Albert-Ludwigs University of Freiburg

Master Thesis

Base pair probabilities of RNA-RNA interactions incorporating seeds and accessibility

Author:
Frank Gelhausen

Supervisor:
Dr. Martin Raden

Examiner:
Prof. Dr. Rolf Backofen
Prof. Dr. Ivo L. Hofacker (University of Vienna)

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science
in the
Bioinformatics Group,
Department of Computer Science

Submitted on 14th November 2019

DECLARATION

I hereby declare, that I am the sole author and composer of my Thesis and that no other sources or learning aids, other than those listed, have been used. Furthermore, I declare that I have acknowledged the work of others by providing detailed references of said work. I hereby also declare, that my Thesis has not been prepared for another examination or assignment, either wholly or excerpts thereof.

Freiburg,	
Place, Date	Signature

First and foremost, I want to thank my supervisor Martin Raden for giving me the opportunity to work on IntaRNA as topic for my Master thesis and for providing me with thorough feedback and help throughout my work.
I also want to thank my brother Rick for his culinary support and his help with bioinformatics related questions.
Furthermore, a special thanks goes to my friends Olivier Puraye and Ben Kap for taking their time to thoroughly proof-read my thesis.

Abstract

Computing base pair probabilities of RNA-RNA interactions allows for a number of useful applications, such as the creation of dot plots, which allow for easy and fast comparison between different base pairing patterns. A number of tools exist that already incorporate base pair probability calculation, such as RNAcofold and NUPACK. However these tools are limited to a specific algorithm for the optimal interaction computation that might lack in precision or computational efficiency depending on the application.

IntaRNA on the other hand is a highly flexible RNA-RNA interaction prediction tool that implements a large number of different prediction algorithms, including very efficient seed-constraint methods.

This thesis explores the benefits and difficulties of introducing the computation of base pair probabilities into a number of IntaRNA predictors, including seed-based predictors.

For this reason IntaRNA was extended with the ability to compute base pair probabilities, depending on the chosen prediction model. The output is provided as a dot plot to allow for easy investigation.

Finally, a number of applications are presented that benefit from base pair probabilities, including the comparison between verified and non-verified RNA-RNA interactions and the detection of multi-site RNA interactions. Based on these results, potential improvements for IntaRNA's prediction model are discussed, including different approaches for the accessibility computation and the incorporation of sequence conservation into the prediction estimation.

Zusammenfassung

Die Berechnung der Basenpaarwahrscheinlichkeiten von RNA-RNA Interaktionen erlaubt eine Reihe nützlicher Anwendungen, wie die Erstellung von Punktdiagrammen, die einen einfachen und schnellen Vergleich zwischen verschiedenen Basenpaarungsmustern ermöglichen. Es gibt eine Reihe von Tools, die bereits die Berechnung der Basenpaarwahrscheinlichkeit beinhalten, unter Anderen RNAcofold und NUPACK. Diese Tools sind jedoch auf einen bestimmten Algorithmus für die optimale Interaktionsberechnung beschränkt, bei dem es je nach Anwendung möglicherweise an Genauigkeit oder Recheneffizienz mangelt.
IntaRNA hingegen ist ein hochflexibles Tool zur Vorhersage von RNA-RNA Interaktionen, das eine große Anzahl verschiedener Vorhersagealgorithmen implementiert, einschließlich sehr effizienter Methoden, die auf der Beschränkung der Interaktionen auf Seed-Regionen basieren.
Diese Arbeit untersucht die Vorteile und Schwierigkeiten bei der Einführung der Berechnung von Basenpaarwahrscheinlichkeiten in eine Reihe von IntaRNA-Prädiktoren, einschließlich Seed-basierter Prädiktoren. Aus diesem Grund wurde IntaRNA um die Fähigkeit erweitert, Basenpaarwahrscheinlichkeiten abhängig vom gewählten Vorhersagemodell zu berechnen. Die Ausgabe wird als Punktdiagramm bereitgestellt, um eine einfache Untersuchung zu ermöglichen.
Schließlich wird eine Reihe von Anwendungen vorgestellt, die von Basenpaarwahrscheinlichkeiten profitieren, einschließlich des Vergleichs zwischen verifizierten und nicht verifizierten RNA-RNA Interaktionen und des Nachweises von Multi-Site-RNA Interaktionen. Basierend auf diesen Ergebnissen werden mögliche Verbesserungen für das Vorhersagemodell von IntaRNA diskutiert, einschließlich verschiedener Ansätze bei der Berechnung der *accessibility* und der Einbeziehung der *sequence conservation* in die Vorhersage.

Contents

Chapter 1

Introduction

IntaRNA is an RNA-RNA interaction (RRI) prediction tool developed by the Bioinformatics group at the University of Freiburg (Busch et al., 2008) with the objective to efficiently compute the optimal interaction between two given RNA sequences. While the predictions are fast and accurate, the output is currently limited to only the best interactions. However, sometimes it might be interesting to also know the probabilities of said interactions occurring, as well as the probabilities of the included base pairs.
In this thesis, I will introduce new IntaRNA predictors that are based on the existing ones, but use partition functions to compute the ensemble energy of all possible interactions. I will then compute the probabilities of the structures as well as individual base pair probabilities between two RNA sequences with the goal of creating dotplots of intermolecular base pair probabilities as an additional output for the IntaRNA predictors.

1.1 Thesis structure

This chapter begins with an overview of related work. It then explains the biological concepts, that are required to understand the thesis, such as RNA structures and interactions and the underlying energy models. It will also detail the algorithms used by IntaRNA. Chapter 2 focuses on the computation of RRI partition functions and explains the difference between standard and seed-constraint algorithms. Chapter 3 explains the differences between intramolecular and intermolecular base pair probability computation and covers the introduction of intermolecular base pair probabilities into IntaRNA, including the implementation details and hurdles. Applications, such as concentration computation and comparison between verified and non-verified interactions using generated dot plots are given in chapter 4. Finally, this thesis is concluded in chapter 5, which also provides an outlook on possible future work.

1.2 Related Work

A large number of different RRI prediction tools exists, that are built using different prediction strategies. Some of these tools already allow the computation of inter-

molecular base pair probabilities. However, they often lack in prediction performance compared to IntaRNA due to differences in their underlying strategies.
A first class of such tools makes predictions based solely on intermolecular base pairs. By ignoring intramolecular base pairs that can affect the availability of some interaction sites, these tools tend to be the fastest, but often lack in precision. Examples of such tools include RNAhybrid (Bernhart et al., 2006b), RNAplex-C (Tafer and Hofacker, 2008), RIsearch (Alkan et al., 2017) and GUUGLE (Gerlach and Giegerich, 2006).
Another class of tools are called *concatenation-based* methods. These algorithms, such as RNAcofold (H. F. Bernhart et al., 2006) and the respective part of NUPACK (Dirks et al., 2007) consider joint structures of both RNA sequences by concatenating them, allowing them to make use of existing techniques used for RNA-secondary structure prediction. They also enable the calculation of partition functions of all joint structures and the computation of intermolecular base pair probabilities using a modification of McCaskill's algorithm.
Due to the inability of *concatenation-based* methods to predict interactions forming on interior, multi- or hairpin loops in the joint structure, *accessibility-based* methods have been introduced which are able to handle this type of interactions by computing the energy necessary to make an interaction site accessible (e.g. free of intramolecular base pairs). Examples of *accessibility-based* approaches are RNAup (Mückstein et al., 2006), IntaRNA (Busch et al., 2008) and RNAplex-a (Tafer and Hofacker, 2008). *Accessibility-based* methods can be further split into *seed* and *non-seed* variants. *Seed regions* are interaction regions with near perfect complementarity, which can be used in order to narrow down the search space for the optimal interaction. Their biological relevance was shown among others by Bentwich (2005) and Brennecke et al. (2004) who have shown that seed regions with seven or eight consecutive bases in animal miRNAs are often sufficient for effective regulation.
IntaRNA implements both variants of *accessibility-based* predictors. The goal of this thesis is to extend the predictors by adding the ability to compute intermolecular base pair probabilities.

1.3 Scientific background

IntaRNA is an interaction prediction tool for Ribonucleic acid (RNA) sequences. RNA results from the transcription of Deoxyribonucleic acid (DNA) and, beside other functions, plays a role in the creation of proteins. An RNA molecule is composed of different organic bases, the distinguishing parts of nucleotides: adenine (A), cytosine (C), guanine (G) and uracil (U). It is defined as a sequence M of n nucleotides, i.e. $M \in \{A, C, G, U\}^n$.
There are two main types of RNA: coding and non-coding RNA. The most well-known type of RNA is the messenger RNA (mRNA), which is a coding RNA and enables the synthesis of proteins (translation). Non-coding RNAs, also called regulatory RNAs, do not take part in the coding of proteins. Small bacterial RNA (sRNA) are an example for ncRNA. They have numerous functions such as the regulation of gene creation or the modification of protein functions (Backofen, 2011).
The nucleotides of an RNA strand fold into a secondary structure via interactive forces

amongst each other, e.g. to form a complementary base pair via hydrogen bonds. These structures define the function of the molecule. According to the models of Watson-Crick and Wobble, the following complementary base pairs are considered: A with U, G with C and G with U.
There are two types of interactions between the bases of a molecule depending on their location. If the bases belong to the same RNA molecule, they form an intramolecular structure. If they belong to different molecules, they form intermolecular structures. The goal of RRI predictors is to find out the optimal intermolecular interaction between two RNA molecules in order to better understand their biological function.

1.3.1 RNA structures and interactions

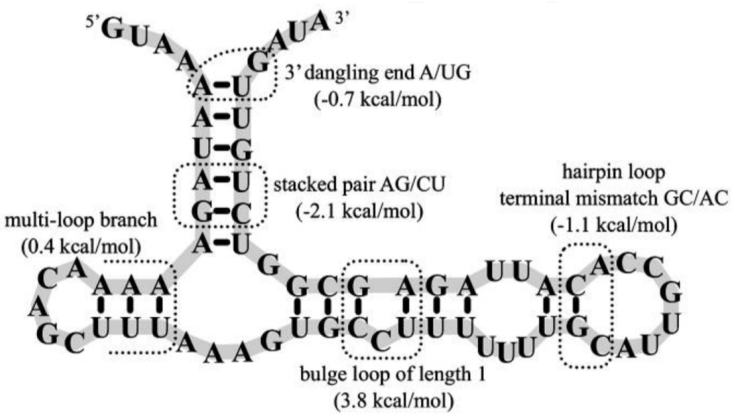


Figure 1.1: *Illustration of the types of loops contained in an RNA secondary structure as well as their energy contributions (Andronescu et al., 2010).*

Given an RNA secondary structure P of size n with $P = \{(i,j)|1 \leq i < j \leq n$ with M_i and M_j forming a base pair$\}$, the following loop types of structure P will be taken into consideration:

1. Hairpin: Hairpin loops have one enclosing base pair:
   $(i,j) \in P$ with $\neg\exists(i',j') \in P|(i < i' < j' < j)$

2. Stack: Stacks are adjacent complementary base pairs:
   $(i,j) \in P \wedge (i+1, j-1) \in P$

3. Internal loop: Interior loops have 2 enclosing base pairs and unpaired nucleotides on both strands: $(i,j) \in P \wedge (i',j') \in P|i+1 < i' < j' < j-1$ with $[i+1, i'-1]$ and $[j'+1, j-1]$ containing only unpaired bases.

4. Bulge: Bulges are a special case of internal loop with unpaired bases on only one strand.

5. Multiloop: Multiloops have 3 or more enclosing base pairs.

An example of an RNA secondary structure can be seen in Fig 1.1. Let $\mathcal{P}$ be the ensemble of all RNA secondary structures: $\mathcal{P} = \bigcup P$.

An interaction I between two RNA sequences M^1 and M^2 is defined by:

$$I = \{(i_1, i_2) | (M^1_{i_1}, M^2_{i_2}) \text{ form a base pair } \wedge \forall (i_1, i_2), (j_1, j_2) \in I : i_1 < j_1 \leftrightarrow i_2 > j_2\}$$

which means that no two distinct base pairs of an interaction can be crossing.
Similar to $\mathcal{P}$, let $\mathcal{I}$ be the ensemble of all RRI: $\mathcal{I} = \bigcup I$.
In order to simplify the notation of interactions, the following definitions are introduced:

$$i(I): \text{ base pair with the smallest } i_1 \text{ in } I$$

$$j(I): \text{ base pair with the largest } i_1 \text{ in } I$$

Throughout the rest of the thesis, i and j stand for either a single sequence position or a base pair, depending on the context.
$isLoop(k, l, I)$ is introduced in Eq 1.1 and indicates whether or not base pairs k and l form a loop inside I.

$$isLoop(k, l, I) = k \in I \wedge l \in I \wedge \nexists k' : k < k' < l \tag{1.1}$$

A seed S of an interaction I is a subinteraction $S = I|_{i(S)..j(S)}$ consisting mostly of stackings:

$$S \in I \leftrightarrow \underset{\substack{k,l \in S \\ isLoop(k,l,S)}}{\forall} : \quad k \in I \wedge l \in I \wedge isLoop(k, l, I)$$

In default IntaRNA, used in the following, seeds consist of seven stacked base pairs.
Considering seed regions for RRI prediction serves two main purposes. Firstly, due to their high energetic stability, they are likely part of the interaction with minimal free energy, hence increasing the quality of the predictors. Secondly they allow the reduction of the search space of predictors which considerably decreases their runtime.
The ensemble of all seed interactions is defined as: $\mathcal{S} = \bigcup S$.

In order to simplify the following formulas in this thesis, the notation of a region is introduced, which is a tuple of base pairs: $R = (l, r)$ where l and r are the region's leftmost and rightmost base pair. The region of an interaction is defined as follows:

$$R(I) = (i(I), j(I))$$

$\mathcal{R} = \bigcup R$ is the ensemble of all interacting regions.

