ALBERT LUDWIGS UNIVERISTY OF FREIBURG

MASTER THESIS

Multi-site RNA-RNA Interaction Prediction

Author:
Yogapriya Ayyanarmoorthy

Supervisor:
Dr. Martin Raden

Examiner:
Prof. Dr. Rolf Backofen
Prof. Dr. Sebastian A. Will
(AMIBio, Laboratoire d'informatique de l'École Polytechnique, IPP, France)

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 April 24, 2020

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.

_____ _____

Place , Date Signature

I would like to thank Prof. Dr. Rolf Backofen so much for offering me this great opportunity to work on my Master Thesis in his group of Bioinformatics in Albert Ludwigs University of Freiburg.

I owe a lot to my supervisor Dr. Martin Raden, this thesis would not have been possible without his continuous guidance and support with both knowledge and ideas throughout the research. I want to thank him a lot for all his efforts, for providing me with thorough feedback and help throughout my work. Working with him was a great pleasure for me.

I am also grateful to my respectful parents for their great support along my stay, without whom I wouldn't have been able to stay abroad.

I want to thank my friends for being helpful and friendly, which contributed positively to my work.

Abstract

Ribonucleic acid (RNA) is an essential biological macromolecule in all biological cells. Computational prediction techniques may be used to determine how two RNA molecules will form an intermolecular base pairing. A variety of biophysical and biochemical approaches are there to test the RNA-RNA interactions (RRIs). At very large computational expense, there are a range of algorithms in the literature that can anticipate a lot of these interactions.

The identification of ncRNA targets is largely regulated by two rules, namely the consistency of the duplex between the two interacting RNAs and the internal structure of both mRNA and ncRNA. Approaches may also be divided between various major categories based on whether they consider the inter-and intramolecular framework. One such approach is accessibility based interaction prediction model, which helps us in finding the single-site interaction of two RNAs. IntaRNA is a tool for rapid and precise prediction of interactions for accessibility based approach. Here, Multi-site RNA-RNA interaction prediction is based on single site interaction prediction.

Within this thesis, we developed a model which predicts the concurrent block of interactions within an accessibility based prediction model and provides us with the prediction of joint structure for the interacting RNAs and total energy. The respective extensions of the IntaRNA package will be included in the main package for external usage and further development. The comparison between various RNAs molecules have been tested and the output is been provided with polygon plots. Further to that, we also discussed about the comparison between the RRIs prediction approaches and advance improvements and drawbacks about the thesis.

Zusammenfassung

Ribonukleinsäure (RNA) ist ein essentielles biologisches Makromolekül in allen biologischen Zellen. Computergestützte Vorhersagetechniken können verwendet werden, um zu bestimmen, wie zwei RNA-Moleküle eine intermolekulare Basenpaarung bilden. Es gibt verschiedene biophysikalische und biochemische Ansätze, um die RNA-RNA-Wechselwirkungen (RRIs) zu testen. Bei sehr hohem Rechenaufwand gibt es in der Literatur eine Reihe von Algorithmen, die viele dieser Wechselwirkungen antizipieren können.

Die Identifizierung von ncRNA-Zielen wird weitgehend durch zwei Regeln reguliert, nämlich die Konsistenz des Duplex zwischen den beiden interagierenden RNAs und die interne Struktur von mRNA und ncRNA. Ansätze können auch in verschiedene Hauptkategorien unterteilt werden, je nachdem, ob sie das inter- und intramolekulare Gerüst berücksichtigen. Ein solcher Ansatz ist das auf Barrierefreiheit basierende Interaktionsvorhersagemodell, das uns hilft, die Single-Site-Interaktion zweier RNAs zu finden. IntaRNA ist ein Werkzeug zur schnellen und präzisen Vorhersage von Interaktionen für einen auf Barrierefreiheit basierenden Ansatz. Die Vorhersage der RNA-RNA-Interaktion an mehreren Stellen basiert auf der Vorhersage der Interaktion an einer Stelle.

In dieser Arbeit haben wir ein Modell entwickelt, das den gleichzeitigen Block von Interaktionen innerhalb eines auf Barrierefreiheit basierenden Vorhersagemodells vorhersagt und uns die Vorhersage der Gelenkstruktur für die interagierenden RNAs und die Gesamtenergie liefert. Die jeweiligen Erweiterungen des IntaRNA-Pakets werden zur externen Verwendung und Weiterentwicklung in das Hauptpaket aufgenommen. Der Vergleich zwischen verschiedenen RNAs-Molekülen wurde getestet und die Ausgabe wurde mit Polygon-Plots versehen. Darüber hinaus diskutierten wir über den Vergleich zwischen den RRI-Vorhersageansätzen und über Verbesserungen und Nachteile der These.

Contents

Chapter 1

Introduction

RNA molecules play important roles in various biological processes. Their regulation and function are mediated by interacting with other molecules, e.g by forming base pairs between two RNAs, called RNA-RNA interactions (RRI). Many RNAs interact via multiple synchronous, non-overlapping subinteractions (M-RRI), e.g. OxyS-fhlA. There are fast and reliable single interaction site (S-RRI) prediction tools like IntaRNA, which helps us in predicting mRNA target sites for given non-coding RNAs (ncRNAs) , also they are capable of modelling all sites individually but not in a joint prediction. The simultaneous prediction of both intra- and inter-molecular base pairing allowing for multiple sites is computationally expensive. Some known approaches are IRIS, piRNA, NUPACK. Here we use a S-RRI prediction tool (namely IntaRNA) for the prediction of M-RRI.

1.1 Biological Background of RNA

In this thesis, I will focus on Ribonucleic acids (RNA). First of all, I would like to provide the basic biological background that is essential for the thesis. Ribonucleic acid, or RNA is one of the three major biological macromolecules that are important for all known forms of life (along with DNA (deoxyribonucleic acid) and proteins). The interaction between two RNAs plays a vital role in the basic cellular activities like transcription, RNA processing and translation. The process by which DNA is copied to RNA is called *transcription*, and that by which RNA is used to produce proteins is called *translation*. RNAs also play an important role in protein synthesis.

DNA is double stranded and RNA is a single-stranded molecule. Each strand of RNA is a sequence of four building blocks called *nucleotides*. Each nucleotide contains sugar, phosphate and nitrogen containing bases. The sugar and phosphate groups form the backbone of RNA strand and the bases bond to each other. The RNA molecules are represented as a sequence $S \in \{A, C, G, U\}^*$, where A (adenine), C (cytosine), G (guanine), U (uracil) are the bases of the nucleotide chain.

According to their potential for coding, RNA's are classified into two major categories i.e., coding RNAs and noncoding RNAs. Coding RNAs mostly refers to mRNA that encodes protein to act as different components including cell structures, signal transductors and enzymes. Noncoding RNAs act as cellular regulators with no protein encoding.
Complementary bases C-G and A-U form stable base pairs with each other using hydrogen bonds. These are called Watson-Crick pairs. Also important are the weaker U-G wobble pairs.

Together they are called *canonical base pairs*. In general, isolated base pairs are unstable. If interacting bases belong to the same molecule of RNA, they form *intra-molecular* structures and if they belong to different molecules of RNA, they form *inter-molecular* structures.

The prediction of RNA-RNA interaction is intended to predict these intermolecular structures between two RNA molecules, an extremely important step in understanding the role of ncRNAs. However, intramolecular and intermolecular structures are not mutually exclusive.

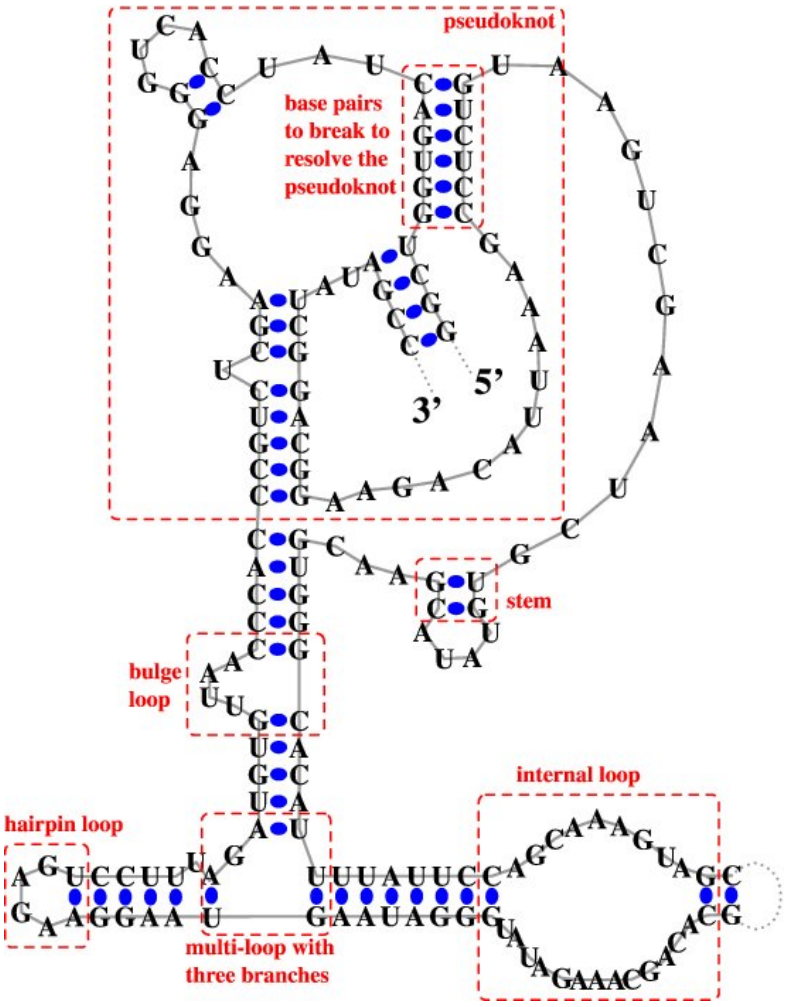


Figure 1.1: Schematic representation of the secondary structure (a set of base pairs) for the RNase P RNA molecule of Methanococcus marapaludis from the RNase P Database. Thick blue dots represents base pairs and red dashed boxes represent structural features such as stacking, bulges, hairpin , interior, multi loops and pseduoknot structure. This figure was taken from the RNAStrand webpage. (Andronescu et al., 2008)

Single stranded nucleic acid sequences contain many complementary regions that can form double helices when the molecule is folded back onto itself. The resulting pattern of double

helical stretches interspersed with loops is called the *secondary* structure of an RNA.

1.2 Formal background of RNA

Here in this section, I would like to bring up the formal definitions of ribonucleic acid.

1.2.1 RNA Structure

The RNA molecules are represented as a sequence $S \in \{A, C, G, U\}^*$. Formally, an RNA secondary structure P of S is a set of base pairs:

$$P \subseteq \{(i,j) | 1 \leq i < j \leq n, Si \text{ and } Sj \text{ are complementary } \},$$

where $n = |S|$ and for all $(i,j), (i',j') \in P$:

$$(i = i' \Leftrightarrow j = j') \text{ and } i \neq j'$$

To form a valid secondary structure, the base pairs must satisfy a number of limitations. Let the bases be numbered from 1 to n in a sequence. If the bases are complementary, a base pair may form between positions i and j , if $|j - i| \geq 4$, since there must usually be at least three unpaired bases in a hairpin loop. Let bases k and l form another allowed pair. The pair (k,l) is said to be compatible with the pair (i,j) if the two pairs can be present in a structure simultaneously. Pairs are compatible if they are non-overlapping (e.g. $i < j < k < l$) or if one is nested within the other (e.g. $i < k < l < j$). The Final case, where the pairs are interlocking or crossing (e.g. $i < k < j < l$) is called pseudo-knot. These pairs are assumed to be incompatible with most programs. An allowed secondary structure is a set of base pairs that are all compatible with each other.

They are different types of RNA secondary structure. i.e. nested and crossing structures. Crossing structures contain pseudo-knots, where two structure parts overlap. Nested structures doesn't have any crossing arcs.

