrom. Predicting the efficacy of short oligonucleotides in antisense and RNAi experiments with boosted genetic programming. *Bioinformatics*, 20(17):3055–3063, Nov 2004.
- [210] H Saini, S Griffiths-Jones, and A Enright. Genomic analysis of human microRNA transcripts. *Proc Natl Acad Sci U S A*, 104(45):17719–17724, Nov 2007.
- [211] D Samarsky, G Ferbeyre, E Bertrand, R Singer, R Cedergren, and M Fournier. A small nucleolar RNA:ribozyme hybrid cleaves a nucleolar RNA target in vivo with near-perfect efficiency. *Proc Natl Acad Sci U S A*, 96(12):6609–6614, Jun 1999.
- [212] A Saraiya and C Wang. snoRNA, a novel precursor of microRNA in giardia lamblia. *PLoS Pathog*, 4(11), Nov 2008.
- [213] P Schattner, W A Decatur, C A Davis, M Ares, M J Fournier, and T M Lowe. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *saccharomyces cerevisiae* genome. *Nucleic Acids Res*, 32(14):4281–4296, 2004.
- [214] D Schmitter, J Filkowski, A Sewer, R Pillai, E Oakeley, M Zavolan, P Svoboda, and W Filipowicz. Effects of dicer and argonaute down-regulation on mRNA levels in human hek293 cells. *Nucleic Acids Res*, 34(17):4801–4815, 2006.
- [215] G Schratt, F Tuebing, E Nigh, C Kane, M Sabatini, M Kiebler, and M Greenberg. A brain-specific microRNA regulates dendritic spine development. *Nature*, 439(7074):283–289, Jan 2006.
- [216] S Schubert, A Gruenweller, V A Erdmann, and J Kurreck. Local RNA target structure influences siRNA efficacy: systematic analysis of intentionally designed binding regions. *J Mol Biol*, 348(4):883–893, 2005.
- [217] B Schwanhausser, M Gossen, G Dittmar, and M Selbach. Global analysis of cellular protein translation by pulsed SILAC. *Proteomics*, 9(1):205–9, 2009.
- [218] D S Schwarz, G Hutvagner, T Du, Z Xu, N Aronin, and P D Zamore. Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, 115(2):199–208, Oct 2003.

- [219] S Seemann, J Gorodkin, and Backofen R. Unifying evolutionary and thermodynamic information for rna folding of multiple alignments. *Nucleic Acids Res.*, 36, 2008.
- [220] S Seemann, A Richter, J Gorodkin, and Backofen R. Hierarchical folding of multiple sequence alignments for the prediction of structures and rna-rna interactions. *Algorithms Mol Biol.*, 5, 2010.
- [221] M Selbach, B Schwanhäusser, N Thierfelder, Z Fang, R Khanin, and N Rajewsky. Widespread changes in protein synthesis induced by microRNAs. *Nature*, 455(7209):58–63, Sep 2008.
- [222] P Sethupathy, B Corda, and A G Hatzigeorgiou. TarBase: A comprehensive database of experimentally supported animal microRNA targets. *RNA*, 12(2):192–197, Feb 2006.
- [223] Y Shao, C Y Chan, A Maliyekkel, C E Lawrence, I B Roninson, and Y Ding. Effect of target secondary structure on RNAi efficiency. *RNA*, 13(10):1631–1640, Oct 2007.
- [224] B Shapiro. An algorithm for comparing multiple RNA secondary structures. *Comput Appl Biosci*, 4(3):387–393, Aug 1988.
- [225] B Shapiro, J Maizel, L Lipkin, K Currey, and C Whitney. Generating non-overlapping displays of nucleic acid secondary structure. *Nucleic Acids Res*, 12(1 Pt 1):75–88, Jan 1984.
- [226] C Sharma, S Hoffmann, F Darfeuille, J Reignier, S Findeisz, A Sittka, S Chabas, K Reiche, J Hackermuller, R Reinhardt, P Stadler, and J Vogel. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, advance on, February 2010.
- [227] C M Sharma, F Darfeuille, T H Plantinga, and J Vogel. A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. *Genes Dev*, 21(21):2804–2817, 2007.
- [228] Y Shimoni, G Friedlander, G Hetzroni, G Niv, S Altuvia, O Biham, and H Margalit. Regulation of gene expression by small non-coding RNAs: a quantitative view. *Mol Syst Biol*, 3:138–138, 2007.
- [229] A Stark, J Brennecke, R Russell, and S Cohen. Identification of *Drosophila* MicroRNA targets. *PLoS Biol*, 1(3), Dec 2003.

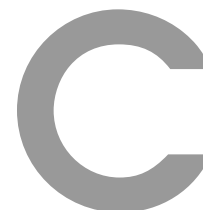
- [230] A Stark, P Kheradpour, L Parts, J Brennecke, E Hodges, G Hannon, and M Kellis. Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. *Genome Res*, 17:1865–1879, 2007.
- [231] C A Stein. Antisense that comes naturally. *Nat Biotechnol*, 19(8):737–738, Aug 2001.
- [232] R Stocsits, H Letsch, J Hertel, B Misof, and P Stadler. Accurate and efficient reconstruction of deep phylogenies from structured RNAs. *Nucleic Acids Res.*, 2009. in press.
- [233] H Tafer, S Ameres, G Obernosterer, C Gebeshuber, R Schroeder, J Martinez, and I Hofacker. The impact of target site accessibility on the design of potent siRNAs. *Nature Biotech.*, 26(5), 2008. in press.
- [234] H Tafer, S L Ameres, G Obernosterer, C A Gebeshuber, R Schroeder, J Martinez, and I L Hofacker. The impact of target site accessibility on the design of effective siRNAs. *Nat Biotechnol*, 26(5):578–583, May 2008.
- [235] H Tafer, J Hertel, I Hofacker, and P Stadler. **RNAsnoop**: Efficient search for H/ACA snoRNA targets. *Bioinformatics*, 2009. Accepted.
- [236] H Tafer and I L Hofacker. **RNAplex**: a fast tool for RNA-RNA interaction search. *Bioinformatics*, Apr 2008.
- [237] H Tafer, S Kehr, J Hertel, I Hofacker, and P Stadler. **RNAsnoop**: efficient target prediction for h/aca snoRNAs. *Bioinformatics*, 26(5):610–616, Mar 2010.
- [238] R Taft, E Glazov, T Lassmann, Y Hayashizaki, P Carninci, and J Mattick. Small RNAs derived from snoRNAs. *RNA*, 15(7):1233–1240, Jul 2009.
- [239] S Takasaki, S Kotani, and A Konagaya. An effective method for selecting siRNA target sequences in mammalian cells. *Cell Cycle*, 3(6):790–795, Jun 2004.
- [240] S Takyar, R Hickerson, and H Noller. mRNA helicase activity of the ribosome. *Cell*, 120(1):49–58, Jan 2005.
- [241] R Thadani and M Tammi. Microtar: predicting microRNA targets from RNA duplexes. *BMC Bioinformatics*, 7 Suppl 5, 2006.