1.3.2 Energy Model

IntaRNA minimizes the free energy of a structure in order to compute the optimal RRI. Dissolving the stabilizing bond defining a stable base pair is an endothermic process whose cost is indicated by a negative energy value. The free energy of a structure indicates the amount of energy required to dissolve all the base pairs that are part of

the structure. The stability of an RNA structure increases with decreasing free energy, and hence the structure itself becomes more likely to occur. Therefore IntaRNA aims at finding the minimum free energy (mfe) of the interaction.

The *Nearest Neighbor energy model* (Borer et al., 1974) is used to determine the stability of an RNA secondary structure. The energy is estimated using a set of parameters based on loops of the secondary structure. The different types of loops are illustrated in Fig 1.1.

The free energy of a secondary RNA structure is the sum of all the energy contributions of its enclosed loops: $E(P) = \sum_{(i,j)\in P} E^P_{i,j}$ with:

$$E^P_{i,j} = \begin{cases} e^H(i,j) & : \text{if hairpin loop} \\ e^S(i,j,i+1,j-1) & : \text{if stack} \\ e^B(i,j,i+1,j') \text{ or } e^B(i,j,i',j-1) & : \text{if bulge} \\ e^I(i,j,i',j') & : \text{if internal loop} \\ e^M(i,j) & : \text{if multi-loop} \end{cases} \tag{1.2}$$

with e^H, e^S, e^B, e^I and e^M the respective energy contributions for each type of loop, as eg. provided by the Turner lab (Turner and Mathews, 2010) or Andronescu (Andronescu et al., 2010). (i', j') describes the enclosed base pair of a stack, bulge or interior loop.

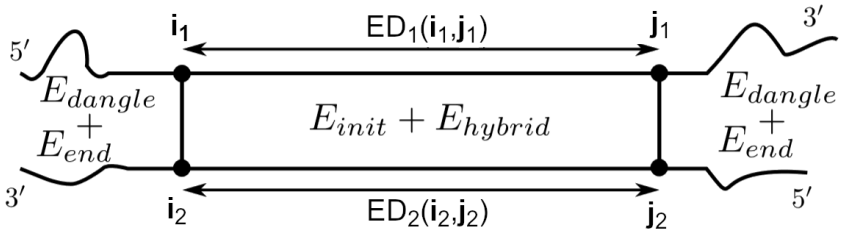


Figure 1.2: *Energy contributions in IntaRNA. The image was taken and modified from Gelhausen (2018)*

The energy of an interaction I, as illustrated in Fig 1.2, is defined in Eq 1.3.

$$E(I) = E_{init} + E_{hybrid}(I) + ED(i(I), j(I)) \tag{1.3}$$

where E_{init} is the *intermolecular initiation energy*, which is a constant value set at 4.1 kcal/mol for the default Turner04 energy parameter set.
In the following the single bases of a base pair x will be denoted by x_1 and x_2 respectively. E_{hybrid} is the *hybridization energy* and computed as in Eq 1.4.

$$E_{hybrid}(I) = \sum_{k<l\in I} \begin{cases} e^S(k_1, k_2, k_1+1, k_2-1) \\ : \text{if stack} \\ e^B(k_1, k_2, k_1+1, l_2) \text{ or } e^B(k_1, k_2, l_1, k_2-1) \\ : \text{if bulge} \\ e^I(k_1, k_2, l_1, l_2) \\ : \text{if internal loop} \end{cases} \tag{1.4}$$

Finally, ED is the *accessibility penalty* and represents the energy required to make the interacting regions accessible by preventing intra-molecular base pairs. It is calculated as the difference of the ensemble energy of all structures $\mathcal{P}$ that can be formed by the RNA sequence M and the ensemble energy of all structures $\mathcal{P}^u$ with accessible interaction site.
In order to compute the accessibility penalty of an interval (i, j), the probability of unpaired regions $Pr_u[k, l]$ cab be used, as shown below:

$$\begin{aligned} ED(i,j) &= -(E(\mathcal{P}) - E(\mathcal{P}^u_{i,j})) \\ &= E(\mathcal{P}^u_{i,j}) - E(\mathcal{P}) \\ &= -\mathsf{RT}log(Z^u_{i,j}) - \mathsf{RT}log(Z) \\ &= -\mathsf{RT}log(\frac{Z^u_{i,j}}{Z}) \\ &= -\mathsf{RT}log(\mathcal{P}^u_{i,j}) \end{aligned}$$

For details on the formula above, please refer to Mückstein et al. (2006) and McCaskill (1990). The total interaction accessibility penalty is then: $ED(i, j) = ED_1(i_1, j_1) + ED_2(i_2, j_2)$, i.e. the sum of the respective terms for both sequences.
The additional energy values to score the unpaired "dangling" neighbored bases of interaction ends (E_{dangle}) and interaction closing base pairs (E_end) from Fig 1.2 are ignored in the rest of the thesis for simplification purposes.

1.3.3 Accessibility-based RRI optimization

RNA molecules fold by intermolecular base pairing, which decreases the free energy by introducing hydrogen bonds. Therefore the most likely structure is the one with the minimum free energy (mfe).

IntaRNA uses the accessibility-based method first introduced in RNAUp (Mückstein et al., 2006) to estimate the RRI with the minimum free energy by taking advantage of existing intramolecular structure prediction tools such as UNAfold (Markham and Zuker, 2008) and ViennaRNA (Lorenz et al., 2011). Additionally, IntaRNA implements, in its recent version 3+, seed-based predictors that reduce the runtime complexity by limiting the computation to a limited amount of seed regions. The interaction with the minimum estimated free energy is likely the most stable structure and therefore the structure fulfilling the function of the RNA molecule. It can be found using Eq 1.5.

$$
\begin{aligned}
I_{mfe} &= \arg\min_{I \in \mathcal{I}} E(I) \\
&\overset{(1.3)}{=} \arg\min_{I \in \mathcal{I}} \Big(ED(I) + E_{hybrid}(I) \Big) \\
&= \arg\min_{R} \Big(ED(R) + \min_{\substack{I \\ R(I)=R}} E_{hybrid}(I) \Big) \\
&\overset{(1.6)}{=} \arg\min_{R} \Big(ED(R) + H_R \Big)
\end{aligned}
\tag{1.5}
$$

where the minimum hybridization energy H_R of a region $R = (i, j)$ between sequences M^1 and M^2 is computed using the Nearest Neighbor model as seen in Eq 1.6.

$$
H_{i,j} = \begin{cases}
E_{init} \\
\quad : \text{if } (M^1_{i_1}, M^2_{i_2}) \text{ can pair, } i = j, \\
\min\limits_{i<k\leq j} \{E_{hybrid}(\{i,k\}) + H_{k,j}\} \\
\quad : \text{if } (M^1_{i_1}, M^2_{i_2}) \text{ and } (M^1_{j_1}, M^2_{j_2}) \text{ can pair, } i \neq j, \\
+\infty \\
\quad : \text{otherwise.}
\end{cases}
\tag{1.6}
$$

1.3.4 Partition functions and probabilities

The aim of this thesis is to compute intermolecular base pair probabilities for RNA-RNA interactions. In comparison intramolecular base pairs of single RNA sequences can be computed using the McCaskill algorithm, which takes advantage of the Boltzmann distribution. The Boltzmann distribution is best suited for computing base pair probabilities according to the maximum entropy principle because it only requires the ensemble to be in an equilibrium, hence providing a large information gain with little information content.
The Boltzmann weight of an energy is defined by the following formula:

$$w(E) = \exp\left(\frac{-E}{\mathsf{RT}}\right)$$

where T is the temperature and R is the gas constant.
The following abbreviations are introduced to symbolize the Boltzmann weights of the free energy of an RNA secondary structure P and an interaction I respectively:

$$w(P) = w(E(P))$$
$$w(I) = w(E(I))$$

In order to find the base pair probabilities of an RRI, we need the partition function Z, which is the sum of the Boltzmann weights of the interaction ensemble $\mathcal{I}$, meaning

all the possible interactions for two given RNA sequences.
Given an ensemble $\mathcal{I}$ of interactions, the total partition function can be computed as follows:

$$Z = \sum_{I \in \mathcal{I}} w(I) \tag{1.7}$$

Using the partition function it is possible to find the total probability of a specific interaction using the following formula:

$$Pr[I] = \frac{w(I)}{Z} \tag{1.8}$$

Likewise, we can calculate the probability of a specific base pair occurring by:

$$Pr[(i,j)] = \sum_{\substack{I \in \mathcal{I} \\ (i,j) \in I}} \frac{w(I)}{Z} \tag{1.9}$$

Similar to interactions, the total partition function can also be found using structures P, by simply replacing all interactions I and ensembles $\mathcal{I}$ in Eq 1.7, Eq 1.8 and Eq 1.9 by P and $\mathcal{P}$ respectively.
In the following $Z_{i,j}$ for interaction region $R = (i,j)$ will be used to denote the partition function of an interaction I. The next chapter discusses how IntaRNA finds the total and partial partition functions with and without considering seed areas.

Chapter 2

Computation of RRI partition functions

Similarly to calculating intramolecular base pair probabilities, the computation of intermolecular base pair probabilities requires the partition function of all the possible structures $\mathcal{P}$ of an interaction I from the set of all possible interactions $\mathcal{I}$.
IntaRNA contains a number of exact predictors as well as heuristic versions aiming to increase the performance. Most of these predictors are implemented both as a seed and a non-seed version. In the following, the partition function computation for both variants are explained in detail.

2.1 No-Seed Z computation

Predictors that are not based on seed-constraint strategies compute the partition function using the following formula:

$$Z = \sum_{I \in \mathcal{I}} w(I) = \sum_{R} \left(w(ED(R)) \cdot \sum_{I \in R} Z^H_{i(I),j(I)} \right) \tag{2.1}$$

where Z^H is the hybridization partition function and can be calculated recursively as follows:

$$\underset{i \leq j}{\forall} Z^H_{i,j} = \begin{cases} 0 \\ \quad : \text{if } j \text{ not complementary} \\ w(E_{init}) \\ \quad : \text{if } i = j \\ \sum\limits_{i \leq k < j} \left(Z^H_{i,k} \cdot w(E_{hybrid}(\{k,j\})) \right) \\ \quad : \text{otherwise.} \end{cases} \tag{2.2}$$

Eq 2.2 has a complexity of $O(N^4)$ under the assumption that the search range of k is limited by a restricted loop length. This assumption holds throughout the rest of the thesis.

2.2 Seed-constraint Z computation

Seed-constraint predictors are based on reducing the overall interaction search space to only a number of predetermined regions containing at least one seed $S \in \mathcal{S}$. Let I_S be the set of all interactions that contain one or more full seeds, as seen in Fig 2.1.

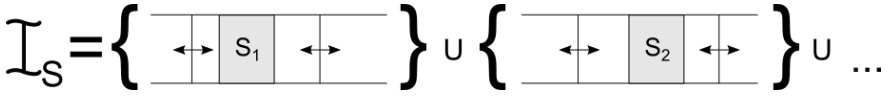


Figure 2.1: *Illustration of the set of all interactions that contain at least one full seed. S_1 and S_2 are seeds. The lines with arrows symbolize all the possible interactions containing the respective full seed.*

The seed-constraint partition function Z^S is computed as follows:

$$Z^S = \sum_{I \in \mathcal{I}_S} w(I) = \sum_{R} \left(w(ED(R)) \cdot \sum_{\substack{I \in \mathcal{I}_S \\ R(I)=R}} Z^{HS}_{i(I),j(I)} \right) \tag{2.3}$$

This means that $Z^S \leq Z$, because $\mathcal{I}_S \subseteq \mathcal{I}$. The following holds:

$$Z^{HS} = \sum_{I \in \mathcal{I}_S} w(I) \quad \leq \quad \sum_{S \in \mathcal{S}} \sum_{I \ni S} w(I) = \sum_{S \in \mathcal{S}} \sum_{i<j} Z^H_{i,i(S)} \cdot w(S) \cdot Z^H_{j(S),j} \tag{2.4}$$

The inequality in Eq 2.4 is a result of possibly counting the same interaction (containing the same seeds) multiple times when using Z^H to compute Z^{HS}. In order to fix this issue, for each seed, the left side must be completely stripped of any other seed. An illustration can be seen in Fig 2.2. This results in the following equation:

$$Z^{HS} = \sum_{I \in \mathcal{I}_S} w(I) \quad = \quad \sum_{S \in \mathcal{S}} \sum_{\substack{I \supseteq S \\ \nexists_{S' \subseteq I} : S' < S}} w(I) = \sum_{S \in \mathcal{S}} \sum_{i<j} Z^{HL}_{i,i(S)} \cdot w(S) \cdot Z^H_{j(S),j} \tag{2.5}$$

For each seed, instead of splitting the interaction into two identical matrices, it is split into a left matrix Z^{HL} containing no seeds and right matrix Z^H which can contain other seeds.
An interaction I_1 is smaller than I_2 if either base of $i(I_1)$ is smaller than the corresponding base of $i(I_2)$, and $i(I_1)$ and $i(I_2)$ are not crossing.

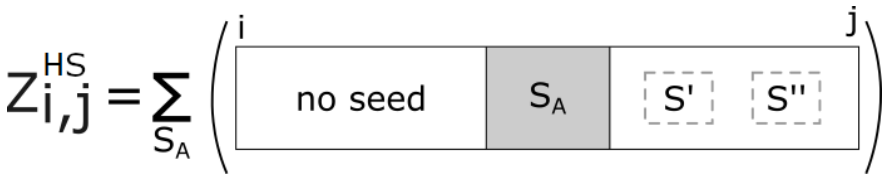


Figure 2.2: *Seed-constraint hybridization partition function computation. S_A is an anchor seed, for which the left side must contain no other seed. The right side of S_A can contain other seeds, illustrated by S' and S''.*

Z^{HL} is computed similarly to Z^H, with a fixed base pair j. However, all energy contributions of seeds must be removed. In order to avoid subtracting possible overlapping parts of seeds multiple times, the notion of *loop-overlapping* seeds is introduced. Two seeds S_1 and S_2 are called loop-overlapping if the last n loops of S_1 are the same as the first n loops of S_2. $loopOverlap(S_1, S_2)$ is defined as the number of overlapping loops between S_1 and S_2. Examples can be seen in Fig 2.3.

