Prediction of non-related RNA-RNA interactions

Rick Gelhausen

April 5, 2017

Introduction

IntaRNA is an algorithm to predict interactions of bacterial sRNA with a target mRNA. An RNA molecule is transcribed from a large part of DNA. There are two classes of RNA, the coding RNA (cRNA) and the non-coding RNA (ncRNA). The cRNA is involved in the translation into proteins. For example the messenger RNA that carries informations to encode proteins. The ncRNA performs different functions in the cell. In the cRNA mainly the composition of the sequence is important, whereas the function of a ncRNA is mainly related to its structure.
The bacterial small RNAs used by IntaRNA, are highly structured small-chained ncRNAs produced by bacteria. They can have multiple functions, such as the modification of the function of proteins or regulate gene creation by binding to mRNA.
The RNA molecules are represented as a sequence $S \in A, C, G, U^*$. $s_i ... s_j$ represents a subsequence of S such that $s \subseteq S$. The sequences are usually ordered from the left $5'$ end to the right $3'$ end.

Figure 1: RNA sequence S

These sequences can be folded into structures that determine the function of an RNA molecule.
An RNA structure P of S is a set of base pairs.

$$P \subseteq \{(i,j) \mid 1 \leq i < j \leq n, S_i \text{ and } S_j \text{ complementary}\},$$

where n = $\mid S \mid$ and the degree of P is atleast one. As the prediction of tertiary structures is a hard problem, secondary structures are used instead. There are different types of RNA structures. Nested structures that contain no pseudoknots and crossing structures containing pseudoknots, as shown in Figure 2. Pseudoknots are, as the name suggests, no real knots in the tertiary structure, but cause problems in the secondary structure representation.

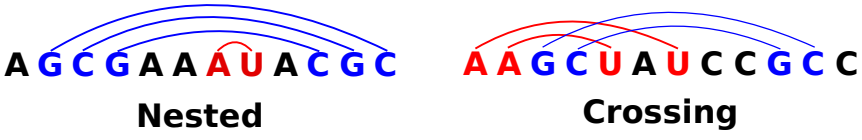


Figure 2: Linear Feynman Diagrams of a nested and a crossing structure.

The structure is created by bases forming base-pairs by hydrogen bonds. Due to their high binding strength G-C, A-U and G-U are the most com-

mon base pairs.

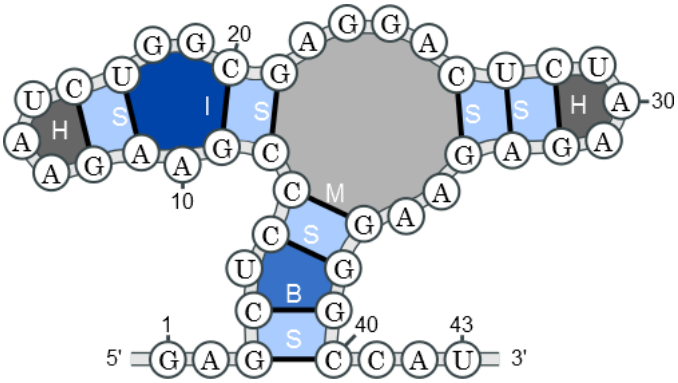


Figure 3: Overview of structure elements. [Mann, 2016]

There are different structural elements that are formed by the binding of base pairs. These elements are the hairpin loops (H), stackings (S), internal/bulge loops (I/B) and multiloops (M), as shown in figure 3. A secondary structure is composed of these structural elements. The ensemble of all structures is denoted $\mathcal{P}_{\text{all}}$.
IntaRNA has different restrictions to reduce the complexity of the algorithm, to make it applicable in practice. One strong restriction is that the target region is single stranded and the loops are no longer than 16 nucleotides.
In this work we remove these restrictions to allow structures in the target region. We will first introduce further notions and the McCaskill algorithm [McCaskill, 1990] used to calculate partition functions and probabilities. Then we will introduce the update recursions of the IntaRNA algorithm to allow structures in the target region.

Energy Model and Probabilities

IntaRNA uses energy minimization to find an optimal solution. The underlying energy model is the Nearests Neighbor Model. It states that the vertical stacking of base pairs gives the largest contribution to the stability of the system. It allows the calculation of a free energy estimate for a RNA secondary structure. The lower the free energy the more energy has to be invested to disrupt the system. A structure is more stable, the lower the free energy. This means that the stablest structure is the minimum free energy structure (mfe). As the free energy is hard to calculate, the usage of energy differences is customary. The free energy is calculated with respect to the unstructured open chain.
The boltzmann distribution is, according to the maximum entropy principal,

the best probability distribution for the calculation of structure or base pair probabilities. It gives us a huge information gain, with a low information content. We therefor calculate probabilities according to their boltzmann weights.

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

where E is a specific energy of a structure P, R is the gas constant used to calculate the energy for a single molecule and T is the temperature.

Using these boltzmann weights we can calculate the partition function Z. Z is the sum over all boltzmann weights for all structures P.

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

Z is required in the calculation of structure and base pair probabilities. These probabilities are calculated in the thermodynamic equilibrium. This means that there are no observable changes on a macroscopic level. Multiple different probabilities can be calculated. Such as the probability of a structure to be formed.

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

This way the k most probable mfe structures can be calculated. As the underlying energy model is an estimation and simplification of the truth, the structure with the highest probability does not have to be the functional structure, but it is safe to assume that the best structure is among the most probable structures.
It is also possible to calculate the probability that a certain base pair appears.

$$Pr[(i,j)|\mathcal{P}] = \sum_{(i,j) \in P} \frac{w(P)}{Z}$$

The probability that base pair (i, j) occurs. These base pair probabilities can be represented in a dot plot and give a good overview how the most probable structure can look like, because the most probable base pairs are likely contained in the most probable structures.
Furthermore the probability that a given region of a structure is unpaired can be determined.

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

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

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

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

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

The unpaired probability is very important, as it allows the calculation of the accessibility of single stranded regions.