- [242] J Thompson, D Higgins, and T Gibson. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, 22(22):4673–4680, Nov 1994.
- [243] I Tinoco, O Uhlenbeck, and M Levine. Estimation of secondary structure in ribonucleic acids. *Nature*, 230(5293):362–367, Apr 1971.
- [244] B Tjaden, S Goodwin, J Opdyke, M Guillier, D Fu, S Gottesman, and G Storz. Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Res.*, 34:2791–2802, 2006.
- [245] C Torchet, G Badis, F Devaux, G Costanzo, M Werner, and A Jacquier. The complete set of h/aca snoRNAs that guide rRNA pseudouridylations in *saccharomyces cerevisiae*. *RNA*, 11(6):928–938, Jun 2005.
- [246] D Tulpan, M Andronescu, S Chang, M Shortreed, A Condon, H Hoos, and L Smith. Thermodynamically based dna strand design. *Nucleic Acids Res*, 33(15):4951–4964, 2005.
- [247] D Turner and N Sugimoto. RNA structure prediction. *Annu Rev Biophys Biophys Chem*, 17:167–192, 1988.
- [248] K Ui-Tei, Y Naito, F Takahashi, T Haraguchi, H Ohki-Hamazaki, A Juni, R Ueda, and K Saigo. Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Res*, 32(3):936–948, 2004.
- [249] S Uliel, X Liang, R Unger, and S Michaeli. Small nucleolar RNAs that guide modification in trypanosomatids: repertoire, targets, genome organisation, and unique functions. *Int J Parasitol*, 34(4):445–454, Mar 2004.
- [250] J H Urban, K Papenfort, J Thomsen, R A Schmitz, and J Vogel. A conserved small RNA promotes discoordinate expression of the *glmUS* operon mRNA to activate *GlmS* synthesis. *J Mol Biol*, 373(3):521–528, Oct 2007.
- [251] P Valentin-Hansen, M Eriksen, and C Udesen. The bacterial Sm-like protein Hfq: a key player in RNA transactions. *Mol Microbiol*, 51(6):1525–1533, 2004.
- [252] F Vankuppeveld, W Melchers, K Kirkegaard, and J Doedens. Structure-function analysis of coxsackie b3 virus protein 2b. *Virology*, 227(1):111–118, Jan 1997.

- [253] R Vareková, I Bradác, M Plchút, M Skrdla, M Wacenovský, H Mahr, G Mayer, H Tanner, H Brugger, J Withalm, P Lederer, H Huber, G Gierlinger, R Graf, H Tafer, I Hofacker, P Schuster, and M Polcák. [www.rnaworkbench.com](http://www.rnaworkbench.com): A new program for analyzing RNA interference. *Comput Methods Programs Biomed*, 90(1):89–94, Apr 2008.
- [254] M C Vella, K Reinert, and F J Slack. Architecture of a validated microRNA::target interaction. *Chem Biol*, 11(12):1619–1623, Dec 2004.
- [255] M Velleca, M Wallace, and J Merlie. A novel synapse-associated noncoding RNA. *Mol Cell Biol*, 14(11):7095–7104, Nov 1994.
- [256] T Vickers, S Koo, C Bennett, S Crooke, N Dean, and B Baker. Efficient reduction of target RNAs by small interfering RNA and RNase H-dependent antisense agents. A comparative analysis. *J Biol Chem*, 278(9):7108–7118, Feb 2003.
- [257] T Vickers, J Wyatt, and S Freier. Effects of RNA secondary structure on cellular antisense activity. *Nucleic Acids Res*, 28(6):1340–1347, Mar 2000.
- [258] N Vo, M Klein, O Varlamova, D Keller, T Yamamoto, R Goodman, and S Impey. A camp-response element binding protein-induced microRNA regulates neuronal morphogenesis. *Proc Natl Acad Sci U S A*, 102(45):16426–16431, Nov 2005.
- [259] J Vogel, L Argaman, E Wagner, and S Altuvia. The small RNA istr inhibits synthesis of an sos-induced toxic peptide. *Curr Biol*, 14(24):2271–2276, Dec 2004.
- [260] J Vogel and C Sharma. How to find small non-coding RNAs in bacteria. *Biol Chem*, 386(12):1219–1238, Dec 2005.
- [261] C Wadler and C Vanderpool. A dual function for a bacterial small RNA: Sgrs performs base pairing-dependent regulation and encodes a functional polypeptide. *Proc Natl Acad Sci U S A*, 104(51):20454–20459, Dec 2007.
- [262] S Washietl, J Pedersen, J Korbelt, A Gruber, J Hackermüller, J Hertel, M Lindemeyer, K Reiche, C Stocsits, A Tanzer, C Ucla, C Wyss, S Antonarakis, F Denoeud, J Lagarde, J Drenkow, P Kapranov, T Gingeras, R Guigó, M Snyder, M Gerstein, A Reymond, I Hofacker, and P Stadler. Structured RNAs in the ENCODE selected regions of the human genome. *Gen. Res.*, 17:852–864, 2007.