1.2.2 Nested secondary structure

Nested secondary structures can be uniquely decomposed into so called loops or secondary structure elements. Depending on the number of enclosed base pairs and unpaired bases, different types of secondary structure elements are distinguished. These are hairpin loop, stacking, bulge loop, internal loop, multi loop.
Let S be a fixed sequence. Further, let P be an RNA structure for S.

- a base pair $(i,j) \in P$ is a *hairpin* loop if
  $\forall i < i' \leq j' < j : (i',j') \notin P$.
- a base pair $(i,j) \in P$ is a *stacking* if
  $(i+1, j-1) \in P$
- two base pairs $(i,j) \in P$ and $(i',j') \in P$ form an *internal* loop (i,j,i',j') if
  $i < i' < j' < j$; $(i' - i) + (j - j') > 2$; no base pair (k,l) between (i,j) and (i',j')

- An internal loop is called left (right, resp.) *bulge* if $j = j' + 1$ or $i' = i + 1$
- A k-*multiloop* consists of multiple base pairs, $(i_1, j_1)...\ (i_k, j_k) \in P$ with a closing base pair $(j_0, i_{k+1}) \in P$ with the property that $\forall 0 \leq l \leq k : (j_l < i_{l+1})$; $\forall 0 \leq l, l' \leq k$ there is no base pair $(i', j') \in P$ with $i' \in [j_l ... i_{l+1}]$ and $j' \in [j'_l ... i_{l'+1}]$.
- $(i_1, j_1)...(i_k, j_k)$ are called the *helices* of the multiloop.

1.2.3 Nearest neighbor model and energy contributions

DeVoe and Tinoco (1962) said vertical stacking of bases gives largest contribution to the stability of the RNA helix. The stacking of unpaired bases is less predictable and stable than the paired bases. Hence, the directly neighboured bases must be taken into account while estimating the energy contribution of a base pair, that results in the *Nearest Neighbor Model* (Borer et al., 1974).

The Nearest Neighbor Model enables the calculation of a free energy estimate for a given RNA secondary structure. The free energy can be taken as the amount of energy stored in a system. The system is more stable when the energy is lower. Hence, for the *most stable structure* of RNA, we go for *minimum free energy (MFE).* The energy difference between the reference state to the system is measured. We have a reference system which we use to understand the stability of the system. The reference is an RNA strcuture with no base pairing(the open chain) ie., $E(\phi) = 0$. Hence, we need to check not only the hydrogen bonds but also the stacking stability. The Nearest Neighbor Model uses a loop-based structure decomposition. To avoid the duplication of stacking, only inner stacking are taken into account.

The terminal mismatch consists of the first unpaired bases immediately after the stacking. The identity of the terminal mismatch provides the energy of the loop. In Bulge or Internal loop also we have the same energy contribution. Energy contributions for external base pairs, which are not enclosed by any other base pairs, are referred to as *dangling end contributions.*
The energy $E(P)$ Eq. 1 of a nested secondary structure P can be estimated by the sum of loop contributions (see Figure 1.2)

$$E(P) = \sum_{(i,j) \in P} \begin{cases} e^H(i,j) & : \text{if hairpin loop,} \\ e^{SBI}(i,j,k,l) & : \text{if stack/bulge/internal loop,} \\ e^M(i,j,x,x') & : \text{if Multi loop,} \end{cases} \qquad \text{(Eq. 1)}$$

Where e^H, e^{SBI} and e^M tells the context sensitive energy contributions of the loops. (k, l) represents the enclosed base pair of stack, bulge or internal and x represent the unpaired bases and x' represents the helices enclosed in the multi loop. We cannot see that there is an exponential number of possible multi loop composition. The energy for them can be calculated as below

$$e^M(i,j,x,x') = e^M_a + e^M_b x + e^M_c x'$$

where the pseudo energy parameter e^M_a scores the multi loop closing base pair (i, j) , e^M_b represents the penalty for x directly enclosed unpaired bases x and e^M_c scores x' enclosed helices .

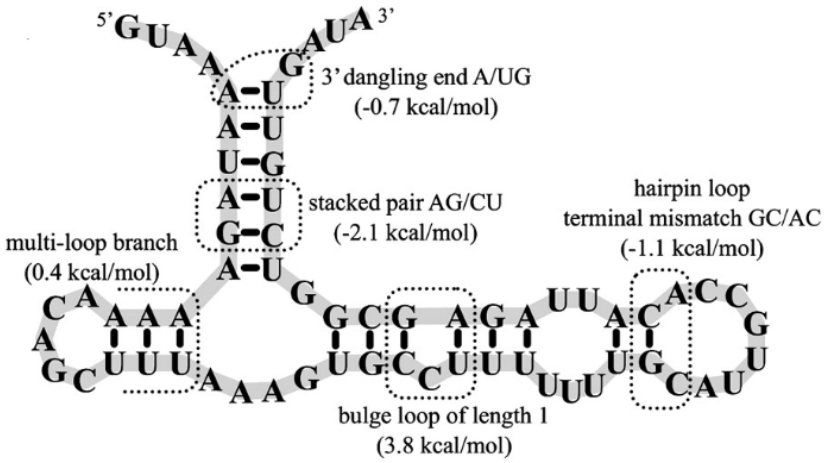


Figure 1.2: Energy contributions of loops. (Andronescu et al., 2010)

Thus the nearest neighbor model gives the energy contributions for the loop types.

From the above energy model, We can define a recursive dynamic programming algorithm to compute the structure which minimizes the energy function, this is called minimum free energy (mfe) structure. This algorithm was introduced by Zuker and Stiegler (1981).

The basic substructures of the secondary structure of the RNA sequence (i.e., stack, hairpin, internal and multi loop) are independent of each other and the energy of the secondary structure is assumed to be the sum of the energies of the substructure. The algorithm is executed in two steps with a single RNA sequence as input. Firstly, the minimum free energy of the input RNA sequences is calculated, then traceback is used to recover the respective secondary structure with the base pairs. Thus given an RNA sequence S, Zuker's algorithm predicts the non-crossing, minimal energy structure P of S in $O(n^3)$ time and $O(n^2)$ space.

1.2.4 Structure probabilities and McCaskill algorithm

Let's discuss about the structural information in terms of probabilities. According to the principal of maximum entropy (Jaynes, 1957) the best probability distribution for the calculation of the structure or base pair probability is the *Boltzmann Distribution*. These probabilities are calculated according to the Boltzmann weight. For RNA structures the unit of the energy value is $\frac{kcal}{mol}$ or $\frac{J}{mol}$. The RNA structure energy is been rescaled for Boltzmann weight computation. i.e., we replace Boltzmann constant k_B with the "mol-scaled" gas constant R to get the Boltzmann weight $w(P)$ of a structure P as:

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

Where $E(P)$ represents the state energy, R represents the gas constant and T is the temperature.
The partition function Z can be calculated using the Boltzmann weights. Z is the sum of the Boltzmann weights of all states within $\mathcal{P}$, which is the set of all possible structures P that can be formed by S.

$$Z = \sum_{P \in \mathcal{P}} w(P)$$

Z is used for the calculation of structure and base pair probabilities. So in the total sum, the distribution does not change from a macroscopic point of view, therefore thermodynamic balance is reached.
The probability of an RNA structure P is given by

$$Pr[P|\mathcal{P}] = \frac{w(P)}{Z}$$

We can also calculate the probabilities of unpaired regions. Formally, we will identify the probability of the subsequences $i..j$ to be unpaired by $\mathcal{P}^u_{i,j}$. This probability depends on the whole ensemble of structures that can be formed by the RNA molecule of interest. Thus, it can be computed by

$$\mathcal{P}^u_{i,j} = \frac{Z^u_{i_j}}{Z}$$

where $Z^u_{i,j}$ is the partition function of all structures where the subsequence $i..j$ is unpaired. i.e.,

$$Z^u_{i,j} = \sum_{P \subset \mathcal{P}^u_{i,j}} w(P) = Z(\mathcal{P}^u_{i,j})$$

where $\mathcal{P}^u_{i,j}$ is the ensemble of all structures that are unpaired between i and j. i.e.,

$$\mathcal{P}^u_{i,j} = \{P \mid \nexists (k,l) \in P \ : i \leq k \leq j \text{ or } i \leq l \leq j\} \subseteq \mathcal{P}$$

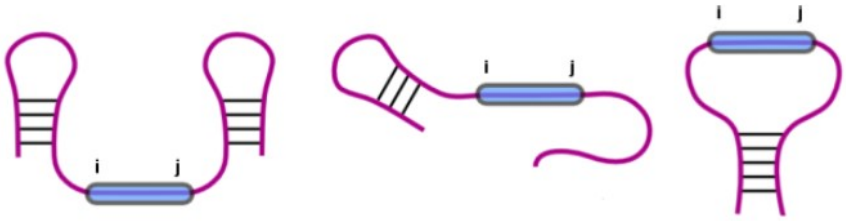


Figure 1.3: Examplary structures that are unpaired in the subsequence i..j. The Figure was inspired by the lecture material of RNA bioinformatics lecture.

The calculation of accessibility of single stranded regions is carried out using unpaired probability (Mückstein et al., 2006), hence it is very important.

Different probabilities can be calculated using McCaskill algorithm. The McCaskill algorithm (McCaskill, 1990) is used to calculate the partition function Z for a given sequence S, which can be used to compute probabilities. It enables efficient computing of the probabilities of the structure of the RNA as well as the probability that a certain base pair is formed. In addition, unpaired probabilities for subsequences can be calculated that reflect the accessibility of RNA parts for other interactions.

1.3 RNA-RNA Interaction

The interaction of RNA molecules is an essential factor for regulatory processes in all organisms. Computational prediction of RNA-RNA interactions (RRI) is a central methodology for the specific investigation of inter-molecular RNA interactions and regulatory effects of non-coding RNAs. RNA–RNA interactions are fast emerging as a major functional component in many newly discovered non-coding RNAs. They are important in many basic cellular activities including transcription, RNA processing, localization, and translation. Many RNA species function is guided by their structure, which is defined by intramolecular base pair formation. Small prokaryotic RNAs display evolutionary unstructured regions that control the expression of their target mRNAs by intermolecular base pairing (Wright et al., 2013). Hence, the prediction of both functional intramolecular and intermolecular structure of RNAs are important bioinformatics tasks.

Let's see about some simple RNA-RNA interactions. In *splicing*, small nuclear RNA's (snRNA) can recognize intronic regions of precursor messenger RNA(mRNA) which is the important step in identifying the RNA splicing products (Modrek and Lee, 2002). In *translation* transfer RNA(tRNA) interact with (mRNA) by "reading" the three letter code and define amino acid sequence (Selmer et al., 2006), (Ibba and Söll, 2000). In RNA modification, small nucleolar RNA(snoRNA) guide the modification of ribosomal RNA(rRNA) (Kiss, 2002). In microRNA (miRNA) targeting, the base pairing between miRNA and mRNA leads to degradation or translation inhibition of the mRNA (Bartel, 2004). For RNA function and regulation these examples show us the importance of the RNA-RNA interaction.

In order to allow highly accurate predictions, state-of-the-art methods not only take into account the stability (energy) of possible RNA–RNA interactions, but they also consider the

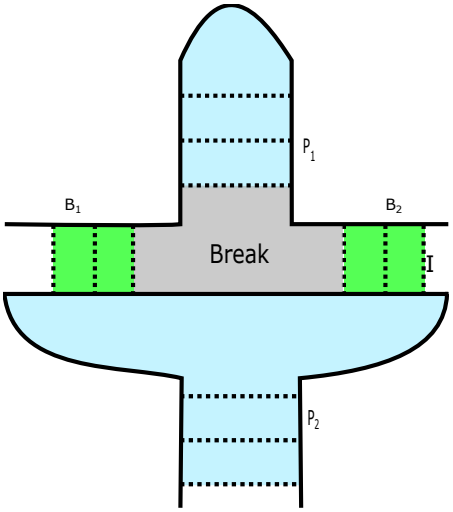


Figure 1.4: The RNA-RNA Interaction is the union of all base pairs in sequence 1 and sequence 2 are denoted as P^1 and P^2 (Blue colour) respectively. I (Green colour) denotes the union of all intermolecular base pairs. B_1, B_2 are the interaction blocks. Break (Grey colour) is the loop enclosed by two inter-molecular base pairs that also contains positions involved in intra-molecular base pairs

accessibility of the interacting subsequences (Umu and Gardner, 2017), i.e., the intramolecular structure pattern.