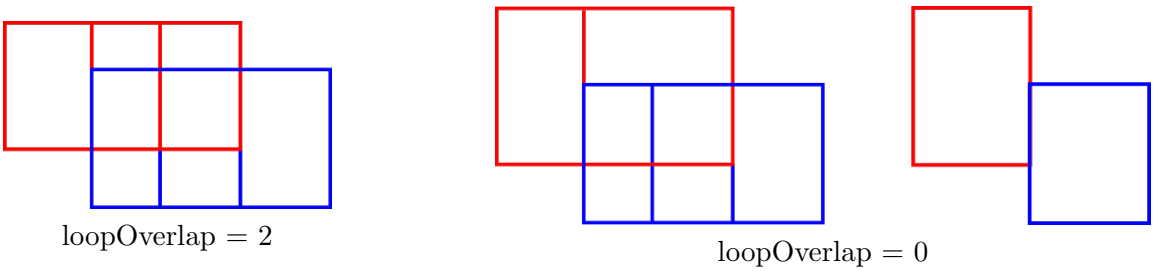
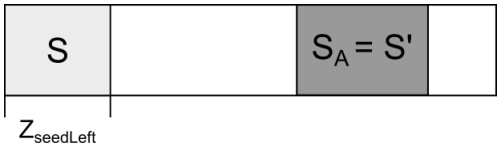


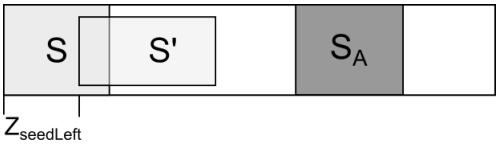
Figure 2.3: *Example of loop-overlapping vs. non loop-overlapping seeds.*

Using the definition of loop-overlap, it can be ensured that $Z^{HL}_{i,j}$ does not contain any seeds by computing $Z_{seedLeft}(i,j)$, which is the ensemble of all partial interactions containing a seed S with $i(S) = i$. All the possible cases for $Z_{seedLeft}(i,j)$ are depicted in Fig 2.4. They depend on the position of S, S_A and S' and whether or not they are loop-overlapping. S' is the leftmost loop-overlapping seed with S and can be found as follows:

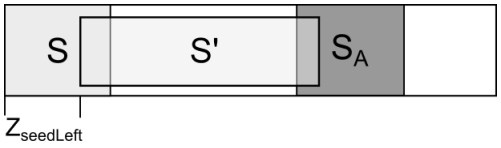
$$S' = \underset{\substack{S^* \\ i(S)<i(S^*)\leq i(S_A)}}{\arg\max} \Big(loopOverlap(S, S^*) \Big) \tag{2.6}$$



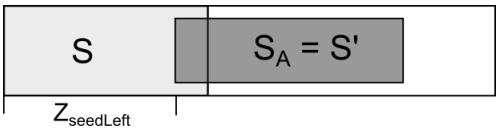
$$loopOverlap(S, S') = 0$$
$$j(S) \leq i(S_A)$$
$$S' = S_A$$



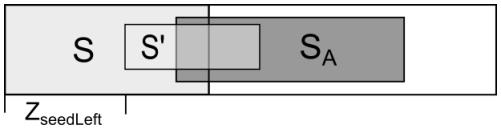
$$loopOverlap(S, S') > 0$$
$$i(S_A) \geq j(S')$$



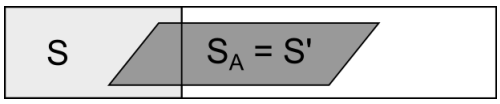
$$loopOverlap(S, S') > 0$$
$$loopOverlap(S', S_A) > 0$$
$$i(S_A) < j(S')$$



$$loopOverlap(S, S') > 0$$
$$S' = S_A$$



$$loopOverlap(S, S') > 0$$
$$loopOverlap(S, S_A) > 0$$
$$loopOverlap(S', S_A) > 0$$
$$i(S') < i(S_A)$$



$$loopOverlap(S, S') = 0$$
$$j(S) > i(S_A)$$
$$Z_{seedLeft} = 0$$
$$S' = S_A$$

Figure 2.4: *Illustration of different cases for $Z_{seedLeft}$ based on the position of the involved seeds and whether or not they are loop-overlapping each other.*

Using Fig 2.4 and Eq 2.6, the partition function $Z_{seedLeft}(i, j)$ can be computed as seen in Eq 2.7.

$$Z_{seedLeft}(i,j) = \begin{cases} 0 \\ \quad : \text{if } i(S') < j(S) \wedge loopOverlap(S, S') = 0 \\ w(S) \cdot Z^{HL}_{j(S),j} \\ \quad : \text{if } j(S) < i(S') \wedge loopOverlap(S, S') = 0 \\ w\left(\sum\limits_{\substack{i \leq k < k' \leq i(S') \\ isLoop(k,k',S)}} (E_{hybrid}(\{k, k'\})) \right) \cdot Z^{HL}_{i(S'),j} \\ \quad : \text{otherwise.} \end{cases} \tag{2.7}$$

Finally, $Z^{HL}_{i,j}$ can be computed as seen in Eq 2.8.

$$\underset{i \leq j}{\forall} Z^{HL}_{i,j} = \begin{cases} 0 \\ \quad : \text{if } i \text{ not complementary} \\ 1 \\ \quad : \text{if } i = j \\ \sum\limits_{i<k\leq j} \left(w(E_{hybrid}(\{i,k\}) \cdot Z^{HL}_{k,j} \right) - Z_{seedLeft}(i,j) \\ \quad : \text{otherwise.} \end{cases} \tag{2.8}$$

Both matrices Z^{HL} and Z^{H} can be computed with a space complexity of $O(N^2)$ each (where N is the maximum sequence length), instead of $O(N^4)$ for a single 4-dimensional matrix. The seed-constraint method also changes the runtime complexity from $O(N^4)$ to $O(|S| \cdot l^4)$ where $|S|$ is the amount of seeds and l is the constant maximum interaction length. Given a reasonable constraint on the interaction length, this results in a drastic reduction in runtime.
In order to simplify the notations in the rest of the thesis, Z^H will be used for both Z^H and Z^S and $\mathcal{I}$ will be used for both $\mathcal{I}$ and $\mathcal{I}_S$ depending on whether or not seeds are involved in the computations.

Chapter 3

Base pair probabilities

The goal of this thesis is to compute the intermolecular base pair probabilities of RRI that were predicted using IntaRNA predictors. This is done by taking advantage of the previously calculated partition functions. The resulting probabilities are gathered and represented using a dot plot. Figure 3.1 shows an example dot plot.

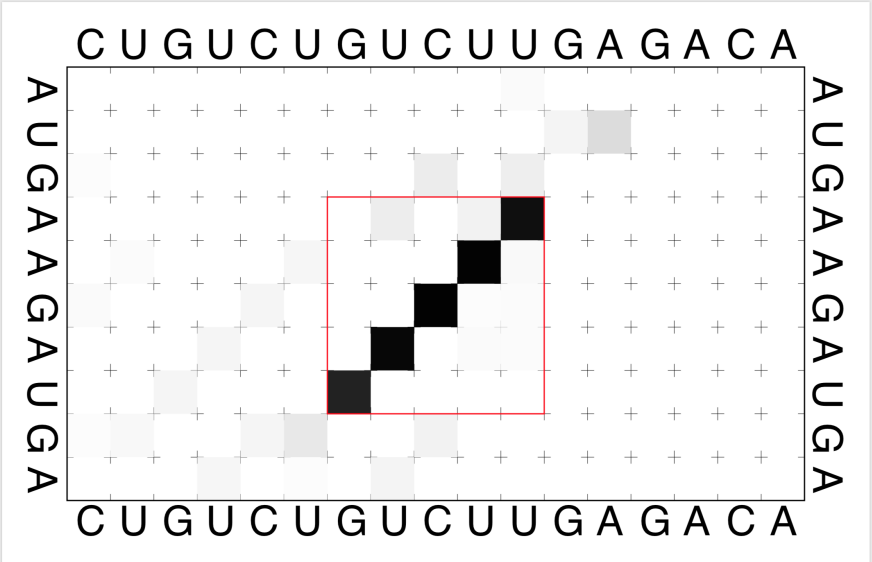
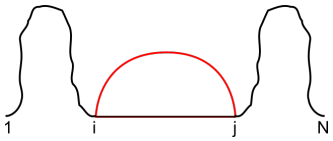


Figure 3.1: *Dot plot for the interaction of the two RNA molecules CUGUCUGUCUUGAGACA and AUGAAGAUGA. Each dot stands for the probability of an intermolecular base pair with darker colors representing higher probabilities. The red rectangle represents the most likely interaction region (mfe) found by IntaRNA.*

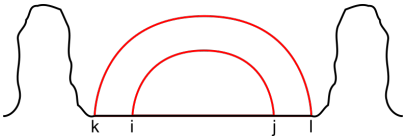
3.1 Single structure intramolecular base pair probabilities

Intramolecular base pair probabilities of single structures can be efficiently computed using a Zuker-like algorithm (Zuker and Stiegler, 1981) where the Boltzmann weights of all possible structures are summed up.
The probabilities for individual base pairs can then be computed using the McCaskill formulas by recursively calculating probabilities of interior base pairs using already computed outer probabilities. There are three different scenarios, depending on the location of base pair (i, j):

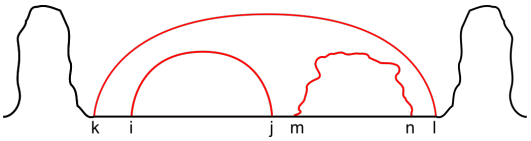


(i, j) is external: $p^E_{i,j}$

Figure 3.2: *Recursive base pair probability computation for external base pair, i.e. it is not spanned by any other base pair.*



(i, j) is inside a stacking, bulge or interior loop: $\sum_{k<i,j<l} p^{SBI}_{i,j}(k, l)$



(i, j) is inside a multiloop: $\sum_{k<i,j<l} p^{M}_{i,j}(k, l)$

Figure 3.3: *Recursive base pair probability computation for internal base pairs* (i, j) *that are enclosed by some other base pair* (k, l).

The overall probability for a base pair (i, j) is found using the following sum:

$$Pr[(i, j)] = p^E_{i,j} + \sum_{k<i,j<l} p^{SBI}_{i,j}(k, l) + \sum_{k<i,j<l} p^M_{i,j}(k, l)$$

where $p^E_{i,j}$, $p^{SBI}_{i,j}$ and $p^M_{i,j}$ are defined as in Gelhausen (2018).

3.2 RRI base pair probabilities

Unlike intramolecular base pair probabilities, their intermolecular counterparts can not be easily computed using a similar, window-based approach. This is due to the requirement of the outer context on both sides of a base pair in order to compute the corresponding accessibility values which in turn are required for the probability calculation.

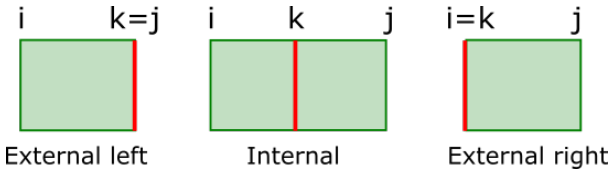


Figure 3.4: *Illustration of partition functions based on position of base pair k.*

Instead, the probability of a base pair k can be found by summing up the Boltzmann-weights of all valid interactions containing k and dividing the result by the total partition function. A visualization of the different positions of k within an interaction is shown in Fig 3.4. To handle situations where k is not external, a new notation Z_k^H is introduced that represents the hybridization partition function of interactions containing base pair k:

$$Z_k^H(i,j) = Z^H(\{I | R(I) = (i,j) \wedge k \in I\}) \cdot Z_{k,k}^H \tag{3.1}$$

Note, we extend the partition function by $Z_{k,k}^H$, ie. considering the interaction of base pair k twice, to simplify notation and presentation in the subsequent chapters.
It is not generally true that $\frac{Z_k^H(i,j)}{Z_{k,k}^H} = Z_{i,j}^H$ which is shown in the following proof:

Hypothesis: $Z_{i,j}^H = \frac{Z_k^H(i,j)}{Z_{k,k}^H}$

Counterexample: Assuming an interaction between the two RNA sequences GGG and CCC with the following counterexample:

a = *(0,0)*
b = *(1,1)* = k
c = *(2,2)*

Then the following partition functions exist:

$Z_{a,b}^H = Z(\{\{a,b\}\})$
$Z_{b,c}^H = Z(\{\{b,c\}\})$
$Z_{b,b}^H = Z(\{\{b\}\})$
$Z_{a,c}^H = Z(\{\{a,c\},\{a,b,c\}\}) \neq \frac{Z_k^H(a,c)}{Z_{k,k}^H} = Z(\{\{a,b,c\}\})$ □

Based on the different external and internal positions of base pair k as seen in Fig 3.4, its probability is then computed using equation 3.2.

$$Pr[k] = \frac{1}{Z} \cdot \sum \begin{cases} \sum_{i \leq k} Z^H_{i,k} \cdot w(ED(i,k)) & : \text{external left} \\ \sum_{i<k<j} \frac{Z^H_k(i,k)}{Z^H_{k,k}} \cdot w(ED(i,j)) & : \text{internal} \\ \sum_{k<j} Z^H_{k,j} \cdot w(ED(k,j)) & : \text{external right} \end{cases} \tag{3.2}$$

3.2.1 No-seed RRI probabilities

If the RRI prediction was performed without seeds, then all combinations of valid interactions have been computed and $Z^H_k(i,j)$ can be calculated as follows:

$$Z^H_k(i,j) = Z^H_{i,k} \cdot Z^H_{k,j}$$

This means that all the required partition functions are available and can be used to apply Eq 3.2 in order to find all base pair probabilities.