- [263] M Waterman and T Smith. Combinatorics of RNA hairpins and cloverleaves. *STUDIES IN APPLIED MATHEMATICS*, 60:91–96, 1978.
- [264] M Waterman and T Smith. RNA secondary structure: a complete mathematical analysis. *MATHEMATICAL BIOSCIENCES*, 42:257–266, 1978.
- [265] G Wayman, M Davare, H Ando, D Fortin, O Varlamova, H Cheng, D Marks, K Obrietan, T Soderling, R Goodman, and S Impey. An activity-regulated microRNA controls dendritic plasticity by down-regulating p250gap. *Proc Natl Acad Sci U S A*, 105(26):9093–9098, Jul 2008.
- [266] S Will, K Reiche, I Hofacker, P Stadler, and R Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol*, 3(4), Apr 2007.
- [267] S Wuchty, W Fontana, I Hofacker, and P Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49:145–165, 1999.
- [268] M Xiao, C Yang, P Schattner, and Y Yu. Functionality and substrate specificity of human box h/aca guide RNAs. *RNA*, 15(1):176–186, Jan 2009.
- [269] Y Xu, H Zhang, D Thormeyer, O Larsson, Q Du, J Elmén, C Wahlestedt, and Z Liang. Effective small interfering RNAs and phosphorothioate antisense DNAs have different preferences for target sites in the luciferase mRNAs. *Biochem Biophys Res Commun*, 306(3):712–717, Jul 2003.
- [270] K Yoshinari, M Miyagishi, and T K. Effects on RNAi of the tight structure, sequence and position of the targeted region. *Nucleic Acids Res.*, 32(2):691–9, 2004.
- [271] Y Zeng, R Yi, and B Cullen. MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. *Proc Natl Acad Sci U S A*, 100(17):9779–9784, Aug 2003.
- [272] J Zhao and G Lemke. Rules for ribozymes. *Mol Cell Neurosci*, 11(1-2):92–97, May 1998.
- [273] D Zorio, K Lea, and T Blumenthal. Cloning of caenorhabditis u2af65: an alternatively spliced RNA containing a novel exon. *Mol Cell Biol*, 17(2):946–953, Feb 1997.

- [274] M Zuker. On finding all suboptimal foldings of an RNA molecule. *Science*, 7(244):48–52, 1989.
- [275] M Zuker. Calculating nucleic acid secondary structure. *Curr Opin Struct Biol*, 10(3):303–310, 2000.
- [276] M Zuker and S P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res.*, 9(1):133–148, 1981.
- [277] C Zwieb, F Müller, and N Larsen. Comparative analysis of tertiary structure elements in signal recognition particle RNA. *Fold Des*, 1(4):315–324, 1996.



# List of figures

## List of Figures

---

2.1	The major types of loops in RNA secondary structures. Taken from [101]	11
2.2	A variety of tree and forest representations of RNA secondary structures have been described in the literature. From left to right: conventional drawing, sequence annotated trees as used e.g. in <i>RNAforester</i> [96], “full tree” [70], Shapiro-style tree [224], and branching structure. For comparison, we also show the “bracket notation”. Taken from [101]	12
2.3	Representations of secondary structures. From left to right: Circular representation, Naview representation, mountain plot, dot plot. Remove the backbone edges from the first two representations leaves the matching $\Omega$ . Below, the structure is shown in “bracket notation”, where each base pair corresponds to a pair of matching parentheses. The structure shown is the purine riboswitch (Rfam RF00167) taken from [101]	13
2.4	<b>R.h.s</b> Minimum free energy (MFE) structure of <i>Crypthecodinium cohnii</i> 5.8S. Its free energy represents $-47.10kcal/mol$ . <b>L.h.s</b> Suboptimal folding of <i>Crypthecodinium cohnii</i> 5.8S sharing no base pair with the MFE structure. This structure has a free energy that differs by only $2.80kcal/mol$ from the MFE.	15
2.5	Plot of the interior loop free energies against the loop length for different loop asymmetries (red: no asymmetry, green: asymmetry of size 1, blue: asymmetry of size 2)	19

- 2.6 Loop decomposition of RNA secondary structure. Hairpin and interior loops are shown in red. Multiloop with more than one component are shown in blue, while multiloop with exactly one component are shown in green. Base Pairs are depicted by arcs. Dotted lines represent unpaired substructures. Taken from [101]. . . . . 20
- 2.7 Comparison of the minimum free energy of structures of dimers, depending on the kind of linker used to concatenate both sequences. Linkers are drawn in cyan, while the interacting sequences are colored in red and black. **Top:** Structure when using a “poly-N” linker. **Middle:** Structure when using a hair-pin structured linker (from [229]). **Bottom:** Structure from RNAcifold. While the structures are in a narrow energy range (-7.4 to -7.3 kcal/mol), they differ substantially. Taken from [17] . . . . . 23
- 2.8 Example of pseudoknotted structures. **l.h.s** Typical H-type pseudoknot fold found i.e. in the catalytic core of various ribozymes. **middle** Kissing hairpin pseudoknot found i.e. in the 3' UTR region of the Coxsackie B Virus [252]. **r.h.s** Kissing hairpin-loop interaction between two RNAs. OxyS-fhlA hybrid in *e. Coli* is a typical example of such an interaction. Strictly speaking this is not a pseudoknot, as it involves two distinct sequences. Still this kind of RNA-RNA interactions are not handled correctly by the standard folding algorithm presented in the previous section as it considers it a pseudoknot. 24
- 2.9 Examples of pseudoknotted RNA-RNA interactions. **R.h.s** H/ACA snoRNA (red) interaction with its target (green). **L.h.s** Bound (right) and unbound (left) conformations of OxyS and fhlA. . . . . 25
- 2.10 Prediction of the hybridization of OmpN 5'-UTR with RybB, a small non-coding RNA found in bacteria, returned by two different RNA-RNA cofolding programs. **L.h.s:** Sequence and structure of the interaction partners. The structure stability of both RNA strands is represented by a color code, where red represents very stable regions and purple very unstable regions. **Bottom:** The hybrid predicted without considering RNA accessibility extends the whole length of RybB. The free energy of interaction of this hybrid is positive with a magnitude of 7.3 kcal/mol. **R.h.s** By considering the target site accessibility the correct hybrid can be retrieved. In this case, the free energy of interaction amounts -16 kcal/mol, making it the favored interaction compared to the previous hybrid. . . . . 27