1.3.1 Formal background of RNA-RNA interactions

Here, we will see the formal background of RNA-RNA interactions.
In general RNA–RNA interaction prediction (RIP) problem, given two RNA sequences S^1 and S^2 (e.g., an antisense RNA and its target), the RIP problem asks one to predict their joint secondary structure. A joint secondary structure between S^1 and S^2 is a set of "pairings" where each nucleotide of S^1 and S^2 is paired with at most one other nucleotide, either from S^1 or S^2 (Alkan et al., 2006).

The RNA-RNA interaction is the combination of the set all of the base pairs in S^1, the set of all base pairs in S^2 and the total intermolecular base pairs between two sequences. Formally, the RRI can be modelled as RRI $= \uplus$ bp $(S^1) \cup \uplus$ bp $(S^2) \cup \uplus$ $(Inter)$. Basically, the set of all base pairs of S^1 is P^1 and S^2 is P^2, then $Inter$ is I which denotes the set of all intermolecular base pairs.

$$\text{RRI} = P^1 \cup P^2 \cup I$$

Now, we further decompose I into the sequence of subsets of consecutive base pairs that form interaction blocks B which is depicted in the Figure 1.4, where $I = (B_1, ..., B_x)$. A block B is the interaction block or interaction site.

Further, the interaction block or interaction site "B" can be represented as,

Figure 1.5: The left side figure shows the Positions are not paired within the loop. This problem starts with the pseudoknot which is shown in the right side figure where the same problem exists for the scoring of crossing structures.

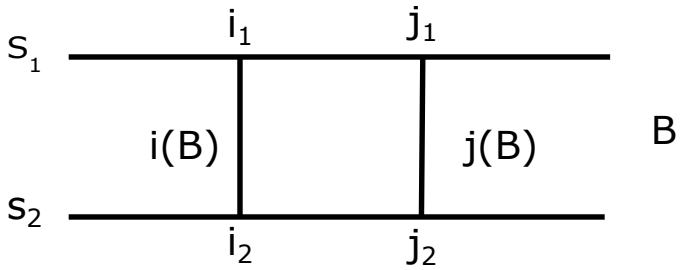


Figure 1.6: The block region $R(B)$ where the left and right most base pairs of B concerning S_1

$$B = \{(i_1, i_2) \mid S^1_{i_1} \text{ complementary to } S^2_{i_2}\} \subseteq [1, n_1] * [1, n_2]$$

Where for all $(i_1, i_2), (j_1, j_2)$ within a block B is

$$(i_1 < i_2) \iff (j_1 > j_2)$$

ie., They should be non-crossing. The block region R(B) is $(i(B), j(B))$ ie., left and right most base pairs of B concerning S^1.

$$i(B) = \underset{i=(i_1,i_2)\in B}{\arg\min}\ (i_1)$$

$$j(B) = \underset{i=(i_1,i_2)\in B}{\arg\max}\ (i_1)$$

Furthermore, no intra molecular base pairs are allowed in block region R(B) of P^1, P^2. ie.,

$$\forall_{B\in I} : \left(\nexists_{(k,l)\in P^1} : i(B)_1 \le k \le j(B)_1 \vee i(B)_1 \le l \le j(B)_1\right) \wedge \left(\nexists_{(k',l')\in P^1} : i(B)_2 \le k' \le j(B)_2 \vee i(B)_2 \le l' \le j(B)_2\right)$$

where I is the union of all blocks (ie., all inter molecular base pairs) We compute the joint structure between S_1 and S_2 through minimizing their total free energy.

The Energy for the block $E(B)$ can be calculated as ,

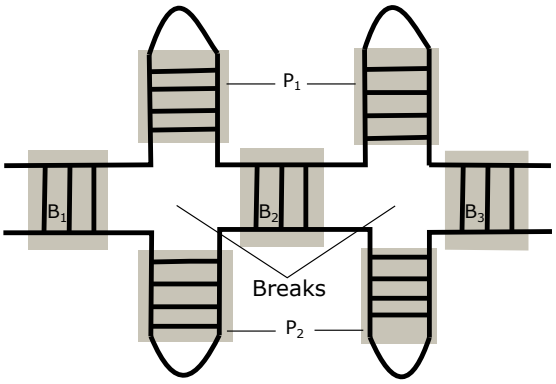


Figure 1.7: The interaction energy of RRI is the energy defined by the loops enclosed by all inter-molecular base pairs. $E(B_1)$ is the energy of block 1 and the $E(breaks)$ can be calculated from sum of all breaks.

$$E(B) = \sum_{i \in B} e^{SBI}(i_1, i_2, j_1, j_2)$$

$$j = \begin{smallmatrix} argmin(i') \\ i' < B \\ i'_1 > i_1 \end{smallmatrix}$$

The $E(I)$ can be calculated as follows,

$$E(I) = E(\uplus B) + E_{init}$$

where, $E(\uplus B) = \sum_B E(B) + E(breaks)$ and E_{init} is fixed init score if $I \neq \phi$

In the Fig 1.4, light violet colour represents the intramolecular loop with the intermolecular base pairs paired. We will need to find out, how to score them. Here, without further knowledge or energy parameters, we score it via standard loop scores ignoring the intermolecular pairings. The problem is similar to pseudoknot scoring. These also contain the loops where the positions are not paired within the loop , see Figure 1.5.

$E(breaks)$ is defined by the sum over all individual breaks between blocks (Fig 1.7). For the $E(breaks)$ it depends on the prediction model which is a tricky part and that will be discussed with the next section along with the approaches idea. Now, we will summarise the formal definition of energy of RRI. By using the above RRI equation, we can write overall energy of RRI as

$$E(RRI) = E(P^1) + E(P^2) + E_{init} + \sum_{B_i \in I} E(B_i) + \sum_{B_i \in I} E_{break}(B_i, B_{i+1}, P^1, P^2) \qquad \text{(Eq. 2)}$$

1.4 RNA-RNA Interaction Prediction Approaches

There are several available methods, that can be classified according to their underlying prediction strategies, each implicating unique capabilities and restrictions often not transparent to the non-expert user.

Mostly for RNA-RNA interaction prediction methods are based on thermodynamic models and provide an efficient computation since Richard Bellman's principle of optimality (Raden et al., 2018) can be applied. RNA–RNA interaction prediction approaches are classified into hybrid-only interaction prediction, general interaction prediction, concatenation-based/cofolding interaction prediction, and accessibility-based interaction prediction.

In the following subsections we will see about the approaches used for predicting the RNA-RNA interactions.

1.4.1 Hybridization-only interaction prediction

In the hybrid-only interaction approach, the identification of RNA-RNA interaction doesn't consider intramolecular base pairs (fig 1.8) and they can be done with $O(nm)$ time and space complexity for two RNA sequences S^1, S^2 of lengths n and m respectively (Tjaden et al., 2006).

Figure 1.8: A full duplex structure where no intramolecular base pairs are assumed. The figure is taken from the paper (Wright et al., 2018)

A dynamic programming approach using a simplified energy model with two dimensional table H is filled via the prefix-based recursion Eq. 3.

$$H_{ij} = max \begin{cases} H_{i-1,j-1} + 1 & : \text{if } S_i^1, \overleftarrow{S_j^2} \text{ are compl. base pair ,} \\ H_{i-1,j} \\ H_{i,j-1}, \end{cases} \qquad \text{(Eq. 3)}$$

Where H_{ij} is the maximal number of intermolecular base pairs for the prefixes $S_1^1..i$ and $\overleftarrow{S_1^2..j}$ the reverse sequnce of S^2. The visual representation of the recursion scheme is given in Fig: 1.9 . The above equation is the variant of the global sequence alignment approach by Needleman and Wunsch (1970) using scoring scheme i.e.,base pair instead of match/mismatch scoring forS_i^1, $\overleftarrow{S_j^2}$ no gap cost. Hence, when initialising $H_{i,0}/H_{0,j}$ with 0, the $H_{n,m}$ gives the maximal number of intermolecular base pairs and we can trace them back. As stated above, this approach has very low runtime, which are preserved when extended to energy minimization.

In order to compute the energy of an RRI using Eq. 3, no intra-molecular structure is considered, i.e. $P^1 = P^2 = \emptyset$.

Thus, eventually, only one block of inter-molecular base pairs is modelled ie., $(I = B)$ and no break is present. They are implemented in tools like TargetRNA, RNAhybrid. The main advantage of these approaches is they are very fast and easy to calculate the significance of hits. Since, intramolecular base pairing is ignored they are used for the identification of short RNA's and overestimate the length of target sites. These disadvantages can be overcome by concatenation and accessibility based approaches.

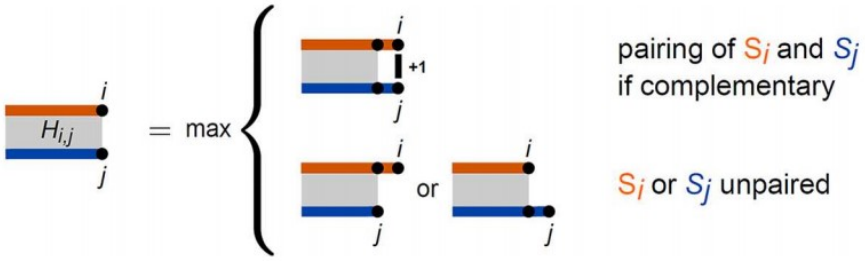


Figure 1.9: Recursion scheme to maximize intermolecular base pairs between two RNAs $S1$ and $S2$ represented in orange/blue, respectively

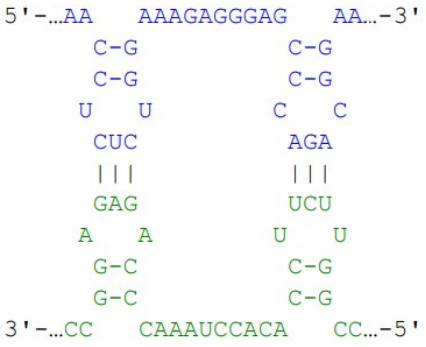


Figure 1.10: Double kissing hairpin interaction. The blue and green denotes the first and second sequence of RNA. Base pairs are denoted by dash. The picture is taken from (Wright et al., 2018)

1.4.2 General RNA–RNA interaction prediction

One of the most general approaches that is used for predicting the two intermolecular RNA molecules is IRIS (Pervouchine, 2004) method. This method is basically implemented by dynamic programming where it is the product of the sequence alignment and two MFOLD type secondary structure prediction algorithms. They can predict *general duplex structures*. This method is applied to some well known interactions such as OxyS with fhlA mRNA which basically forms a double kissing hairpin interactions as shown Fig 1.10.

It shares most common features with pseudoknots, but is less computationally intensive. The input is made up of two sequences of RNA. Each sequence can form its own nested secondary structure and hybridize into the other molecule. The time and space complexities are $O(n^3m^3)$ and $O(n^2m^2)$, where n and m are the lengths of the sequences. The configuration of the OxyS-fhlA complex proposed in (Argaman and Altuvia, 2000) consists of four neighbouring stem loops, two in each of the molecules which connect, forming two stable kissing complexes. In this method, the main goal is the simultaneous optimization of intra- and inter-molecular base pairing.