McCaskill

Preliminaries

The McCaskill algorithm is used to calculate the partition function Z for a given sequence S, which can be used to compute probabilities.
The basic idea is to use an algorithm, similar to the Zuker algorithm, to sum up the boltzmann weights for all possible structure. The important part is to count every structure only once, which requires the creation of a special multiloop case.
There are four matrices required in the algorithm.

$$Q_{i,j} = Z_{\mathcal{S}_{i,j}} \qquad Q^m_{i,j} = Z^m_{\mathcal{S}^{1bd}_{i,j}}$$
$$Q^b_{i,j} = Z^b_{\mathcal{S}_{i,j}} \qquad Q^{m1}_{i,j} = Z^m_{\{P \in \mathcal{S}^{1bd}_{i,j} | \text{only one exterior bp in P}\}}$$

$Q_{i,j}$ contains the summed boltzmann weights for all structures which only contain bonds in range $[i,j]$. $Q^b_{i,j}$ has the additional property that (i,j) forms a base pair. $Q^m_{i,j}$ has the additional property that there is atleast one base pair in range (i,j). $Q^{m1}_{i,j}$ has the property that there is exactly one base pair within (i,j). The final result $Z = Z_{\mathcal{S}_{all}}$ is contained in $Q_{1,|N|}$.
The decomposition has to be disjoint and independent to ensure that every structure is only counted once. This requires a change in the multiloop split case of the Zuker algorithm, as it is an ambiguous decomposition. Therefore the matrix $Q^{m1}_{i,j}$ is introduced, where the multiloop split is always done at the last bond. This ensures an unambiguous decomposition.

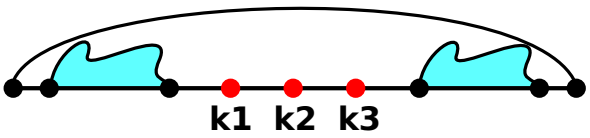
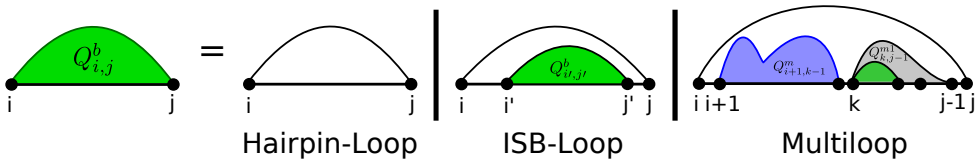


Figure 4: Ambiguous multiloop split. The split could be done at k1, k2 or k3.

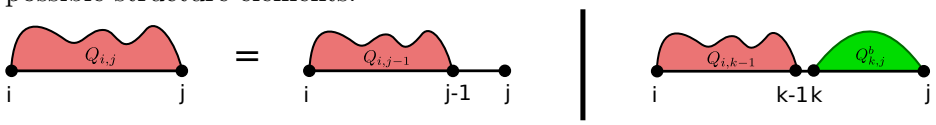
Matrices

In the following equations a, b and c represent energy contributions of multiloops. This is a simplification of the real energy contributions of multiloops, because the real energy contribution computation is too complex. a denotes the energy contribution for closing the multiloop. b denotes the contribution for enclosed helices and c denotes the contribution for enclosed unpaired bases. The energy contribution for hairpin loops is denoted as $eH(i,j)$ and the contribution for stacking, internal and bulge loops are denoted as $eSBI(i,j,i',j')$.

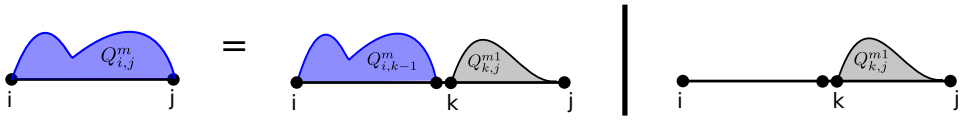


$$Q^b_{ij} = \sum \begin{cases} e^{\frac{-eH(i,j)}{RT}} \\ \sum\limits_{i<i'<j'<j} (Q^b_{i',j'} \cdot e^{\frac{-eSBI(i,j,i',j')}{RT}}) \\ \sum\limits_{i<k<j} (Q^m_{i+1,k-1} \cdot Q^{m1}_{k,j-1} \cdot e^{\frac{-a}{RT}}) \end{cases} \tag{1}$$

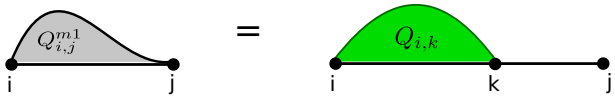
In the $Q^b_{i,j}$ matrix, we add up the energy contributions for the different possible structure elements.



$$Q_{i,j} = Q_{i,j-1} + \sum_{i \leq k < j} Q_{i,k-1} \cdot Q^b_{k,j} \tag{2}$$



$$Q^m_{i,j} = \sum_{i \leq k < j} (Q_{i,k-1} + e^{\frac{-(k-i)c}{RT}}) \cdot Q^{m1}_{kj} \tag{3}$$



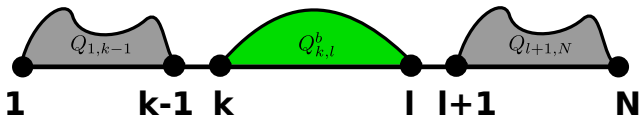
$$Q^{m1}_{kj} = \sum_{i < k \leq j} Q^b_{i,k} \cdot e^{\frac{-b}{RT}} \cdot e^{\frac{-(j-k)c}{RT}} \tag{4}$$

The initialization for the matrices $Q^b_{i,j}$, $Q^m_{i,j}$ and $Q^{m1}_{i,j}$ is 0. For the single-stranded sequence there is no base, therefore there will not be an entry in these matrices. $Q_{i,j}$ is initialized with 1 as it covers the single-stranded sequence.

Base pair probabilities:

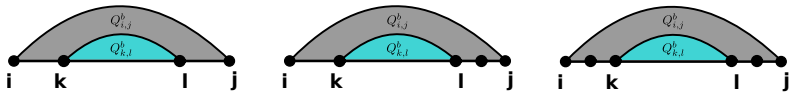
The probabilities for unpaired bases can be calculated using the McCaskill recursions. There are three possible locations for a base pair (k, l).

