ALBERT LUDWIGS UNIVERISTY OF FREIBURG

MASTER THESIS

Multisite RNA-RNA Interaction Prediction

Yogapriya Ayyanarmoorthy

January 27, 2020

Contents

Chapter 1

Introduction

RNA molecules play important roles in various biological processes.Their regulation and function are mediated by interacting with other molecules. Forming base pairs between two RNAs, called RNA-RNA interactions (RRI). There are fast and reliable single interaction site (S-RRI) prediction tools like 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. 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. 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 "central dogma" of molecular biology states that the flow of genetic information in a cell is from DNA through RNA to proteins: “DNA makes RNA makes protein” (as first suggested by Jean Brachet in 1960)(Brachet and Ficq, 1956). 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. Non-coding 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 two interacting bases belonging to the same molecule of RNA form *intra-molecular* structures and if they belong to different molecules of RNA form *inter-molecular* structures, as seen in figure 1.1

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, Intra and intermolecular structures are not mutually exclusive.

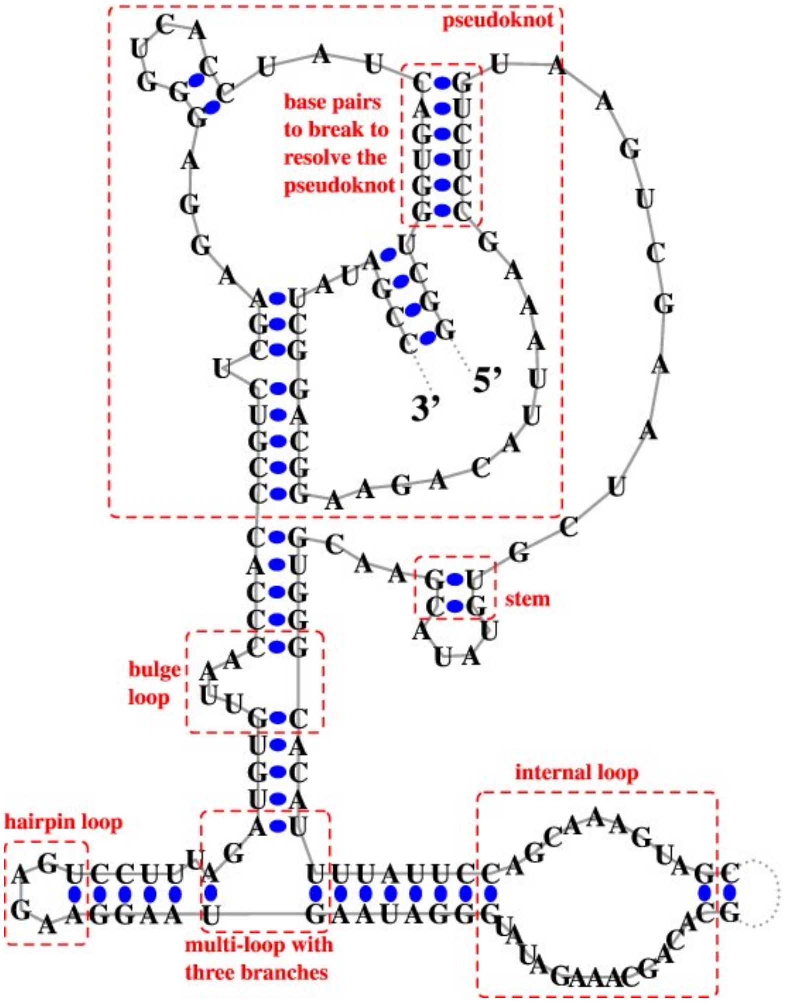

pseudoknot
base pairs to break to resolve the pseudoknot
5'
3'
stem
bulge loop
internal loop
hairpin loop
multi-loop with three branches

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

Formally, an RNA secondary structure P of S is a set of base pairs:

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

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

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

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

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 , and 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 (e.g. $i < k < j < l$) is called pseudo-knot.These pairs are assumed to be incompatible with most dynamic programs. An allowed secondary structure is a set of base pairs that are all compatible with each other.

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 (BP) and unpaired bases (UB), different types of secondary structure elements are distinguished.They 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
  $\forall 0 \leq l \leq k : (j_l < i_{l'+1})$; $\forall 0 \leq l, l' \leq k$ is true that 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.

DeVoe and Tinoco discovered that 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).

