Journal of Bioinformatics and Computational Biology
© Imperial College Press

IntaRNAhelix - Composing RNA-RNA interactions from stable inter-molecular helices boosts bacterial sRNA target prediction

Rick Gelhausen[1], Sebastian Will[2], Ivo L. Hofacker[2], Rolf Backofen[1,3] and Martin Raden[1]

[1] *Bioinformatics Group, University of Freiburg, Georges-Koehler-Allee 106, 79110 Freiburg, Germany,*
gelhausr@informatik.uni-freiburg.de

[2] *Institute for Theoretical Chemistry, University of Vienna, Waehringer Strasse 17, 1090 Wien, Austria*

[3] *Signalling Research Centres BIOSS and CIBSS, University of Freiburg, Schaenzlestr. 18, 79104 Freiburg, Germany*

Received (Day Month Year)
Revised (Day Month Year)
Accepted (Day Month Year)

Efficient computational tools for the identification of putative target RNAs regulated by prokaryotic sRNAs rely on thermodynamic models of RNA secondary structures. While they typically predict RNA–RNA interaction complexes accurately, they yield many highly-ranked false positives in target screens. One obvious source of this low specificity appears to be the disability of current secondary-structure-based models to reflect steric constraints, which nevertheless govern the kinetic formation of RNA–RNA interactions. For example, often—even thermodynamically favorable—extensions of short initial kissing hairpin interactions are kinetically prohibited, since this would require unwinding of intra-molecular helices as well as sterically impossible bending of the interaction helix. Another source is the consideration of instable and thus unlikely subinteractions that enable better scoring of longer interactions. In consequence, the efficient prediction methods that do not consider such effects show a high false positive rate.
To increase the prediction accuracy we devise IntaRNAhelix, a dynamic programming algorithm that length-restricts the runs of consecutive inter-molecular base pairs (perfect canonical stackings), which we hypothesize to implicitely model the steric and kinetic effects. The novel method is implemented by extending the state-of-the-art tool IntaRNA. Our comprehensive bacterial sRNA target prediction benchmark demonstrates significant improvements of the prediction accuracy and enables more than 40-times faster computations. These results indicate—supporting our hypothesis—that stable helix composition increases the accuracy of interaction prediction models compared to the current state-of-the-art approach.

Keywords: RNA–RNA Interaction Prediction; Steric Constraints; Constrained Helix Length; Canonical Helix; Seed.

1. INTRODUCTION

Small RNAs (sRNAs) are central regulators in prokaryotic cells[26]. For instance, they can trigger mRNA decay[15] or modulate translation[12] via direct inter-molecular base pairing. Different mechanisms are known (detailed e.g. in[22]) like the blocking of the ribosomal binding site causing translation inhibition or the (de-)stabilization of mRNAs by covering (or providing) binding sites of RNAases. The sRNA–RNA interactions typically contain a small nearly perfect subinteraction of about 7 base pairs (known as *seed region*)[14] and have been shown to be located at accessible regions that are mainly unpaired[25]. Thus, beside general RNA–RNA interaction prediction approaches (reviewed e.g. in[32]), dedicated prediction tools like IntaRNA[8] or RNApredator[10] have been developed and applied[16]. Recently, fast heuristics for genome-wide screens have been implemented, e.g. sTarPicker[34], RIblast[11] or RIsearch2[1].

For elucidating the regulatory network of sRNAs, target prediction is applied[2] to guide experimental validation. While the essential bioinformatics machinery for this task is available[3], computational methods still predict a high number of false positive targets. The latter can be reduced when individual target predictions of homologous sequences[18] are combined in comparative approaches like CopraRNA[31]. Unfortunately, this technique is only applicable for the identification of evolutionary conserved targets. Another option is to incorporate experimental structure probing data to amend the RNAs' accessibility information[20]. Integrating probing data, which can be obtained from high-throughput experiments[9], can significantly alleviate the problem of inaccurate accessibility prediction. However, it does not touch—and is even orthogonal to—the here discussed issues of target prediction.

In this work, we study means to efficiently improve sRNA target prediction by restricting the admissible interaction patterns. This is hypothesized to incorporate steric and kinetic aspects going beyond the thermodynamic secondary structure-based models. Specifically, our method is motivated by two observations.

Observation 1: Interacting sites of sRNAs are either not enclosed by any base pairing (exterior) or located within loop regions. For loop regions, the formation of (long) inter-molecular helices (i.e. the entangling of the RNA molecules) requires the 'unwinding' of intra-molecular helices, which imposes additional constraints on the substantial steric rearrangements (rotating large parts of the molecules through space) while the interaction grows. Consequently, the formation of long inter-molecular duplexes seems to be prohibited, even if it would be expected in the currently used thermodynamic models due to high hybridization stability and sufficient accessibility. This well-known phenomenon has been studied in the context of other loop-initiated RNA–RNA interactions[7,13].

Observation 2: Thermodynamic interaction prediction is based on summing inter-molecular loop terms. Consequently, more loops, i.e. base pairs, enable lower energies, especially if the overall interaction pattern is formed by stable parts (with no or few bulges) that are linked by very instable subinteractions, which are composed of a sequence of bulges and interior loops.

Here, we test whether interaction prediction can be improved by composing interactions from helices rather than individual loops, which should amend for Observation 2. Furthermore, we explicitly constrain the maximal length of considered inter-molecular helices, which—as we conjecture—indirectly considers steric and kinetic constraints. While preserving tractability, limiting this length ensures that long helices must be interrupted by interior loops—which is thought to relax the 'winding tension' following Observation 1.