1. (k,l) is an external base pair:



$$p^E_{kl} = \frac{Q_{1,k-1} \cdot Q^b_{k,l} \cdot Q_{l+1,n}}{Q_{1,n}}$$

2. (k,l) limits a stacking, bulge- or interior loop closed by bp (i,j), where $i < k < l < j$.



$$p^{SBI}_{kl}(i,j) = p_{ij} \frac{\exp(\frac{-eSBI(i,j,k,l)}{RT}) Q^b_{k,l}}{Q^b_{i,j}}$$

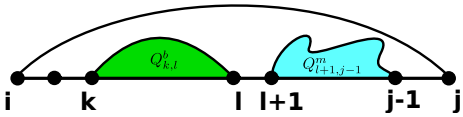
The probility that base pair (i,j) is formed $p_{i,j}$ is computed in an outside recursion, before the computation of p^{SBI}_{kl}. $p_{i,j}$ is then corrected by taking the additional constraint that the loop i,j,k,l is formed. The numerator of the fraction is the partition function of all structures containing bp (k,l). It is corrected by the denominator $Q^b_{i,j}$ which is the partition function of all bp containing bp (i,j).

3. (k,l) closes an inner helix of a multiloop closed by bp (i,j), where $i < k < l < j$.

$p^M_{kl}(i,j) = p_{ij} \cdot Pr[\text{Multiloop with inner bp } (k,l) \text{ closed by } (i,j) \mid (i,j)]$

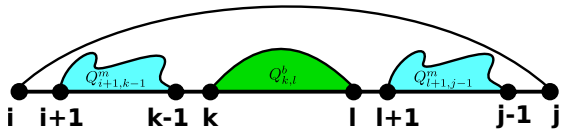
There are again three locations for (k,l) inside the multiloop:

(a) (k,l) is the leftmost base pair:



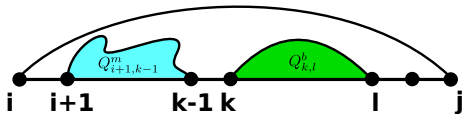
$$\frac{Q^b_{k,l} \cdot Q^m_{l+1,j-1} \cdot \exp(\frac{-(a+b+(k-i-1)c)}{RT})}{Q^b_{i,j}}$$

(b) (k,l) is the middle base pair:



$$\frac{Q^m_{i+1,k-1} \cdot Q^b_{k,l} \cdot Q^m_{l+1,j-1} \cdot \exp(\frac{-(a+b)}{RT})}{Q^b_{i,j}}$$

(c) (k,l) is the rightmost base pair:



$$\frac{Q^m_{i+1,k-1} \cdot Q^b_{k,l} \cdot \exp(\frac{-(a+b+(j-l-1)c)}{RT})}{Q^b_{i,j}}$$

$$\begin{aligned} p^M_{kl}(i,j) = \frac{p_{ij}}{Q^b_{i,j}} \cdot (&Q^b_{k,l} \cdot Q^m_{l+1,j-1} \cdot \exp(\frac{-(a+b+(k-i-1)c)}{RT}) \\ &+ Q^m_{i+1,k-1} \cdot Q^b_{k,l} \cdot Q^m_{l+1,j-1} \cdot \exp(\frac{-(a+b)}{RT}) \\ &+ Q^m_{i+1,k-1} \cdot Q^b_{k,l} \cdot \exp(\frac{-(a+b+(j-l-1)c)}{RT})) \end{aligned}$$

The overall probability for a base pair (k,l) is denoted:

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

Probabilities of unpaired regions:

A very important concept for RNA-RNA interaction prediction is the calculation of the probability of unpaired regions, as unpaired regions are possible targets for interactions.
These probabilities $Pr_u[i,j]$ can again be calculated using the McCaskill recursions.

$$Pr_u[i,j] = \frac{Z_{\mathcal{P}^u_{i,j}}}{Z_{\mathcal{P}_{all}}}$$

The probability that region $[i, j]$ is unpaired, where $\mathcal{P}^u_{i,j}$ is the set of structures where region $[i, j]$ is unpaired and $\mathcal{P}_{all}$ the set of all structures for a given sequence.
There are two different locations for an unpaired region it is either exterior or enclosed. An exterior region is enclosed by no base pairs. The enclosed region is either enclosed by a hairpin, an interior/bulge loop or a multiloop. Using a disjoint decomposition of $\mathcal{P}^u_{i,j}$ the different cases can be viewed independantly.

Case I $[i, j]$ is exterior:

$$Pr_u[i, j \mid exterior] = \frac{Q_{1,i-1} \cdot 1 \cdot Q_{j+1,N}}{Q_{1,N}}$$

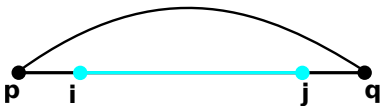
where N is the length of the sequence and 1 is the boltzmann weight of the unpaired region. The multiplication is justified as this is an independent decomposition of the sequence.

Case II $[i, j]$ is enclosed by bp (p, q):

$$Pr_u[i, j \mid enclosed] = \sum_{p<i,j<q} \frac{Pr[(p, q)]}{Q^b_{p,q}} \cdot Q^{pq}_{i,j}$$

where $Pr[(p, q)]$ is the probability that (p, q) forms a base pair. $Q^b_{p,q}$ is the partition function of all structures that form are enclosed by (p, q) and $Q^{pq}_{i,j}$ is the partition function of all structures that have a base pair (i, j) enclosed by a bp (p, q).
$Q^{pq}_{i,j}$ is composed of the sum over the different cases of structural elements.

1. Hairpin Loop:

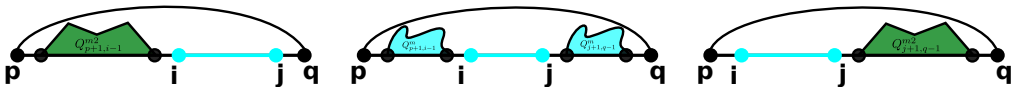


$$e^{\frac{-eH(p,q)}{RT}}$$

2. Stacking, Internal- and Bulge Loop:

$$\sum_{\substack{p<i\leq j<k \\ or \\ l<i\leq j<q}} e^{\frac{-eSBI(p,q,k,l)}{RT}} \cdot Q^b_{k,l}$$

3. Multiloop:



$$\begin{aligned}\sum_{p<i\leq j<q} \quad & Q^{m2}_{p+1,i-1} \cdot e^{\frac{-(q-i)c}{RT}} \\ & + Q^{m}_{p+1,i-1} \cdot e^{\frac{-(j-i+1)c}{RT}} \cdot Q^{m}_{j+1,q-1} \\ & + e^{\frac{-(j-p)c}{RT}} \cdot Q^{m2}_{j+1,q-1} \\ \text{with } Q^{m2}_{i,j} = & \sum_{p<k<q} Q^{m}_{p,q} \cdot Q^{m1}_{k+1,q}\end{aligned}$$

where Q^{m2} ensures that we have atleast two helices.