1.2.3 Nearest neighbor model and energy contributions

The Nearest Neighbor Model enables the calculation of a free energy estimate for a given RNA secondary structure. For the performance of work, the free energy can be taken as the amount of energy stored in a system. The positive energy is in the form of heat and the negative energy is used to destroy the system. Always, lower the energy gives more stable the system. 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. 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 textitdangling end contributions.
The energy $E(P)$ 1.1 of a nested secondary structure P can be estimated by the sum of loop contributions (see Figure 1.2)

$$E(P) = \sum_{(i,j)\in P} \begin{cases} e^H(i,j) & \text{: if hairpin loop,} \\ e^{SBI}(i,j,k,l) & \text{: if stack/bulge/internal loop,} \\ e^M(i,j,x,x') & \text{: if Multi loop,} \end{cases} \tag{1.1}$$

Where e^H, e^{SBI} and e^M tells the context sensitive energy contributions of the loops. Where (k, l) represents the enclosed base pair of stack,bulge or internal and x represents the unpaired bases and x' represents the helices enclosed in the multi loop. We can 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_a^M scores the multi loop closing base pair (i, j) , e_b^M represents the penalty for directly enclosed unpaired bases x and e_c^M represents the number of enclosed helices x'. Thus the nearest neighbor model gives the energy contributions for the loop types.

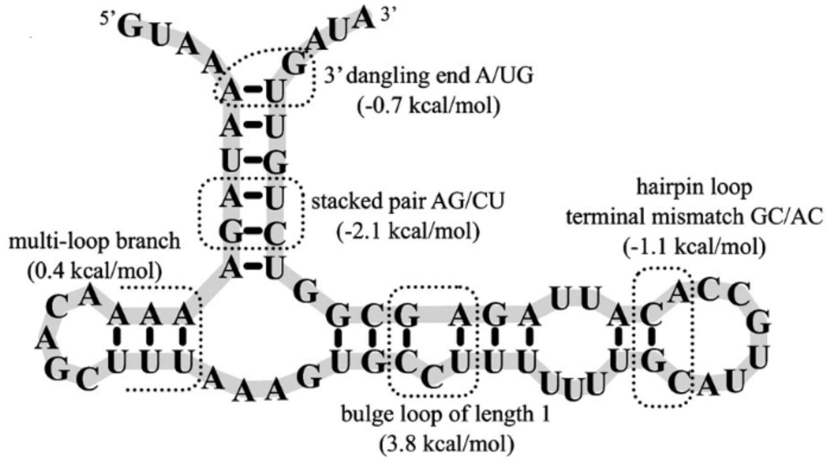


Figure 1.2: Energy contributions of loops. (Andronescu et al., 2010)

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 has been calculated , then traceback is used to recover the 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 "mol-scaled" gas constant R

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

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 P.

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

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

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

and normalising with the partition function Z for the structure ensemble $\mathcal{P}$.

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 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.,

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

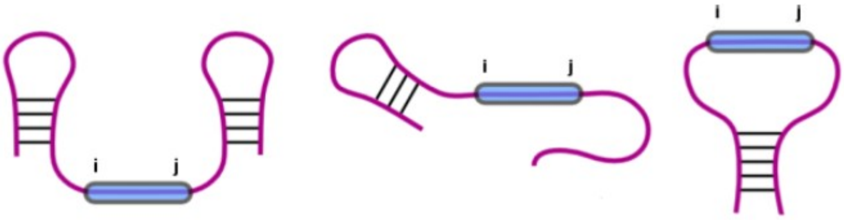


Figure 1.3: Examplary structures that are unpaired in the subsequence i..j.

Where $\mathcal{P}_{all}$ is the ensemble of all structures that can be formed from a sequence. The calculation of accessibility of single stranded regions is carried out using unpaired probability (Mückstein et al., 2006), hence it is very important.

The below figure 1.3 was inspired by the lecture material of RNA bioinformatics lecture .

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. Interacting RNA strands is classified into two types. ie., Intermolecular and Intramolecular. 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 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 RNAs(mRNA) which is the important step in identifying the RNA splicing products (Modrek and Lee, 2002). In *translation* transfer RNAs(tRNA) interact with messenger RNAs(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 RNAs(snoRNA) guide the modification of ribosomal RNAs(rRNA) (Kiss, 2002). In microRNA (miRNA) targetting, the base pairing between an miRNA and mRNA leads to degradation or translation inhibition of the mRNA (Bartel, 2004). For RNA function and regulation these examples gives 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 take the accessibility of the interacting subsequences (Umu and Gardner, 2017).

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 all base pairs in S^1 , 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.

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

The same problem we have for the scoring of crossing structure in a pseudoknot.
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 **??** , 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,

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

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

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

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

For a valid RRI, lets consider bases k and l form another allowed base pair. Then,

