ALBERT LUDWIGS UNIVERISTY OF FREIBURG

MASTER THESIS

Multisite RNA-RNA Interaction Prediction

Yogapriya Ayyanarmoorthy

November 12, 2019

Contents

1 Introduction **3**
1.1 Biological Background of RNA 3
1.2 Formal background of RNA 5
1.2.1 RNA Structure 5
1.2.2 Nested secondary structure 5
1.2.3 Nearest neighbor model and energy contributions 6
1.2.4 Structure probabilities and McCaskill algorithm 7
1.3 RNA-RNA Interaction 10
1.4 RNA-RNA Interaction Prediction Approaches 10
1.4.1 Hybrid 10
1.4.2 General 11
1.4.3 Concatenation 11
1.4.4 Accessibility 12
1.4.5 Adv. and disadv. 13

2 Multisite Accessibility Based **14**

3 Results **15**

4 Discussion and conclusion **16**

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.

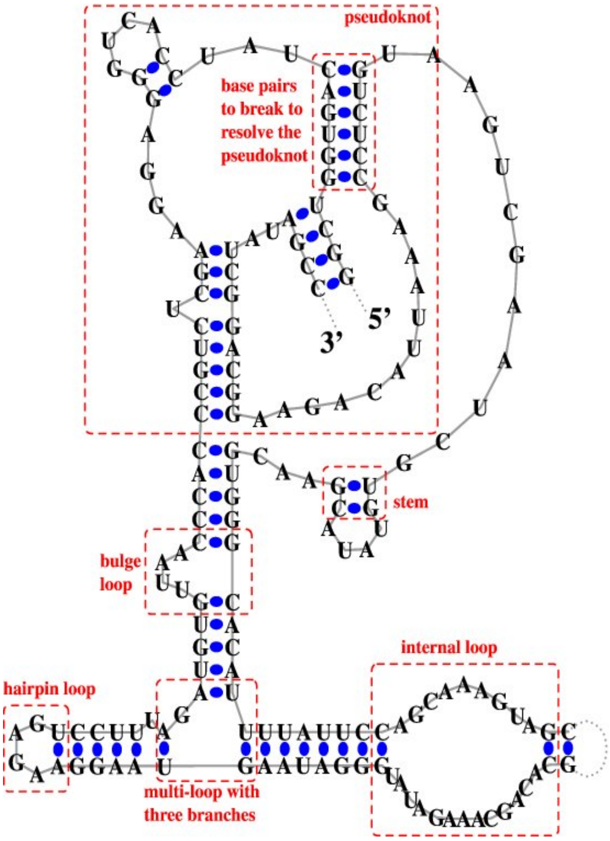
pseudoknot
base pairs to break to resolve the pseudoknot
5'
3'
stem
bulge loop
internal loop
hairpin loop
multi-loop with three branches

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, 1962) 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 $E(P)$ of a nested secondary structure P can be estimated by the sum of loop contributions (see Figure 1.2)

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

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^M_a scores the multi loop closing base pair (i,j) , e^M_b represents the penalty for directly enclosed unpaired bases x and e^M_c represents the number of enclosed helices x'. Thus the nearest neighbor model gives the energy contributions for the loop types.

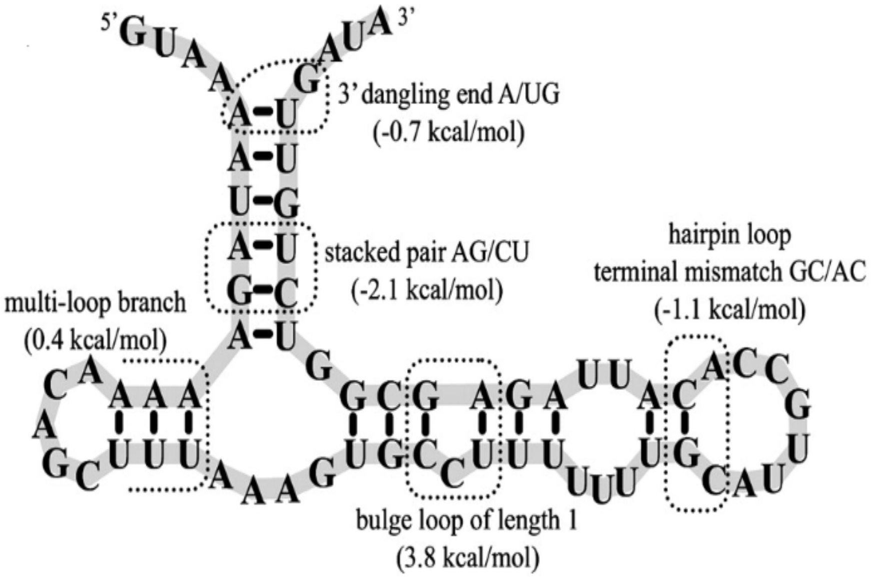

Figure 1.2: Energy contributions of loops. (Andronescu et al., 2010)

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.