IRIS also supports crossing of consecutive blocks treated by the last recursion case in the lower

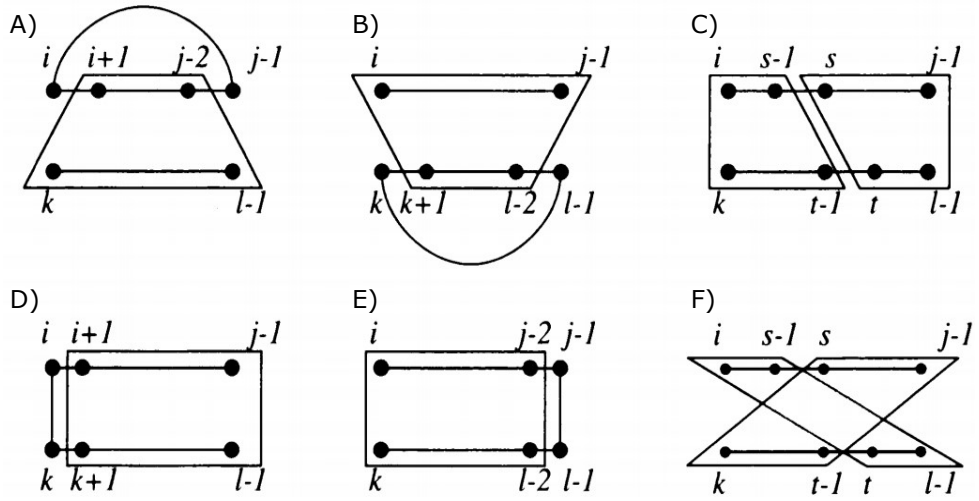


Figure 1.11: Depiction of the recursion $M_{j..l}^{i..k}$ which handles intramolecular (a,b) and intermolecular (d,e) base pair extensions as well as a general decomposition (c) and crossing (f) case. Figure is taken from the paper Pervouchine (2004)

right of Fig 1.11, which further complicates energy scoring of breaks. The energy contribution of general approach doesn't follow the interaction energy model instead they have pseudoknot energy. The energy associated with exterior pseudoknot can be given as (Xu and Chen, 2015)

$$G^{Pseudo} = \beta_1 + \beta_2 B^p + \beta_3 U^p \qquad \text{(Eq. 4)}$$

where β_1 represents penalty for introducing a pseudoknot, B^p is the number of base pairs that border the interior of the pseudoknot (i.e. number of paired positions), and U^p is the number of unpaired bases inside the pseudoknot 1.5(right). If a pseudoknot is inside a multiloop then they can be represented as β_1^m(by replacing the β_1) and if pseudoknot is inside another pseudoknot they can be represented as β_1^p (by replacing β_1).

As an approximation, one could use $E(PK-loop)$ with such pseudoknot energy terms based on G^{Pseudo} to score breaks. Note, to get an even more accurate overall energy scoring of an interaction, one would have to use pseudoknot energy terms also for such loops formed by intramol base pairs (refer 1.5 (left)). For simplicity, Eq. 2 uses only nested energy terms to assess intra-molecular energies. Thus, the exact energy computation of the general approach is not covered by the formalizations used within this thesis.

The time and space usage of IRIS are $O(n^6)$ and $O(n^4)$, respectively. The partition function version of RNA-RNA interaction prediction allows computation of probabilities of intermolecular interactions, which is used to access the stability. Due to its high complexity, several methods for reducing the requirements have been introduced. One such approach introduced by Chitsaz et al. (2009) called *biRNA* is used to predict the multiple simultaneous binding site.

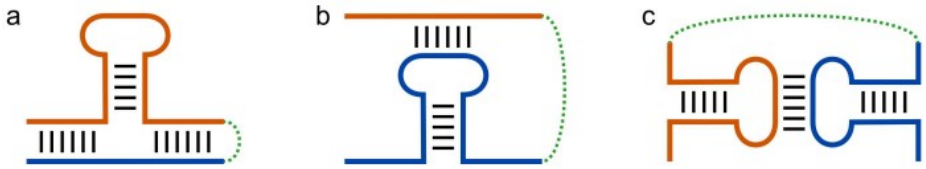


Figure 1.12: a) Pattern that can be predicated by Concatenation b)Kissing stem-loop and c) kissing hairpin interaction. Both (b) and (c) cannot be predicted as they form a crossing structure in the concatenated model. The blue and orange are the two different RNA's and the dotted green is the linker , black lines represents the base pairs. Figure inspired from paper (Raden et al., 2018)

1.4.3 Concatenation-based RNA-RNA interaction prediction

Concatenation or co-folding approach is used for predicting the interacting base pairs of two RNA molecules based on intramolecular structure prediction methods. Here,two or more sequences are concatenated into a single sequence with special inter-spacing linker sequences. The final sequence is used within an adaptation of a standard structure prediction that takes care of the linker sequences. *mfold* was the first Concatenation-based prediction tool using the Nearest-Neighbor energy model, later implemented in *MutliRNAfold* and *RNAcofold*.

RNA sequences are concatenated by a linker of length $l + 1$, where l is the minimal loop size, to ensure the concatenated sequence ends can form a base pair. We don't need any special energy treatment because the intra- and inter- molecular loops are treated equally. Hence the breaks are considered as multi loop and scored accordingly.

Concatentation-based approaches overcome the disadvantage of hybrid-only approach by incorporating the competition of intra- and intermolecular base pairing. Still they cannot predict all interaction patterns because both intra- a d inter- molecular base pairs have to be nested. For example, interactions like kissing stem-loop or kissing hairpin-loop (as seen in fig 1.12) cannot be predicted by standard tools because they form a pseudoknot in the concatenation model.

NUPACK which is a pseudoknot prediction tool that solves the problem but with the higher runtime. They are based on dynamic programming approaches for specific classes of pseudoknot structures, but does not seem to be a significant drawback in terms of the accuracy of predictions for shorter sequences. (Dirks and Pierce, 2003).

1.4.4 Accessibility-based interaction prediction

To overcome the drawbacks of concatenation approaches, Accessibility approaches have been introduced. The main aim of this approach is to integrate ensemble properties of the single sequences that are necessary for the interaction. It can predict single site interaction patterns of two respective RNA subsequences. Tools like RNAup and IntaRNA implement this strategy. Here we have to dissolve intra molecular structure before the intermolecular interaction is formed. That is, in order to form a stable interaction of intermolecular base pairs, the intra molecular base pairs have to be opened/broken.

We can classify single-site RNA-RNA interactions based on the structural context of the respective subsequences, which are

- exterior - not enclosed by any base pair.
- hairpin loop - directly enclosed by a base pair.
- non-hairpin loop - subsequence enclosed by two base pairs forming a bulge, interior or multi-loop.

IntaRNA can predict single-site interactions within any structural context of the respective subsequences, but concatenation-based approaches can only predict exterior-exterior context combinations. Energy scoring differs from normal $E(RRI)$, since intra-molecular structure is only considered implicitly via ensemble energies.

The term *ensemble* refers to the set of all secondary structures which can be formed through an RNA sequence. In an RNA sequence S, the accessibility energy of a region $[i,k]$ is determined by the energy difference (referred to as ED):

$$ED(i,k) = -(E^{all} - E^{u}_{i,k})$$

Where E^{all} denotes the energy of the set of all possible secondary structures that can be generated by sequence S and $E^{u}_{i,k}$ denotes the energy of the ensemble of structures which have a single stranded area $[i,k]$.
The partition function is the total of all states $\mathcal{P}$ over the Boltzmann factors. The energy of the ensemble E^{all} is

$$E^{all} = -RTln(Z)$$

The probability of unpaired regions can be used for calculating the accessibility penalty for an interval $[i,k]$, as shown below:

$$\begin{aligned} ED(i,k) &= -(E(\mathcal{P}) - E(\mathcal{P}^{u}_{i,k})) \\ &= E(\mathcal{P}^{u}_{i,k}) - (E(\mathcal{P}) \\ &= -RTln(Z^{u}_{i,k}) - -RTln(Z) \\ &= -RTln(\frac{Z^{u}_{i_j}}{Z}) \\ &= -RTln(\mathcal{P}^{u}_{i,k}) \end{aligned}$$

$$\begin{aligned} ED^{1}_{i,k} &= -RT \cdot \log(\mathcal{P}^{u1}_{i,k}), \\ ED^{2}_{j,l} &= -RT \cdot \log(\mathcal{P}^{u2}_{j,l}) \end{aligned}$$

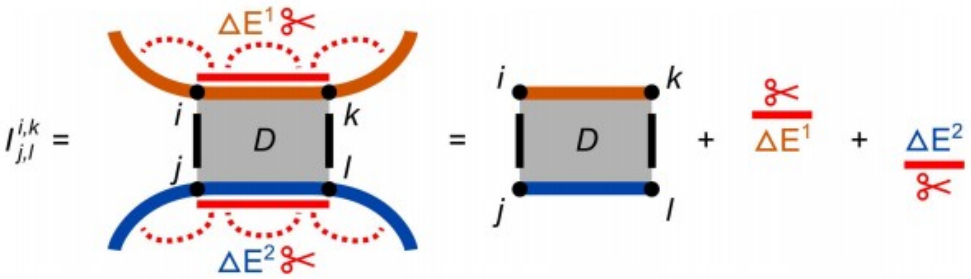


Figure 1.13: Depiction how accessibility-based approaches score an interaction of two RNAs $S1$ and $S2$ in orange and blue respectively. $\triangle E^1 + \triangle E^2$ are the energy needed to break the intramolecular base pairs and D is hybridization/duplex energy. Figure is taken from the paper (Raden et al., 2018)

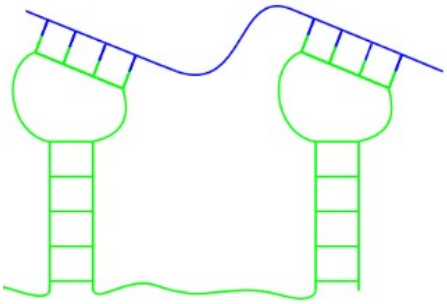


Figure 1.14: Double stem loop interaction cannot be handled by standard accessibility-based approaches as they have two binding sites (blocks) separated by intra- molecular structure. the figure is taken from COAT PhD summer school 2012

Therefore, the alternative $E(RRI)$ formula for a single interaction block B is:

$$E(RRI) = E(B) + E_{init} + Eens(S1) + ED^1 + Eens(S2) + ED^2 \qquad \text{(Eq. 5)}$$

Since $ED = E^u - E^{all}$
substituting ED value in $ED + E^{all}$ gives
$E^u - E^{all} + E^{all}$
$= E^u$
$E(RRI) = E(B) + E_{init} + E^{u1} + E^{u2}$

Since E^{all} is constant for a given sequence, accessibility-based approaches only optimize $(E(B) + E_{init} + ED^1 + ED^2)$. To this end P^u values are precomputed.

The energy of accessibility approach has no break , since the interaction I forms only one interaction block. Approaches like RNAup and IntaRNA use precalculated ED values for all possible interaction regions. They gives us how much energy is needed to free of intramolecular base pairs.

The main drawback of accessibility approach is, it can handle only one non-crossing block.

These approaches cannot be modelled correctly for the double kissing hairpin interaction which has more than one crossing blocks of interaction 1.14

1.4.5 Comparison of approaches for RRI prediction

In this subsection, we will see the comparison between the approaches for some distinct interaction pattern. Below Table 1.1 gives an overview which interaction pattern can be predicted by which approaches.

Table 1.1: Comparison of RNA-RNA interaction prediction approaches for different interaction pattern

Comparison of RRI approaches						
RRI Pattern			**RRI prediction approaches**			
Figures	RRI description	No.of blocks	Hybrid-only	General	Concatenation (nested)	Accessibility
1.8	Full duplex structure	1	yes	yes	yes	yes
1.12 (a)	Nested joint structure without pseudoknots	2	no	yes	yes	no
1.12 (b)	Stem loop interaction	1	no	yes	no	yes
1.12 (c)	Kissing hairpin loop	1	no	yes	no	yes
1.10	Double kissing hairpin loop	2	no	yes	no	no
1.4	Kissing stem interaction	2	no	yes	no	no
1.14	Double kissing stem loop	2	no	yes	no	no
NA	Best Time complexity for each approach	NA	$O(nm)$	$O(n^3m^3)O(n^3)$		$O(n^2)$

The understanding of RNA structure and RNA-RNA interaction prediction approaches is important to ensure correct result interpretation and an knowing of their limitations are necessary to avoid wrong conclusions. Here we give a concise overview of the relevant theoretical history to the most general algorithmic approaches. We could say that the accessibility-based approach is the best approach for single site RNA-RNA interaction. To handle two or multi crossing blocks of interaction, we are introducing multisite accessibility based approach. The Multi-site RRI optimization is based on single-site IntaRNA predictions. Hence, we are going for the multisite accessibility based approach in the next chapter.