The probability that $[i, j]$ is unpaired is:

$$\begin{aligned}Pr_u[i,j] &= Pr_u[i,j \mid exterior] + Pr_u[i,j \mid enclosed] \\ &= \frac{Q_{1,i-1} \cdot 1 \cdot Q_{j+1,N}}{Q_{1,N}} + \sum_{p<i,j<q} \frac{Pr[(p,q)]}{Q^b_{p,q}} \cdot Q^{pq}_{i,j}\end{aligned}$$

IntaRNA

IntaRNA is an algorithm to predict RNA-RNA interactions of bacterial small RNAs and a target mRNA. The original version of IntaRNA is limiting the loop sizes in the target regions to 16 nucleotides and it allows neither inter, nor intra molecular base pairs in the target region.
There are two major components in IntaRNA that determine the quality of RNA-RNA interactions between two subsequences of sequences S^1 and S^2, the hybridisation energy $H(i, j, k, l)$ and the accessibility of the interaction site.
The hybridisation energy is calculated using the nearest neighbour energy model. It represents the hybridisation minimum free energy of two subsequences, where the leftmost positions of both subsequences form a base pair. For simplification purposes $E_{3'}^{dangle}$, $E_{5'}^{dangle}$ and E_{mm}^{term} will not be considered in the following recursions.
For subsequences $S_i^1...S_k^1$ and $S_j^2...S_l^2$, 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.

$$H(i, j, k, l) = \min \begin{cases} E_{init} \\ \quad : \text{if } S_i^1, S_j^2 \text{ can pair, } i = k \text{ and } j = l, \\ \min\limits_{r,s}\{E^{loop}(i, j, r, s) + H(r, s, k, l)\} \\ \quad : \text{if } S_i^1, S_j^2 \text{ can pair, } i \neq k \text{ and } j \neq l, \\ \infty \\ \quad : \text{otherwise.} \end{cases}$$

where $E^{loop}(i, j, k, l)$ represents the energy contribution of the loop formed by the base pairs (i, j) and (k, l).
The accessibility represents the energy required to make the interaction site single-stranded. It is calculated as the energy difference between the energy of the ensemble of all structures that can be formed by S and the energy of the ensemble of all structures, where the interaction site is single-stranded. This energy difference $ED(i, j)$ is computed using a partition function approach as introduced by [McCaskill, 1990].
The free energy of the ensemble $\mathcal{P}$ is

$$E^{ens}(\mathcal{P}) = -RT \cdot ln(Z_{\mathcal{P}})$$

It follows that $ED(i, k)$ is

$$ED(i, k) = E^{ens}(\mathcal{P}_{i,k}^{unpaired}) - E^{ens}(\mathcal{P})$$

Figure 5 shows the energy contributions considered in the original IntaRNA recursions.

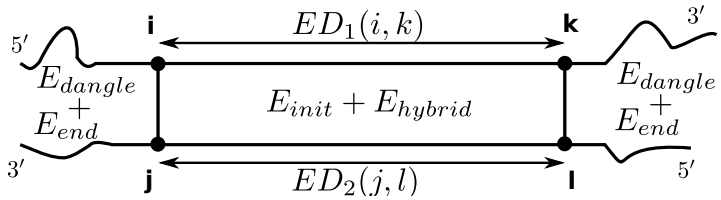


Figure 5: The energy contributions needed in the original IntaRNA recursions.

Both the accessibility and the hybridisation energy are combined to form the extended hybridisation energy. The extended hybridisation energy of a specific hybridisation between $S_i^1...S_k^1$ and $S_j^2...S_l^2$ is defined by:

$$C(i,j,k,l) = \begin{cases} H(i,j,k,l) + ED_1(i,k) + ED_2(j,l) \\ \quad : \text{if } S_i^1, S_j^2 \text{ can pair, } i \neq k \text{ and } j \neq l, \\ \infty \\ \quad : \text{otherwise.} \end{cases}$$

Motivation

Due to the strong restrictions to the interaction site by IntaRNA, many different structures, that exist in nature as seen in Figure 6, cannot be predicted.

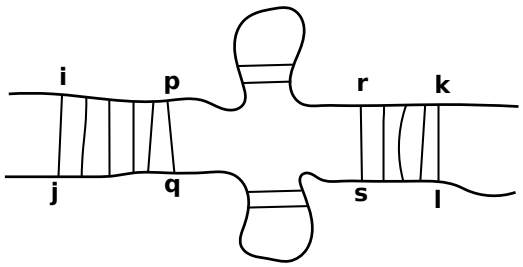


Figure 6: Predicted structures containing multi-loops.

This is because we disallow intra-molecular base pairs in the target regions. Our aim is to change the IntaRNA recursions to allow structure in the interaction site. This will create different multi-loop cases and leave us with a new matrix $H^m(i,j,k,l)$.
By incorporating structure in the interaction site, we have to introduce a new energy $ES(i,k)$ which denotes the energy contribution of the structure

present in the interval $]i, k[$.

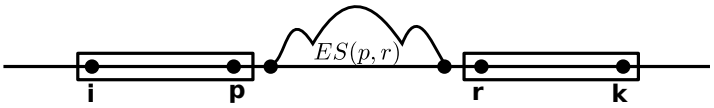


Figure 7: The energy contribution of the structure on the interaction site.

The $ES(p, r)$ is calculated similar to the free energy of the structure ensemble, under the condition that the structure is in the interval $]p, r[$ excluding the bases p and r that are forming base pairs with q and s respectively.

$$ES(p, r) = -RT \cdot ln(Q^m_{p+1,r-1})$$

where Q^m [3] denotes the same matrix used in the McCaskill recursion [McCaskill, 1990]. It ensures atleast one base pair in the interaction site.

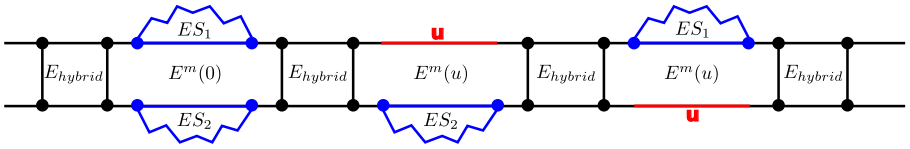


Figure 8: The different multi-cases with their associated energy contributions.