This is by far the most computationally expensive alternative as all regions in the interaction have to be considered, resulting in a complexity of $O(N^6)$ for unconstrained interaction lengths and $O(N^2 \cdot l^4)$ when constraining interaction regions to a length of $max(j-i) \leq l$.

3.2.2 Seed-constraint RRI base pair probabilities

If the seed extension strategy is used to make interaction predictions, the computation of intermolecular base pair probabilities becomes much more complicated. On the other hand, by looping over seeds with restricted length instead of all possible interactions, the complexity is drastically reduced.
As seen chapter 2.2, the partition function can be calculated with respect to the seeds in an interaction. This results in missing partition function values for some regions that are needed to compute base pair probabilities. More specifically, all the values for regions that are not including a seed are missing:

$$Z^{HS}(R) \neq 0 \leftrightarrow \exists_{S \in \mathcal{S}} : \exists_{\substack{I \in \mathcal{I} \\ R(I)=R}} : I|_{i(S)..j(S)} = S$$

Z^{HN} is introduced for all the missing hybridization partitions that do not contain a full seed.

An overview of all the possible Z^{HS} and Z^{HN} values can be seen in Fig 3.5.

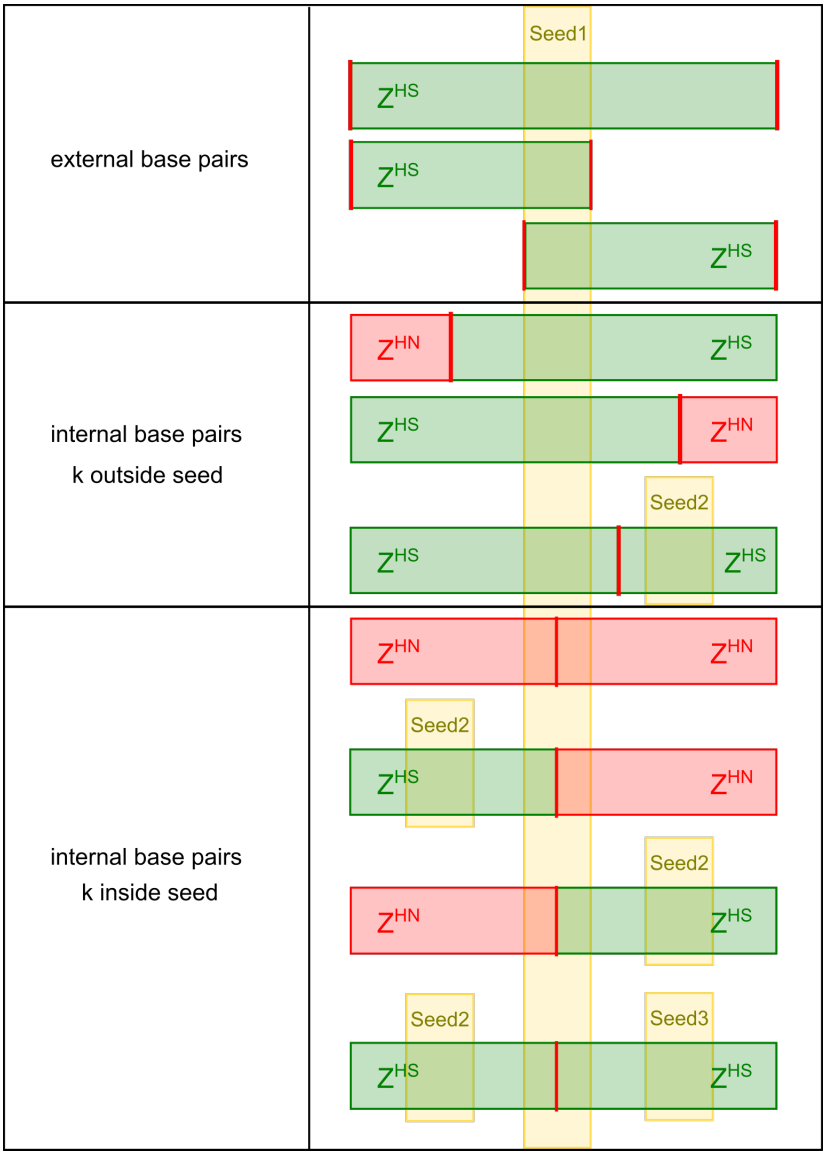


Figure 3.5: *Illustration of the possible positions for base pair k indicated as red lines and possible partition functions given the seeds marked in yellow. Green areas mark non-zero partition functions Z^{HS} as computed by Eq 2.5. Red areas mark partition functions Z^{HN} that are not defined so far for seed-constraint partition function computation.*

Probabilities with seed length below three

In the case of seeds with a length below three, the last case in Fig 3.5 can be ignored because there are no internal seed base pairs. The new formula for $Z_k^H(i,j)$ is then:

$$\begin{aligned} Z_k^H(i,j) = & Z_{i,k}^{HS} \cdot Z_{k,j}^{HS} \\ & + Z_{i,k}^{HN} \cdot Z_{k,j}^{HS} \\ & + Z_{i,k}^{HS} \cdot Z_{k,j}^{HN} \end{aligned} \tag{3.3}$$

for any given base pairs $i < k < j$. The case of $Z_{i,k}^{HN} \cdot Z_{k,j}^{HN}$ is omitted as per definition of Z^{HS} it can not occur. The following is a proof for Eq 3.3:

Hypothesis:

Given that $\forall i : Z_{i,i}^H = w(E_{init})$ *and a seed length of less than three, Eq 3.3 holds* $\forall i < k < j$.

Proof:

Let I^S *be the the set of all interactions* $I \in \mathcal{I}$ *that contain a seed and* I^N *the set of all interactions that do not contain a seed.*
$Z_{i,j}^H$ *contains the hybridization partition function of all interactions* $I_{i,j} \in \mathcal{I}$ *that either contain a seed or not.*
Therefore: $Z_{i,k}^H = Z(I_{i,k}^S \cup I_{i,k}^N) = Z_{i,k}^{HS} + Z_{i,k}^{HN}$.
Analogously, $Z_{k,j}^H = Z(I_{k,j}^S \cup I_{k,j}^N) = Z_{k,j}^{HS} + Z_{k,j}^{HN}$.

Hence:

$$\begin{aligned} Z_k^H(i,j) &= Z(I_{i,k}^S \cup I_{i,k}^N) \cdot Z(I_{k,j}^S \cup I_{k,j}^N) \\ &= \left(Z_{i,k}^{HS} + Z_{i,k}^{HN}\right) \cdot \left(Z_{k,j}^{HS} + Z_{k,j}^{HN}\right) \\ &= Z_{i,k}^{HS} \cdot Z_{k,j}^{HS} \\ &+ Z_{i,k}^{HS} \cdot Z_{k,j}^{HN} \\ &+ Z_{i,k}^{HN} \cdot Z_{k,j}^{HS} \\ &+ Z_{i,k}^{HN} \cdot Z_{k,j}^{HN} \end{aligned}$$

Finally:

$$\begin{aligned} Z_k^H(i,j) = & Z_{i,k}^{HS} \cdot Z_{k,j}^{HS} \\ & + Z_{i,k}^{HS} \cdot Z_{k,j}^{HN} \\ & + Z_{i,k}^{HN} \cdot Z_{k,j}^{HS} \end{aligned}$$

□

as $Z_{i,k}^{HN} \cdot Z_{k,j}^{HN}$ *must be 0 per definition of* Z^{HS}.

With the absence of inner seed base pairs, the formula can be rewritten as:

$$Z_k^H(i,j) \overset{(2.2)}{\underset{(2.8)}{=}} Z_{i,k}^{HS} \cdot Z_{k,j}^{H} + Z_{i,k}^{HL} \cdot Z_{k,j}^{HS} \tag{3.4}$$

where Z^{HL} and Z^{H} are the intermediate left and right partition functions that are generated during the computation of Z^{HS} in chapter 2.2, $Z_{i,k}^{HL} = Z_{i,k}^{HN}$ and $Z_{k,j}^{H} = Z_{k,j}^{HN} + Z_{k,j}^{HS}$. Some values might be missing, because so far they were only needed for seed boundaries. The missing values can however be computed using the same equations from chapter 2.2.

Finally all the required partition functions are available to apply Eq 3.2. For seeds with a length greater than two, the above assumptions do not hold anymore, which will be detailed in the next chapter.

Probabilities with arbitrary seed length

The main difference between allowing seeds of length two and longer seeds is that the latter contain inner seed base pairs. This introduces a number of challenges:

1. Eq 3.3 and Eq 3.4 do not hold anymore because no inner seed base pairs were considered as seen in Fig 3.5.
2. Inner seed base pairs introduce the possibility of overlapping seeds. This means that it has to be ensured that no partition functions are counted multiple times in the decompositions and probability computations.

In order for the decomposition in Eq 3.4 to apply for arbitrary seed lengths, all inner seed base pairs have to be considered. There are three different cases to handle when computing missing partition functions given a base pair k:

Cases 1+2: k at left or right of seed

If the base pair k lies at either side of a seed, the computation of missing partition functions remains the same as for seed lengths below three.

Case 3: k is an inner seed base pair

Due to larger seeds, it is now possible that k is an inner seed base pair which means that Eq 3.3 and Eq 3.4 do not generally hold anymore. An example with a seed S overlapping k can be seen in Fig 3.6.
$Z_{i,j(S)}^{HS}$ and $Z_{i(S),j}^{HS}$ exist because they both contain a complete seed. Therefore, given that k overlaps a single seed S, the decomposition can be done using Eq 3.5.

$$Z_k^H(i,j) = \frac{Z_{i,j(S)}^{HS} \cdot Z_{i(S),j}^{HS}}{w(S)} \tag{3.5}$$

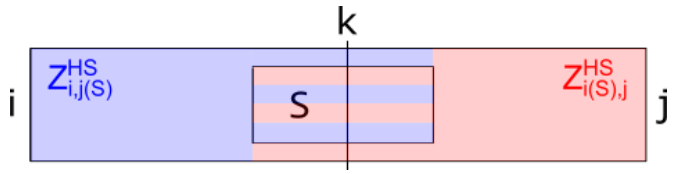


Figure 3.6: *Illustration of an interaction with a seed containing an inner seed base pair.*

However, in general there can be multiple overlapping seeds at k as seen in Fig 3.7. In this scenario the computation of the missing partition function depends on whether or not the seeds are loop-overlapping. This is due to the fact that loop-overlapping seeds have been subtracted during the seed-constraint partition function computation in chapter 2.2.

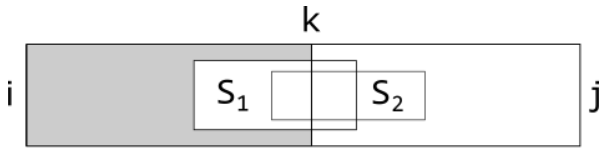


Figure 3.7: *Region (i,k) and (k,j) contain no full seeds. One or more full seeds in (i,j).*

Given the seeds S_1 and S_2 in Fig 3.7, their contribution to the decomposition of missing partition functions can be computed using Eq 3.6.

$$Z^H_k(i,j) = \begin{cases} \frac{Z^{HS}_{i,j(S_1)} \cdot Z^{HS}_{i(S_1),j}}{w(S_1)} = \frac{Z^{HS}_{i,j(S_2)} \cdot Z^{HS}_{i(S_2),j}}{w(S_2)} \text{: if } S_1 \text{ loop-overlapping } S_2 \\ \frac{Z^{HS}_{i,j(S_1)} \cdot Z^{HS}_{i(S_1),j}}{w(S_1)} + \frac{Z^{HS}_{i,j(S_2)} \cdot Z^{HS}_{i(S_2),j}}{w(S_2)} \text{: if } S_1 \text{ not loop-overlapping } S_2 \end{cases} \tag{3.6}$$

Finally, the interaction can contain multiple clusters of loop-overlapping seeds at base pair k at the same time, as seen in Fig 3.8.

In order to handle all possible cases of different numbers of loop-overlapping clusters at base pair k, the notation $C(k)$ of a set of leftmost loop-overlapping seeds containing base pair k is introduced in Eq 3.7.

$$C(k) = \{S | S \in \mathcal{S} \wedge k \in S \wedge \nexists_{S'} : (i(S') < i(S) \wedge k \in S' \wedge loopOverlap(S, S') > 0)\} \tag{3.7}$$

By combining 3.6 and 3.7, all decomposition cases with an inner seed base pair k can be computed as seen in Eq 3.8 under the assumption, that there is no seed left or right of k.

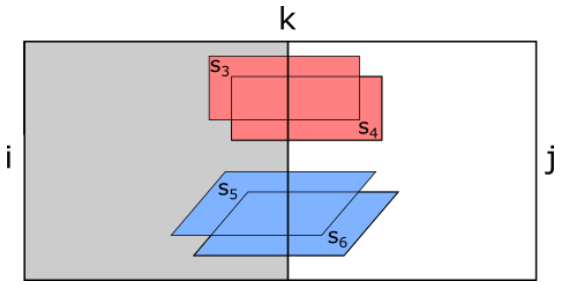


Figure 3.8: *Illustration of two clusters of loop-overlapping seeds. S_3 is loop-overlapping S_4 and S_5 is loop-overlapping S_6.*

$$Z_k^H(i,j) \overset{(3.6)}{\underset{(3.7)}{=}} \sum_{S \in C(k)} \frac{Z_{i,j(S)}^{HS} \cdot Z_{i(S),j}^{HS}}{w(S)} \tag{3.8}$$

The overall formula to compute the probability of a base pair within an interaction is thus given by Eq 3.9.

$$\begin{aligned} Z_k^H(i,j) = & Z_{i,k}^{HS} \cdot Z_{k,j}^{H} \\ & + Z_{i,k}^{HL} \cdot Z_{k,j}^{HS} \\ & + \sum_{S \in C(k)} \frac{Z_{i,j(S)}^{HS} \cdot Z_{i(S),j}^{HS}}{w(S)} \end{aligned} \tag{3.9}$$

Now equation Eq 3.4 and Eq 3.9 can be used to find all the missing partition functions depending on the context of the base pair.
This allows all intermolecular probabilities for a seed-constraint interaction prediction to be computed using Eq 3.2.