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

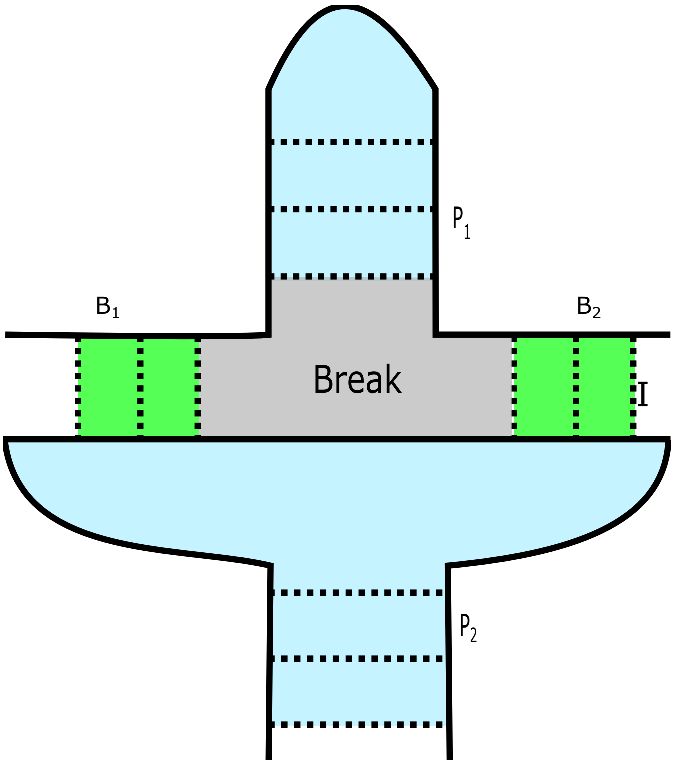


Figure 1.4: The RNA-RNA Interaction is the union of all base pairs in sequence 1 and sequence 2 are denoted as P^1 and P^2 respectively. I denotes the union of all intermolecular base pairs. B_1, B_2 are the interaction blocks. Break is the loop enclosed by two inter-molecular base pairs that also contains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

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 ,

$$E(B) = \sum_{\substack{i \in B \\ j \in B \\ \exists i < j}} E^{SBI}(i, j, k, l)$$

where $j = argmin_{i' \in B \wedge i' > i}(i'_1)$.

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 starts with the pseudoknot, in the loops where the positions are not paired within

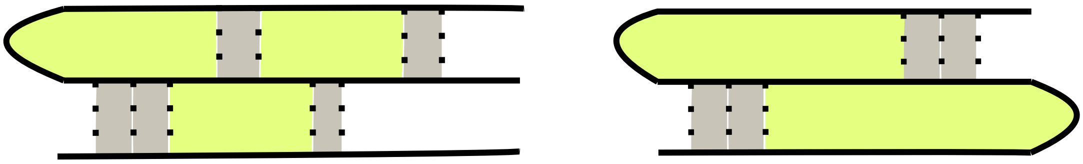


Figure 1.5: The left side figure shows the Positions are not paired within the loop. This problem starts with the pseudoknot which is shown in the right side figure where the same problem exists for the scoring of crossing structures.

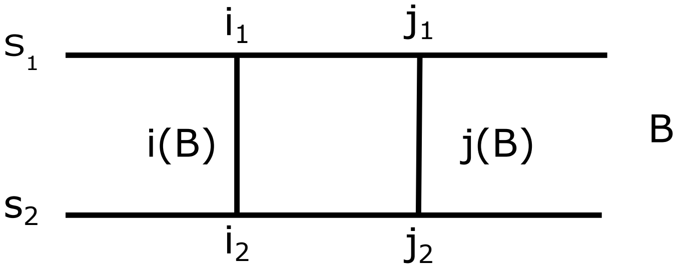


Figure 1.6: The block region $R(B)$ where the left and right most base pairs of B concerning S_1

the loop 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

$$E(RRI) = E(P^1) + E(P^2) + E_{init} + \sum_{B_i \in I} E(B_i) + \sum_{B_i \in I} E_{break}(B_i, B_{i+1}, P^1, P^2) \tag{1.2}$$

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 according to their algorithmic idea into hybrid-only interaction prediction,General interaction prediction, concatenation-based/cofolding interaction prediction , and accessibility-based interaction prediction.

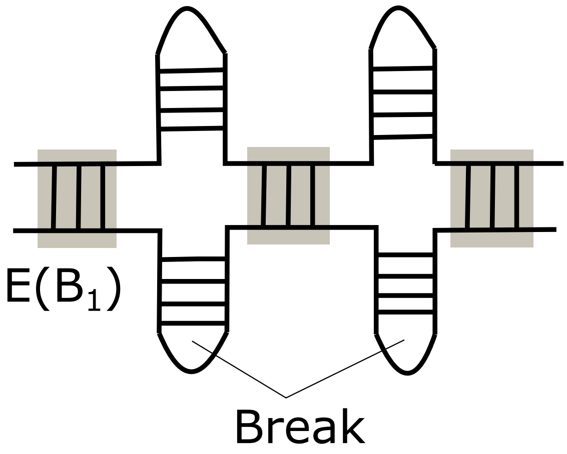
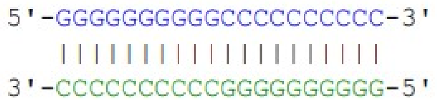


Figure 1.7: The interaction energy of RRI is the energy defined by the loops enclosed by all inter-molecular base pairs. $E(B_1)$ is the energy of block 1 and the $E(breaks)$ can be calculated from sum of all breaks.