- 2.11 Overview of the miRNA maturation process. First miRNAs are transcribed from their loci into pri-miRNAs **top**. pri-miRNAs are then processed by Drosha and Pasha proteins into pre-miRNAs. Dicer processes pre-miRNAs into short double stranded miRNA/miRNA\* duplexes. These duplexes get loaded into RNA-induced silencing complex. Generally the strand with the less stable 5' end is introduced into RISC, while the other strand (passenger strand) is degraded. Once loaded into RISC, miRNAs are ready to recognize their targets through base pairing, leading to the mRNA degradation and/or translation disruption . . . . . 29
- 2.12 Structure of a 19 nucleotides RNA duplex bound to Afpiwi. Afpiwi is an archeal PIWI domain-containing protein which is used to model eukaryotic Argonaute. The guide strand is depicted in green, while the target RNA is in yellow. The region on the mRNA that is cleaved by Argonaute is shown in red. Adapted from [192] . . . . . 35
- 2.13 Impact of siRNA characteristics along the silencing pathway. The innate immune system may be activated by dsRNAs. dsRNAs with specific sequence patterns or high "U" contents are recognized by Toll Like Receptors (TLRs) inducing inflammatory cytokines and interferon of type I ( $\text{IFN}-\alpha$ ,  $\text{IFN}-\beta$ ). Large dsRNAs (>30nts) are sensed by PKR (double-stranded RNA-activated protein kinase) which can induce interferon response, expression of inflammatory cytokines and cell death. dsRNAs with 2nts overhangs escape the RIG-1 triggered cytokines and interferon response. Once into RISC, the passenger strand is separated from the guide strand. The strand with the lower 5'–end stability is incorporated into RISC, while the other strand is degraded. A wrong asymmetry results in the selection the bad siRNA strand, leading to no on-target effect. siRNAs that are highly structured are not able to hybridize to their target. Reciprocally siRNAs targeting highly structured region can not bind to their target. Finally sequence specific off-target effects makes it more difficult to gain information from RNAi experiments. . . . . 37

- 2.14 Canonical C/D and H/ACA snoRNAs structures. **L.h.s** CD-Box snoRNA, made of a small stem and a large loop. The loop region contains either one or two sets of C/D boxes. The region directly upstream of the D boxes is responsible for the correct target recognition. **R.h.s** HACA-Box snoRNA, made of two target stems separated by a unpaired region containing the H Box. The interior loop in each stem is responsible for the correct target recognition. . . . . 40
- 3.1 A base pair  $(s_p, s_q)$  can close various loop types. According to the loop type different contributions have to be considered. a A hairpin loop is depicted in blue. b In case of an interior loop, which is shown in red, two independent contributions to  $Z_{pq}[i, j]$  are possible: The unstructured region  $s[i..j]$  can be located on either side of the stacked pairs  $(s_p, s_q)$  and  $(s_k, s_l)$ . c If region  $s[i..j]$  is contained within a multiloop we have to account for three different conformations, indicated in the green structures, a more detailed description is given in the text. . . . . 45
- 3.2 Alternative representation of figure 3.1 for multiloops only. Base pair  $(s_p, s_q)$  that includes the unpaired region  $(s_i, s_j)$  is drawn as an arc connecting bases  $s_p$  and  $s_q$ . The unpaired region  $s[i..j]$  is drawn as a bold black line. In the one-sided multiloop case (A) a structured region containing *at least* two structure components is on one side of the unpaired region. In case (B) the unpaired region  $s[i..j]$  is between two structured regions. In case (B) we have to take care to make a unique decomposition of the multiloop into a 3' part that contains exactly one component and a 5' part with at least one component. . . . . 47

- 3.3 Probability of being unpaired  $P_u[i, i]$  (dashed line), probability of binding to siRNA at position  $i$ ,  $P_i^*$ , (thick black line) and  $\Delta G_i$ , the optimal free energy of binding in a region including position  $i$  (thick red line) near the known target site of VsiRNA1. The scale for the probabilities is indicated on the left side, the scale for the minimal free energy of binding on the right side. At the bottom the protein expression levels in experimental data [216] are indicated. The isolated 21mer target sequence, displaying the same activity as the wild type mRNA, and 3 mutants are shown. A decreasing optimal free energy of binding is correlated with increasing expression. In the case of the HP5\_6 mutant an alternative binding site becomes occupied as the optimal free energy of binding due to this alternative interaction nearly equals  $\Delta G_i$  at the proposed target site. . . . . 51
- 3.4 Breaking energy profile and pair probability profile around the start codon of all mRNA in four different bacteria species. Boundaries of region with increased accessibility are shown by vertical blue lines. The orange dotted line represents the position of the start codon. The red dotted line represents the mean accessibility measure of the shuffled regions. **R.h.s** Mean breaking energy. **L.h.s** Mean base pair probability. . . . . 56
- 3.5 Mean Breaking Energy and Pairing Probability distribution around the start codon for all genes in *E. coli*. The black curve represents the density distribution, the red line represents the values for *rpoS* before binding, the blue lines delimits the quartiles of both distributions, the green line represents the values for *rpoS* after hybridization with *RprA*, while the orange line represents the value for *rpoS* after hybridization with *DsrA*. *rpoS* is among the most inaccessible mRNAs before binding, while after binding the local breaking energy belongs to the lower half. For the pairing probability, an even stronger trend is seen, as *rpoS* after binding belongs to the 25% most open targets. . . . . 57
- 3.6 Opening energy,  $\Delta G_u$  and single nucleotide base pairing probability plotted around the start codon of *RpoS* versus sequence position for the interaction of *DsrA* and *RprA*. The red area represents regions of higher than average structural stability before sRNA binding on *RpoS*, while the green region represents regions of lesser than average structural stability after sRNA binding to *RpoS*. The blue line represents the position of the start codon. . . . . 58