3.3 Interaction propensity profiles

By taking advantage of the RRI base pair probabilities computed in this chapter, it is also possible to calculate position-specific probabilities to be involved in a base pair. This can be done by simply adding the respective column or row of the base pair probability matrix. Given two RNA sequences M_1 and M_2, the probability of a base taking part in a basepair P_{bp} can be computed as seen in Eq 3.10.

$$\underset{i \in M_1}{\forall} P_{bp}[i] = \sum_{j \in M_2} P[(i,j)] \tag{3.10}$$

This is the intermolecular equivalent to intramolecular base pairs which can be computed by a number of different tools. An example of such a program is RNAplfold (Bernhart et al., 2006a) which provides the probability $P_{unpaired}$ that a base is not

engaged in a base pair. The intramolecular base pair probability is then simply $1 - P_{unpaired}$. An example comparison between the two types of base pairs can be found in Fig 3.9.

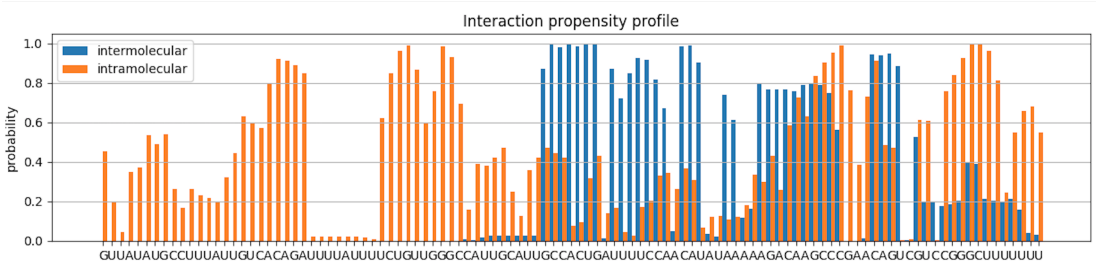


Figure 3.9: *Propensity profile of the sRNA MicC in the context of an interaction with RNA mraZ and containing probabilities of bases to engage in intermolecular and intramolecular base pairs. Intermolecular base pairs were computed using IntaRNA while the intramolecular probabilities were calculated in RNAplfold.*

Chapter 4

Results and Discussion

The introduction of ensemble-based RNA-RNA interaction computation as well as base pair probability calculation methods allow for a number of applications that would be difficult or impossible with different techniques. In the following some of these applications are presented.

4.1 Concentration computation

Using the partition functions introduced in the RRI computation, it is possible to calculate equilibrium concentrations for arbitrary species of complexes in a dilute solution, like for example different interacting RNA. This is already being used by a number of different tools such as *RNAcofold* (H. F. Bernhart et al., 2006) and *NUPACK* (Dirks et al., 2007).
In order to compute the concentration dependence between two nucleic acid sequences A and B, the free energy of the relevant molecular species are required, which are the monomers A and B, the homodimers AA and BB, as well as the heterodimer AB. Their ensemble hybridization energies can be computed using the respective partition functions.

$$E^H_{ens}(X) = -\mathsf{RT} \cdot log(Z^H(X)) \tag{4.1}$$

where X represents the ensemble of the set of possible interactions. Since the free energies of dimers F_{AA}, F_{BB}, F_{AB} provided by IntaRNA do not contain intra-molecular energies, they have to be added manually as seen in Eq 4.2.

$$\begin{aligned} F_{AA} &= E^H_{ens}(AA) + 2 \cdot F_A \\ F_{BB} &= E^H_{ens}(BB) + 2 \cdot F_B \\ F_{AB} &= E^H_{ens}(AB) + F_A + F_B \end{aligned} \tag{4.2}$$

The free energies can then be used to calculate the free binding energy:

$$\Delta F = F_{AB} - F_A - F_B = E^H_{ens}(AB) \tag{4.3}$$

where the free energies of monomers F_A and F_B are computed by a tool like RNAfold using McCaskill's algorithm. The final computation of the concentration dependency

is detailed in H. F. Bernhart et al. (2006) and not part of IntaRNA itself. In order to compare the accuracy between both RNAcofold and IntaRNA, the concentration dependency plot from the RNAcofold paper has been reproduced using the latest version of each tool (IntaRNA 3.1.1 and RNAcofold 2.4.14 at the time of writing) and the same energy model (Turner99). The resulting plots can be seen in Fig 4.1.

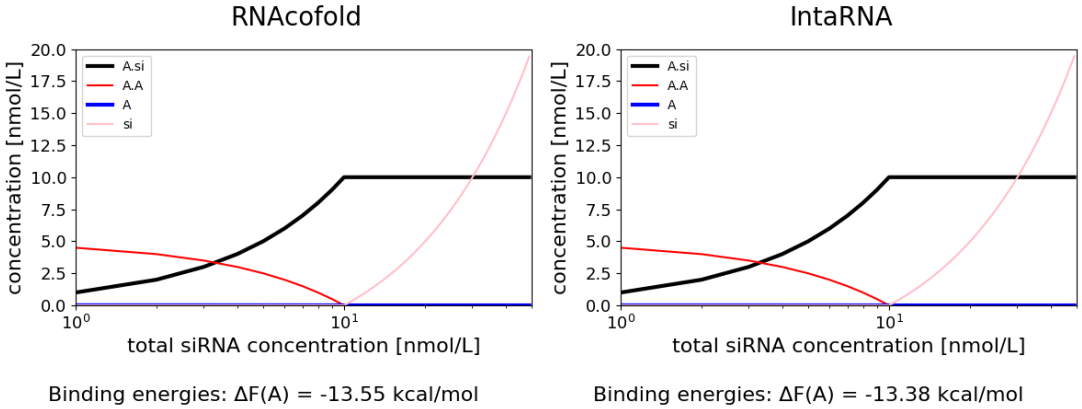


Figure 4.1: *Concentration dependency for mRNA-siRNA binding. Left: RNAcofold. Right: IntaRNA. si is the siRNA "VsiRNA". A is the part of mRNA "VR1 straight". The concentration of the mRNA is fixed at 10 nmol/L. Both RNA sequences were taken from Schubert et al. (2005).*

It can be seen that both plots are practically indistinguishable with only an insignificant difference in the binding energy $\Delta F(A)$. Therefore, it is shown that the ensemble-based predictors of IntaRNA can be used to compute concentration equilibriums similar to tools like RNAcofold. Using the formulas above, it is now possible do perform concentration dependency studies if loops are involved in the interactions. This is not possible in RNAcofold.

4.2 Verified vs non-verified sRNA-mRNA interactions

The partition functions of ensemble-based predictors allow for the computation of base pair probabilities which can be represented in dot plots. These plots are a very convenient tool for comparing the base pairing patterns of different RNA-RNA interactions. One application for such a use case is the comparison between experimentally verified and non-verified sRNA-mRNA interactions in order to find potential differences that can be used to further optimize the predictors. For the purpose of this thesis, non-verified interactions taken from Wright et al. (2013) were compared to verified interactions containing the same sRNA using top ranking targets taken from the IntaRNA benchmark (Gelhausen et al., 2019). Table 4.1 contains the verified and non-verified interactions used in the following comparisons.

sRNA	verified target	non-verified target	Figure
MicA	ompA (b0957) (Udekwu et al., 2005)	ftsB (b2748)	Fig A.1
MicC	ompC (b2215) (Chen et al., 2004)	mraZ (b0081)	Fig A.2
GcvB	cycA (b4208) (Pulvermacher et al., 2009)	mraZ (b0081)	Fig A.3
RprA	csgD (b1040) (Mika et al., 2012)	phoU (b3724)	Fig A.4
ChiX	chbC (b1737) (Overgaard et al., 2009)	opgG (b1048)	Fig A.5

Table 4.1: *Table of sRNA with corresponding verified and non-verified target mRNA.*

4.2.1 Computational results

Unfortunately, at the time of writing this thesis, the seed-based base pair computation was not yet fully implemented in IntaRNA. Therefore the following experiments and observations were all performed on dot plots generated using memory efficient (-m) ensemble-based (--model=P) no-seed (--noseed) predictors of IntaRNA. Table 4.2 shows the runtime and maximum memory consumption of each example from Table 4.1. Fig 4.2 illustrates the influence of the sRNA sequence length on the overall runtime and memory consumption.

Experiment	sRNA seq. lengths	Runtime	Max. memory
MicA-ompA	72	1.04 h	0.837 G
MicA-ftsB	72	1.12 h	0.871 G
MicC-ompC	109	5.98 h	2.434 G
MicC-mraZ	109	6.18 h	2.476 G
RprA-csgD	105	4.90 h	2.165 G
RprA-phoU	105	4.32 h	1.904 G
ChiX-chbC	84	1.44 h	1.130 G
ChiX-opgG	84	1.57 h	1.109 G
GcvB-cycA	201	44.29 h	8.897 G
GcvB-mraZ	201	51.15 h	9.633 G

Table 4.2: *Table containing the interaction RNA pairs, their corresponding sRNA sequence length, runtime and maximum used memory*

It can be seen that due to the high runtime and space complexity of the non-seed-based prediction and base pair computation discussed in 2.1 and 3.2.1 respectively, the calculation quickly becomes unfeasible for larger sequence lengths. As an example, the interaction between *Gcvb* and *mraZ* took over two full days to finish and had a peak memory requirement of $9.6GB$.

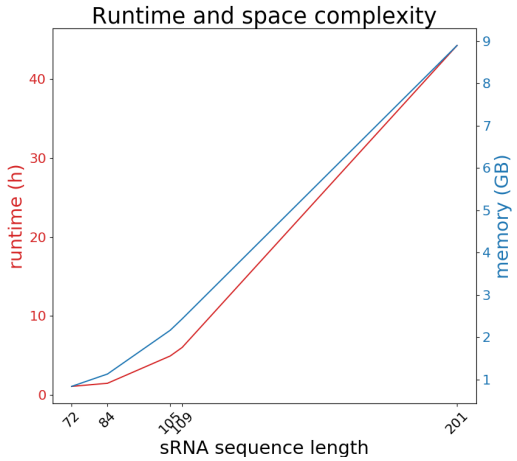


Figure 4.2: *Illustration of the influence of the sRNA sequence length on runtime and memory consumption using IntaRNAs no-seed interaction prediction with base pair probability computation.*

4.2.2 Observations

The following observations are performed on the experiments found in Table 4.1. In order to maximize the probability of identifying differences between the verified and non-verified interactions, multiple metrics have been included in the plot of each experiment, namely the probability dot plot and propensity profile of the interacting RNAs, the sequence conservation plot of the sRNA (framed in green) and the mfe structure plot of the sRNA (framed in orange).
Before discussing the observations made using all these metrics, it is worth noting that too small sRNAs severely limit the ability to find noticeable differences between verified and non-verified interactions. Therefore, especially the experiments involving sRNAs *ChiX* and *MicA* provided little new information.

Observation 1: summation-based scoring issues

A first observation that is very well represented using probability dot plots is the existence of two subhelices within the optimal interaction region found by IntaRNA, which are linked by a region of high uncertainty. The most noticeable example of this phenomenon can be seen in the verified interaction GcvB-cycA in Fig A.3. Two strong, overlapping subhelices can be seen that are connected with a high amount of low-probability base pairs. Similar examples are found in the verified interaction MicC-ompC in Fig A.2 and the non-verified interaction Rpra-phoU in Fig A.4.

The high uncertainty between two strong subhelices is likely an artifact resulting from the summation-based scoring of the minimum free energy used in IntaRNA. If the score of the total region is better than that of each single subhelix, then IntaRNA considers the overall region to be the more likely interaction. However, in reality either one of

the subhelices might be the more reasonable interaction and will be overlooked due to IntaRNAs scoring method. This problem increases with the length of the respective region as the negative impact of the uncertainty in between reduces. It is related to a similar flaw in local sequence alignment approaches like the Smith-Waterman algorithm, as discussed in Arslan et al. (2001). Designed to find highly conserved sequence fragments, the algorithm suffers from sometimes connecting well-conserved fragments by poorly-conserved parts.

Observation 2: influence of sequence conservation

More observations were made by combining the base pair probability dot plots with sequence conservation plots of the involved sRNAs. Examples include the interactions with MicC in Fig A.2 where it can be clearly seen that the first subhelix of the predicted interaction with ompC covers the highly conserved 5' end of the sRNA sequence. However the same conserved region is not part of the prediction of the non-verified interaction with mraZ. Another example of subhelices covering areas of high conservation is the verified interaction GcvB-cycA in Fig A.3. Here both subhelices of the predicted interaction are covering areas of high conservation. Lastly, the first subhelix of the non-verified interaction Rpra-phoU in Fig A.4 covers an area with little to no conservation.

Given these observations, it seems that the sequence conservation of sRNAs can be used as a metric to differentiate between verified and non-verified interactions. It looks like the predicted regions of verified interactions are more likely covering highly conserved areas than the non-verified counterparts. This suggests that it might be worth investigating possibilities to incorporate conservation values into IntaRNAs prediction model in order to further improve the quality of its predictions. The TargetRNA2 web server (Kery et al., 2014) already includes the conservation of regions of an sRNA sequence as a feature to identify regulatory targets.