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

1.4.1 Hybrid

In 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 $S1$, $S2$ of lengths n and m respectively (Tjaden et al., 2006).



```
5'-GGGGGGGGGGCCCCCCCCCC-3'
   ||||||||||||||||||||
3'-CCCCCCCCCCGGGGGGGGGG-5'
```

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 1.3 for the nussinov like interaction prediction.

$$H_{ij} = max \begin{cases} H_{i-1,j-1} + 1 & : \text{if } S^1_i,\ \overleftarrow{S^2_j} \text{ are compl. base pair ,} \\ H_{i-1,j} \\ H_{i,j-1}, \end{cases} \tag{1.3}$$

Where H_{ij} is the maximal number of intermolecular base pairs for the prefixes $S^1_1..i$ and $\overleftarrow{S^2_1..j}$ the reverse sequnce of S^2. The visual representation of the recursion scheme 1.9 . The above

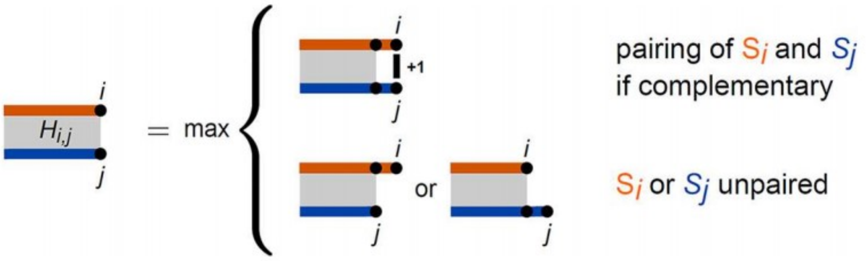


Figure 1.9: Recursion scheme to maximize intermolecular base pairs between two RNAs $S1$ and $S2$ represented in orange/blue, respectively

equation is the s a 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 back them. As stated above, this approach has very low runtime.

In order to compute the energy of an RRI using equation 1.3 , no intra-molecular structure is considered, i.e. $P^1 = P^2 = \emptyset$.

Thus, eventually, only one block of inter-mol base pairs is modelled ie., ($I = B1$) and no break is present. They are implemented in tools like TargetRNA, RNAhybrid, RNAplex. The main advantages of this approach is they allows temperature to taken into account, 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.

1.4.2 General

One of the most common general approach 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 in below fig 1.10.

It shares most common features with pseudoknots, but is less computationally intensive, secondary structure prediction system. 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 computing time and space are $O(n^3m^3)$ and $O(n^2m^2)$, where n and m are the length of the sequence sizes. The configuration of the oxyS-fhlA complex proposed in (Argaman and Altuvia, 2000) consists of four neighboring stem loops, two in each of the molecules which connect, forming two stable kissing complexes. In this method, the main goal is the simulatinous optimization of intra and inter-molecular base pairing.

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

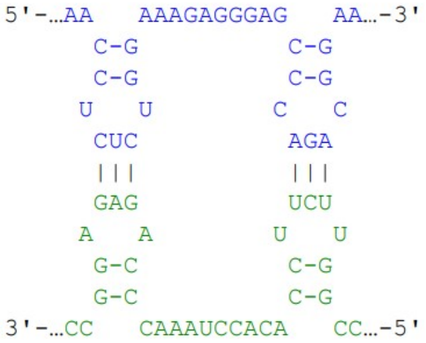


Figure 1.10: Double kissing hairpin interaction. The blue and green denotes the first and second sequnence of RNA. Base pairs are denoted by dash. The picture is taken from the paper (Wright et al., 2018)

lower right, 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)

$$G^{Pseudo} = \beta_1 + \beta_2 B^p + \beta_3 U^p \tag{1.4}$$

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. reference to **??** (right), here Bp=3 for the hairpin-PKs. If 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) such pseudoknot energy terms based on $G^P seudo$ 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 intra-mol base pairs (refer **??** (left)). for simplicity, formula (refer 1.3) 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 to predict the suboptimal interaction and its probabilities also the computation of probabilities of intermolecular interactions, which is used to access the stability. Due to its high complexity, several approaches for reducing the requirements of this method have been introduced.

1.4.3 Concatenation

Concatenation or co-folding approach is used for predicting the interacting base pairs of two RNA molecules. 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