We have to consider the hybridisation energy, the structural energy and the multi-loop energies for each part. Furthermore, we will introduce a new energy function $E^m(u)$, which represents the energy contributions for the multi-loop.

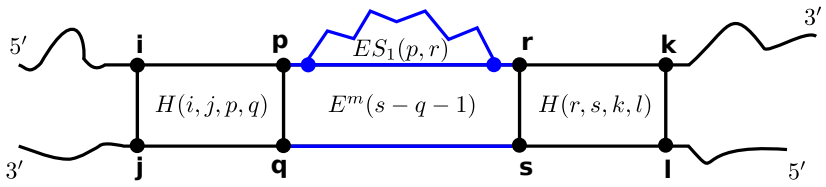
$$E^m(u) = a + b + c \cdot u$$

where u is the number of unpaired bases enclosed by the multi-loop, a is the energy contribution for closing the multi-loop, b is the energy contribution for the number of helices and c is the contribution for unpaired bases u.
The usage of the simple E^m energy function is justified, as all further energy contributions resulting from the structure are contained in the ES values, due to the Q^m matrix term.
Moreover, we have to analyse whether it makes more sense to add the initial energy for each structural part as a penalty or to add it once for the entire sequence.
Another idea is to replace the energy contribution for the closing pair of the multi-loop by the initial energy term. In the Turner 2004 parameters the a term is given by 9.3 kcal/mol and the initial energy E_{init} is 4.1 kcal/mol.

Multiloop cases

To introduce structure in the interaction site, we have to distinguish the different multiple cases. The structure is either in S^1, in S^2 or in both.

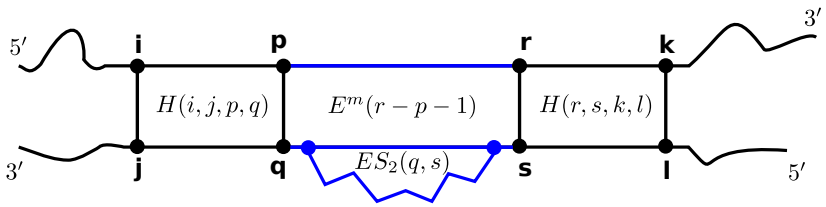
1. **Structure in $S^1_{p+1}...S^1_{r-1}$:**
   The interval $]p, r[$ is structured while the intervals $[i, p]$, $[r, k]$ and $[j, l]$ are free, with $i \leq p < q \leq k$ and $j \leq r < s \leq l$, under the condition that (i, j), (k, l), (p, q) and (r, s) form a base pair.
   To reduce the complexity we limit the length of the unpaired region $1 \leq s - q - 1 \leq 16$. This ≤ 16 is the same as in the interior loop length restriction.



   The energy is then computed using the hybridisation energy, the ED and ES values and the contribution for closing base pairs, helices and unpaired bases.

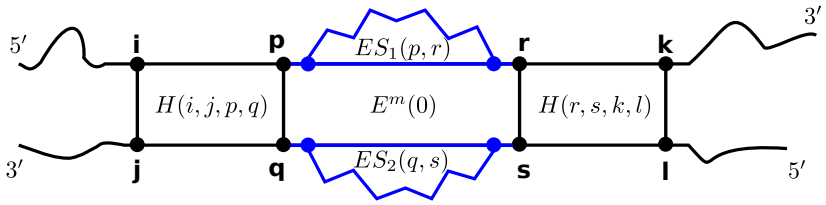
$$\begin{aligned} E = & H(i,j,p,q) + ES_1(p,r) \\ & + E^m(s-q-1) + H(r,s,k,l) \\ & + ED_1(i,k) + ED_2(j,l) \end{aligned}$$

2. **Structure in $S^2_{q+1}...S^2_{s-1}$:**
   This case in analogous to the first case. The structure is now in the interval $]q, s[$.



$$\begin{aligned} E = & H(i,j,p,q) + E^m(r-p-1) \\ & + ES_2(q,s) + H(r,s,k,l) \\ & + ED_1(i,k) + ED_2(j,l) \end{aligned}$$

3. **Structure in both subsequences:**
   The interval $]p, q[$ and $]r, s[$ are both structured and (i, j) and (p, q), as well as (r, s) and (k, l) form base pairs, with $i \leq p < q \leq k$ and $j \leq r < s \leq l$.



$$\begin{aligned} E = {} & H(i, j, p, r) + ES_1(p, r) + E^m(0) \\ & + ES_2(q, s) + H(q, s, k, l) \\ & + ED_1(i, k) + ED_2(j, l) \end{aligned}$$

Combining these cases leaves us with the new $H^m(i, j, k, l)$ matrix, which replaces the old $H(i, j, k, l)$ matrix.

$$H^m(i,j,k,l) = \min \begin{cases} E_{init} \\ \quad : \text{Initialization energy} \\ \min\limits_{\substack{i<r\leq k \\ j<s\leq l}} E^{loop}(i, j, r, s) + H^m(r, s, k, l) \\ \quad : \text{Interior loop case} \\ \min\limits_{\substack{i<r\leq k \\ j<s\leq l}} \begin{cases} H^m(r, s, k, l) + ES_1(i, r) + ES_2(j, s) + E^m(0) \\ H^m(r, s, k, l) + ES_1(i, r) + E^m(r - i - 1) \\ H^m(r, s, k, l) + E^m(s - j - 1) + ES_2(j, s) \end{cases} \\ \quad : \text{Multi-loop cases} \\ \infty \end{cases}$$

The $C^m(i, j, k, l)$ is analogous to the $C(i, j, k, l)$ matrix. The $H(i, j, k, l)$ entries are simply replaced by $H^m(i, j, k, l)$ entries.

Complexity

The newly created matrix $H^m(i, j, k, l)$ has a space complexity of $O(n^4)$ and a time complexity of $O(n^6)$ when allowing ES values in both the query and the target sequence. IntaRNA offers the option to chose between different modi, we can allow ES values in either the query sequence (–predMulti=Q), the target sequence (–predMulti=T), both sequence simultanously (–predMulti=B) or in both sequences independently (–predMulti=X). When allowing ES values in only the query or target structure or allowing ES values in both independently the complexity is reduced to $O(n^5)$. Allowing ES values in both sequences simultaneously causes $O(n^6)$ time complexity.