Chapter 2

Multisite Accessibility Based

In this chapter, we will introduce an accessibility-based approach that can be used for multisite RNA-RNA interaction prediction. In simple words we could say, it is Multi-site RRI optimization based on single-site IntaRNA predictions. Accessibility-based RNA-RNA interaction prediction methods are typically modelling a single block of consecutive inter-molecular base pairs. Thus, interaction pattern that consists of multiple concurrently formed blocks can not be predicted. Within this thesis, we are developing and testing possibilities to efficiently predict concurrent blocks of interaction within an accessibility-based prediction model. The approach will be based on IntaRNA, which is one of the state-of-the-art programs for RNA-RNA interaction prediction.

IntaRNA, developed by (Busch et al., 2008) and (Raden et al., 2018) at the bioinformatics group at Freiburg University, is a general and fast approach to the prediction of RNA-RNA interactions incorporating both the accessibility of interacting sites as well as the existence of a user-definable seed interaction. IntaRNA uses energy minimisation to find the best possible interaction.

Many RNAs interact via multiple synchronous, non-overlapping subinteractions (M-RRI), e.g. OxyS-fhlA. The simultaneous prediction of both intra- and inter-molecular base pairing allowing for multiple sites is computationally expensive. Current approaches in include IRIS, NUPACK, piRNA- , etc. There are fast and reliable single interaction site (S-RRI) prediction tools like RNAup and IntaRNA, that often show the additional sites within their suboptimal list, ie. are capable of modelling all sites individually but not in a joint prediction. To overcome this, we use the iterative method in this thesis for finding the interaction between multiple blocks. Beforehand, details of the S-RRI accessbility-based tools RNAup and IntaRNA are introduced.

2.1 RNAup - Exact Recursion for single site

In the following, I will first introduce the RNAup-like exact recursions (Mückstein et al., 2006) and then give an overview of IntaRNA heuristic version. The total energy score of the interaction is measured as the sum of the free hybridization energy and the free energy required to make the interaction sites available.

Thus, Scoring an interaction in IntaRNA is dependent on two energy contributions:

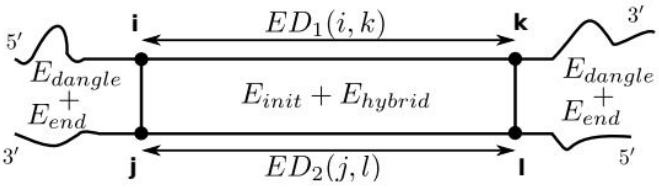


Figure 2.1: The energy contribution of IntaRNA. The image is taken from (Gelhausen, 2018)

- **Hybridization energy** : energy value E_hybrid from intermolecular base pairings in the form of stackings, bulges or internal loops, i.e., energy is typically a negative value.
- **Accessibility energy** : An amount of energy ED needed to single-strand the interacting region, i.e. not include intramolecular pairings, i.e., energy is a positive value.

The energy of an ensemble of structures is calculated using a partition function (McCaskill, 1990). Similarly, we get $ED(i,k)$ by calculating the partition function, $Z^u_{i,k}$ covering the ensemble of all structures which can be formed by a sequence S, with a single stranded region $[i,k]$. As introduced in the section 1.4.4. Therefore,

$$ED(i,k) = -RTlog(\mathcal{P}^u_{i,k}$$

Mückstein et al. (2006) gives more detailed information on the same. The hybridization energy is measured using the Nearest Neighbor Energy Model. This represents the minimum free hybridization energy of two subsequences, where a base pair is generated by the left and right most positions of both subsequences. For sub-sequences $S^1_i...S^1_k$ and $S^2_j...S^2_l$, where S^1 is ordered from 5' to 3' and S^2 in the reverse order:

$$H(i,j,k,l) = min\{E(P) \mid (i,j) \in P \wedge (k,l) \in P\}$$

The hybridization energy is calculated with a Zuker-like recursion.

$$H(i,j,k,l) = min \begin{cases} E_{init} \\ \quad \text{: if } (S^1_i, S^2_j) \text{ can pair and } i = k, j = l, \\ min_{r,s}\{e^{SBI}(i,j,r,s) + H(r,s,k,l)\} \\ \quad \text{: if } (S^1_i, S^2_j) \text{ and } (S^1_k, S^2_l) \text{ can pair } i < k \text{ and } j < l, \\ \infty \\ \quad \text{: otherwise,} \end{cases} \qquad \text{(Eq. 1)}$$

Here e^{SBI} is the energy contribution of stack, bulge and internal loop. The traceback helps us to find the base pairs of optimal interaction with energy $H(i,j,k,l)$. Both the accessibility and hybridisation energy forms the extended hybridisation energy which is the specific hybridisation between $S^1_i...S^1_k$ and $S^2_j...S^2_l$ given by,

$$C(i,j,k,l) = \begin{cases} H(i,j,k,l) + ED_1(i,k) + ED_2(j,l) \\ \quad \text{: if } (S^1_i, S^2_j) \text{ and } (S^1_k, S^2_l) \text{ can pair } i \neq k \text{ and } j \neq l, \\ \infty \\ \quad \text{: otherwise,} \end{cases} \qquad \text{(Eq. 2)}$$

We get the time and space complexity of $O(n^2m^2)$ by limiting the loop size, which is still very high. When we limit the interaction length to l , it has a complexity of $O(nml^2)$ time and $O(nml^2)$ space. The interaction with the minimum estimated free energy (mfe) is probably the most stable structure and thus the structure fulfills the RNA molecule function. Therefore, we are interested in

$$mfe = \arg\min_{i,j,k,l} C(i,j,k,l)$$

2.2 IntaRNA - Heuristic recursion for single site

The exact recursions of RNAup are not suitable for larger genome wide studies due to its high time and space complexity ie., $O(n^2m^2)$ where n represents the length of query and m is the length of target sequence. In order to overcome the time and space complexity problem, IntaRNA introduced the heuristic recursion. This recursion is based on sparsification technique. Hence, we consider only one right end of interactions with left end (i,j) which is single and locally optimal, instead of all the possible interaction. This will help us to reduce the space and time complexity to $O(nm)$, as introduced for IntaRNA version 1 & 2. The heuristic version is defined as:

$$C(i,j) = \begin{cases} E_{init} + ED_1(i,i) + ED_2(j,j) \\ \quad : \text{in the case of new interaction} \\ min_{p,q}\{e^{SBI}(i,j,p,q) + C(p,q) - ED_1(p,K(p,q)) - ED_2(q,L(p,q)) \\ \qquad +ED_1(i,K(p,q)) + ED_2(j,L(p,q))\} \\ \quad : \text{if } (S_i^1, S_j^2) \text{ can pair }, \\ \infty \\ \quad : \text{otherwise}, \end{cases} \quad \text{(Eq. 3)}$$

Here $K(p,q), L(p,q)$ are newly introduced matrices which provide the right end of the best interaction with left end (p,q). Since ED values are not additive, we have to subtract the old ED values before we add the new ED value.

Seed regions are a feature found in many RNA-RNA interactions. The seed region is an interaction region of almost complete complementarity. For animal microRNAs the seed region were first discovered (Bentwich, 2005) , (Brennecke et al., 2005). Then, Tjaden et al. (2006) discovered it for many bacterial sRNAs.

RNAUP does not use any seed condition. The general approach to the RRI prediction includes target site accessibility and user-definable seeds. At least one seed is needed at the interaction site. The length of the seed can be set by the user. The following seed features are can be controlled by the user (Busch et al., 2008).

- P: the number of bases perfectly paired in the seed region.
- $b^{max}, b_m^{max} and b_s^{max}$: the maximal number of unpaired bases in the seed region.

2.3 Iterative scheme for double-side RRIs

We use a Single-RRI prediction tool (namely IntaRNA) for the prediction of Multi-RRI. For simplicity, the approach is first introduced for two sites B_1 and B_2. To this end, an iterative scheme is to be applied, which are described in the following steps.

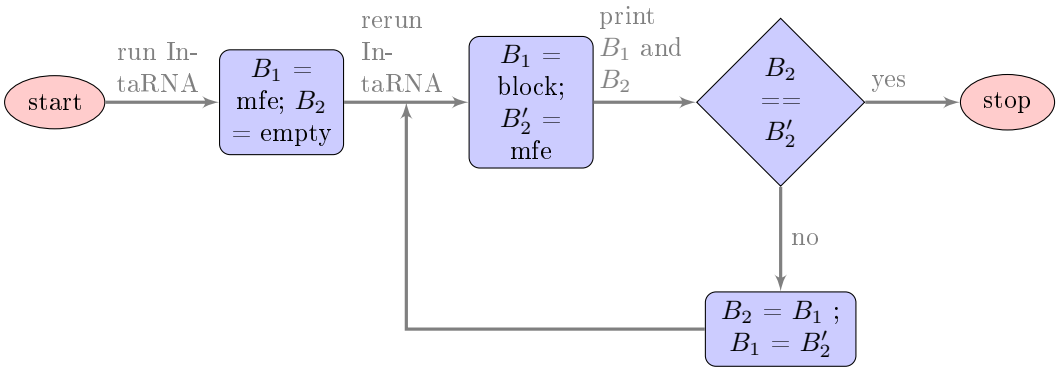
- Step 1: Firstly, we have to run IntaRNA and store minimum free energy and boundaries of respective B_1.

- Step 2: Then we get the 'blocking' constraint from step 1 to rerun IntaRNA and predict the conditional minimum free energy and site B_2. Here we block B_1 (Constrainted IntaRNA) both for intra- and inter molecular base pairing and we get B_2 as minimum free energy. Then, the energy of a respective M-RRI can be computed from the two energies, ie., the energy of the conditional call can be added.

$$E(B_1 \wedge B_2) = E(B_1) + E(B_2|B_1) \ .$$

- Step 3: Since prediction of B_2 is conditional, the existence of B_2 can have effects B_1. Thus, one starts to iterate the procedure from (2) but swaps the conditional site and check for convergence: is the site from two-steps before retained ($B_2 == B_2'$)? If yes: convergence and stop iteration by printing the B_1, B_2. If no: repeat constraint prediction until convergence by swapping .

Below is the flowchart representation for the same procedure.



Below is the proof of why the energy of the conditional call can be just added. The joint site has energy

$$\begin{aligned} E(B_1 \wedge B_2) & \\ &= E_{hyb}(B_1 \wedge B_2) + ED(B_1 \wedge B_2) \\ &= E_{hyb}(B_1) + E_{hyb}(B_2) - RTlog(\mathcal{P}^u(B_1 \wedge B_2)) \end{aligned} \quad \text{(Eq. 4)}$$

The first block B_1 is scored by

$$\begin{aligned} E(B_1) &= E_{hyb}(B_1) + ED(B_1) \\ &= E_{hyb}(B_1) - RTlog(\mathcal{P}^u(B_1)) \end{aligned} \quad \text{(Eq. 5)}$$

and the conditional prediction of B_2 by

$$E(B_2|B_1) = E_{hyb}(B_2|B_1) + ED(B_2|B_1)$$
$$= E_{hyb}(B_2|B_1) - RTlog(\mathcal{P}^u(B_2|B_1)) \quad \text{(Eq. 6)}$$

Now, we add right end side values of Eqn Eq. 6 + Eq. 5, we get,

$$E_{hyb}(B_1) + E_{hyb}(B_2|B_1) - RTlog(\mathcal{P}^u(B_1)) - RTlog(\mathcal{P}^u(B_2|B_1)) \quad \text{(Eq. 7)}$$

As we know $log(A) + log(B) = log(A * B)$, we apply this condition for log values in Eq. 7

$$-RTlog(\mathcal{P}^u(B_1)) - RTlog(\mathcal{P}^u(B_2|B_1))$$
$$= -RTlog(\mathcal{P}^u(B_1) * \mathcal{P}^u(B_2|B_1))$$

Since $P(A \wedge B) = P(A) * P(B|A)$ and $E_hyb(B_2|B_1)$ is independent of B_1, WE GET,

$$E_{hyb}(B_1) + E_{hyb}(B_2) - RTlog(\mathcal{P}^u(B_1 \wedge B_2)) \quad \text{(Eq. 8)}$$

Now, we see the equations Eq. 8 and Eq. 4 are equal.