- 4.1 Application of `RNAplfold` to separate functional from non-functional siRNAs. (a) The RNA is folded locally in a sliding window approach (window size  $W$ ). Within  $W$ , base pairing is restricted to a maximum distance  $L$ .  $u$  represents the stretch of consecutive nts within a siRNA target site starting at its 3' end for which the accessibility is computed. Green lines represent possible base pairs. Interactions outside the span size of  $L$  or the flanking window  $W$  are not allowed (dotted green lines). (b) Box-plot diagram comparing the accessibility of functional and non-functional siRNAs. The dataset was divided into functional siRNAs (repression efficiency  $> 75\%$ ) and non-functional siRNAs (repression efficiency  $< 25\%$ ); black horizontal lines within the boxes depict medians. The circles represent outliers and dotted lines show the standard deviation. The Wilcoxon p-value is  $5 \cdot 10^{-4}$ . Cutoffs for the accessibility to discriminate functional and non-functional siRNAs was set at 0.01157 (red horizontal line). The parameters  $W$ ,  $L$  and  $u$  are indicated. (c) Accessibility distributions of functional and non-functional siRNAs are best differentiated for a length of 8 and/or 16 nts (according to p-values). p-values, were determined from a Wilcoxon test and are plotted against the length of the analyzed region starting at the 3' end of the target site. . . . . 64
- 4.2 Box-plot diagram of functional and non-functional siRNAs for different target site G/C content. Functional and Non-functional siRNAs are partitioned into five groups according to their G/C content. A wilcoxon test was applied showing a significant separation for all G/C windows analyzed. . . . . 66
- 4.3 Correlation plots for different design criteria for 2433 siRNAs from dataset 1. The siRNAs were grouped into bins, each of them containing 36 siRNAs. The binning was done according to the design criteria. Correlation plots of the novartis repression score against accessibility (a), asymmetry (b) and self-folding (c) are shown. (d) Ranking of siRNAs for the combination of all design criteria including accessibility plotted against the normalized inhibitory activity. . . . . 67

- 4.4 Box-plot diagrams comparing asymmetry, self-folding and free-end for functional and non-functional siRNAs. The dataset 1 consisting of 474 siRNAs was used to determine the single criteria thresholds (red line). The dataset was divided into 363 functional siRNAs of (white boxes,  $>0.900$  repression score) and 109 non-functional siRNAs (grey boxes,  $<0.354$  repression score). The quartiles are represented by the edges of the rectangles, which contain 50% of the data, black horizontal lines within the boxes depict medians. The circles represent outliers and dotted lines show the standard deviation. Thresholds were chosen conservatively, such that at least 75% of the working siRNAs were kept. Note, the same was done for dataset 2 (data not shown). 69
- 4.5 Performance of RNAXs on a set of 360 siRNAs targeting the four genes firefly luciferase, human cyclophilin B, ALPPL2 and DBI. SiRNAs were grouped into functionality classes of less than 50% mRNA repression  $<F50$ , repression of at least 50%  $\geq F50$ , 70%  $\geq F70$ , 80%  $\geq F80$  or 90%  $\geq F90$ . The random distribution is depicted in black. Functional class enrichments for (a) asymmetry, (b) accessibility, (c) the combination of asymmetry with self-folding plus free-end and (d) all parameters including accessibility (RNAXs) are shown in light gray. The three top ranked siRNAs are all contained in  $\geq F50$  (dark gray). (e) Comparison of RNAXs to other design tools. OptiRNA , Ambion (siRNA Target Finder), Qiagen (siRNA Design Tool), Invitrogen (Block-iT RNAi Designer), oligowalk21 and Sirna (using total score threshold; score  $> 12$ ) were compared to RNAXs for the four functional classes ( $<F50$ ,  $\geq F50$ ,  $\geq F80$ ,  $\geq F90$ ). All tools were used with default parameters using the available web servers. For each tool, the repression efficiency of the three best-ranked siRNAs was assessed. RNAXs performed better than the other design tools for all functional classes. (f) Western blot analysis of extracts prepared from Eph4 cells, transiently transfected with scrambled siRNA, Dharmacon mmLEF1 SMARTpool (a combination of four siRNAs) or the single top ranked siRNA designed with RNAXs. Relative LEF1 expression levels are indicated. Actin protein levels show equal loading. . . . 71

- 4.6 RNAs input page. The input page is divided into three areas: a sequence input area, where a FASTA formatted sequence is pasted. A design area where thresholds on different parameters as well as base preferences can be set and the output area which allows to set the number of siRNAs candidates. For each siRNA candidate a plot of the accessibility is generated. 73
- 4.7 Typical output of RNAs session. A user defined number of siRNA are shown with their features scores as well as a plot of the accessibility profile around the target site. For each siRNA, a link to NCBI blast allows to search for putative off-targets . . . . . 74
- 4.8 Functional siRNA distributions of randomly selected siRNAs (black bars) and rationally designed siRNAs (dark gray bars), as well as the 3 top RNAs predicted siRNAs (light gray bars) targeting: (A) Firefly luciferase, (B) human cyclophilin B, (C) human ALPPL2, and (D) human DBI . . . . . 75
- 4.9 Accessibility plots for all four murine LEF1 siRNAs from a commercial siRNA pool (A)-(D) and for the RNAs designed siRNA (E). The sequence of the respective siRNA duplex is indicated. siRNAs A and B would be rejected by RNAs because of poor target accessibility. siRNA C would be rejected based on the asymmetry rule. . . . . 77
- 4.10 **L.h.s** Translational repression of the *Renilla* luciferase (RL), normalized the firefly luciferase (FL), was measured for accessible as well as for non-accessible let-7 reporter constructs.**R.h.s** Distribution of opening energies for human miRNAs. The continuous line represents the density distribution of opening energies corresponding to 3'-UTRs complementary to the seeds for all known human miRNAs. The dotted line shows the density of the shuffled sequences. . . . . 78

- 5.1 Comparison of the ompA-micA hybrids predicted with and without considering intramolecular structures. (a) Hybrid structure predicted with **RNAplex** without considering the intramolecular structures of the RNA sequences. The hybrid extends over 67 and 69 nucleotides on ompA and micA, respectively and has an hybridization energy of  $-42.3$  kcal/mol. Still the energy needed to unfold both binding regions on ompA and micA amounts  $22.7 + 26.8 = 49.5$  kcal/mol, larger than the energy gained through binding. (b) ompA-micA interaction predicted by **RNAup**. OmpA-micA hybrid is shown on the right hand side, with the micA sequence represented by a bold line. Even though the hybrid is much smaller than the interaction in (a), it has a lower total interaction energy (ddG) of  $-12.25$  kcal/mol, due to the fact that the interacting regions are less structured. . . . . 83
- 5.2 Comparison of the **RNAplex** energy model against the Turner energy model for bulges and interior loops a) Plot of the interior loop penalty against the total loop size for three different values of asymmetry. The model used in **RNAplex** slightly overestimates the loop energies. b) Plot representing the bulge loop penalty against the bulge size. Our model agree exactly with the Thurner model for bulge size up to 6 nts. . . . . 84
- 5.3 Simplified representation of the structure decomposition used in **RNAplex**. For clarity only the decomposition of the closed structure terms (see equation (5.1)) is shown. Black dots represent paired bases. White dots denote unpaired bases. Given that  $x_i$  and  $y_j$  are paired,  $C$  stores the best energy of interaction between  $x_1..x_i$  and  $y_j..y_m$ .  $\mathcal{S}$  is the stacking energy of two pairs of nucleotides.  $P$  is the bulge penalty to add to 1x0 bulges.  $I$  is the matrix holding the best energy of interaction given that  $x_i$  and  $y_j$  are in an interior loop.  $I_1^1$  is the destabilizing energy of a 1x1 interior loop (1x2, 2x1 and 2x2 cases not shown) and  $B^x$  represents the matrix storing the best energy of interaction given that residue  $y_j$  is aligned to a bulge. The cases where  $x_i$  and  $y_j$  do not pair (interior loop and bulge extension and/or creation) are not shown . . . . . 87