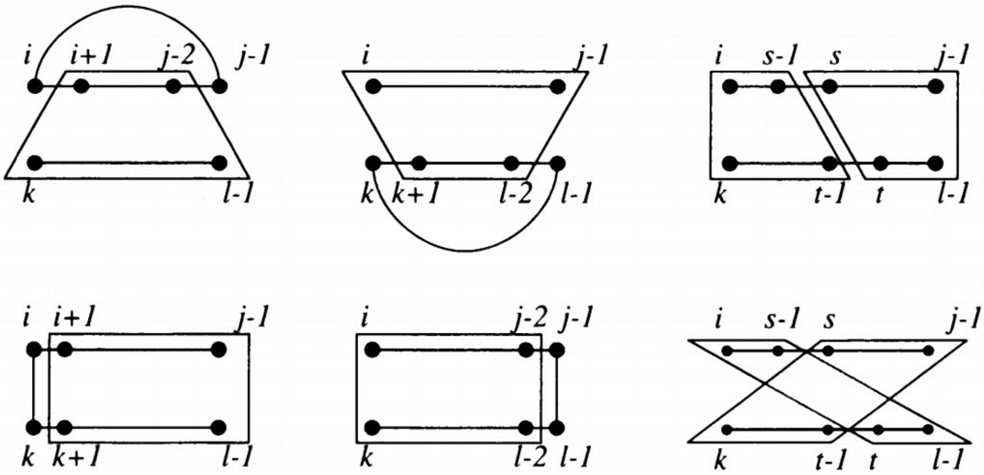


Figure 1.11: Depiction of the recursion $M^{i..k}_{j..l}$. Figure is taken from the paper (Pervouchine, 2004)

structure prediction that takes care of the linker sequences. The first implementation of this approach was by using Nearest neighbor model which was used for mfold and then implemented in tools like MutliRNAfold and RNAcofold.

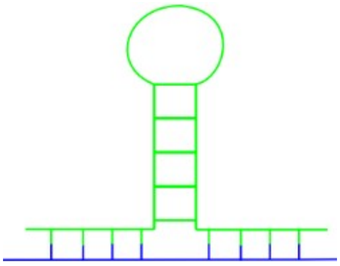


Figure 1.12: Green and blue are the two different RNA's that are interacting.

This is extension of single structure prediction recursion with a special handling of linker sequence. Here,the input is restricted to two RNA sequences that are concatenated by a linker of length $l + 1$ to ensure the presence of a linker and that the concatenated sequence ends can form a base pair. We don't need any special energy treatment because the intra and inter molecular base pairs are treated equally.

The energy for the concatenation approach, here the inter and intra molecular base pairs have same score and they are treated as a single sequence. 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 the interaction patterns because hybrid structures are nested. For example, interactions like

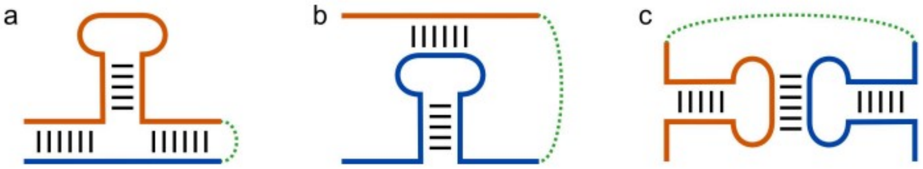


Figure 1.13: a) Pattern that can be predicated by Concatenation b)Kissing stem-loop c) kissing hairpin interaction. The blue and orange are the two different RNA's and the dotted green is the linker , black lines represents the base pairs.

kissing stem-loop or kissing hairpin-loop (as seen in fig 1.13) cannot be predicted because they form a pseduoknot by them. To predict these patterns we go for Accessibility based approaches.

1.4.4 Accessibility

Concatenation approaches cannot predict the structures which contains pseudoknots. To overcome the drawback of concatenation approach, Accessibility approaches has been introduced. The main aim of this approach is to ensemble properties of the single sequences that are necessary for the interaction. It can predict single site an interaction pattern of two respective RNA subsequences. Tools like RNAup and IntaRNA can be used to predict such approaches. Here we have to neglect the intra molecular structure before the intermolecular interaction is formed. In order to form a stable interaction of intermolecular base pairs, the intra molecular base pairs has to be opened /broken.

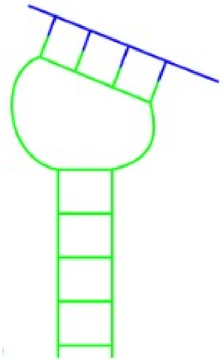


Figure 1.14: Green and blue are the two different RNA's that are interacting and forms pseudoknots.

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-mol structure 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 Q over the Boltzmann factors. The energy of the ensemble E^{all} is

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

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

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

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

Therefore, the alternative $E(RRI)$ formula :

$$E(RRI) = E(B1) + E_{init} + Eens(S1) + ED1 + Eens(S2) + ED2 \quad (1.5)$$

Since $ED = E^{u} - E^{all}$
substituting ED value in $ED + E^{all}$ gives
$E^{u} - E^{all} + E^{all}$
$= E^{u}$
$E(RRI) = E(B1) + E_{init} + E^{u1} + E^{u2}$