Results

We tested the new IntaRNA recursion on several datasets. First, we used a structure from the Accessfold paper [DiChiacchio et al. [2016]], shown in figure 9, and we tried to reproduce the structure using our recursions.

Figure 9: tmRNA structure. Taken from DiChiacchio et al. [2016].

The representation below shows the structure of figure 9 the way IntaRNA returns results. The other depictions show the output of IntaRNA using the different modi. The green regions were predicted correctly, the blue regions are missing in every result and the red regions were wrongly predicted, when comparing the output structures with the desired tmRNA structure. A base pair containing two colors indicates that the right query base formed a base pair with the wrong target base or vice versa.

```
>query Betaproteobacteria_1 1-47
GGGGGUGUACUGGUCUCGACAGGGCGGACAAAGGUGCGCAGGCAACU
>target Betaproteobacteria_2
AUGGUUGUAAUUCCGGCAAUCUGGAGGCGUUCUGGACAGGGGUUCGAUUCCCCUCACCUCCACCA
```

tmRNA structure:

```
target
                3                                                                          61
                |                                                                           |
         5'-AU       UAAUU  G    A      GGAGG        GACAGGGGUUCGAUUCCCCU           ACCA-3'
              GGUUG        CC GC  AUCU       CGUUCUG                     CACCUCC
              :||||        || ||  |:||       ||::|||                     ||||:||
              UCAAC        GG CG  UGGA       GCGGGAC                     GUGGGGG
           3'-              A  CG     AACAG         AGCUCUGGUCAU                   -5'
              |                                                                     |
              47                                                                    1
query
```

Results IntaRNA without multiloop recursion:

```
target
             28      34
              |       |
5'-AUG...GAGG          GACA...CCA-3'
              CGUUCUG
              ||::|||
              GCGGGAC
3'-UCA...ACAG          AGCU...GGG-5'
              |       |
             26      20
query
```

Result of IntaRNA --pred=M --mode=E --predMulti=T --noSeed:

```
target
                3                                                                          61
                |                                                                           |
         5'-AU        AAUUCCG  A     GGAGG                GGGUUCGAUUCCCCU              ACCA-3'
              GGUUGU          GC AUCU     CGUUCUG     GA CAG                 CACCUCC
              :||||:          || |:||     ||::|||     || |||                 ||||:||
              UCAACG          CG UGGA     GCGGGAC     CU GUC                 GUGGGGG
           3'-      GACG             AACAG       AGCU   G   AU                      -5'
              |                                                                     |
              47                                                                    1
query
```

Result of IntaRNA --pred=M --mode=E --predMulti=Q --noSeed:

```
target
             28      34
              |       |
5'-AUG...GAGG          GACA...CCA-3'
              CGUUCUG
              ||::|||
              GCGGGAC
3'-UCA...ACAG          AGCU...GGG-5'
              |       |
             26      20
query
```

Result of IntaRNA --pred=M --mode=E --predMulti=X --noSeed:

```
target
                3                                                                          61
                |                                                                           |
         5'-AU        AAUUCCG  A     GGAGG                GGGUUCGAUUCCCCU              ACCA-3'
              GGUUGU          GC AUCU     CGUUCUG     GA CAG                 CACCUCC
              :||||:          || |:||     ||::|||     || |||                 ||||:||
              UCAACG          CG UGGA     GCGGGAC     CU GUC                 GUGGGGG
           3'-      GACG             AACAG       AGCU   G   AU                      -5'
              |                                                                     |
              47                                                                    1
query
```

Result of IntaRNA --pred=M --mode=E --predMulti=B --noSeed:

```
target
                3                                                                   61
                |                                                                    |
         5'-AU        AAUUCCGGCAAUCUGGAGGCGUUCUGGACAGGGGUUCGAUUCCCCU          ACCA-3'
              GGUUGU                                              CACCUCC
              :||||:                                              ||||:||
              UCAACG                                              GUGGGGG
           3'-      GACGCGUGGAAACAGGCGGGACAGCUCUGGUCAU                     -5'
              |                                                      |
              47                                                     1
query
```

The best result is achieved by using the "–predMulti=T" mode, allowing ES values for the target sequence only or "–predMulti=X" allowing ES values for the target or the query sequence. In both cases, there is one region with missing interactions and the region before the intramolecular structure should not interact.
The query sequence used in the Accesfold paper [DiChiacchio et al. [2016]] was restricted to the first 47 nucleotides. We tried using the full query sequence.

```
>query Betaproteobacteria_1
GGGGGUGUACUGGUCUCGACAGGGCGGACAAAGGUGCGCAGGCAACUCGUCAGGCGAUCGACGUUAAUGAAGCAAAUCCAUAAUUGCCAAUGAUGAGCAAUUCGCUAUUGCCGCCUAAAAACG
GUUAGCCGGGGCUGCUAGAGCCUUGUUACCAAAGAUAGCCGGCGGGGACUUCGGUCCCCGUCGUCA
>target Betaproteobacteria_2
AUGGUUGUAAUUCCGGCAAUCUGGAGGCGUUCUGGACAGGGGUUCGAUUCCCCUCACCUCCACCA


tmRNA structure (full sequence):
target
              3                                                                  61
              |                                                                  |
        5'-AU     UAAUU  G    A     GGAGG         GACAGGGGUUCGAUUCCCCU          ACCA-3'
             GGUUG      CC GC   AUCU       CGUUCUG                    CACCUCC
             :||||      || ||   |:||       ||::|||                    ||||:||
             UCAAC      GG CG   UGGA       GCGGGAC                    GUGGGGG
3'-AUA...UGC               A  CG     AACAG         AGCUCUGGUCAU                  -5'
              |                                                                  |
              47                                                                 1
query

Result of IntaRNA without multiloop recursion:
target
              55      61
              |       |
5'-AUG...CCCU          ACCA-3'
             CACCUCC
             ||||:||
             GUGGGGG
3'-AUA...UCAU          -5'
              |       |
              7       1
query

Result of IntaRNA --pred=M --mode=E --predMulti=T --noSeed:
target
              4                                                                  61
              |                                                                  |
        5'-AUG     UAAUU     CAAUCUGGAGGCGUUCUGGACAGGGGUUCGAUUCCCCU          ACCA-3'
             GUUG       CCGG                                          CACCUCC
             ||:|       ||:|                                          ||||:||
             CAGC       GGUC                                          GUGGGGG
3'-AUA...GGGA     UCU        AU                                              -5'
              |                                                                  |
              20                                                                 1
query
```