Observation 3: influence of accessibility

Using propensity profile and minimum free energy plots in conjunction with the base pair probability dot plots allowed for some further observations. The predicted region for the non-verified interaction GcvB-mraZ lies in a highly structured area where the sRNA is already involved in intramolecular base pairings. This is even more noteworthy as there is an alternative, somewhat weaker region visible in the upper left of the dot plot. On the other side, the verified interaction has its predicted region located in a much more accessible area. Experiments including smaller sRNAs such as MicA and ChiX can hardly be used to make observations, because their accessibility values tend to be close to zero.

The fact that IntaRNA predicts minimum free energy regions in highly structured areas while there are alternatives in more accessible regions suggests that the influence of accessibility in the overall prediction model might be too small. A similar problem affects the interactions with short sRNAs, where the accessibility values are too low.

A possible countermeasure could be the use of local accessibility constraints instead of global region accessibility constraints.

4.3 Multi-site RNA interactions

Beside the easy comparison of different RRI interactions, base pair probability dot plots also provide a fast way of detecting multiple potential interaction sites of a single RRI. Even though IntaRNA already offers information on a given amount of suboptimal interactions, a dot plot allows for much faster investigation, as all the potential interaction sites are presented in a single image, making their relative position inside each RNA sequence immediately visible.
For testing purposes the sRNA-mRNA pair $OxyS - fhlA$ was used in order to generate the base pair probability dot plot seen in Fig 4.3 using IntaRNA. The result was then compared to the same experimental results that were also used in Salari et al. (2012). The two regions with the highest base pair probabilities in Fig 4.3 coincide with the binding sites from the literature.
However it is important to note that IntaRNA finds the different interactions sites independently from each other, meaning that in general the comparison to experimental results is of limited usefulness, because in reality they might not be independent.

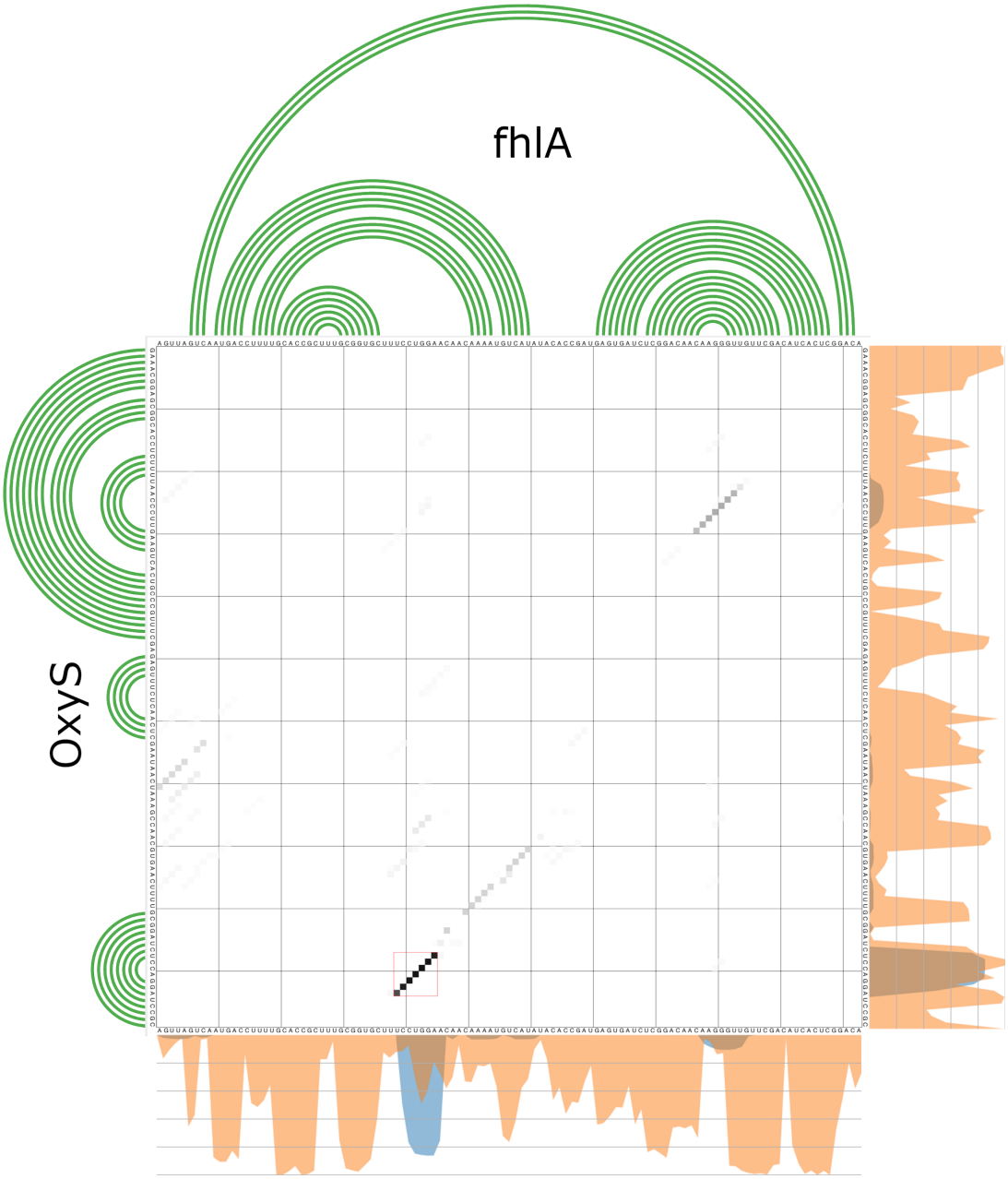


Figure 4.3: *Dot plot for the interaction of RNA molecules OxyS and fhlA. The green lines show intramolecular base pairs as computed by RNAfold. Orange areas show the probability of a nucleotide to be involved in an intramolecular base pair and blue areas the probability to form an intermolecular base pair.*

Chapter 5

Conclusion and outlook

In this thesis, the idea of seed-constraint predictors was extended by the ability to compute intermolecular base pair probabilities. The motivation was to improve the computational performance of already existing solutions by taking advantage of the drastically reduced search space in seed-based predictions. The theoretical foundation was established and all the involved difficulties were discussed for both standard and seed-based intermolecular base pair probability computation. Furthermore, the non-seed-based variant was implemented in IntaRNA. Unfortunately at the time of writing this thesis, the seed-constraint variant was not yet fully implemented and therefore unavailable for testing and comparison purposes.
Using the implemented variant, a number of experiments were conducted to showcase the possible use cases of both ensemble-based predictions and base pair probability computation. It was shown that they can be used to create concentration dependency plots by using the calculated partition functions. The dot plots generated from base pair probabilities can be used to easily compare verified and non-verified interactions in order to find metrics that might allow to further improve IntaRNAs prediction model. To a limited extend, the dot plots can also be used to identify multi-site RNA interactions.

Based on the results from this thesis a number of potential improvements of IntaRNAs prediction model could be detected. First of all, it has been seen that in some situations, IntaRNA puts too little weight on the influence of accessibility by predicting highly structured regions. Possible countermeasures for this problem could be an alignment-based accessibility estimation, additional constraints as discussed by Raden et al. (2019) or a different weighting of accessibility values versus hybridization energies. The latter could possibly be done using a pareto optimization.
Another metric useful for the differentiation between non-verified and verified interactions is the sequence conservation. This could potentially be used to further improve IntaRNAs estimations by introducing the conservation level into the prediction model itself. Possible variants of this could be the restriction of conservation-based effects to seed regions.

For some experiments, namely the ones containing smaller sRNAs such as MicA, the intermolecular base pair probability dot plots proved to be insufficient to make meaningful observations between a verified and non-verified interaction. Here more information might be required, such as an intramolecular base pair dot plot or a weighted base pair dot plot as provided by the webserver output of CARNA (Sorescu et al., 2012).

When computing the results for this thesis, the non seed-based method of calculating base pair probabilities proved to be impractical for larger interactions. The largest experiment in this thesis was done using RNA sequences of length 201 and 300 respectively and already took over two full days do compute, as seen in Table 4.2. The logical next step is therefore the implementation of the seed-based method into IntaRNA and the comparison of runtime complexity and memory consumption between both versions as well as the qualitative differences in their results.

Overall the seed-based partition function and base pair probability computation represent highly performant alternatives to existing solutions and offer exciting opportunities for further improvements by incorporating additional metrics.

Bibliography

Alkan, F., Wenzel, A., Palasca, O., Kerpedjiev, P., Rudebeck, A. F., Stadler, P. F., Hofacker, I. L., and Gorodkin, J. (2017). RIsearch2: suffix array-based large-scale prediction of RNA-RNA interactions and siRNA off-targets. *Nucleic Acids Res.*, 45(8):e60.

Andronescu, M., Condon, A., Hoos, H. H., Mathews, D. H., and Murphy, K. P. (2010). Computational approaches for RNA energy parameter estimation. *RNA*, 16(12):2304–2318.

Argaman, L. and Altuvia, S. (2000). fhlA repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. *Journal of molecular biology*, 300:1101–12.

Arslan, A. N., Eğecioğlu, m., and Pevzner, P. A. (2001). A new approach to sequence comparison: normalized sequence alignment . *Bioinformatics*, 17(4):327–337.

Backofen, R. (2011). Bioinformatics of Bacterial sRNAs and Their Targets. pages 221–239.

Bentwich, I. (2005). Prediction and validation of microRNAs and their targets. *FEBS Letters*, 579(26):5904 – 5910. RNAi: Mechanisms, Biology and Applications.

Bernhart, S. H., Hofacker, I. L., and Stadler, P. F. (2006a). Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22 5:614–5.

Bernhart, S. H., Tafer, H., Mückstein, U., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2006b). Partition function and base pairing probabilities of RNA heterodimers. *Algorithms for Molecular Biology*, pages 1748–7188.

Borer, P. N., Dengler, B., Tinoco, I., and Uhlenbeck, O. C. (1974). Stability of ribonucleic acid double-stranded helices. *Journal of Molecular Biology*, 86(4):843 – 853.

Brennecke, J., Stark, A., Russell, R., and Cohen, S. (2004). Principles of MicroRNA–Target Recognition. *PLoS Biol.*, 3:e85.

Busch, A., Richter, A. S., and Backofen, R. (2008). IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–56.

Chen, S., Zhang, A., Blyn, L. B., and Storz, G. (2004). MicC, a Second Small-RNA Regulator of Omp Protein Expression in Escherichia coli. *Journal of Bacteriology*, 186(20):6689–6697.

Dirks, R. M., Bois, J. S., Schaeffer, J. M., Winfree, E., and Pierce, N. A. (2007). Thermodynamic analysis of interacting nucleic acid strands. *SIAM Rev*, pages 65–88.

Gelhausen, R. (2018). Constrained RNA-RNA interaction prediction. Master's thesis, Albert-Ludwigs University of Freiburg.

Gelhausen, R., Will, S., Hofacker, I. L., Backofen, R., and Raden, M. (2019). IntaRNA-helix - composing RNA-RNA interactions from stable inter-molecular helices boosts bacterial sRNA target prediction. *Journal of Bioinformatics and Computational Biology*. (accepted for publication).

Gerlach, W. and Giegerich, R. (2006). GUUGle: a utility for fast exact matching under RNA complementary rules including G–U base pairing. *Bioinformatics*, 22(6):762–764.

H. F. Bernhart, S., Tafer, H., Mückstein, U., Flamm, C., Stadler, P., and Hofacker, I. (2006). Partition Function and Base Pairing Probabilities of RNA Heterodimers. *Algorithms for molecular biology : AMB*, 1:3.

Kery, M. B., Feldman, M., Livny, J., and Tjaden, B. (2014). TargetRNA2: identifying targets of small regulatory RNAs in bacteria. *Nucleic Acids Research*, 42(W1):W124–W129.

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

Markham, N. R. and Zuker, M. (2008). *UNAFold*, pages 3–31. Humana Press, Totowa, NJ.

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

Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S. H., Stadler, P. F., and Hofacker, I. L. (2006). Thermodynamics of RNA–RNA binding. *Bioinformatics*, 22(10):1177–1182.

Mika, F., Busse, S., Possling, A., Berkholz, J., Tschowri, N., Sommerfeldt, N., Pruteanu, M., and Hengge, R. (2012). Targeting of csgD by the small regulatory RNA RprA links stationary phase, biofilm formation and cell envelope stress in Escherichia coli. *Molecular Microbiology*, 84(1):51–65.

Overgaard, M., Johansen, J., Møller-Jensen, J., and Valentin-Hansen, P. (2009). Switching off small RNA regulation with trap-mRNA. *Molecular Microbiology*, 73(5):790–800.

Pulvermacher, S. C., Stauffer, L. T., and Stauffer, G. V. (2009). Role of the sRNA GcvB in regulation of cycA in Escherichia coli. *Microbiology*, 155(1):106–114.

Raden, M., Müller, T., Mautner, S., Gelhausen, R., and Backofen, R. (2019). The impact of various seed, accessibility and interaction constraints on srna target prediction - a systematic assessment. *BMC Bioinformatics*. (under review).

Salari, R., Sahinalp, C., and Backofen, R. (2012). A partition function algorithm for RNA-RNA interaction.

Schubert, S., Grünweller, A., Erdmann, V. A., and Kurreck, J. (2005). Local RNA Target Structure Influences siRNA Efficacy:Systematic Analysis of Intentionally Designed BindingRegions. *Journal of molecular biology*, 348:883–893.

Sorescu, D. A., Möhl, M., Mann, M., Backofen, R., and Will, S. (2012). CARNA—alignment of RNA structure ensembles. *Nucleic Acids Research*, 40(W1):W49–W53.

Tafer, H. and Hofacker, I. L. (2008). RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, 24(22):2657–2663.

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

Udekwu, K., Darfeuille, F., Vogel Prof. Dr, J., Reimegård, J., Holmqvist, E., and Wagner, E. G. H. (2005). Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA. *Genes and development*, 19:2355–66.