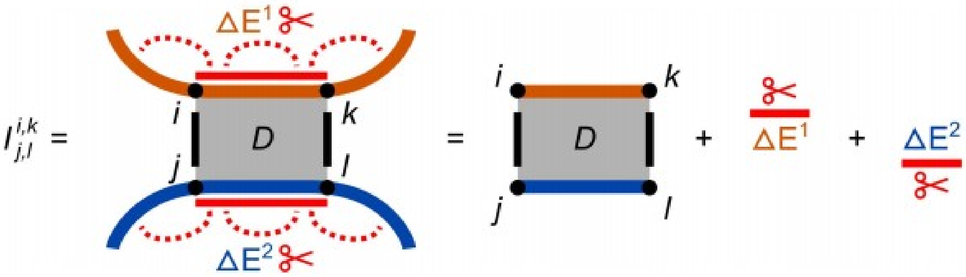


Figure 1.15: Depiction how accessibility-based approaches score an interaction of two RNAs $S1$ and $S2$ in orange and blue respectively

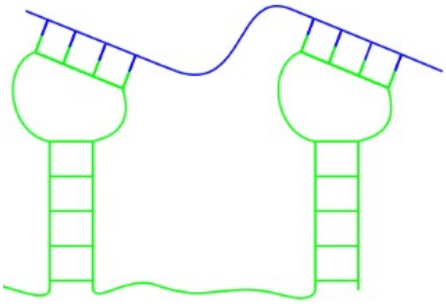


Figure 1.16: Double stem loop interaction cannot be handled by accessibility as they have two binding site.

The maximal number of base pairs $D^{i,k}_{j,l}$ for all interaction sites are computed using hybrid approach. Then the unpaired probabilities P^{u1} and P^{u2} are tabularized for both sequences S^1 and $\overleftarrow{S}^2$, respectively using mcckaskill approach.

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.

1.4.5 Comparison with approaches

In this subsection, we will see the comparison between the approaches for few interaction pattern. Below table 1.1 gives the overview for which interaction pattern , the approaches can be used.

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

Comparison of RRI approaches						
RRI Pattern			**RRI prediction approaches**			
Figures	RRI description	No.of blocks	Hybrid	General	Concatenation	Accessibility
1.8	Full duplex structure	1	yes	yes	yes	yes
1.13 (a)	Nested joint structure without pseudoknots	2	no	yes	yes	yes
1.12	Nested joint structure without pseudoknots	2	no	yes	yes	yes
1.13 (b)	Stem loop interaction	1	no	yes	no	yes
1.14	Stem loop interaction	1	no	yes	no	yes
1.13 (c)	Kissing hairpin loop	1	no	yes	no	yes
1.10	Double kissing hairpin loop	2	no	yes	no	no
??	Kissing stem interaction	2	no	yes	no	no
1.16	Double kissing stem loop	2	no	yes	no	no

We could conclude that the accessibility based approach is the best approach for single site RNA-RNA interaction. As, they can't predict the multisite RRI because in IntaRNA model, we remove the base pairs while predicting, when they are no intramolecular base pairs are in between, it is considered to be a wrong model (ie., double kissing hairpin loop 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 see about the approach used for multisite interaction prediction for RNA. 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) 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. Some approaches in include IRIS, NUPACK, piRNA , etc., There are fast and reliable single interaction site (S-RRI) prediction tools like 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.

2.1 RNAup - Exact Recursion for single site

In the following, I will first introduce the RNAup-like exact recursions and then give an overview of IntaRNA heuristic version. IntaRNA is an interaction prediction tool developed for the prediction of mRNA target sites for small regulatory bacterial RNAs. 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.

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

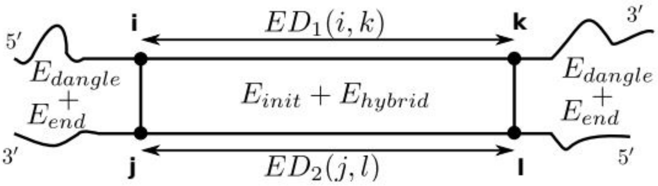


Figure 2.1: The energy contribution of IntaRNA. The image is taken from (Gelhausen, 2018)

- **Hybridization energy** : energy value from intermolecular base pairings in the form of stackings, bulges or internal loops. i.e., energy is a negative value.
- **Accessibility energy** : An amount of energy needed to single-strand the interacting region, i.e. not include in 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, $Z_{s^u_{i,k}}$ where $s^u_{i,k}$ is an ensemble of all structures which can be formed by a sequence S, with a single stranded region $[i,k]$. Refer to the section 1.4.4. Therefore,

$$ED(i,k) = -RTln(Z_{s^u_{i,k}})$$

(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 energy hybridization of two subsequences, where a base pair is generated by the leftmost 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.

$$H(i,j,k,l) = min \begin{cases} E_{init} \\ \quad : \text{if } (S^1_i, S^2_j) \text{ can pair } i = k, j = l, \\ min_{r,s}\{e^{SBI}(i,j,r,s) + H(r,s,k,l)\} \\ \quad : \text{if } (S^1_i, S^2_j) \text{ and } (S^1_k, S^2_l) \text{ can pair } i \neq k \text{ and } j \neq l, \\ \infty \\ \quad : \text{otherwise,} \end{cases} \tag{2.1}$$