- 5.4 Error representation of our accessibility model. **l.h.s** Boxplot representation of the the distribution of the relative breaking energy between our approximated model and the standard energy model for different  $\Delta$  size and a fixed target size of 20 nts. The larger  $\Delta$  the smaller the error in our approximation. **RNAplex** uses  $\Delta = 4$ . At this level of approximation, the pearson correlation coefficient between the approximated model and the real model reaches 0.92 . . . . . 99
- 5.5 Bar plots representing the time necessary to complete the target search for 19 bacterial sRNAs in 100 random sequences of length 1200 nts for different RNA-RNA interaction tools. **RNAplex -c** is the fastest application with a completion time of 27[s]. **RNAplex -a** needs 36[s] to achieve the same task. This grows to 90[s] if one considers the time necessary to compute the accessibility profile. **RNAplex -a** is 1000 times faster than **IntaRNA** and 2422 times faster than **RNAup**. . . . . 104
- 5.6 Runtime of **RNAplex** with alignment and accessibility against the number of sequences in alignments for a set of 9 query and 100 target sequences. The runtime of **RNAplex** increases proportionally to  $\sqrt{N}$ , where  $N$  is the number of sequences in the alignments. . . . . 105
- 5.7 Density distribution of interaction energy as computed by **RNAplex** for miR-134 against all mouse 3'UTRs. The vertical line represents the energy of the experimentally confirmed miR-134/Limk1 interaction as computed by **RNAplex**. The black area represents the proportion of 3'UTRs having a higher energy of interaction than the experimentally predicted one. The number in parenthesis represents the percentage of the entire distribution falling below the threshold defined binding energy, as computed by **RNAplex**, of the experimentally verified interaction. The inset shows the density of distribution of interaction energy as computed by **RNAup** for miR-134 against the mouse 3'UTRs. The vertical line represents the energy of binding as computed by **RNAup** for the experimentally confirmed miR-134/Limk1 interaction. The number in parenthesis represents the percentage of the entire distribution falling below the threshold defined by the binding energy, as computed by **RNAup**, of the experimentally verified interaction. . . . . 107

5.8	Procedure used to select high-binding, highly similar target binding site in multiple sequence alignments. a) Sequences are sorted based on their sequence similarities with <code>clustalw</code> . b) RNAplex is ran on each sequences in order to select the n-best hits for each sequences. In this study $n = 3$ was used. c) A recursive approach based on the sequence similarities and the strength of interaction of target sites is used to find the best set of target sites among the m-species. d) Starting from the target site with the minimum score, the best set of target sites is retrieved through backtracking. e) The set of target sites is realigned. It is used to compute the multiple-alignment interaction between the sRNAs and the selected target sites. Accessibility information are retrieved thanks to the coordinates found in the multiple alignments, e.g. 253-270 for gene ColiAPE_APEC01_62 and 181-198 for eColi_K12_b0957. . . . .	112
5.9	Boxplots showing the interaction energy distribution as a function of the number of sequences in the alignments for sRNA <i>GcvB</i> . Well conserved interactions have in average a higher interaction energy than interactions involving less sequences. . . . .	113
6.1	Decomposition of interaction structure. . . . .	121
6.2	Boxplots showing the accessibility distribution for all known uridines in human (top) and yeast (bottom) 28S and 18S rRNAs. The target accessibility was computed by using <code>RNAup</code> on the whole length sequences of 28S and 18S rRNAs. The target size was varied between 3 and 19 nts in steps of 2 nts and was centered around the (pseudo)uridine site. . . . .	123

- 6.3 Features considered in the SVM model. Structural (black bold lines) and energy features (shaded regions). **TE**: lower stem energy, **LE**: 5' interaction energy, **DE**: upper stem energy, **RE**: 3' interaction energy, For each nucleotide in the target, its local opening energy is represented by a gray circle, where light gray represents low local opening energy and dark gray high local opening energy. The target total opening energy (**OE**) is the sum of all local opening energies, **YE**:  $YE = LE + RE + TE + DE$ , **XE**:  $XE = LE + RE + DE$ , **dYE**:  $dYE = YE + OE$ , **t\_i\_gap**: number of nucleotides between the 5' end of the upper stem and the 3' end of the 5' interaction on the snoRNA, **U\_gap**: number of nucleotides between the 3' end of the 5' interaction and the 5' end of the 3' interaction on the mRNA, **i\_b\_gap**: number of nucleotides between the end of the lower stem and the 3' end of the 5' interaction on the snoRNA, **i\_t\_gap**: number of nucleotides between the 5' end of the 5' interaction and the 5' end of the snoRNA stem, **stem\_length**: length of the upper stem, **stem\_asymmetry**: difference in the number of nucleotides located in loops between the 5' and 3' side of the upper stem. **gap\_right**: number of gaps in the 3' interaction on the mRNA . . . . . 125
- 6.4 SnoRNA-target features . . . . . 129
- 6.5 Time dependency of RNAsnoop on the target size (top) and snoRNA size (bottom). The target size was varied between 500 and 25000 nts while the snoRNA sizes were varied between 20 and 500 nts. The runtime of RNAsnoop grows linearly with the target size and grows more rapidly than linear with the snoRNA size. . . . . 131
- 6.6 Ratio of the time dependency of RNAsnoop against RNAhybrid and snoGPS. (left) Dependence of the ratio on the target size. (right) Dependence of the ratio on the snoRNA size. All three programs were run so that only the best interaction was returned. Under these conditions RNAsnoop has a runtime similar to that of snoGPS (red curve), while RNAhybrid is about 15 times slower than RNAsnoop (black curve). Due to the higher than linear runtime dependency this difference becomes smaller for larger snoRNA (right, black curve). . . . . 132