$$exp\left(\frac{-E(x)}{k_B T}\right)$$

Where $E(x)$ represents the state energy , k_B represents the Boltzmann 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 X.

$$Z = \sum_{x' \in X} exp\left(\frac{-E(x')}{k_B T}\right)$$

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

$$Pr[P|p] = exp\left(\frac{\frac{-E(P)}{RT}}{k_B T}\right)$$

and normalising with the partition function Z for the structure ensemble p

$$Z = \sum_{P' \in p} exp\left(\frac{-E(P')}{RT}\right)$$

The structure with the highest probability is not necessarily the function structure. This means they are biologically correct one.
we can study the probabilities of individual structural elements. The probability of structural elements can be summarised by using base-pair probabilities. We can also calculate them. The probability of a base pair (i,j) for S :

$$Pr[(i,j)|S] = \sum_{P \ni (i,j)} Pr[P|S]$$

Base pair probabilities enable new view at structure ensemble both visually and algorithmically. They are represented by dot plot.
Similar to the base pair probabilities, we can also calculate the probabilities of unpaired regions. Formally, we will identify the probability of the subsequences i..j to be unpaired by $Pr_u[i,j]$. 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

$$Pr_u[i,j] = \frac{Z^u_{i_j}}{Z}$$

where $Z^u_{i_j}$ is the partition function of all structures where the subsequence i..j is unpaired. The below figure was inspired by the lecture material of RNA bioinformatics lecture

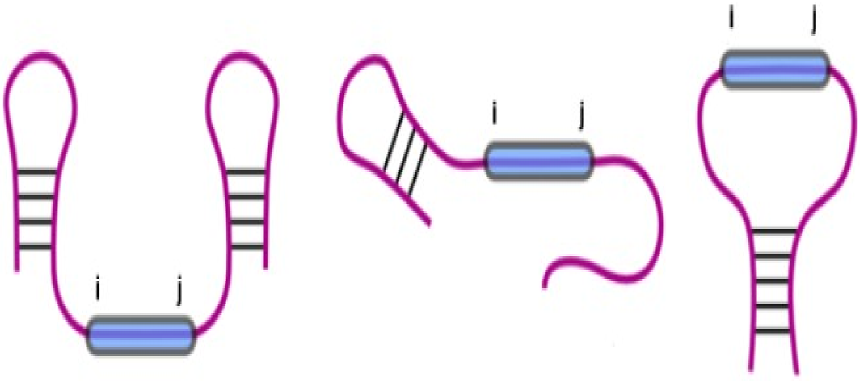

Figure 1.3: Examplary structures that are unpaired in the subsequence i..j.

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.What is RNA
- 2.RNA representation a,c,g,u
- 3.classes of rna
- 4.base pairs of RNA
- 5.RNA secondary structure
- 6.types of rna secondary structure
- 7.nearest neighbor model
- 8.unpair probabilities

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.

- Why RRI

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).

1.4 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.
Most computational methods for RNA structure or RNA-RNA interaction prediction are based on thermodynamic models and provide an efficient computation since Richard Bellman's principle of optimality (Raden et al., 2018) can be applied.
In the following subsections we will see about the various approaches that can predict 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 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). A dynamic programming approach using a simplified energy model with two dimensional table H is filled via the prefix-based recursion for the nussinov like interaction prediction.

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

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 1.4 . The above 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^1_i, $\overleftarrow{S^2_j}$ 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.

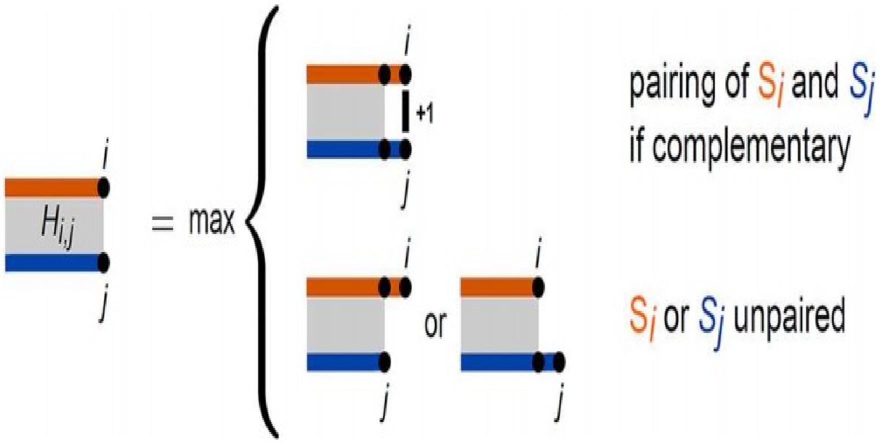

Figure 1.4: Recursion scheme to maximize intermolecular base pairs between two RNAs $S1$ and $S2$ represented in orange/blue, respectively

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 adn 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 overcomed by concatenation and accessibility based approaches.

1.4.2 General

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 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.