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 $S^2_j ... S^2_l$ is given by,

$$C(i,j,k,l) = \begin{cases} H(i,j,k,l) + ED_1(i,k) + ED_2(j,l) \\ \quad : \text{if } (S^1_i, S^2_j) \text{ and } (S^1_k, S^2_l) \text{ can pair } i \neq k \text{ and } j \neq l, \\ \infty \\ \quad : \text{otherwise,} \end{cases} \tag{2.2}$$

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. RNA molecules fold by intermolecular base pairing, by incorporating hydrogen bonds reduces free energy. Hence the system with the minimum free energy (mfe) is most likely the structure. The interaction with the minimum estimated free energy is probably the most stable structure and thus the structure fulfills the RNA molecule function.

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

2.2 IntaRNA - Heuristic recursion for single site

The exact recursions are not suitable for the 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 where the matrix $c(i,j,k,l)$ many entries has same value and those values are not used often for the recursion. Hence, we consider only the right end of interaction i, j which is single and locally optimal, instead of all the possible interaction. Similarly, we don't need to consider all possible ranges for time complexity. This will help us to reduce the space and time complexity to $O(nm)$. This is based on IntaRNA version 1 & 2. The heuristic version is defined as:

$$C(i,j) = \begin{cases} E_{init} + ED_1(i,i) + ED_2(j,j) \\ \quad \text{: in the case of new interaction} \\ min_{p,q}\{e^{SBI}(i,j,p,q) + C(p,q) - ED_1(p, K(p,q)) - ED_2(q, L(p,q)) \\ \qquad +ED_1(i, K(p,q)) + ED_2(j, L(p,q))\} \\ \quad \text{: if } (S^1_i, S^2_j) \text{ can pair ,} \\ \infty \\ \quad \text{: otherwise,} \end{cases} \tag{2.3}$$

Here $K(p,q), L(p,q)$ are the newly introduced matrices which are optimal. Since ED values are not additive, we have to subtract the old ED values before we add the new ED value.

2.3 Iterative scheme for multisite

We use a Single-RRI prediction tool (namely IntaRNA) for the prediction of Multi-RRI. To this end, an iterative scheme is to be applied.

- Step 1: Firstly, we have to run IntaRNA and store minimum free energy as B_1 and the respective interaction site as B_2 which is empty.
- Step 2: Then we get the 'blocking' constraint from step 1 by rerunning IntaRNA and predict conditional minimum free energy and site. Here we block B_1 (Constrainted IntaRNA) and we get B_2' as minimum free energy. Then, the energy of an respective M-RRI can be computed from the two energies. ie., the energy of the conditional call can be added using the conditional probability.

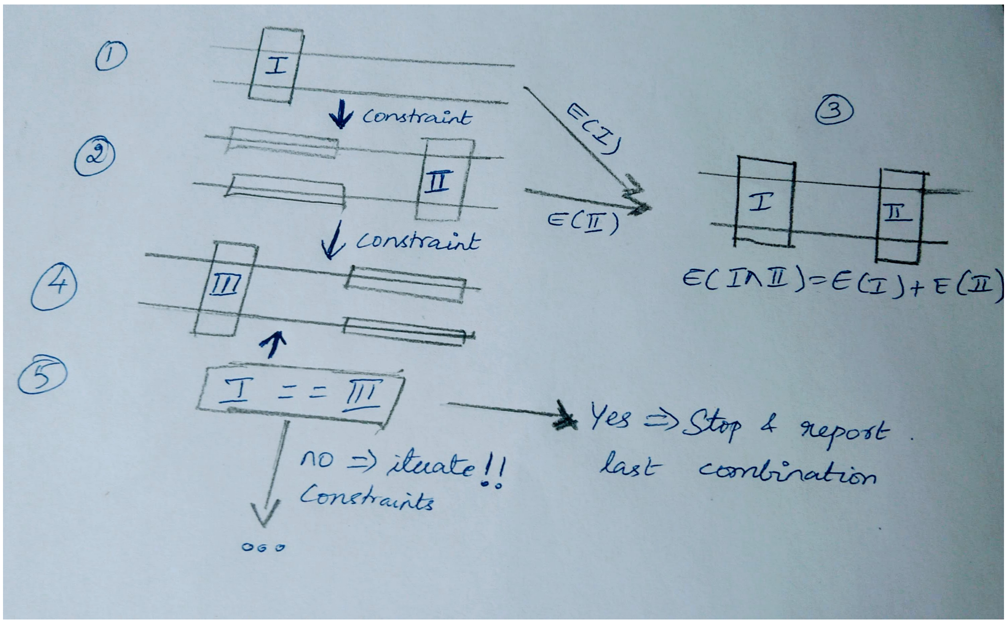
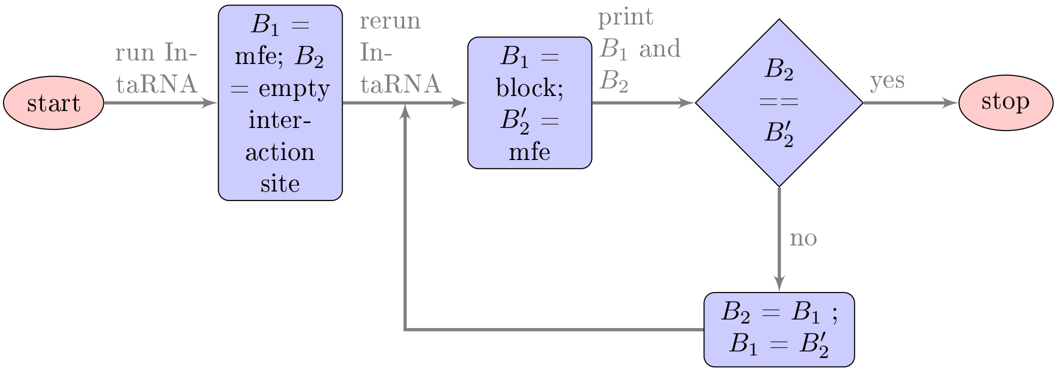