Wright, P., Richter, A., Papenfort, K., Mann, M., Vogel Prof. Dr, J., Hess, W., Backofen, R., and Georg, J. (2013). Comparative genomics boosts target prediction for bacterial small RNAs. *Proceedings of the National Academy of Sciences of the United States of America*, 110.

Zuker, M. and Stiegler, P. (1981). Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Research 9.1*, pages 133–48.

Appendices

Appendix A

Interaction dot plots

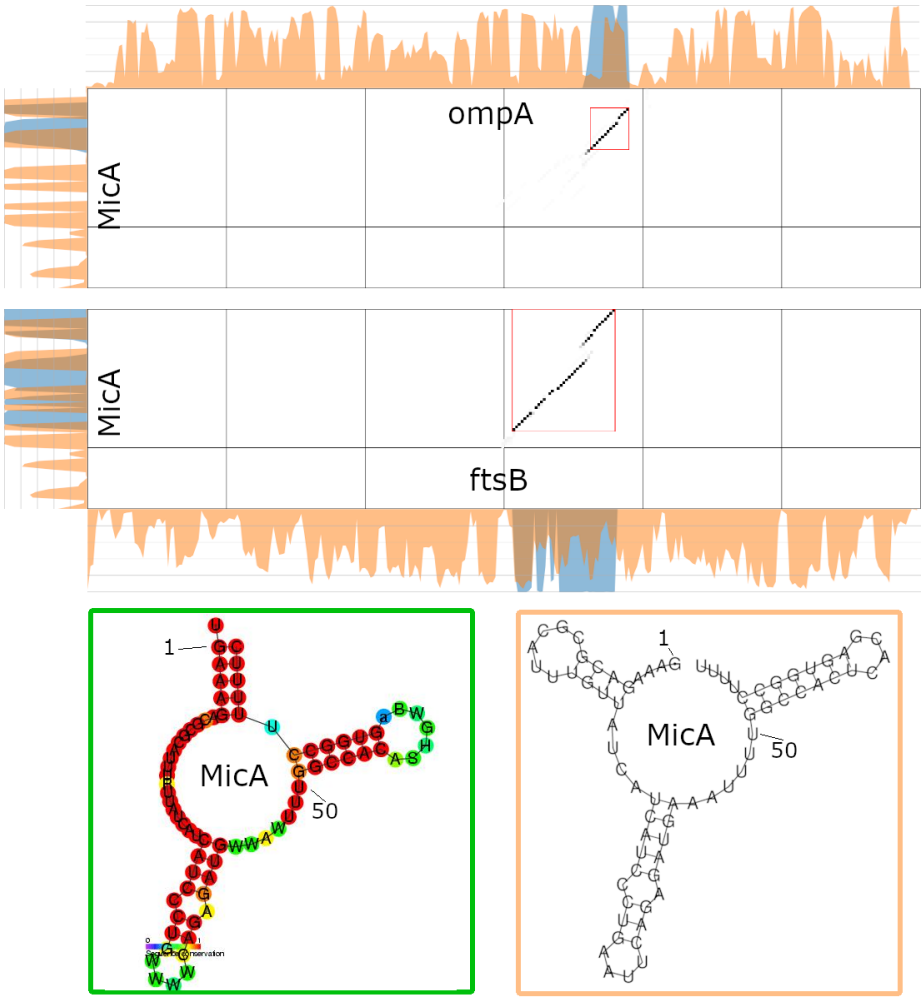


Figure A.1: *Dot plots for interactions between sRNA "micA" and mRNAs "ompA" (top) and "ftsB" (bottom). The dot plots have a grid spacing of 50 nucleotides and their probabilities are indicated using a gray scale. Darker dots indicate a higher probability. The red outline represents the minimum free energy interaction predicted by IntaRNA. The dot plots are flanked by propensity profile plots for each involved RNA which indicate the probability that the given nucleotide is involved in an intramolecular (orange) or intermolecular (blue) base pair. Framed in a green box is the Rfam RNAalifold alignment for the sRNA where colors indicate an increasing level of conservation from blue to red. The minimum free energy structure plot of the sRNA is shown in the box framed in orange. Here the same color coding indicates an increasing probability to engage in an intramolecular base pair.*

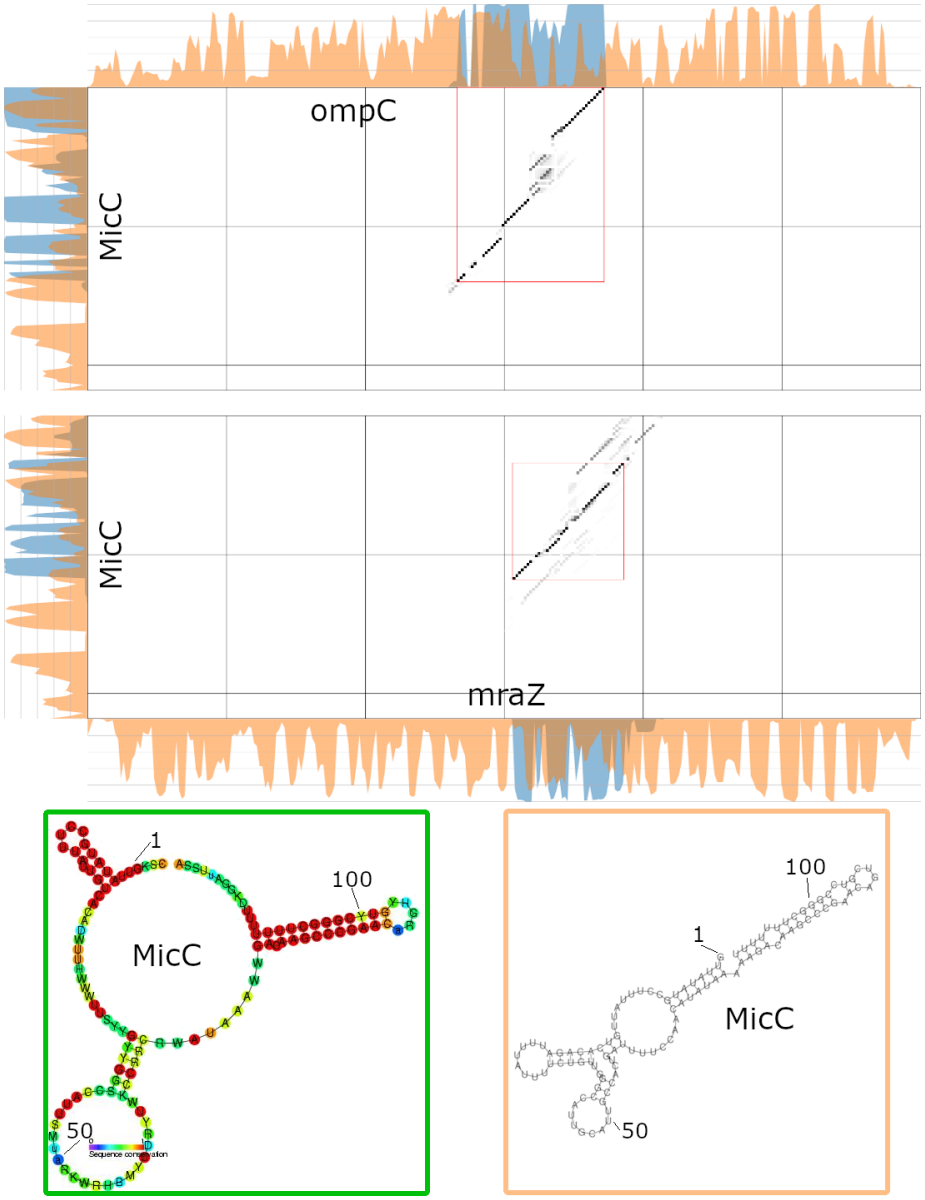


Figure A.2: *Dot plots for interactions between sRNA "micC" and mRNAs "ompC" (top) and "mraZ" (bottom). Informations on how to read the plot are provided in the caption of Fig. A.1*

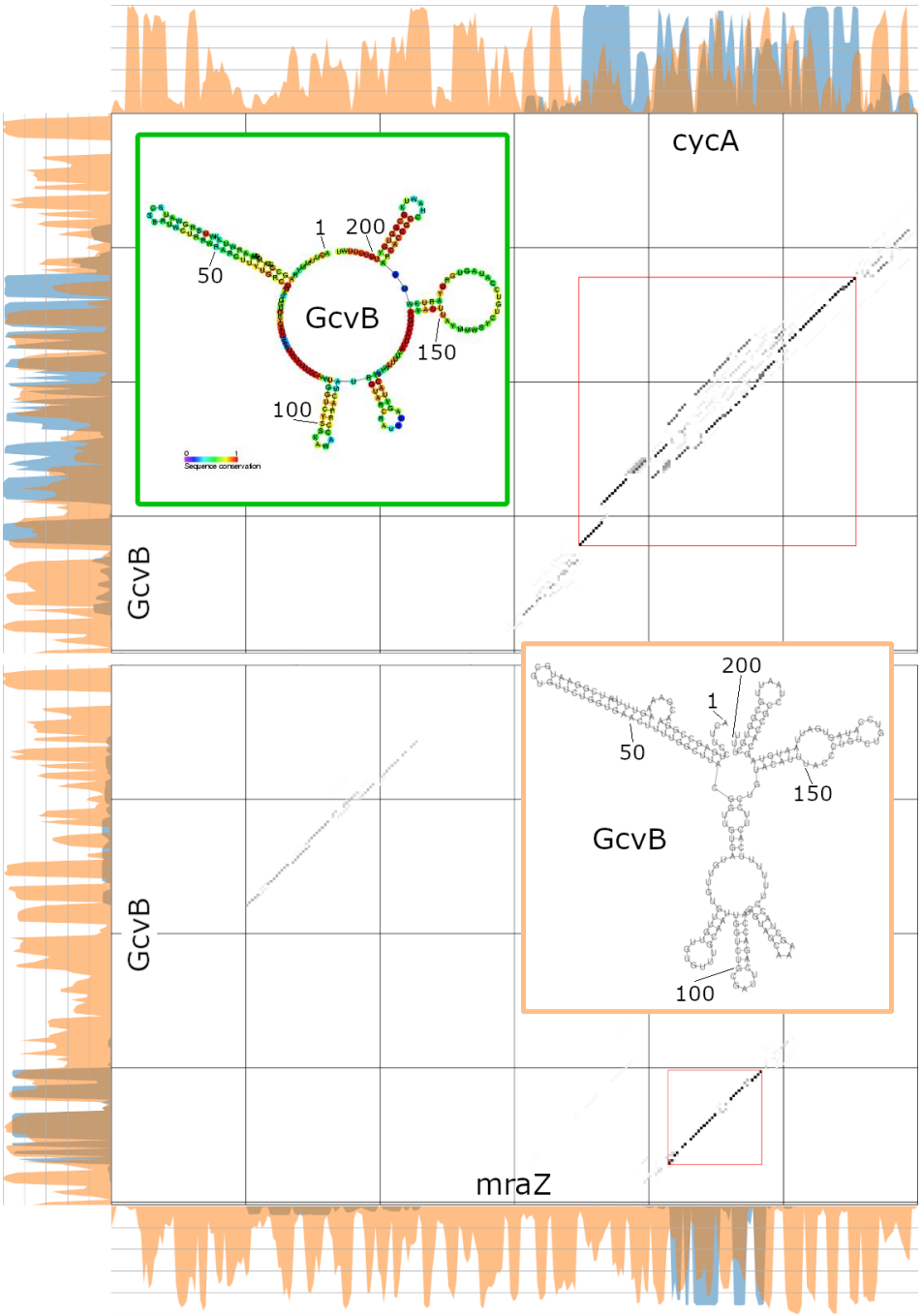


Figure A.3: *Dot plots for interactions between sRNA "gcvB" and mRNAs "cycA" (top) and "mraZ" (bottom). Informations on how to read the plot are provided in the caption of Fig. A.1*