We provide efficient dynamic programming algorithms both for exact as well as heuristic helix-based interaction prediction, incorporating the new helix-length constraint (in addition to the well-established seed constraint of previous approaches). The approach is incorporated into IntaRNA[19], a state-of-the-art RNA–RNA interaction prediction tool[30], and available as individual tool IntaRNAhelix. Finally, we assess the effect of the helix-length constraints on a large prokaryotic sRNA target prediction data set extending[33]. In this benchmark, the helix length limitation reduces the overall runtime and, supporting our conjecture, improves the prediction quality.

2. METHODS

In the following, we will first present the recursions used by the current state-of-the-art prediction approaches like RNAup[21] or IntaRNA[19]. Subsequently, we introduce the new recursions for helix-length restricted prediction. First, all recursions are given for exhaustive/optimal interaction prediction, followed with a discussion how they can be turned into efficient heuristic variants. To ease readability, we provide graphical recursion depictions and provide respective formulas in the Appendix.

2.1. *Accessibility-based Interaction Prediction*

Given two RNAs S_1, S_2 of length n, m, resp., we want to find the interaction sites $i..k \in [1, n]$ of S_1 and $j..l \in [1, m]$ of S_2 that minimize the interaction energy $E(i, j, k, l)$. That is, we are interested in the most stable interaction of an sRNA with a given putative target. This interaction energy can then be used for target ranking and the selection of the most promising candidates.

The interaction sites are considered free of intra-molecular base pairs and can only form inter-molecular base pairs. Two positions of the RNAs can form a base pair if the respective nucleotides are complementary (i.e. AU, GC, or GU). We consider only sites where the boundaries are forming two inter-molecular base pairs $(i, j), (k, l)$. No two inter-molecular base pairs $(x, y), (x', y') \in [1, n] \times [1, m]$ are allowed to be crossing, i.e. it holds $x \leq x' \leftrightarrow y \leq y'$, nor allowed to share a position within the same RNA. Following the Nearest Neighbor energy model[27], the hybridization or duplex formation energy of a site is thus given by the sum of the loop energies[29] defined by consecutive base pairs. Here, we distinguish between stacked (directly neighbored) base pairs, scored by E_S terms, and neighbored base pairs that enclose unpaired positions, evaluated by E_{IL} terms. The hybridization energy also contains a general energy penalty term E_{init} that, to some extent, reflects the probability of interaction initiation. The optimal (minimal) hybridization energy among

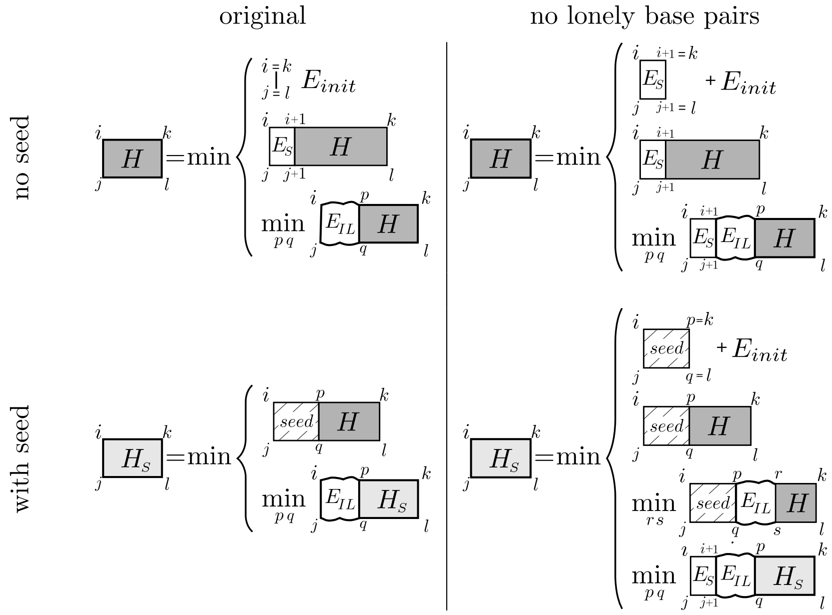

Fig. 1. Sketch of (left) the original IntaRNA recursions to compute the optimal interaction energy without and with seed constraints and (right) the new variants that exclude lonely base pairs.

all possible interactions of the sites is given by $H(i, j, k, l)$. The energy penalty ED needed to break all intra-molecular base pairs within the individual sites is used to incorporate the sites' accessibility for interaction formation. The overall energy is thus given by

$$E(i, j, k, l) = H(i, j, k, l) + ED_1(i..k) + ED_2(j..l). \tag{1}$$

All energy terms presented in the following are given in $kcal/mol$ unit and are computed using the Vienna RNA package[17] version 2.4.12. For simplicity, we exclude dangling-end and helix-end contributions within Eq. 1. For formalisms, we refer to the detailed introduction provided in[24].