Figure 2.2: Iterative scheme that is used in this thesis.

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

- Step 3: Since prediction (2) is conditional, the existence of the interaction from (2) can have effects on (1). Thus, one starts to iterate the procedure from (2) but fixes the conditional site. Then , we print the respective the B_1 and B_2. Finally, check for convergence: is the site from two-steps before retained ($B_2 == B_2'$)? If yes: convergence and stop iteration by printing the new B_1, B_2. If no: repeat constraint prediction until convergence by swapping (ie.,$B_2 = B_1; B_1 = B_2'$).

Below is the flowchart representation for the same.


start
run IntaRNA
B_1 = mfe; B_2 = empty interaction site
rerun IntaRNA
B_1 = block; B_2' = mfe
print B_1 and B_2
B_2 == B_2'
yes
stop
no
$B_2 = B_1$; $B_1 = B_2'$

Below is the proof of why the energy of the conditional call can be just added.

$$\begin{aligned}E(B_1 \wedge B_2) &= E(B_1) + E(B_2|B_1)\\ &= E_{hyb}(B_1 \wedge B_2) + ED(B_1 \wedge B_2)\\ &= E_{hyb}(B_1) + E_{hyb}(B_2) - RTlog(\mathcal{P}^u(B_1 \wedge B_2))\end{aligned} \tag{2.4}$$

$$\begin{aligned}E(B_1) &= E_{hyb}(B_1) + ED(B_1)\\ &= E_{hyb}(B_1) - RTlog(\mathcal{P}^u(B_1))\end{aligned} \tag{2.5}$$

$$\begin{aligned}E(B_2|B_1) &= E_{hyb}(B_2|B_1) + ED(B_2|B_1)\\ &= E_{hyb}(B_2) - RTlog(\mathcal{P}^u(B_2|B_1))\end{aligned} \tag{2.6}$$

Now, we add right end side values of Eqn 2.6 + 2.5, we get,

$$E_{hyb}(B_1) + E_{hyb}(B_2) - RTlog(\mathcal{P}^u(B_1)) - RTlog(\mathcal{P}^u(B_2|B_1)) \tag{2.7}$$

As we know $log(A)+log(B) = log(A.B)$, we apply this condition for log values in the Eqn 2.7

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

Since $P(A \wedge B) = P(A) * P(B|A)$ we get,

$$-RTlog(\mathcal{P}^u(B_1 \wedge B_2))$$

After adding and simplifying the right end side values of Eqn 2.6 + 2.5 we get,

$$E_{hyb}(B_1) + E_{hyb}(B_2) - RTlog(\mathcal{P}^u(B_1 \wedge B_2)) \tag{2.8}$$

Now, we see the equations 2.8 and 2.4 are equal ,

$$E_{hyb}(B_1)+E_{hyb}(B_2)-RTlog(\mathcal{P}^u(B_1 \wedge B_2)) \Leftrightarrow E_{hyb}(B_1)+E_{hyb}(B_2)-RTlog(\mathcal{P}^u(B_1 \wedge B_2)) \tag{2.9}$$

We also know that

$$\begin{aligned}E_{hyb}(B_1) &= E_{hyb}(B_1)\\ E_{hyb}(B_2|B_1) &= \sum_{\forall loopsinB_2} E_{hyb}(B_2)\end{aligned}$$

Hence proved.

Chapter 3

Results

Chapter 4

Discussion and conclusion

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.

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.

J Brachet and A Ficq. Remarks on the biological role of nucleic acids. *Archives de biologie*, 67 (3-4):431–446, 1956.

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.

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

Rick Gelhausen. *Constrained RNA-RNA interaction prediction.* PhD thesis, Master's thesis, Albert-Ludwigs University of Freiburg, 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.

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.

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.

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.