- 6.7 Structure of the interactions between the human orphan  $\Psi$  sites and their predicted snoRNAs as returned by RNAsnoop. From left to right: ACA55-2:18S-681, ACA13-1:18S-1248, SNORA38B-1:28S-1523, ACA52-2:28S-3747, U71c-2:28S-3863, ACA64-1:28S-4266, ACA51-2:28S-4323, ACA10-1:28S-4501, where i.e. ACA51-2:28S-4323, means that the second stem of ACA51 binds to position 4323 on rRNA 28S. All structures were generated by RNAsnoop. The accessibility for each nucleotide is color-coded, with a red representing accessible and green inaccessible nucleotides. . . . . 135
- 6.8 . Structure of the interactions between Or-aca4 and its putative target. **L.h.s.:** Single sequence structure. **R.h.s:** Multiple sequence structure. **Below:** Alignment of the target (up to the & column) and the snoRNA. For the multiple sequence and alignment figures, the color in the order red,ocher,green indicate 1 through 3 different type of base pairs. The consensus structure is represented in dot bracket format on top of the alignment. The angle brackets represent intermolecular base pairs and the braces represent intramolecular base pairs. . . . . 137



# D

## List of tables

### List of Tables

---

- |     |   |    |
|-----|---|----|
| 2.1 | Free energies for stacked pairs in kcal/mol. Note that both base-pairs have to be read in 5'-3' direction. . . . .  | 18 |
| 2.2 | Summary of widely used miRNA target prediction tools. The first columns contains the name of the tools. The second column indicates the method used by the tools. Conservation means that conservation of the seed/target site is important. Hybridization means that the energy of interaction between the miRNA and its target is relevant. Accessibility means that the structuredness of the target site is taken into account. Besides the conservation of the target site, TargetScan further considers the hybrid structure, the position of the target site on the 3'-UTR as well as the AU content around the target site. The third column lists how target information can be accessed. The last column reports the corresponding literature citation. . | 32 |
| 3.1 | Binding site summary for the 10 functional interactions published by Urban et.al [250]. Column $\Delta\Delta G$ shows the optimal binding energy calculated with RNAup. Column Position gives the binding position relative to the start codon. Column Position lit. gives the binding position found in the literature.  | 54 |
| 5.1 | Additional “energy” parameters for alignment folding . . . . .  | 96 |

- 5.2 Binding site summary for 27 functional miRNA–mRNA interactions in Human, taken from TarBase [222]. Columns 1 and 2 contain the name of the mRNA and miRNA, respectively. The column 3 to 5 contain the interaction energy for the reported miRNA mRNA interactions as computed by **RNA duplex**, **RNAplex** and **RNAhybrid**, respectively. The number in parenthesis represent the rank of the experimental target site where 1 stands for the most stable interaction and 10 for the 10th best interaction. NF means that the reported target site was not found among the 10 best interaction sites and are shown in red. . . . . 98
- 5.3 Binding site summary for a set of 17 functional interactions from [38]. The first and second columns contain mRNAs- and sRNAs-ID, respectively. We compared **biRNA**, **RNAup** and to **RNAplex**. **biRNA** and **RNAup** were run using the default parameters, while **RNAplex** was run with either an extension penalty of 0.3 [kcal/mol] (**RNAplex -c**) or the accessibility files produced by **RNAup** (**RNAplex -a**). All predictions made by **RNAup**, **biRNA** and **RNAplex -a** overlapped with the experimentally reported interactions, while **RNAplex -c** missed four interactions. The last row reports the average deviation between the experimentally found locations and the predicted ones . . . . . 101
- 5.4 Speedup and memory improvement of the accessibility based **RNAplex** against **IntaRNA** and **RNAup** for different random query and target sequences as measured by the *time* application. The first two columns show the target and query length, respectively. The third and fifth column show the runtime improvement of **RNAplex** against **IntaRNA** and **RNAup**, respectively. The difference between **RNAplex** and the two other tools slightly grow with increasing target length, but diminish with increasing query length. On this dataset, **RNAplex** is between 600 and 1600 times faster than **IntaRNA** and from 1500 up to 65400 times faster than **RNAup**. Note that for very short sequences (less than 400 nts), **RNAplex** needs less than 1/100th second to compute the hybrid, too fast to be precisely measured by time, hence the NA for the first entries. The memory consumption of **RNAplex** is 17.4 to 1330 times smaller than that of **IntaRNA** and 15 to 626 times smaller than that of **RNAup**. Note that larger sequences could not be used because the memory need of both **RNAup** and **IntaRNA** exceeded the available RAM (4G) 103

- 5.5 Binding site summary for the 9 functional interactions from [250]. The number in parenthesis represents the quantity of predicted interactions involving the same ncRNA, overlapping with a 401 nts long region centered around the start codon and having a higher interaction energy than the functional hybrid. Positions in red indicate target sites that were misspredicted by the respective tool. For **RNAplex -c** a per nucleotide penalty of 0.3 kcal/mol was used. . . . . 116
- 5.6 Summary of the predicted binding sites for the 9 functional interactions reported by [250]. The first and second columns show the name of interaction partners. Column 3 and 4 give the predicted and experimentally reported binding regions, respectively. Column 5 and 6 report the binding  $\Delta G$  computed by **RNAup** and **RNAplex**, respectively. The numbers in parenthesis in the sixth column represent the number of interactions, located within a window of 80 nts centered around the start codon, with a lower interaction energy than the experimentally reported interaction for the predictions made by **RNAplex** with and without considering the opening energy, respectively. Column 7 gives the interaction energy for the multiple sequences interactions. The numbers in parenthesis in column 7 represent the number of sequences in the final alignments. Column 8 shows the rank of the interaction when looking only at the interaction energy. Column 9 shows the rank of the interactions based on the  $Z$ -score corrected for the number of sequences in the alignment. Finally column 10 shows the rank of the interaction based on the  $Z$ -score, given that only interactions with a greater or equal number of sequences in the alignment are taken into account. The number in parenthesis in the last column represent the number of better scoring elements in the case of alignment when no accessibility information are taken into account. . . . . 117