ED terms can be efficiently computed via dynamic programming[5]. This leaves the computation of the optimal interaction energy H, also accessible via dynamic programming[21]. Figure 1 (top-left) visualizes the central recursion that either scores an initial base pair (E_{init}) or extends a shorter optimal interaction with a stacked base pair (E_S) or an interior loop contribution (E_{IL}). All individual energy contributions E_{init}, E_S and E_{IL} are $+\infty$ if the respective boundary indices are non-complementary, i.e. can not form a base pair. Note, interior loop sizes (p-i and q-j) are typically restricted to a fixed maximal length $w \ll n, m$, which results in a runtime complexity of $O(n^2m^2)$. The base pairs of an optimal interaction with energy $H(i, j, k, l)$ can be obtained via traceback if needed. A heuristic variant of this recursion available in IntaRNA with $O(nm)$ runtime was introduced in[8], which also

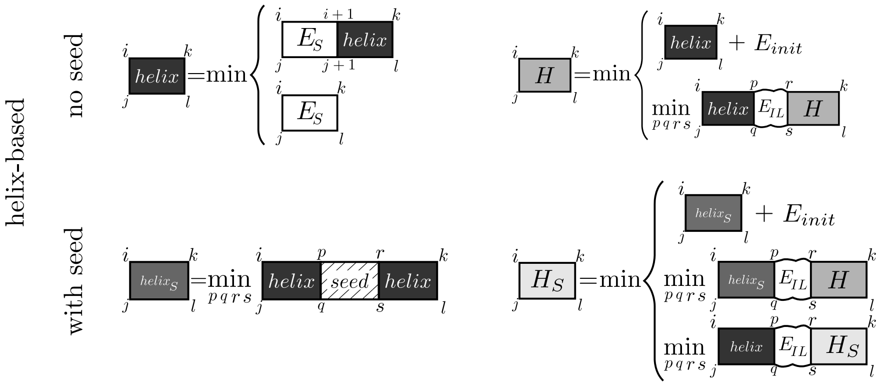

Fig. 2. Recursion depictions (top-left) to compute canonical helix energies *helix* using energy terms E_S for stacked base pairs used for (top-right) to optimize energy H for a given interaction site using the energy terms E_{IL} for interior loops. (bottom-left) seed-containing helix $helix_S$ computation to enable (bottom-right) seed-constraint helix-based predictions via H_S.

incorporates seed constraints (see Fig. 1 (bottom-left) for a respective non-heuristic variant).

Within this work, we also investigate a new prediction strategy that considers only interactions *without lonely base pairs* (see Fig. 1 (right) for recursion depictions). A base pair is lonely if it is not stacked on either side. In contrast to the unconstraint variant, the seed-incorporating recursion of H_S has to explicitely consider interaction initialization via seeds (case 1) and the interior-loop-spaced seed extension of an interaction (case 3). The recursions do not alter time nor space complexity of the prediction.

2.2. *Helix-length Restricted Prediction*

In order to restrict the length of inter-molecular helices to a predefined constant $c_B \geq 2$, referred to as *helix length*, we decompose the prediction process into two steps: (a) the energy pre-computation of possible helices composed of at most c_B base pairs, and (b) their assembly in order to find the optimal interaction energy for a given site.

For simplicity, we first consider canonical helices, i.e. perfect helices composed of stacked base pairs only. Adaptions to non-canonical helices containing small bulges and interior loops are discussed in a subsequent section. Figure 2 shows the recursion to compute the energy of canonical helices with the left-/right-most inter-molecular base pair $(i, j) < (k, l)$, resp., stored in $helix(i, j, k, l)$. The length constraint c_B is ensured for canonical helices by setting all entries to $+\infty$ if the helix is too long, i.e. $\max(k - i, l - j) \geq c_B$. Note, for non-canonical helices, $helix(i, j, k, l)$ will contain the optimal energy of any helix fulfilling the relaxed constraints.

Given this, the optimal hybridization energy $H(i, j, k, l)$ for the given interaction sites $i..k$ and $j..l$ can be computed via the recursion depicted in Fig. 2. That is, we either consider a full helix (if possible for the given boundaries) or compose an interaction via the addition of a new helix (on the left) to extend a smaller optimal interaction. The composition inserts an interior loop between the helix and the next interaction to ensure that no two helices are combined into a longer one. Thus, the interior loop has to span at least one unpaired position, i.e. $(p-r)+(q-s)>2$, and is constrained in length as for the recursions discussed before. Since both helix length as well as interior loop length are constrained by respective constants c_B and w, the overall runtime complexity is still $O(n^2m^2)$.

2.3. *Enforcing Seeds*

As already discussed, seed-constraints are a central tool to reduce false positive sRNA target predictions[28,6]. Within IntaRNA, possible seed interactions and respective energies are efficiently computed via dynamic programming analogously to the presented helix energy pre-processing; please refer to[8] for details. In the following, the optimal energy for the seed with left-/right-most base pairs $(i,j),(k,l)$, resp., are stored in $seed(i,j,k,l)$.

In order to ensure that a reported interaction contains a seed region, we follow the approach presented in[8]. Therein, a second dynamic programming table H_S is computed based on H that provides the optimal energy for a site given that the considered interaction contains a seed region. The optimal energy of a site with seed is then given by

$$E(i,j,k,l) = H_S(i,j,k,l) + ED_1(i..k) + ED_2(j..l) \tag{2}$$

replacing Eq. 1.

Since seeds are valid parts of helices, which are the building blocks for our introduced H computation, we use a second auxiliary matrix $helix_S$ that provides the optimal helix energy given that the helix contains a seed. If the region contains no valid seed or this would lead to too many base pairs, the energy is set to $+\infty$. Figure 2 depicts the recursion in order to fill $helix_S$ based on the already introduced $helix$ information that is combined with the $seed$ energy. To this end, all possible locations of a seed combined with flanking helices are evaluated. Due to the independence of the seed and helix constraints, it is possible to allow unpaired bases in the seed, even when not allowing unpaired bases in the helix constraints and vice versa.