2.4 Generalization to multi-site RRI prediction

Chapter 3

Results & Discussion

In order to evaluate the multi site interaction, I am comparing the results of the new approach for same RNA interactions with single site interaction tool IntaRNA and results from the literature. Details of the reported RRIs are given at the end of the section in **??**

3.1 Setup

I have used the IntaRNA-3.1.3-windows-64bit version for my thesis. The following parameter has been set. The –outMode=C a flexible interface to generate RNA-RNA interaction output in CSV format (using ; as separator). The argument -n 1 or –outNumber=1 can be used to generate up to N interactions for each query-target pair. IntaRNA provides the possibility to constrain the accessibility computation using the –qAccConstr and –tAccConstr parameters. In this I have used "b" blocked to indicate the positions are occupied by some other interaction (implies single-strandedness). It is possible to restrict the overall length an interaction is allowed to have. This can be done independently for the query and target sequence using –qIntLenMax and –tIntLenMax, respectively. We can alter indexing (independently for query and target) using the –qIdxPos0 and –tIdxPos0 parameters, respectively.

With the above setup, we were able to test multi-site RNA-RNA interactions. The used sequences are listed in Appendix Table.

3.2 OxyS – fhlA

The small RNA OxyS binds to a short sequence inside the fhlA mRNA coding region. This is one of the classic examples of multi-site RRIwhere OxyS forms a stable kissing hairpin complex with fhlA.

The first interaction is between sRNA OxyS and its mRNA target fhlA. It is taken from the paper (Argaman and Altuvia, 2000). The pairing mechanism between the two RNAs is dramatically influenced by their structure. For this purpose, a full comprehension of the pairing process involves thorough knowledge of the individual RNA structures.

OxyS binds wild-type fhlA mRNA shows that kissing complex forming between OxyS and fhlA results in a healthy anti-sense-target structure. The secondary structure of the 5 'end region of fhlA mRNA was predicted to include two stemloop structures. The findings of the study of the structure confirm the presence of the two structure.

When we run with IntaRNA first, we observed that the 104:-15& 98:-9 is the block that is predicted. Then, we run them with our tool, we find no difference in the Energy parameters and also with the hybridDB (ie., hybrid in dot-bar notation). The total energy for both blocks is -7.99. Here the sequence length for OxyS starts from 1 to 109 and for fhla it is from -53 to 60.We can clearly see from the 3.2 where the interaction that has been predicted and the original prediction from the same are almost the same.

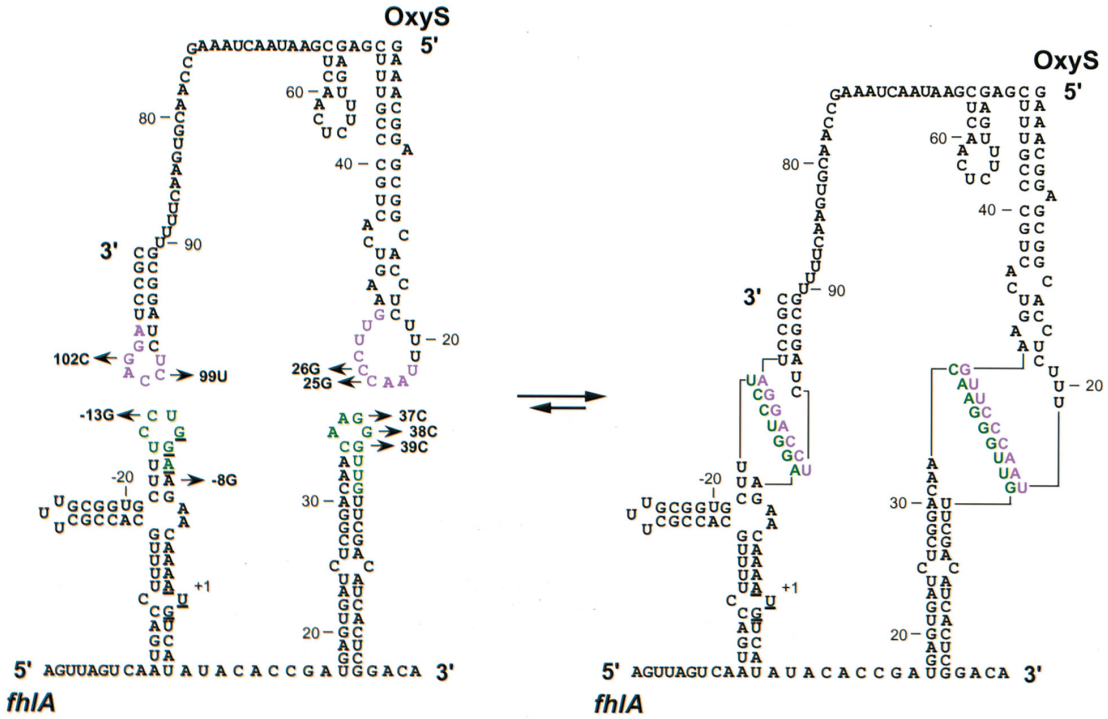


Figure 3.1: Interaction of OxyS - fhlA

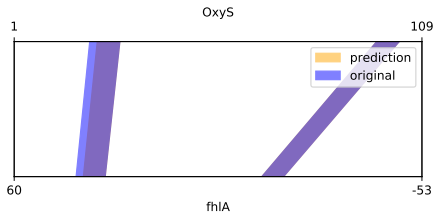


Figure 3.2: OxyS:fhlA

3.3 Spot42 – sthA

The Spot42 Interaction with sthA, Base-pairing RNA Spot42 plays a large role in the suppression of catabolites in Escherichia coli (E.coli) by the direct suppression of genes involved in primary and secondary metabolism.

This example is taken from (Beisel and Storz, 2011). Three consecutive nucleotides in each single-stranded region were mutated to block expected base-pair interactions with target mRNAs (I – III) refer **??**. To validate base pairing through the three single-stranded regions II of Spot 42, compensatory mutations have been made in sthA. Such findings suggest that the base pairing with target mRNAs can accrue at at three single-stranded regions of Spot 42. The mutation in region III influenced the repression of fusion of sthA the most (refer Fig **??**)

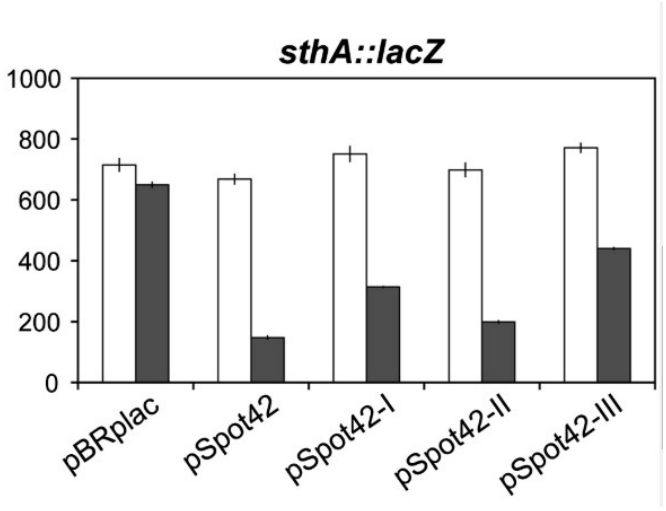


Figure 3.3: Interaction of Spot42 – sthA

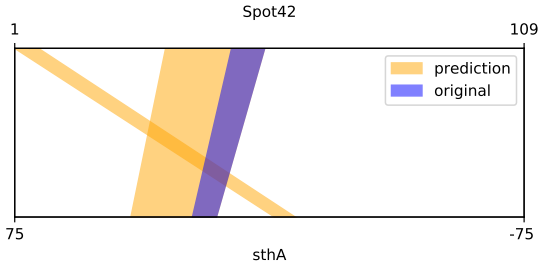


Figure 3.4: spot42:stha

When we run with IntaRNA first, we observed that the 55:15&34:40 is the block that is predicted. Then, we run them with our tool here when Spot42 interacts with sthA, we observed that the ED value changes. This is because, the unconstrained has been added to constrained value. Though there is no much difference in the total energy value, we are ignoring it. The total

Figure 3.5: a) Three single-stranded regions of spot42. b) Mutation in Region III

energy value is -26.38. In order to have the same ED values, we can use the –energyNoDangles option. Here for Spot42, we couldn't find the original interaction, hence we are comparing with the bar chart from the figure 2 from paper (Beisel and Storz, 2011). From the Figure 3.3 we can clearly see . There is huge impact on pSpot32-III and pSpot32-I whereas less during the pSpot42-II. This is because, I and III are close in structure though they are far away from the sequence.

3.4 GcvB – oppA

The gcvB gene encodes two small, nontranslated RNAs that regulate OppA and DppA. The structure of the GcvB-oppA complex consists of two intermolecular helices that precede and follow the putative terminator. This is the example where there are four blocks are presented as per the (Pervouchine, 2004) .

oppA is the periplasmic-binding protein portion of the OppA oligopeptide transport system. The functional consequence of deleting the gcvB gene is a derepression of the oppA. The mechanism of GcvB regulation of oppA is likely to be translational (Urbanowski et al., 2000). In addition to their roles in the transport of dietary peptides, oppA have functions.

Study of the GcvB sequence identified a complementarity area near the ribosome-binding sites of oppA mRNAs. The findings from (Pulvermacher et al., 2008) indicate that various regions of GcvB have specific functions in the control of oppA mRNA. The Shine-Dalgarno sequence in the GcvB-oppA complex is obstructed (Pervouchine, 2004), whereas the GcvB-oppA complex structure is located in the upstream region.This is very much in accordance with the assumption that the oppA control seems to be at the translational stage,

When we run with IntaRNA first, we observed that the 14:67&-9:89 is the block that is predicted. Then, we run them with our tool as, we predict the multi block interaction, we have predicted the two sites of the interaction out of four. The total interaction energy for two sides of block is 26.88 kcal/mol.