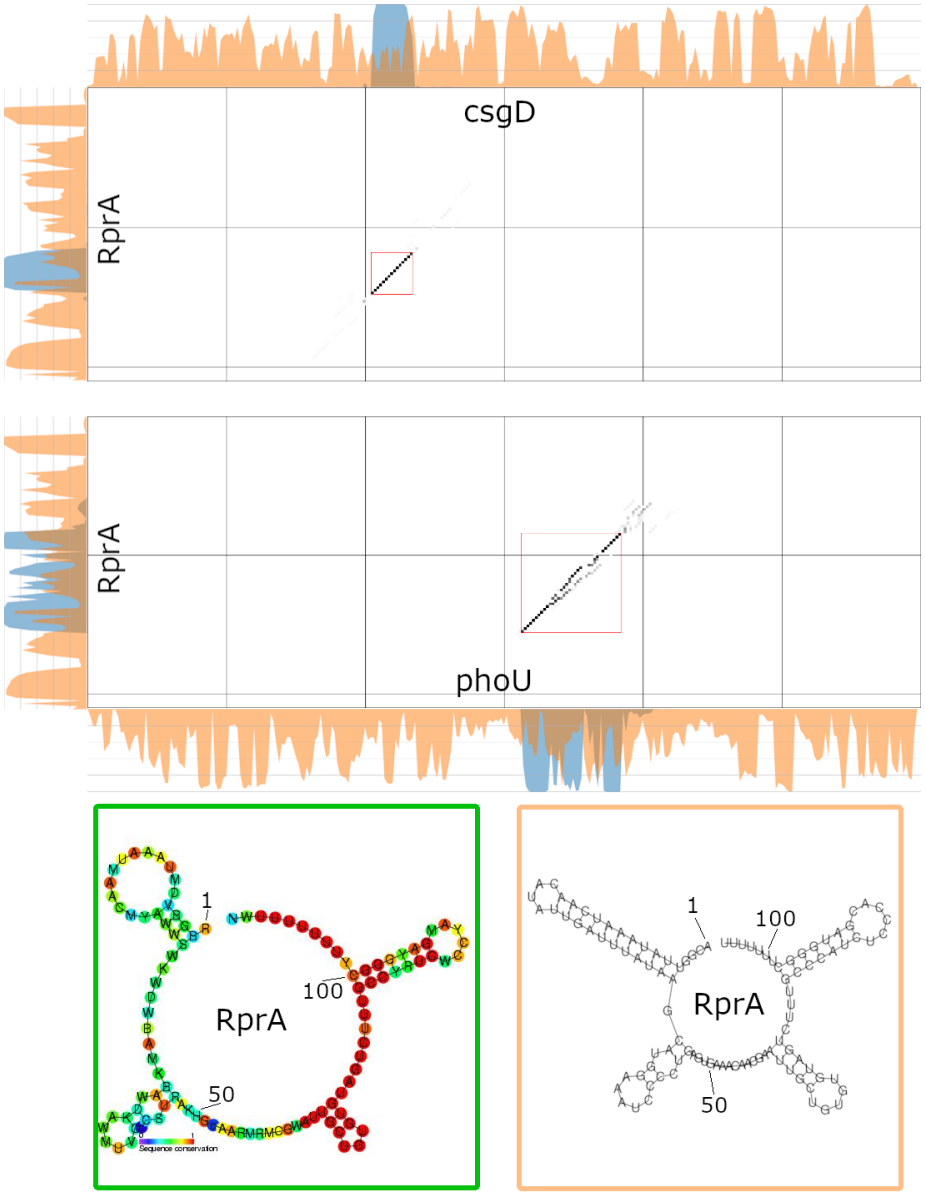


Figure A.4: *Dot plots for interactions between sRNA "rprA" and mRNAs "csgD" (top) and "phoU" (bottom). Informations on how to read the plot are provided in the caption of Fig. A.1*

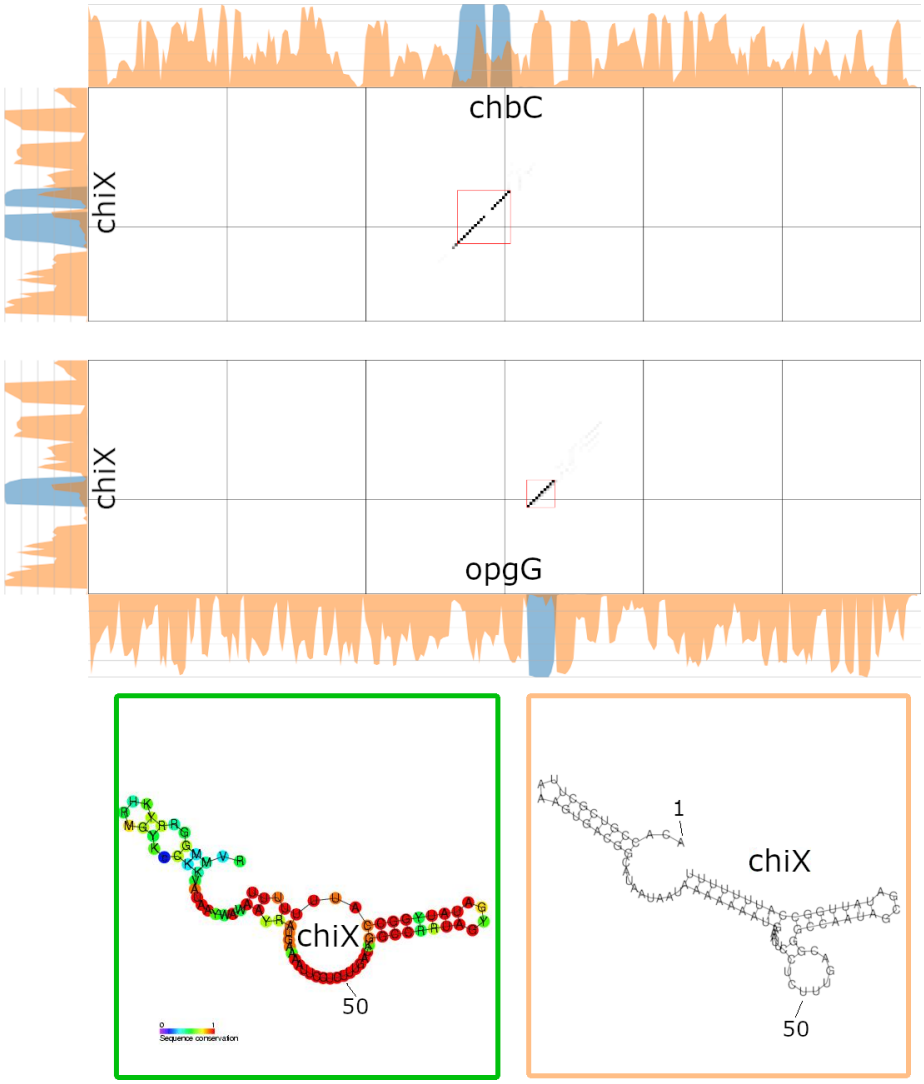


Figure A.5: *Dot plots for interactions between sRNA "chiX" and mRNAs "chbC" (top) and "opgG" (bottom). Informations on how to read the plot are provided in the caption of Fig. A.1*

Appendix B

RNA Sequences

RNA	Sequence
VsiRNA	UUCGCGUAGAAGAUGAAGUUG
VR1 straight	AGCUUGGUUGGUACAAGCGAUCUUCUACUUCAACUUCUAGAA
GcvB	ACUUCCUGAGCCGGAACGAAAAGUUUUAUCGGAAUGCGUGUUCUGGUGAACUUUUGGCUUAC GGUUGUGAUGUUGUGUUGUUGUGUUUGCAAUUGGUCUGCGAUUCAGACCAUGGUAGCAAAGC UACCUUUUUUUCACUUCCUGUACAUUUACCCUGUCUGUCCAUAGUGAUUAAUGUAGCACCGCCU AAUUGCGGUGCUUU
MicA	GAAAGACGCGCAUUUGUUAUCAUCAUCCCUGAAUUCAGAGAUGAAAUUUUGGCCACUCACGAG UGGCCUUUU
MicC	GUUAUAUGCCUUUAUUGUCACAGAUUUUAUUUUCUGUUGGGCCAUUGCAUUGCCACUGAUUUU CCAACAUAUAAAAAGACAAGCCCGAACAGUCGUCCGGGCUUUUUUU
RprA	ACGGUUAUAAAUCAACAUAUUGAUUUAUAAGCAUGGAAAUCCCCUGAGUGAAACAACGAAUUG CUGUGUGUAGUCUUUGCCCAUCUCCCACGAUGGGCUUUUUUU
ChiX	ACACCGUCGCUUAAAGUGACGGCAUAAUAAUAAAAAAAUGAAAUUCCUCUUUGACGGGCCAAU AGCGAUAUUGGCCAUUUUUUU
OxyS	GAAACGGAGCGGCACCUCUUUUAACCCUUGAAGUCACUGCCCGUUUCGAGAGUUUCUCAACUC GAAUAACUAAAGCCAACGUGAACUUUUGCGGAUCUCCAGGAUCCGC
fhlA	AGUUAGUCAAUGACCUUUUGCACCGCUUUGCGGUGCUUUCCUGGAACAACAAAAUGUCAUAUA CACCGAUGAGUGAUCUCGGACAACAAGGGUUGUUCGACAUCACUCGGACA

Table B.1: *Table of RNAs with their corresponding sequence as used in this thesis. VsiRNA and VR1 straight were taken from Schubert et al. (2005), fhla from Argaman and Altuvia (2000) while the other sequences are from the IntaRNA benchmark (Gelhausen et al., 2019).*

RNA	Sequence
ompA	CTTTTTTTTCATATGCCTGACGGAGTTCACACTTGTAAGTTTTCAACTACGTTGTAGACTTTAC ATCGCCAGGGGTGCTCGGCATAAGCCGAAGATATCGGTAGAGTTAATATTGAGCAGATCCCCC GGTGAAGGATTTAACCGTGTTATCTCGTTGGAGATATTCATGGCGTATTTTGGATGATAACGA GGCGCAAAAAATGAAAAAGACAGCTATCGCGATTGCAGTGGCACTGGCTGGTTTCGCTACCGT AGCGCAGGCCGCTCCGAAAGATAACACCTGGTACACTGGTGCTAAAC
ompC	ATCTTAAAAAGTTCCCTTGCATTTACATTTTGAAACATCTATAGCGATAAATGAAACATCTTAA AAGTTTTAGTATCATATTCGTGTTGGATTATTCTGCATTTTTGGGGAGAATGGACTTGCCGAC TGATTAATGAGGGTTAATCAGTATGCAGTGGCATAAAAAAGCAAATAAAGGCATATAACAGAG GGTTAATAACATGAAAGTTAAAGTACTGTCCCTCCTGGTCCCAGCTCTGCTGGTAGCAGGCGC AGCAAACGCTGCTGAAGTTTACAACAAAGACGGCAACAAATTAGATC
csgD	TTAGTTACATGTTTAACACTTGATTTAAGATTTGTAATGGCTAGATTGAAATCAGATGTAATC CATTAGTTTTATATTTTACCCATTTAGGGCTGATTTATTACTACACACAGCAGTGCAACATCTG TCAGTACTTCTGGTGCTTCTATTTTAGAGGCAGCTGTCAGGTGTGCGATCAATAAAAAAAGCG GGGTTTCATCATGTTTAATGAAGTCCATAGTATTCATGGTCATACATTATTGTTGATCACTAA ATCTTCTTTGCAGGCGACAGCTCTCTTGCAGCACCTTAAACAATCGC
chbC	GTTTGTTACCCAACAAACCGGTTGAAGTAATTGACTCGCTGCTTTATGGCAAAGTCGATGGTT TAGGCGTGCTTAAGGCTGCGGTTGCAGCGATTAAAAAAGCCGCAGCAAATTAATTTATTTTAA ATTTTCCCGTCAAAGAGTTATTTCATAAATCAATACCGCAATATTTAAATTGCGGTTTTTAAGG GTATTTTTCTATGAGTAATGTTATTGCATCGCTTGAAAAGGTACTCCTCCCTTTTGCAGTTAAA ATAGGAAAGCAGCCACACGTTAATGCAATCAAAAATGGCTTTATTC
mraZ	GATTTTTTCTTACAGCTATTCATAACGTTAATTTGCTTCGCACGTTGGACGTAAAATAAACAAC GCTGATATTAGCCGTAAACATCGGGTTTTTTACCTCGGTATGCCTTGTGACTGGCTTGACAAG CTTTTCCTCAGCTCCGTAAACTCCTTTCAGTGGGAAATTGTGGGGCAAAGTGGGAATAAGGGG TGAGGCTGGCATGTTCCGGGGAGCAACGTTAGTCAATCTCGACAGCAAAGGGCGCTTATCAGT GCCTACCCGTTATCGGGAACAGCTGCTTGAGAACGCTGCCGGTCAAA
ftsB	GGGCTAATTTGTACTTTCCCGTCTCCTCTGTTCATAATTCAAACCGTAACTAATAATGAGATTA TGTTCTGCACGCCCTGGGTATACGTAACAATGGACAAATGTGGTACATTTGCCCGCGTTGTCG CGGTATCCCCAACAGAGGATGTAGAGTCGTCTTCGGATGCATGGGATGATGATGCCGTTTTTC AGGGGGCAGGATGGGTAAACTAACGCTGCTGTTGCTGGCTATTCTGGTCTGGCTACAGTATTC GCTGTGGTTCGGTAAGAACGGTATACATGACTATACCCGCGTCAATG
phoU	ACCGAACTGAAGCAGGATTACACCGTGGTGATCGTCACCCACAACATGCAGCAGGCTGCGCGT TGTTCCGACCACACGGCGTTTATGTACCTGGGCGAATTGATTGAGTTCAGCAACACGGACGAT CTGTTCACCAAGCCAGCGAAGAAACAAACAGAAGACTACATCACCGGTCGTTACGGTTGATTC AGGAGTGCGTTATGGACAGTCTCAATCTTAATAAACATATTTCCGGCCAGTTCAACGCCGAAC TGGAAAGTATCCGCACGCAGGTGATGACCATGGGCGGCATGGTGGAGC
opgG	ACACGAAGTCGATGCTTCTGTCTTTAGGAGAAGCACGGAAAGTGAAAACGGTTGCAATCAGGT GCTTAATCCATGAGCCAGCGTGCTGAACGATACCGGGATTCTGTTGTCGGAATGGCTGGTTAT CCATTAAAATAGATCGGATCGATATAAGCACACAAAGGGGGAAGTGCTTACTAATTATGAAAC ATAAACTACAAATGATGAAAATGCGTTGGTTGAGTGCTGCAGTAATGTTAACCCTGTATACAT CTTCAAGCTGGGCTTTCAGTATTGATGATGTCGCAAAGCAAGCTCAAT
cycA	ATTTTTGTGAGCTGTTTCGCGTTATCACCGTGATATGACACTCACTTTAAACATAAAATTAACA TTAGATCTAAATCTTAGTATTCATCCCGCGTATTGTTACCTAATATCGATGAGTCCCGATACAG ATTCGTCGTATCATAGACTGACTAAAGGCCGTAGAGCCTGAACAACACAGACAGGTACAGGAA GAAAAAAACATGGTAGATCAGGTAAAAGTCGTTGCCGATGATCAGGCTCCGGCTGAACAGTCG CTACGGCGCAATCTCACAAACCGACATATTCAGCTTATTGCCATTG

Table B.2: *Table of genomic subsequences as used in this thesis. All sequences are positions -200 and +100 of a genome sequence around the start codon. They are taken from the IntaRNA benchmark (Gelhausen et al., 2019).*