Given this, the optimal hybridization energy H_S for a given site containing a seed and only helices with at most c_B base pairs, can be computed using a recursion as depicted in Fig. 2. That is, either (i) the site can be filled with a single helix containing a seed (plus accounting for interaction initiation), or (ii) a helix-length-constrained interaction site is extended with a seed-containing helix, or (iii) we extend an interaction that contains already a seed with a helix that is not constrained to contain a seed.

2.4. *Enforcing a Minimal Helix Stability*

Given our focus on helices, we can easily enforce additional constraints on the helices that are considered for interaction composition. As the first step, we introduce a minimal stability notion via an upper hybridization energy bound $E^{helix}_{\max}$ for individual helices. Since energy is inversely related to stability, our approach will produce interaction patterns of stable subhelices connected by interior loop regions.

The energy threshold can be easily incorporated into the presented recursions by extending the computation of H and H_S from Fig. 2 with side conditions. That is, entries from *helix* or $helix_S$ are only considered, if the respective *overall* energy value is below the given threshold $E^{helix}_{\max}$, i.e. it holds for a helix with boundary base pairs (i, j) and (k, l)

$$helix(i, j, k, l) + ED_1(i..k) + ED_2(j..l) < E^{helix}_{\max}. \tag{3}$$

2.5. *Consideration of Non-canonical Helices*

So far, we only considered canonical helices for the computation of *helix*. While this models the most stable helices that can be formed, minor variance of this ideal, i.e. allowing for bulges or interior loops spanning only single or very few unpaired bases, will still resemble a stable helix. But considering stable helices only (using $E^{helix}_{\max}$) would likely exclude such helices if the canonical subhelices are too short. Thus, we next discuss how the hybridization energy *helix* for helices including minor bulges of at most c_U unpaired bases can be computed. We consider an interior loop as minor if $c_U \leq 2$.

To this end, we introduce the auxiliary matrix $helix_U(i, j, k, l, B)$ that provides the optimal helix hybridization energy for the given site boundaries and the number of base pairs B while allowing minor bulges of size c_U. The optimal helix hybridization energy $helix(i, j, k, l)$ is thus given by $helix_U(i, j, k, l, c_B)$. Note, enforcing the helices to be stable (via $E^{helix}_{\max}$) will without further constraints exclude helices composed of bulges only.

In addition to the altered *helix* computation, we also have to ensure that the helices assembled within the H and H_S computation are spaced by interior loops exceeding c_U. That is, it holds for Fig. 2 that $(r - p) + (s - q) > c_U + 2$. Note that setting $c_U = 0$ will provide the same results as if using canonical helices only.

2.6. *Heuristic Helix-length Restricted Prediction*

Due to the high time and space complexity of the exact approach, we implemented heuristic variants of the recursions following the ideas from[8] introduced for IntaRNA. That is, instead of considering all interaction ranges for a given left-most base pair (i, j), only the optimal right boundary (k, l) together with the respective hybridization energy is stored in H and H_S. Thus, for a given left-most base pair (i, j), the recursions from above are not confined to a specific right k, l bound but use the right end of the respective optimal recursion case. Please refer to[8] for further details. This heuristic reduces the space and time complexity to $O(nm)$, provides almost the same prediction quality[30], and makes the approach feasible for the needed large-scale target screens also discussed in the Result section.

Here, we apply this strategy not only to H and H_S but also to the *helix* and $helix_S$ matrices. That is, we only memorize the best helix energy (and right boundary) for each left-most helix base pair (i, j). Note, both matrices have to be computed using small auxiliary matrices that replace the respective recursions. Note further, the computation of H and H_S becomes more simple and faster, since we do not consider different helix lengths (via p and q) but only use the right-most base pair of the best helix with left-most base pair (i, j). This approach is incorporated into IntaRNA version 3, instantiated as standalone tool IntaRNAhelix, available from and documented at

https://github.com/BackofenLab/IntaRNA

3. RESULTS

3.1. *Data Set for sRNA Target Prediction Benchmark*

To assess the impact of the helix-based interaction prediction, we created an sRNA target prediction benchmark extending the ideas and data from[8,28,33]. The whole benchmark data set including respective scripts is available at

https://github.com/BackofenLab/IntaRNA-benchmark

The benchmark consists of a large set of bacterial sRNA queries and potential target sequences. We restrict our analysis to sRNA regulation based on the blocking of the ribosomal binding site (see[22] for a discussion). Thus, the targets are genomic sub-regions around the start codon of the respective mRNA including 200 nucleotides upstream and 100 nucleotides downstream. The dataset comprises 4,319 target regions from the *Escherichia coli* genome (GenBank[4] accession number NC_000913) and 4,552 target regions from the *Salmonella typhimurium* genome (NC_003197). The query data set consists of 30 sRNAs, 15 from each organism, that have been shown experimentally to act as post-transcriptional regulators by base-pairing to at least one of the targets. For these, we extracted 149 such verified sRNA-target pairs from the literature.

Within the benchmark, we test how well these verified pairs can be separated from all possible sRNA-target pairs. To this end, we identify for each sRNA the top-100 targets with lowest overall interaction energy, i.e. putative targets with most stable interactions. Subsequently, we count how many of the verified sRNA-target pairs can be recovered within all top-100 predictions.