This is extension of nussinov 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 optimal hybrid structures can be listed by the sub optimal traceback implementation. The parenthesis "()" are used for representing the intramolecular base pairs and square brackets "[]" represents the intermolecular base pairs, "X" represents the linker.

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 kissing stem-loop or kissing hairpin-loop (as seen in 1.5) cannot be predicted because they form a pseduoknot by them. To predict these patterns we go for Accessibility based approaches.

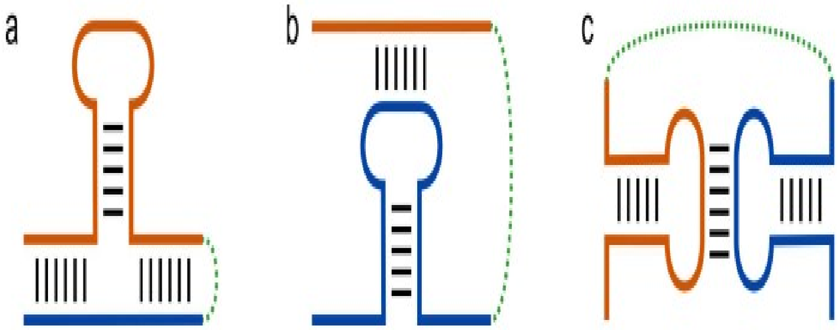

Figure 1.5: a) Pattern that can be predicated by Concatenation b)Kissing stem-loop c) kissing hairpin interaction. The blue and orange are the two different RNA's and the dotted green is the linker , black lines represents the base pairs.

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. 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.

The accessibility-incorporating interaction scorings are computed and stored in table I. A non-zero entry $I^{i,k}_{j,l}$ represents the combined scoring for an interaction of $S^1_{i..k}$ with $\overleftarrow{S^2_{j..l}}$ with left/right most base pairs $(S^1_i, \overleftarrow{S^2_j})/(S^1_k, \overleftarrow{S^2_l})$, respectively.

$$I^{i,k}_{j,l} = max \begin{cases} (E^{bp} \cdot D^{i,k}_{j,l} + ED^1_{i,k} + ED^2_{j,l}) & : \text{if } D^{i,k}_{j,l} > 0, \\ -\infty & : \text{otherwise} \end{cases}$$

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

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.

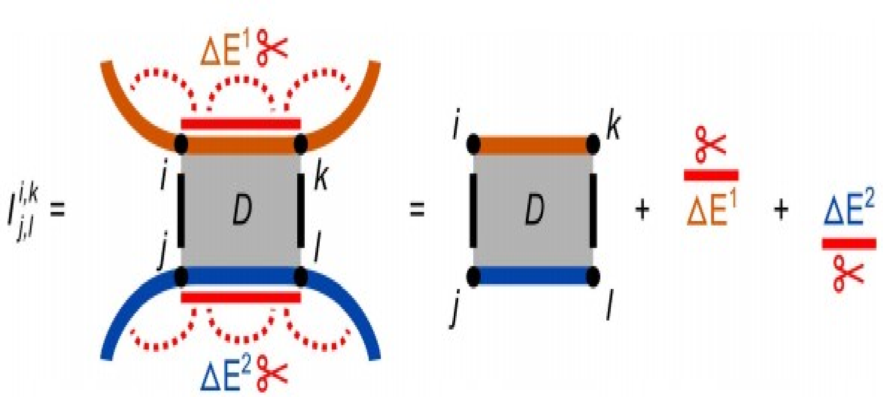

Figure 1.6: Depiction how accessibility-based approaches score an interaction of two RNAs $S1$ and $S2$ in orange and blue respectively

To further simplify the recursions, we use dedicated calculations for the duplex energy 19) and the overall interaction energy. This is a 4-d matrix, in which $D^{i,k}_{j,l}$ provides the the duplex energy of the interacting sites. 19Þ The first case represents the initiation of a new interaction that covers only the intermolecular base pair (i,j), second case is the extension of already computed interaction of $S^1_{p..k}, \overleftarrow{S^2_{q..l}}$ with a new base pair (i, j) and the third case is used if base pairs cannot be formed.

$$D^{i,k}_{j,l} = max \begin{cases} 1 & : \text{if } S^1_i, \overleftarrow{S^2_j} \text{ compl.}, i = k, j = l \\ \max\limits_{\substack{i<p\leq k,\ j<q \\ leql}} \left(1 + D^{p,k}_{q,l}\right) & : \text{if } S^1_i, \overleftarrow{S^2_j} \text{ compl.}, i < k, j < l, \\ -\infty & : \text{otherwise} \end{cases}$$

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. These approaches cannot be modelled correctly for the double kissing hairpin interaction.

- 1. S-RRI, M-RRI
- 2. problems with S-RRI

1.4.5 Adv. and disadv.

Hence we go for, Multi-site RRI optimization based on single-site IntaRNA predictions.

Chapter 2

Multisite Accessibility Based

Chapter 3

Results

Chapter 4

Discussion and conclusion

Bibliography

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.

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.

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

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

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.

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.

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.

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.

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.