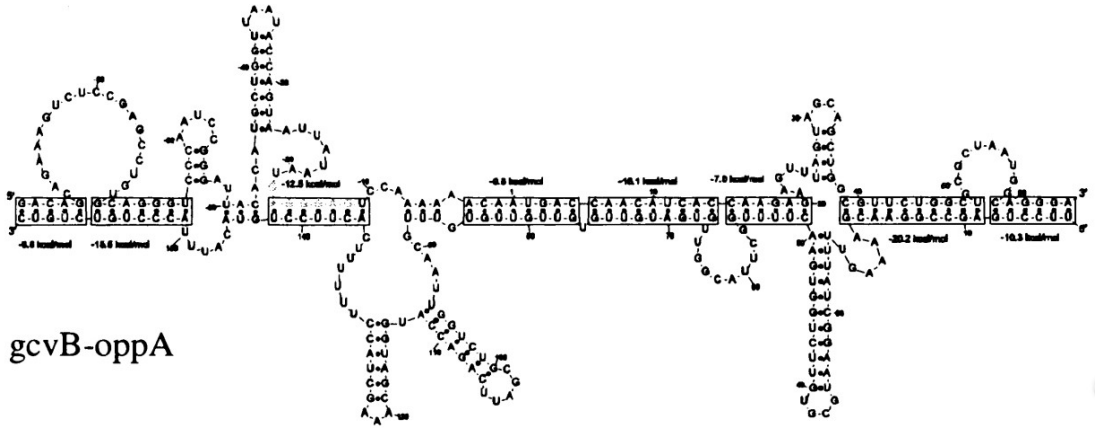


Figure 3.6: Interaction of GCVB – oppA

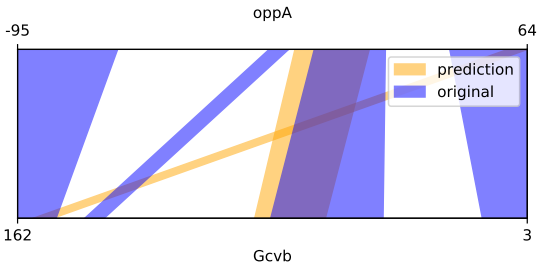


Figure 3.7: Interaction of GCVB – oppA

3.5 DicF - ftsZ

This is an example of a pseudknot (crossing interaction). We note that the DicF-ftsZ complex admits a area of complementarity, which gives rise to a generalized pseudo-knot. The composition of the dicF-ftsZ complex is identical to that of the dsrA-hns in which the dsrA RNA communicates with both the 5 'and 3' ends of the mRNA.

The function of dicF was predicted on the basis of the complementarity of DicF RNA with the ftsZ mRNA binding region. DicF RNA is a 53-nucleotide gene formed in certain E.coli Mutants. Here, DicF RNA is an antisense regulator of ftsZ translation.

This example is taken from the paper (Pervouchine, 2004). They say that DicF RNA has substantial complementarity with ftsZ mRNA in the area surrounding the ShineDalgarno sequence, which is compatible with the finding that DicF controls ftsZ through interaction with the ribosome binding. The DicF-ftsZ relationship tends to be downstream of the start codon. As in the case of GcvB-dppA, it is proposed that DicF-mediated ftsZ control occurs at mRNA level, whereas ftsZ control occurs at translational level. Due to region of complementarity which leads to generalized pseudoknot .

When we run with IntaRNA first, we observed that the 52:55&35:73 is the block that is predicted. Then, we run them with our tool. For this example we used the parameter file and set the tIdxPos0=-73 , tIntLenMax=20 (for restricting the overall length an interaction).The total interaction energy for two sides of block that has been predicted by the tool is -13.86

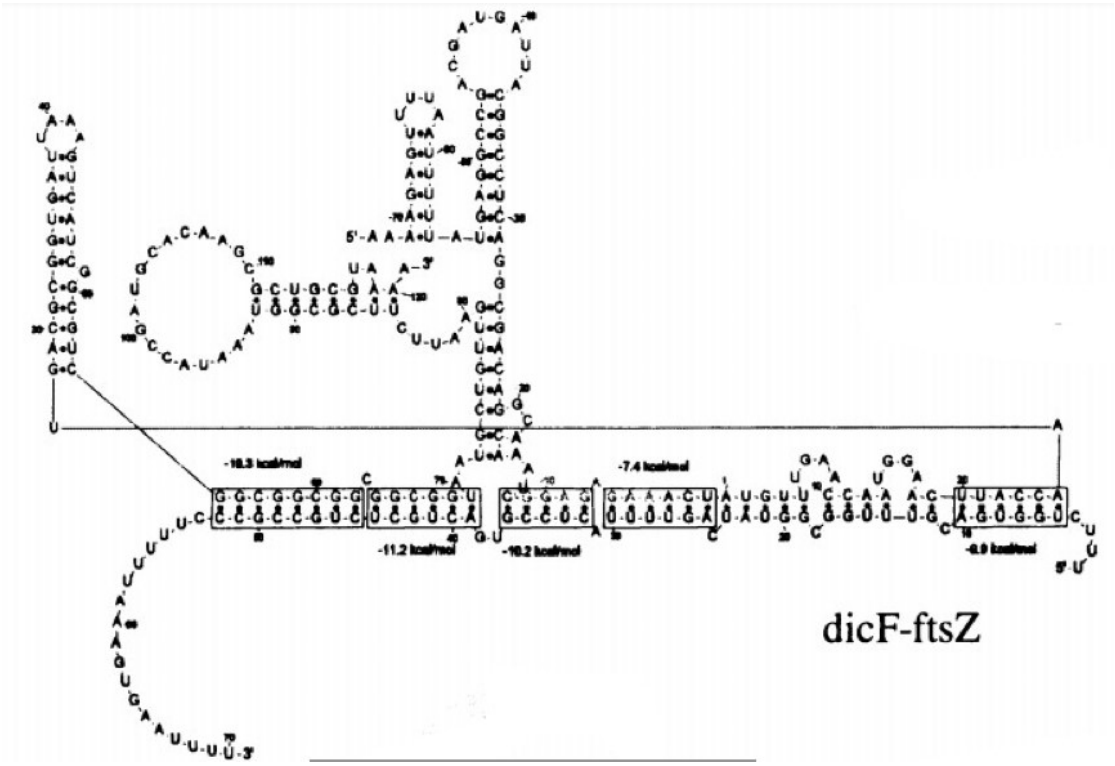


Figure 3.8: Interaction of Dicf- ftsz

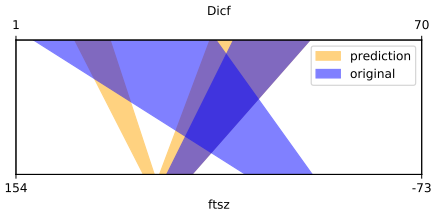


Figure 3.9: Dicf:ftsz

3.6 S-mRNA - EGS

RNase P can be recruited to cleave any mRNA using a modified external reference sequence (EGS) that hybridizes with the target mRNAto shape a structure resembling the tRNA substrate. Just the exact sequence of the S mRNA around the targeting region is displayed (in red) and the EGS sequence is shown in blue. The RNase P cleavage site is labelled with an arrowhead. The sequences of S-SER and S-SER-C relative to T-stem and loop and the variable area of the tRNA molecule were derived from tRNASer, while those of S-C386 and S-C386-C were derived from the EGS variant C386.

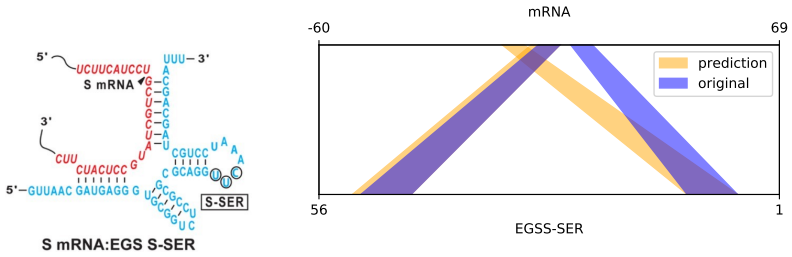
External guide sequences (EGSs) are RNA molecules that consist of a sequence that complements the target mRNA and recruits intracellular ribonuclease P (RNase P), a tRNA processing enzyme, for the precise degradation of the target mRNA. EGS RNAs derived from natural tRNA sequences can be good in blocking gene expression in bacteria. It is possible that an improvement in the RNase P cleavage rate could be attributed to additional tertiary interactions that theoretically stabilize the mRNA-EGS complex.

Variant C386 was chosen for this analysis because the EGS RNAs derived from this version are among the most efficient EGS's. EGS S-C386 was built by connecting the EGS domain of C386 to targeting sequences complementary to the S mRNA. The EGS, S-SER, originating from the normal tRNASer series, was also built. If this is the case, the binding affinity of the EGS variant (i.e. S-C386) to the target S RNA sequence might be greater than that of the EGS (i.e. S-SER) derived from the normal tRNA sequence.

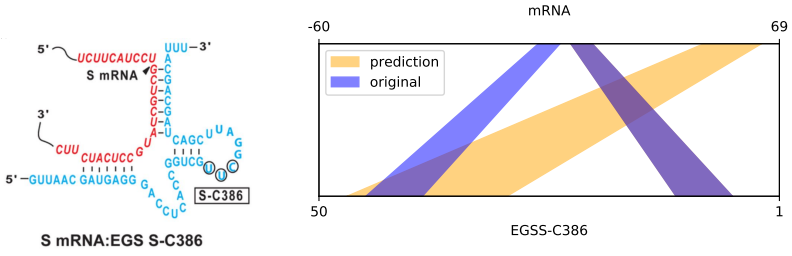
S-C386-C and S-SER-C were derived from S-C386 and S-SER, respectively, and incorporated simple substitutions (5'-UUC-3'- $>$ AAG) at the three closely conserved locations in the T-loop of these EGSs. Nucleotides in these three positions are highly conserved among tRNA molecules and are essential for the folding and recognition of tRNA molecules by RNase P, so mutations in these positions are involved in the EGS process. S-C386-C and S-SER-C had the same anti-sense pattern to the targetS RNA series as S-C386 and S-SER 3.10 and had identical binding affinities to S38 as S-C386 and S-SER, respectively. S-C386-Cand S-SER-C can also be used as anti-sense regulation of such EGSs.

This example is taken from the paper(Zhang et al., 2013). Here, the start of the S-mRNA is complementary to the end of the S-SER, which led to the crossing pattern. We can also clearly see that there is a shortage of the sequence. Also, the tails are not the actual, they are just the artificial one's. Due to its very short sequence, we took the 50nt long on the left side and right side of UCUUCAUCCUGCUGCUAUGCCUCAUCUUC of S-mRNA. The index position is taken from the site of cleavage.

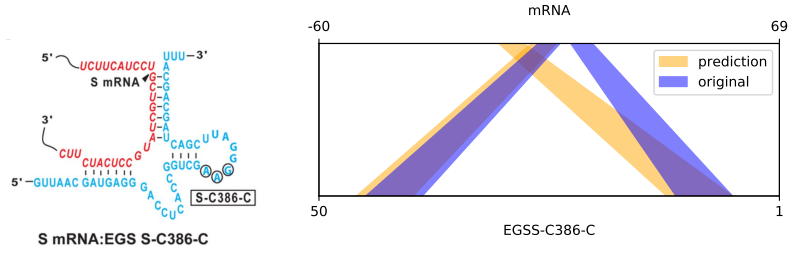
Hence we add the parameterfile and set qIdxPos0=-10. When we run with IntaRNA first, we observed that the -1:53&7:46 is the block that is predicted for both mRNA:EGS S-SER and mRNA:EGS S-SER-C. For the mRNA:EGS S-C386 and mRNA:EGS S-C386C -1:47&7:40 blocks are predicted. The total energy for mRNA:EGS S-C386 is -14.44, mRNA:EGS S-C386C is -13.24, mRNA:EGS S-SER and mRNA:EGS S-SER-C is -16.0



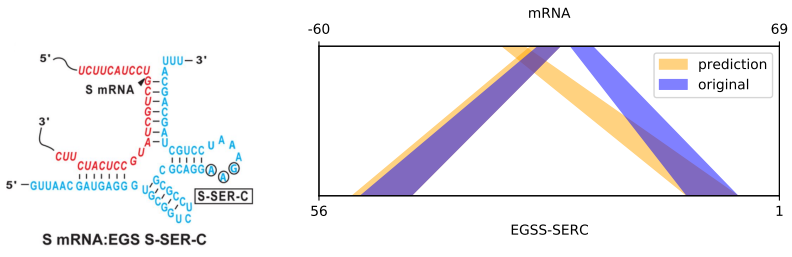
(a) mRNA:EGS S-SER



(b) mRNA:EGS S-C386



(c) mRNA:EGS S-C386-C



(d) mRNA:EGS S-SERC

Figure 3.10: An EGS resembling the structure of a tRNA.