As reference and "gold standard" for the evaluation of our helix-length restricted approach, we use the prediction performance of IntaRNA version 3.alpha (compiled with ViennaRNA v2.4.12) using default values (i.e. heuristic predictions in-

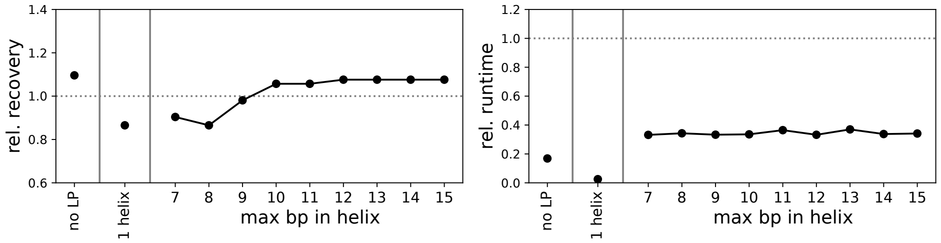

Fig. 3. Effect of different maximal canonical helix lengths c_B in terms of the recovery of verified sRNA-target pairs among the top ranked predictions relative to the original IntaRNA results (dotted line).

cluding seeds with 7 base pairs) on the introduced benchmark data set. In the following, we refer to this version with ”original” and provide qualitative and computational performance measures in relation to it.

The original IntaRNA version recovers 52 verified targets among the top-100 targets predicted for each sRNA. Thus, division by this value provides the reported relative recovery rates for the tested interaction prediction variants. Analogously, relative overall runtimes for the target screens of the benchmark are stated to abstract from the hardware used.

3.2. *The Impact of Lonely Base Pairs*

A first step towards helix-focused RNA-RNA interaction prediction is the exclusion of lonely base pairs, since they represent instable subinteractions composed of two bulges or interior loops. Using this heuristic, we improve the recovery of verified targets within our benchmark set by 10% while reducing the overall runtime by a factor of about 6 to 17%.

While this strategy provides already strong improvements, it does not enable further constraints on the interaction pattern. Thus, we tested the helix-based prediction approach evaluated next.

3.3. *Helix-length Constraints Enable Faster Predictions and Improve Prediction Quality*

Effect of Maximal Helix Length

Following our hypothesis, we tested the effect of limiting the length of intermolecular helices to incorporated steric constraint that can hinder long helix formations. We constrained the maximal number of base pairs within helices c_B with values from 7 to 15. Results are depicted in Fig. 3 for canonical helices without bulges. For low c_B values (< 10), we observe a reduced prediction accuracy compared to the original recursions. In contrast, when relaxing the bound on the maximal helix lengths to values of 10 or higher, improved recovery rates are found. This shows the

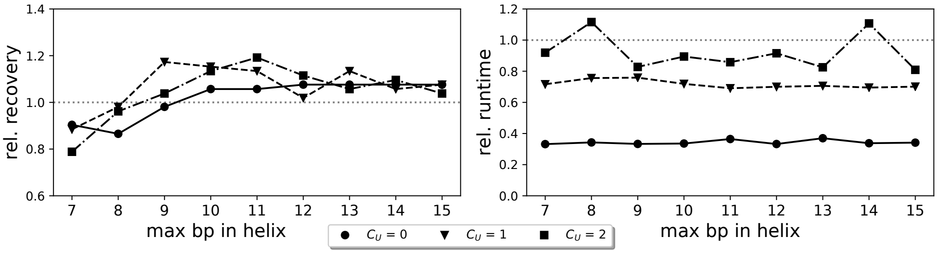

Fig. 4. Effect of maximal helix-interior-loop size c_U for different maximal helix lengths c_B.

potential of a prediction strategy based on helix composition compared to standard loop-composing approaches.

In addition to the improved prediction accuracy, the constrained version is about 3-times faster compared to the original version, while maintaining the same memory consumption.

Considering only a Single Helix

When investigating the helix-length-restricted results, we found that most predicted interactions are composed of a single helix. Thus, we investigated whether helix composition is needed at all or if it is sufficient to extend a seed into a single helix. While this strategy is very fast (about 40-times faster compared to the original approach), it offers only a reduced prediction accuracy of 87%. We therefore conclude that composing individual helices to larger interactions is central to provide the improved results reported above.

Effect of Non-Canonical Helices

When relaxing the helix definition to non-canonical helices that are allowed to contain minor bulges or interior loops with up to c_U unpaired bases, the overall prediction performance can be further improved. Figure 4 compares the results for maximal helix-interior-loop size $c_U \in \{0, 1, 2\}$ for maximal helix lengths $c_B \in \{10, 11, 12\}$ that performed well for canonical helices. Here, best results are observed for $c_U = 2$ and a maximal helix length $c_B = 11$.

Generally, relaxing the helix definition slightly improves prediction accuracy but enables only reduced ($c_U = 1$) or even no runtime improvement ($c_U = 2$). Thus, we focus in the following on canonical helices only.

Effect of Minimal Helix Stability

Next, we investigated whether the restriction to stable helices can further improve the prediction quality. To this end, we tested different overall helix energy thresholds $E^{helix}_{\max}$ from -10 to 0 for the well performing maximal helix lengths $c_B \in \{10, 11, 12\}$.