5.7	Summary of the number of false positives under different condition. In the third column, for each confirmed interaction, the number of better scoring interactions involving the corresponding dinucleotide shuffled sRNA and any dinucleotide shuffled <i>e.coli</i> mRNAs is reported. It should be noted that the interaction should take place in the region located 50 nts upstream and 30 nts downstream of the start codon. In the last column, the number of expected hits per nucleotide is reported. In this case there is no location restriction. . . . .	118
6.1	Prediction comparison of <b>RNASnoop</b> (abbreviated <b>RNASn.</b> ), <b>snoGPS</b> and <b>fisher</b> for the known snoRNA-rRNA interactions in yeast. <b>RNASn. A</b> stands for the accessibility version of <b>RNASnoop</b> . . . . .	127
6.2	Prediction performance in human for <b>snoGPS</b> , <b>RNASnoop</b> ( <b>RNASn.</b> ), <b>RNASnoop</b> with accessibility ( <b>RNASn. A</b> ) and the SVM in human. The numbers represent the rank of the interaction for the corresponding snoRNA stem. In column Type, +, − represent experimentally confirmed or rejected interactions, respectively. When using the human interactions for testing, we trained the SVM exclusively on the yeast dataset. . . . .	128
6.3	Predicted snoRNAs targeting the orphan pseudouridines in human ribosomal RNAs. No snoRNAs were found for position 1849, 3674 and 3749 on rRNA 28S. ACA51 and SNORA38B are orphan snoRNAs while ACA52-2 and ACA64-1 are orphan stems . . . . .	136
A.1	List of Symbols . . . . .	145

E

Resume

# HAKIM TAHER

## PERSONAL DATA

Born in Geneva, 6 February 1976

email

[htafer@bioinf.uni-leipzig.de](mailto:htafer@bioinf.uni-leipzig.de)

phone

+49 151 233 17 34 2

## WORK EXPERIENCE

### Summary

#### Summarized Work Experience

- DEVELOPMENT AND IMPLEMENTATION OF ALGORITHMS IN C
- DATA MANIPULATION IN PERL/PYTHON
- STATISTICAL ANALYSIS OF LARGE DATASETS WITH R

#### RNA annotation pipeline

2009–Current      Research assistant, Leipzig

non coding RNA (functional) annotation. This covers the development, review and benchmarking of tools to annotate ncRNA genes. Development of an annotation pipeline to automate and accelerate ncRNA annotations in new species. Scoring the effect of single nucleotide polymorphisms on the annotation elements. Development of a new approach for microarray design in cooperation with the University of Bodenkultur, Vienna. Supervision of 3 diploma works

#### siRNA design

Winter 2009      Research assistant, Copenhagen

Development of new approaches to design more potent and more specific small interfering RNAs. Development encompassed data analysis, selection of machine learning applications as well as development of a webserver. Work done in cooperation with the University of Copenhagen and financed by LUNDBECK

#### RNA-RNA algorithm development

2005–2009      PhD student, Vienna

Study of RNA-RNA interactions. Development of fast algorithms to predict interactions between two RNA strands. Applications for the functional annotation of ncRNAs. Development of a new method to efficiently design siRNAs named RNAXs. Development of a webserver for RNAXs. Work financed by SIEMENS.

#### Teaching

2002      Substitute teacher, Geneva

Teacher for Mathematics in Geneva at the Middle-School Level

#### NMR spectroscopy

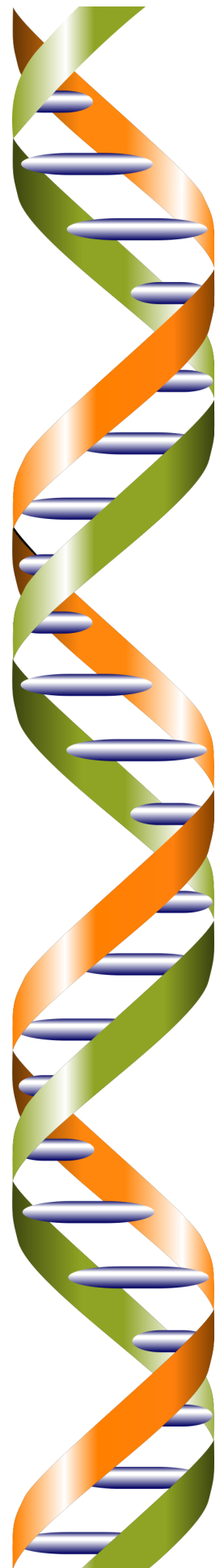
2001–2002      Research assistant ETH, Zurich

Study of protein structures with NMR spectroscopy. Development of tools to facilitate protein structural annotation.

#### Particle Physics

Summer 1999      Internship CERN, Geneva

Monitoring the luminosity of lead-tungsten crystals used in the cms



experiment.

#### EDUCATION

##### *Diploma in Physics*

1996-2001      ETH, Zurich  
5.14/6 · Major: Physics Minor: Computer Aided Physics  
1998-1999 Erasmus at the Università degli Studi di Pisa  
Thesis: Nonrandom structure in the urea-unfolded Escherichia coli  
outer membrane protein X (OmpX)  
Advisor: Prof. Kurt WÜTHRICH

#### ADDITIONAL INFORMATION

2003      Ili, Cairo  
Learning of colloquial and modern standard arabic at the international  
language institute, Cairo

2003      Moqattam, Cairo  
Volunteer for the Soeur Emmanuelle Charity Organization in the  
Moqattam District as teacher for french/mathematics

1990-2007      Rowing: Switzerland, Italy, Egypt, Austria  
Participation at rowing regatta at the international (swiss team, won  
Coupe de la Jeunesse), national (2x swiss champion, 1x italian  
universitary champion), regional (10x Romandy Champion, diverse  
regatta in Vienna, Cairo)

##### *Language*

Native      · FRENCH  
Fluent      · ENGLISH—GERMAN—ITALIAN—SPANISH  
Basics      · ARABIC

February 10, 2011