The depictions above show the best result for the full query sequence, when assuming that the result for the full sequence is equal to the result of the sliced one. It is achieved by allowing only ES values in the target sequence. The unpaired region is predicted, but it is too large. Some of the target bases pair correctly but, due to the large unpaired region, with the wrong bases in the query sequence. all these problems could eventually be fixed by including seed constraints. The next page gives an overview over the results of the different modi.

tmRNA structure (full sequence):

```
target
            3                                                           61
            |                                                           |
       5'-AU     UAAUU  G   A    GGAGG        GACAGGGGUUCGAUUCCCCU       ACCA-3'
            GGUUG     CC GC  AUCU      CGUUCUG                    CACCUCC
            :||||     || ||  |:||      ||::|||                    ||||:||
            UCAAC     GG CG  UGGA      GCGGGAC                    GUGGGGG
3'-AUA...UGC            A  CG    AACAG        AGCUCUGGUCAU               -5'
            |                                                           |
            47                                                          1
query
```

Result of IntaRNA without multiloop recursion:

```
target
            55     61
             |      |
5'-AUG...CCCU       ACCA-3'
             CACCUCC
             ||||:||
             GUGGGGG
3'-AUA...UCAU       -5'
             |      |
             7      1
query
```

Result of IntaRNA --pred=M --mode=E --predMulti=T --noSeed:

```
target
             4                                                         61
             |                                                         |
       5'-AUG    UAAUU    CAAUCUGGAGGCGUUCUGGACAGGGGUUCGAUUCCCCU        ACCA-3'
             GUUG     CCGG                                       CACCUCC
             ||:|     ||:|                                       ||||:||
             CAGC     GGUC                                       GUGGGGG
3'-AUA...GGGA    UCU      AU                                        -5'
             |                                                         |
            20                                                         1
query
```

Result of IntaRNA --pred=M --mode=E --predMulti=Q --noSeed:

```
target
            24                                                                                                                                                    61
             |                                                                                                                                                    |
5'-AUG...UCUG     GU     A   GGGU                                                                                                                   UU  CCU        ACCA-3'
             GAGGC  UCUGG CAG                                                                                                                   UCGA  CC    CACCUCC
             :||||  |||:| |||                                                                                                                   ||||  ||    ||||:||
             UUCCG  AGAUC GUC                                                                                                                   AGCU  GG    GUGGGGG
3'-AUA...AUUG                GGGGCCGAUUGGCAAAAAUCCGCCGUUAUCGCUUAACGAGUAGUAACCGUUAAUACCUAAACGAAGUAAUUGCAGCUAGCGGACUGCUCAACGGACGCGUGGAAACAGGCGGGAC    CU  UCAU       -5'
             |                                                                                                                                                    |
           147                                                                                                                                                    1
query
```

Result of IntaRNA --pred=M --mode=E --predMulti=X --noSeed:

```
target
             3                                                                                                                                                            61
             |                                                                                                                                                            |
        5'-AU          A            UGG        U                                                                                             U       GGGUUCGAUUCCCCU       ACCA-3'
             GGUU   GUA UUCCGGC AAUC      AGGCG                                                                                            UC GGA CAG               CACCUCC
             ||:|   |:| ::||||| ||:|      |||||                                                                                            || :|| |||               ||||:||
             CCGA   CGU GGGGCCG UUGG      UCCGC                                                                                            AG UCU GUC               GUGGGGG
3'-AUA...UGUU    GAU   C       A    CAAAAA     CGUUAUCGCUUAACGAGUAGUAACCGUUAAUACCUAAACGAAGUAAUUGCAGCUAGCGGACUGCUCAACGGACGCGUGGAAACAGGCGGGAC  C   G   AU                    -5'
             |                                                                                                                                                            |
           145                                                                                                                                                            1
query
```

Result of IntaRNA --pred=M --mode=E --predMulti=B --noSeed:

```
target
            24                                                                                                                                                    61
             |                                                                                                                                                    |
5'-AUG...UCUG     GU     A   GGGUUCGAUUCCCCU                                                                                                                       ACCA-3'
             GAGGC  UCUGG CAG                                                                                                                               CACCUCC
             :||||  |||:| |||                                                                                                                               ||||:||
             UUCCG  AGAUC GUC                                                                                                                               GUGGGGG
3'-AUA...AUUG                GGGGCCGAUUGGCAAAAAUCCGCCGUUAUCGCUUAACGAGUAGUAACCGUUAAUACCUAAACGAAGUAAUUGCAGCUAGCGGACUGCUCAACGGACGCGUGGAAACAGGCGGGACAGCUCUGGUCAU       -5'
             |                                                                                                                                                    |
           147                                                                                                                                                    1
query
```

Two other structures that we analysed are shown in figure 10. They were taken from the *Cross-Catalytic Replication of an RNA Ligase Ribozyme* article by Kim and Joyce [2017]. These structures have intramolecular structure in both sequences and are therefore ideal for testing the new IntaRNA recursion.

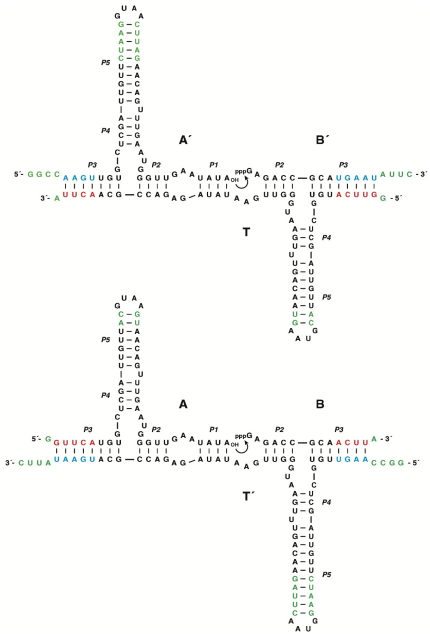


Figure 10: Sequence and Secondary Structure of
the Ribozymes and Substrates used to Carry Out
Cross-Catalytic Replication.
Taken from Kim and Joyce [2017].