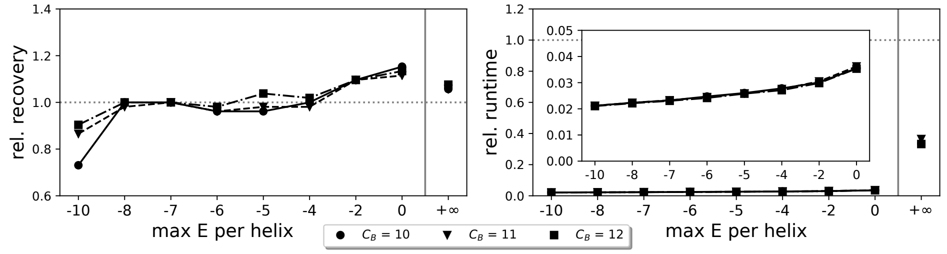

Fig. 5. Effect of maximal helix energies $E^{helix}_{\max}$ ($+\infty$ = unconstraint) on the prediction performance for different maximal helix lengths c_B. The inset provides details for low runtime values.

Figure 5 summarizes the results. While $E^{helix}_{\max} = 0$ eventually defines the least constraint, it shows best performance with a recovery improvement of about 15% for $c_B = 10$. This is about three-times higher compared to $c_B = 10$ results without stability constraint (see above).

The most significant impact of the helix-stability constraint was observed when comparing runtime with the original approach. We observe speedups from 28 up to 47 for $E^{helix}_{\max} = 0$ and -10, resp., since runtime is reduced to about 3.5 to 2.1%, resp., which results in the substantial decrease of putative helices that are considered for interaction prediction.

Note, all tested $E^{helix}_{\max}$ values provide improved prediction performance such that we conclude that combining stable helices results in the best prediction strategy.

4. DISCUSSION & CONCLUSION

Helix-based predictions of RNA-RNA interactions can be done with the same time and space complexity as known for loop-based methods. In fact, the helix-length constraint already reduces the search space such that we observe on average a 3-times faster target prediction for our benchmark data set. When further constraining the search to stable helices only (via enforcing upper energy bounds $E^{helix}_{\max}$), target screens become more than 40 times faster. The reduced runtime results from the following: compared to the current state-of-the-art recursion from Fig. 1, the helix-composing approach from Fig. 2 faces the same search space for the interior loop sizes but appends full helices instead of individual base pairs. This becomes even more striking for the heuristic variant, which does not considered all possible helix lengths but only the optimal helix for the left-most base pair (i, j).

Furthermore, we observe enriched target prediction accuracy measured in terms of increased recovery rates of verified sRNA-target pairs known from the literature. Maximal helix lengths c_B of 10-12 base pairs show the best prediction quality, while shorter drastically reduced the recovery rate. This is mainly achieved by disregarding low-energy subinteractions composed of many bulge and interior loops (putative false positives) rather than altering the interaction details of the verified sRNA-target pairs. Results can be further improved when only stable helices with an overall energy below a given threshold $E^{helix}_{\max}$ are considered for prediction. For helices of a maximal length of 10-12 base pairs, the loose upper energy bound of 0 provides the best target prediction performance. Notably, the consideration of a relaxed helix definition allowing for small bulge or interior loops improves the results compared to perfect canonical helices but has much higher runtime requirements.

Our observations support our hypothesis that long inter-molecular helices are less likely due to steric and kinetic constraint of the interaction formation process. That is, we think the 'helix-composed interaction model'—where short stable helices are interrupted by single flexible interior loops—a more realistic model compared to unconstrained loop-composing predictions. The latter can be significantly improved when lonely base pairs are excluded, but it is still inferior to stable-helix composition. Within the helix-based model, interaction formation can be seen either as a parallel process where each helix is formed independently or a serial interaction extension via subsequent helix formations that are paused to overcome the energetically unfavorable spacing interior loops.

One way to further improve the model would be to confine helix-length constrained predictions to regions mainly unpaired in loop regions while applying the unconstrained approach for exterior unpaired regions. This can be efficiently distinguished during ED computation from the underlying partition functions[21,5]. Another planned direction is to apply further constraints on the helices considered within the prediction. For instance, we will incorporate and test the effect of excluding lonely base pairs in the unconstrained helix model to reduce its computational cost and further improve prediction accuracy. Furthermore, we will investigate the correlation of helix base pair number c_B and optimal upper energy bounds $E^{helix}_{\max}$, since they are most likely linked by the average stacking energy or similar terms. An integration of IntaRNAhelix into the Freiburg RNA tools webserver[23] is under construction.

Acknowledgements

This work was supported by the German Research Foundation (DFG) grant [*BA2168/16-1*], under Germany's Excellence Strategy (CIBSS - EXC-2189 - Project ID 390939984) and the Austrian Science Fund (FWF) [*I 2874-N28*].

Appendix - Formal Recursions

$$E_S(i,j,k,l) = \begin{cases} \begin{pmatrix} \text{energy contribution for} \\ \text{stacking base pairs } (i,j),(k,l) \end{pmatrix} & : \text{if } k-i=1 \text{ and } l-j=1, \\ +\infty & : \text{otherwise} \end{cases} \tag{4}$$

$$E_{IL}(i,j,k,l) = \begin{cases} \begin{pmatrix} \text{energy contribution for} \\ \text{stack or interior loop } (i,j),(k,l) \end{pmatrix} & : \text{if } i<k \text{ and } j<l, \\ +\infty & : \text{otherwise} \end{cases} \tag{5}$$

$$\mathit{helix}(i,j,k,l) = \begin{cases} \min \begin{cases} E_S(i,j,i+1,j+1) + \mathit{helix}(i+1,j+1,k,l) \\ E_S(i,j,k,l) \end{cases} & : \text{if canonical helix} \\ \\ \mathit{helix}_U(i,j,k,l;c_U,c_B) & : \text{otherwise} \end{cases} \tag{6}$$