3.7 Details of studied RRIs

Below is the table for the comparison between the different interacting RNAs that are used for the study purpose within this thesis. The table provides us with the original interaction from the paper and the prediction's from the tool. The table also clearly tells us that $E(B1 + B2)$ the multi-site energy is almost double the single site energy $E(B1)$. 1st Index gives us with the index positions and the length determines the sequence length.

query:target	original	prediction	E(B1)	E(B1+B2)	1st index	length
OxyS:fhlA	104:-15&98:-9 30:34&22:42	104:-15&98:-9 24:40&30:34	-4.37	-7.99	1:-53	109:113
Spot42:sthA	55:15&48:22 7:NA&5:NA	55:15&34:40 7:-8&1:-2	-7.85	-13.05	1:-75	109:150
oppA:gcvB	-95:163&-64:151 -17:142&-11:136 -3:84&19:49 39:18&64:1	14:67&-9:89 63:152&57:158	-14.57	-26.38	-95:3	159:160
DicF:ftsZ	52:55&39:69 36:-12&5:25	52:55&35:73 12:82&18:76	-6.89	-13.86	1:-73	70:227
mRNA:EGSS-SER	16:7&10:13 7:46&1:52	-1:53&7:46 -9:13&-3:7	-8.41	-16.0	-60:1	129:156
mRNA:EGSS-SER-C	16:7&10:13 7:46&1:52	-1:53&7:46 -9:13&-3:7	-8.41	-16.0	-60:1	129:156
mRNA:EGS S-C386	1:46&7:40 10:13&16:7	47:48&63:31 10:13&16:7	-8.15	-14.44	-60:1	129:50
mRNA:EGS S-C386-C	1:46&7:40 10:13&16:7	-1:47&6:41 -10:14&-3:7	-8.14	-13.24	-60:1	129:50

Table 3.1: Collections of multisite RNA interaction

Chapter 4

Summary

The motivation of the thesis is to predict the multi-site interaction prediction tool by using the single-site interacting tool IntaRNA. We have implemented the double side interaction here by blocking the interaction site and fixing the conditional site while running the IntaRNA tool using the iterative scheme. The study says that interaction energy for two blocks B1 and B2 together doubles. The overall energy has two times of $E(init)$. We used the python polygon plots for plotting the blocks that are predicted. We compared the predicted blocks with the original blocks from the paper with different interacting RNAs. Different types of cases have been discussed in this thesis, such as with the crossing interaction (pseudoknot) , with two intermolecular helices, kissing interaction. Each examples shows its own structure, but most of them were very well close to the original one's. The detailed study of such examples are given in a table for the easy comparison.

RNA interaction prediction is still an expanding field. The future work that can extended beyond this thesis, one possible idea would be the multi-site prediction. For multi-site interaction , we need to fix two conditional site and iterate it until it converges.

Bibliography

Can Alkan, Emre Karakoc, Joseph H Nadeau, S Cenk Sahinalp, and Kaizhong Zhang. RNA–RNA interaction prediction and antisense RNA target search. *Journal of Computational Biology*, 13 (2):267–282, 2006.

Mirela Andronescu, Vera Bereg, Holger H Hoos, and Anne Condon. RNA STRAND: The RNA secondary structure and statistical analysis database. *BMC bioinformatics*, 9(1):340, 2008.

Mirela Andronescu, Anne Condon, Holger H Hoos, David H Mathews, and Kevin P Murphy. Computational approaches for RNA energy parameter estimation. *RNA*, 16(12):2304–2318, 2010.

Liron Argaman and Shoshy Altuvia. fhla repression by oxys RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. *Journal of molecular biology*, 300 (5):1101–1112, 2000.

David P Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *cell*, 116(2): 281–297, 2004.

Chase L Beisel and Gisela Storz. The base-pairing RNA spot 42 participates in a multioutput feedforward loop to help enact catabolite repression in escherichia coli. *Molecular cell*, 41(3): 286–297, 2011.

Isaac Bentwich. Prediction and validation of microrna and their targets. *FEBS letters*, 579(26): 5904–5910, 2005.

Philip N Borer, Barbara Dengler, Ignacio Tinoco Jr, and Olke C Uhlenbeck. Stability of ribonucleic acid double-stranded helices. *Journal of molecular biology*, 86(4):843–853, 1974.

Julius Brennecke, Alexander Stark, Robert B Russell, and Stephen M Cohen. Principles of microrna–target recognition. *PLoS biology*, 3(3), 2005.

Anke Busch, Andreas S Richter, and Rolf Backofen. IntaRNA: efficient prediction of bacterial srna targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24): 2849–2856, 2008.

Hamidreza Chitsaz, Rolf Backofen, and S Cenk Sahinalp. biRNA: Fast RNA-RNA binding sites prediction. *International Workshop on Algorithms in Bioinformatics*, pages 25–36, 2009.

Howard DeVoe and Ignacio Tinoco. The stability of helical polynucleotides: base contributions. *Journal of molecular biology*, 4:500–17, 1962.

Robert M Dirks and Niles A Pierce. A partition function algorithm for nucleic acid secondary structure including pseudoknots. *Journal of computational chemistry*, 24(13):1664–1677, 2003.

Rick Gelhausen. Constrained RNA-RNA interaction prediction, 2018.

Michael Ibba and Dieter Söll. Aminoacyl-tRNA synthesis. *Annual review of biochemistry*, 69(1): 617–650, 2000.

Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

Tamás Kiss. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, 109(2):145–148, 2002.

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

Barmak Modrek and Christopher Lee. A genomic view of alternative splicing. *Nature genetics*, 30(1):13, 2002.

Ulrike Mückstein, Hakim Tafer, Jörg Hackermüller, Stephan H Bernhart, Peter F Stadler, and Ivo L Hofacker. Thermodynamics of RNA–RNA binding. *Bioinformatics*, 22(10):1177–1182, 2006.

Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3): 443–453, 1970.

Dmitri D Pervouchine. Iris: intermolecular RNA interaction search. *Genome Informatics*, 15(2): 92–101, 2004.

Sarah C Pulvermacher, Lorraine T Stauffer, and George V Stauffer. The role of the small regulatory RNA gcvb in GcvB/mRNA posttranscriptional regulation of oppA and dppA in escherichia coli. *FEMS microbiology letters*, 281(1):42–50, 2008.

Martin Raden, Mostafa Mahmoud Mohamed, Syed Mohsin Ali, and Rolf Backofen. Interactive implementations of thermodynamics-based RNA structure and RNA–RNA interaction prediction approaches for example-driven teaching. *PLoS computational biology*, 14(8):e1006341, 2018.

Maria Selmer, Christine M Dunham, Frank V Murphy, Albert Weixlbaumer, Sabine Petry, Ann C Kelley, John R Weir, and Venki Ramakrishnan. Structure of the 70s ribosome complexed with mRNA and tRNA. *Science*, 313(5795):1935–1942, 2006.

Brian Tjaden, Sarah S Goodwin, Jason A Opdyke, Maude Guillier, Daniel X Fu, Susan Gottesman, and Gisela Storz. Target prediction for small, noncoding RNAs in bacteria. *Nucleic acids research*, 34(9):2791–2802, 2006.

Sinan Uğur Umu and Paul P Gardner. A comprehensive benchmark of RNA–RNA interaction prediction tools for all domains of life. *Bioinformatics*, 33(7):988–996, 2017.

Mark L Urbanowski, Lorraine T Stauffer, and George V Stauffer. The gcvb gene encodes a small untranslated RNA involved in expression of the dipeptide and oligopeptide transport systems in escherichia coli. *Molecular microbiology*, 37(4):856–868, 2000.

Patrick R Wright, Andreas S Richter, Kai Papenfort, Martin Mann, Jörg Vogel, Wolfgang R Hess, Rolf Backofen, and Jens Georg. Comparative genomics boosts target prediction for bacterial small RNAs. *Proceedings of the National Academy of Sciences*, 110(37):E3487–E3496, 2013.

Patrick R Wright, Martin Mann, and Rolf Backofen. Structure and interaction prediction in prokaryotic RNA biology. *Microbiol Spectrum*, 6(2):10–1128, 2018.

Xiaojun Xu and Shi-Jie Chen. Physics-based RNA structure prediction. *Biophysics reports*, 1 (1):2–13, 2015.

Zhigang Zhang, Gia-Phong Vu, Hao Gong, Chuan Xia, Yuan-Chuan Chen, Fenyong Liu, Jianguo Wu, and Sangwei Lu. Engineered external guide sequences are highly effective in inhibiting gene expression and replication of hepatitis b virus in cultured cells. *PloS one*, 8(6), 2013.

Michael Zuker and Patrick Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic acids research*, 9(1):133–148, 1981.

Appendices

Appendix A

RNA Sequences

RNA	Sequence
OxyS	GAAACGGAGCGGCACCTCTTTTAACCCTTGAAGTCACTGCCCGTTTCGAGAGTTTCTCAACTCGAATAACTAAAGC CAACGTGAACTTTTGCGGATCTCCAGGATCCGC
fhlA	AGTTAGTCAATGACCTTTTGCACCGCTTTGCGGTGCTTTCCTGGAAGAACAAAATGTCATATACACCGATGAGTGA TCTCGGACAACAAGGGTTGTTCGACATCACTCGGACA
Spot42	GUAGGGUACAGAGGUAAGAUGUUCUAUCUUUCAGACCUUUUACUUCACGUAAUCGGAUUUGGCUGAAUAUUUUAGC CGCCCCAGUCAGUAAUGACUGGGGCGUUUUUUA
sthA	GGGATCAATTGGCTTACCCGCGATAAAATGTTACCATTCTGTTGCTTTTATGTATAAGAACAGGTAAGCCCTACCA TGCCACATTCCTACGATTACGATGCCATAGTAATAGGTTCCGGCCCCGGCGGCGAAGGCGCTGCAATGGGCCTG
gcvB	TTCCTGAGCCGGAACGAAAAGTTTTATCGGAATGCGTGTTCTGATGGGCTTTTGGCTTACGGTTGTGATGTTGTGT TGTTGTGTTTGCAATTGGTCTGCGATTCAGACCACGGTAGCGAGACTACCCTTTTTCACTTCCTGTACATTTACCC TGTCTGTC
oppA	GACAGCAGAAAGUCUCCGAGCCUGUGCAGGGUCCCAAUCCGGGAUUACACAUGCUGGUUAAUACCAGUAAUUAUAA UGAGGGAGUCCAAAAAACAAUGACCAACAUCACCAAGAGAAGUUUAGUAGCAGCUGGCGUUCUGGCUGCGCUAAUG GCAGGGA
DicF	TTTCTGGTGACGTTTGGCGGTATCAGTTTTACTCCGTGACTGCTCTGCCGCCCTTTTTAAAGTGAATTTT
ftsZ	AAAAGAGTTTTAATTTTTATGAGGCCGACGATGATTACGGCCTCAGGCGACAGGCACAAATCGGAGAGAAACTATG TTTGAACCAATGGAACTTACCAATGACGCGGTGATTAAAGTCATCGGCGTCGGCGGCGGCGGCGGTAATGCTGTTG AACACATGGTGCGCGAGCGCATTGAAGGTGTTGAATTCTTCGCGGTAAATACCGATGCACAAGCGCTGCGTAAAA
EGS	AACTTGTCCTGGTTATCGCTGGATGTGTCTGCGGCGTTTTATCATCTTCCTCTTCATCCTGCTGCTATGCCTCATC TTCTTGTTGGTTCTTCTGGACTATCAAGGTATGTTGCCCGTTTGTCCTCTAAT
S-SER	GTTAACGATGAGGGTGCGGTCTCCGCGCGCAGGTTCAAATCCTGCTAGCAGCATTT
S-SER-C	GTTAACGATGAGGGTGCGGTCTCCGCGCGCAGGAAGAAATCCTGCTAGCAGCATTT
S-C386	GTTAACGATGAGGGACCTCACCGGTCGTTCGGATTCGACTAGCAGCATTT
S-C386C	GTTAACGATGAGGGACCTCACCGGTCGAAGGGATTCGACTAGCAGCATTT

Table A.1: Table of RNAs with their corresponding sequence as used in this thesis.