As the resulting structure has intramolecular structure in both sequences but independant from each other, we choose mode "–predMulti=X", for ES values in query or target, to recreate the structure shown in Figure 10. Using this mode IntaRNA predicts the exact structure given in the article, see page 21. Furthermore, we tried to run IntaRNA using only parts of the entire sequences to check whether we can still predict the intramolecular structures. The results are shown on page 22. The results for the first part of the target sequence are very satisfying, as the intramolecular structure in the target sequence is nearly correct. The second part of the target sequence is likely too short to predict the intramolecular structure in the query.
The other structure shown in figure 10 is the mirrored structure of the first one. Using IntaRNA this structure is also perfectly predicted. The different results for the mirrored structure are shown on page 23.

```
>A
GGUUCAUGUGCUCGAUUGUUACGUAAGUAACAGUUUGAAUGGGUUGAAUAUA
>B
GAGACCGCAACUUA
>T = A + B
GGUUCAUGUGCUCGAUUGUUACGUAAGUAACAGUUUGAAUGGGUUGAAUAUAGAGACCGCAACUUA
>A'
GGCCAAGUUGUGCUCGAUUGUUCUAAGGUAACUUAGAACAGUUUGAAUGGGUUGAAUAUA
>B'
GAGACCGCAUGAAUAUUC
>T' = A' + B'
GGCCAAGUUGUGCUCGAUUGUUCUAAGGUAACUUAGAACAGUUUGAAUGGGUUGAAUAUAGAGACCGCAUGAAUAUUC
>T = A + B

A' + B' / T structure:
A' + B'
```


```
T

Prediction of IntaRNA using --pred=M --mode=E --predMulti=X --noSeed
A' + B'
```


```
T
```

```
A' + B' / T structure:
A' + B'
            5                                                                                                       74
            |                                                                                                       |
      5'-GGCC      GCUCGAUUGUUCUAAGGUAACUUAGAACAGUUUGAAUG    GAA     GA                                            AUUC-3'
            AAGUUGU                                   GGUU   UAUA   GACC                                     GCAUGAAU
            |||||||                                   ||||   ||||   ||||                                     ||||||||
            UUCAACG                                   CCAG   AUAU   UUGG                                     UGUACUUG
         3'-A                                             AG     AAG    GUAAGUUUGACAAUGAAUGCAUUGUUAGCUCG         G-5'
            |                                                                                               |
           65                                                                                               2
T

Prediction of IntaRNA using -q=T -t=A' --pred=M --mode=E --predMulti=X --noSeed
A'
            5                                              59
            |                                              |
      5'-GGCC      GCUCGAUUGUUCUAAGGUAACUUAGAACAGUUUGAAUG    GAA   A-3'
            AAGUUGU                                   GGUU   UAU
            ||||||:                                   |||:   |||
            UUCAACG                                   CCAG   AUA
         3'-A                                             AG   UAAG...UGG-5'
            |                                              |
           65                                              50
T

Prediction of IntaRNA using -q=T -t=B' --pred=M --mode=E --predMulti=X --noSeed
B'
            1           14
            |           |
         5'-    AC      AUUC-3'
            GAG  CGCAUGAAU
            |||  |:|||||||:
            CUC  GUGUACUUG
3'-AUU...UUAG            G-5'
            |           |
           13           2
T
```

```
A + B / T' structure:
A + B
          5'-G          GCUCGAUUGUUACGUAAGUAACAGUUUGAAUG      GAA        GA                                                         A-3'
              GUUCAUGU                                    GGUU    UAUA    GACC                                      GCAACUU
              ||||||||                                    ||:|    |:||    ||||                                      |||||||
              UAAGUACG                                    CCAG    AUAU    UUGG                                      UGUUGAA
       3'-CUUA                                                AG       AAG     GUAAGUUUGACAAGAUUCAAUGGAAUCUUGUUAGCUCG      CCGG-5'
T'

Prediction of IntaRNA using --pred=M --mode=E --predMulti=X --noSeed
A + B
              2                                                                                                            65
              |                                                                                                            |
          5'-G          GCUCGAUUGUUACGUAAGUAACAGUUUGAAUG      GAA      GA                                                   A-3'
              GUUCAUGU                                    GGUU    UAUA    GACC                                      GCAACUU
              :||||||:                                    |||:     ||||    :|||                                      :||||||
              UAAGUACG                                    CCAG    AUAU    UUGG                                      UGUUGAA
       3'-CUUA                                                AG       AAG     GUAAGUUUGACAAGAUUCAAUGGAAUCUUGUUAGCUCG      CCGG-5'
              |                                                                                                            |
             74                                                                                                            5
T'

Prediction of IntaRNA using -q=T' -t=A --pred=M --mode=E --predMulti=X --noSeed
A
              2                                                     52
              |                                                     |
          5'-G          GCUCGAUUGUUACGUAAGUAACAGUUUGAAUG      GAA      -3'
              GUUCAUGU                                    GGUU    UAUA
              :||||||:                                    |||:     ||||
              UAAGUACG                                    CCAG    AUAU
       3'-CUUA                                                AG       AAGU...CGG-5'
              |                                                     |
             74                                                     57
T'

Prediction of IntaRNA using -q=T' -t=B --pred=M --mode=E --predMulti=X --noSeed
B
              1          13
              |          |
          5'-    AC      A-3'
             GAG  CGCAACUU
             |||  |:||||||
             CUC  GUGUUGAA
3'-CUU...UUAG         CCGG-5'
              |          |
             15          5
T'
```

Bibliography

Laura DiChiacchio, Michael F. Sloma, and David H. Mathews. Accessfold: predicting rna–rna interactions with consideration for competing self-structure. *Bioinformatics*, 32(7):1033, 2016. doi: 10.1093/bioinformatics/btv682. URL +http://dx.doi.org/10.1093/bioinformatics/btv682.

Dong-Eun Kim and Gerald F Joyce. Cross-catalytic replication of an rna ligase ribozyme. *Chemistry and Biology*, 11:1505–12, 2017. doi: 10.1016/j.chembiol.2004.08.021. URL +http://dx.doi.org/10.1016/j.chembiol.2004.08.021.

Martin Mann. RNA Bioinformatics 2016. https://ilias.uni-freiburg.de/goto.php?target=crs_565591&client_id=unifreiburg, 2016.

J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers*, 29(6-7):1105–1119, 1990. ISSN 1097-0282. doi: 10.1002/bip.360290621. URL http://dx.doi.org/10.1002/bip.360290621.