$$\mathit{helix}_U(i,j,k,l;U,B) = \min \begin{cases} \min\limits_{\substack{p,q \text{ with} \\ \min(k-p,l-q)\geq B-1 \\ (p-i)+(q-j)\leq U+2}} \begin{pmatrix} E_{IL}(i,j,p,q) \\ +\mathit{helix}_U(p,q,k,l;U,B-1) \end{pmatrix} & : \text{if } \ B>2, \\ E_{IL}(i,j,k,l) \qquad : \text{if } \ ((k-i)+(l-j)\leq U+2) & \wedge \text{ B=2}, \\ +\infty & : \text{otherwise.} \end{cases} \tag{7}$$

$$\mathit{helix}_S(i,j,k,l) = \min_{\substack{p,r \\ q,s}} \big(\mathit{helix}(i,j,p,q) + \mathit{seed}(p,q,r,s) + \mathit{helix}(r,s,k,l)\big) \tag{8}$$

$$H(i,j,k,l) = \min \begin{cases} \mathit{helix}(i,j,k,l) + E_{init} \\ \min\limits_{\substack{p,r,q,s \text{ with} \\ (r-p)+(s-q)>c_U+2 \\ \mathit{helix}(i,j,k,l)\leq E^{\mathit{helix}}_{\max}}} \Big(\mathit{helix}(i,j,p,q) + E_{IL}(p,q,r,s) + H(r,s,k,l)\Big) \end{cases} \tag{9}$$

$$H_S(i,j,k,l) = \min \begin{cases} \mathit{helix}_S(i,j,k,l) + E_{init} \\ \\ \min\limits_{\substack{p,r,q,s \text{ with} \\ (r-p)+(s-q)>c_U+2 \\ \mathit{helix}_U(i,j,k,l)\leq E^{\mathit{helix}}_{\max}}} \Big(\mathit{helix}_S(i,j,p,q) + E_{IL}(p,q,r,s) + H(r,s,k,l)\Big) \\ \min\limits_{\substack{p,r,q,s \text{ with} \\ (r-p)+(s-q)>c_U+2 \\ \mathit{helix}(i,j,k,l)\leq E^{\mathit{helix}}_{\max}}} \Big(\mathit{helix}(i,j,p,q) + E_{IL}(p,q,r,s) + H_S(r,s,k,l)\Big) \end{cases} \tag{10}$$

References

1. Alkan F, Wenzel A, Palasca O, Kerpedjiev P, Rudebeck A, Stadler PF, Hofacker IL, Gorodkin J, RIsearch2: suffix array-based large-scale prediction of RNARNA interactions and siRNA off-targets, *Nucleic Acids Research* **45**(8):e60, 2017. doi: 10.1093/nar/gkw1325.
2. Backofen R, Amman F, Costa F, Findeiss S, Richter AS, Stadler PF, Bioinformatics of prokaryotic RNAs, *RNA Biol* **11**(5), 2014.
3. Backofen R, Engelhardt J, Erxleben A, Fallmann J, Grüning B, Ohler U, Rajewsky N, Stadler PF, RNA-bioinformatics: Tools, services and databases for the analysis of RNA-based regulation, *J Biotechnol* **261**:76–84, 2017. doi: 10.1016/j.jbiotec.2017.05.019.

4. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL, GenBank, *Nucleic Acids Res* **36**(Database issue):25–30, 2008.
5. Bernhart SH, Mückstein U, Hofacker IL, RNA accessibility in cubic time, *Algorithms for Molecular Biology* **6**(1):3, 2011. doi:10.1186/1748-7188-6-3.
6. Bouvier M, Sharma CM, Mika F, Nierhaus KH, Vogel J, Small RNA binding to 5' mRNA coding region inhibits translational initiation, *Mol Cell* **32**(6):827–837, 2008.
7. Brunel C, Marquet R, Romby P, Ehresmann C, RNA loop–loop interactions as dynamic functional motifs, *Biochimie* **84**(9):925 – 944, 2002. doi:10.1016/S0300-9084(02)01401-3.
8. Busch A, Richter AS, Backofen R, IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions, *Bioinformatics* **24**(24):2849–56, 2008. doi:10.1093/bioinformatics/btn544.
9. Choudhary K, Deng F, Aviran S, Comparative and integrative analysis of RNA structural profiling data: current practices and emerging questions, *Quantitative Biology* **5**(1):3–24, 2017.
10. Eggenhofer F, Tafer H, Stadler PF, Hofacker IL, RNApredator: fast accessibility-based prediction of sRNA targets, *Nucleic Acids Res* **39**(Web Server issue):W149–54, 2011. doi:10.1093/nar/gkr467.
11. Fukunaga T, Hamada M, RIblast: an ultrafast RNARNA interaction prediction system based on a seed-and-extension approach, *Bioinformatics* **33**(17):2666–2674, 2017. doi:10.1093/bioinformatics/btx287.
12. Hoe CH, Raabe CA, Rozhdestvensky TS, Tang TH, Bacterial sRNAs: Regulation in stress, *International Journal of Medical Microbiology* **303**(5):217 – 229, 2013. doi: 10.1016/j.ijmm.2013.04.002.
13. Kolb FA, Malmgren C, Westhof E, Ehresmann C, Ehresmann B, Wagner EG, Romby P, An unusual structure formed by antisense-target RNA binding involves an extended kissing complex with a four-way junction and a side-by-side helical alignment., *RNA* **6**(3):311–324, 2000.
14. Künne T, Swarts DC, Brouns SJ, Planting the seed: target recognition of short guide RNAs, *Trends in Microbiology* **22**(2):74 – 83, 2014. doi:10.1016/j.tim.2013.12.003.
15. Lalaouna D, Simoneau-Roy M, Lafontaine D, Mass E, Regulatory RNAs and target mRNA decay in prokaryotes, *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms* **1829**(6):742 – 747, 2013. doi:10.1016/j.bbagrm.2013.02.013.
16. Li W, Ying X, Lu Q, Chen L, Predicting sRNAs and their targets in bacteria, *Genomics, Proteomics & Bioinformatics* **10**(5):276 – 284, 2012. doi: 10.1016/j.gpb.2012.09.004.
17. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, Hofacker IL, ViennaRNA Package 2.0, *Algorithms for Molecular Biology* **6**(1):26, 2011. doi:10.1186/1748-7188-6-26.
18. Lott SC, Schäfer RA, Mann M, Backofen R, Hess WR, Voss B, Georg J, GLASSgo - automated and reliable detection of sRNA homologs from a single input sequences, *Frontiers in Genetics* **9**:124, 2018. doi:10.3389/fgene.2018.00124.
19. Mann M, Wright PR, Backofen R, IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions, *Nucleic Acids Res* **45**(W1):W435–W439, 2017. doi: 10.1093/nar/gkx279.
20. Miladi M, Montaseri S, Backofen R, Raden M, Integration of accessibility data from structure probing into RNA-RNA interaction prediction, *Bioinformatics* , 2019. doi: 10.1093/bioinformatics/bty1029, (epub ahead of print).
21. Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF, Hofacker IL, Thermodynamics of RNA–RNA binding, *Bioinformatics* **22**(10):1177–1182, 2006. doi:

10.1093/bioinformatics/btl024.

22. Nitzan M, Rehani R, Margalit H, Integration of bacterial small RNAs in regulatory networks, *Annual Review of Biophysics* **46**(1):131–148, 2017. doi:10.1146/annurev-biophys-070816-034058.
23. Raden M, Ali SM, Alkhnbashi OS, Busch A, Costa F, Davis JA, Eggenhofer F, Gelhausen R, Georg J, Heyne S, Hiller M, Kundu K, Kleinkauf R, Lott SC, Mohamed MM, Mattheis A, Miladi M, Richter AS, Will S, Wolff J, Wright PR, Backofen R, Freiburg RNA tools: a central online resource for RNA-focused research and teaching, *Nucleic Acids Research* **46**(W1):W25–W29, 2018. doi:10.1093/nar/gky329.
24. Raden M, Mohamed MM, Ali SM, Backofen R, Interactive implementations of thermodynamics-based RNA structure and RNA-RNA interaction prediction approaches for example-driven teaching, *PLOS Comput Biol* **14**(8):e1006341, 2018. doi: 10.1371/journal.pcbi.1006341.
25. Richter AS, Backofen R, Accessibility and conservation: General features of bacterial small RNA-mRNA interactions?, *RNA Biol* **9**(7):954–65, 2012. doi:10.4161/rna.20294.
26. Storz G, Vogel J, Wassarman K, Regulation by small RNAs in bacteria: Expanding frontiers, *Molecular Cell* **43**(6):880 – 891, 2011. doi:10.1016/j.molcel.2011.08.022.
27. Tinoco Jr I, Borer P, Dengler B, Levin M, Uhlenbeck O, Crothers D, Bralla J, Improved estimation of secondary structure in ribonucleic acids, *Nature New Biology* **246**(150):40–41, 1973. doi:10.1038/newbio246040a0.
28. Tjaden B, Goodwin SS, Opdyke JA, Guillier M, Fu DX, Gottesman S, Storz G, Target prediction for small, noncoding RNAs in bacteria, *Nucleic Acids Res* **34**(9):2791–2802, 2006.
29. Turner DH, Mathews DH, NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure, *Nucleic Acids Res* **38**(Database issue):D280–2, 2010. doi:10.1093/nar/gkp892.
30. Umu SU, Gardner PP, A comprehensive benchmark of RNA-RNA interaction prediction tools for all domains of life, *Bioinformatics* **33**(7):988–996, 2017.
31. Wright PR, Georg J, Mann M, Sorescu DA, Richter AS, Lott S, Kleinkauf R, Hess WR, Backofen R, CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains, *Nucleic Acids Res* **42**(Web Server issue):W119–23, 2014. doi:10.1093/nar/gku359.
32. Wright PR, Mann M, Backofen R, Structure and interaction prediction in prokaryotic RNA biology, *Microbiol Spectrum* **6**(2), 2018. doi:10.1128/microbiolspec.RWR-0001-2017.
33. Wright PR, Richter AS, Papenfort K, Mann M, Vogel J, Hess WR, Backofen R, Georg J, Comparative genomics boosts target prediction for bacterial small RNAs, *Proceedings of the National Academy of Sciences* **110**(37):E3487–96, 2013. doi: 10.1073/pnas.1303248110.
34. Ying X, Cao Y, Wu J, Liu Q, Cha L, Li W, sTarPicker: A method for efficient prediction of bacterial sRNA targets based on a two-step model for hybridization, *PLOS ONE* **6**(7):1–12, 2011. doi:10.1371/journal.pone.0022705.