Structure and interaction prediction in prokaryotic RNA biology

Patrick R. Wright[1,*], Martin Mann[1,*] and Rolf Backofen[1,2,*]

[1]Bioinformatics Group, University of Freiburg, Freiburg, Germany

[2]Center for Biological Signaling Studies (BIOSS), University of Freiburg, Germany

[*]contributed equally

Abstract

Many years of research in RNA biology have soundly established the importance of RNA based regulation far beyond most early traditional presumptions. Importantly, the advances in "wet" laboratory techniques have produced unprecedented amounts of data that require efficient and precise computational analysis schemes and algorithms. Hence, many *in silico* methods that attempt topological and functional classification of novel putative RNA based regulators are available. In this review we technically outline thermodynamics-based standard RNA secondary structure and RNA-RNA interaction prediction approaches that have proven valuable to the RNA research community in the past and present. For these, we highlight their usability with a special focus on prokaryotic organisms and also briefly mention recent advances in whole genome interactomics and how this may influence the field of predictive RNA research.

1 Introduction

For over a decade, prokaryotic and eukaryotic RNA biology exploration has unveiled the multi-faceted and central contribution of RNA based control in all domains of life. RNA interactions are at the core of many regulative processes and have hence been heavily studied by wet-lab and biocomputational researchers alike. Within this review, we focus in biocomputational methods and outline the technical details of standard algorithms for RNA secondary structure and RNA-RNA interaction prediction. Furthermore, we highlight their application in the context of prokaryotic RNA biology.

Similar to DNA, RNA molecules can undergo stable base pairing when stretches of complementary nucleotides are present and form a so called duplex. The generalized model assumes that adenine (`A`) can form base pairs with uracil (`U`) while guanine (`G`) can pair with cytosine

(C) or U. *In vivo*, further interactions are possible [1, 2, 3], which are, however, not considered for the methods presented in this review. Base pairs in RNA or DNA can be established due to complementary nucleotides forming hydrogen bonds [4]. Two types of base pairing are conceivable for RNA molecules. Firstly, base pairing can occur within a single RNA molecule. These intramolecular interactions give rise to RNA secondary structures that are often important for an RNA molecule's function or regulation and are thus central to cellular physiology.

The second type of base pairing is referred to as intermolecular interaction or RNA-RNA interaction. These interactions occur between RNAs that are present as individual molecules and play a key role in processes that employ RNA molecules as regulators of other RNA molecules. Examples are prokaryotic *trans* acting small RNAs (sRNAs) or eukaryotic microRNAs (miRNAs). Both sRNAs and miRNAs are posttranscriptional regulators that often bind target RNAs and thereby modulate the target's function in a positive or negative manner [5, 6, 7]. Laboratory based identification and verification of RNA-RNA interactions is a cumbersome task. In accordance with this, and taking the pivotal role of RNA-RNA interactions in the regulatory network of cellular systems into account, *in silico* prediction of such interactions has been intensely studied and several predictive approaches are available.

2 Intramolecular RNA structure prediction

RNA molecules are usually chain-like polymers of nucleotides that differ in their base composition and length. They are therefore typically represented by their base sequence in $5' \rightarrow 3'$ direction, where 5' denotes the five prime phosphate group and 3' denotes the three prime hydroxyl group of the first and last nucleotide, respectively. Thus, an RNA molecule of length n is encoded by its sequence $\mathcal{R} \in \Sigma^n$ where $\Sigma = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{U}\}$ encodes the possible bases. A structure $\mathcal{P}$ for a given RNA $\mathcal{R}$ can be encoded by its set of base pairs $\mathcal{P} = \{\ (i,j) \mid 1 \leq i < j \leq n\ \}$.

In the basic RNA secondary structure model, each base can form only a single base pairing within the molecule. A valid secondary structure $\mathcal{P}$ fulfills the following criteria: i) unique base pairing ($\forall (i,j) \neq (k,l) \in \mathcal{P} : i \notin \{k,l\} \wedge j \notin \{k,l\}$), ii) sequence complementarity ($\forall (i,j) \in \mathcal{P} : \{\mathcal{R}_i, \mathcal{R}_j\} \in \{\{\mathtt{A},\mathtt{U}\},\{\mathtt{C},\mathtt{G}\},\{\mathtt{G},\mathtt{U}\}\}$), and iii) minimal base pair span s_l ($\forall (i,j) \in \mathcal{P} : i + s_l < j$). Therefore, an empty secondary structure $\mathcal{P}^{\mathrm{oc}} = \emptyset$ corresponds to the open chain without intramolecular base pairings. The minimal base pair span s_l, also called loop size, incorporates steric bending constraints into the structure model and is usually set to $s_l = 3$. If the base pairs

within a structure are nested, i.e. it holds $\nexists_{(i,j)\neq(k,l)\in\mathcal{P}} : i < k < j < l$, then we call $\mathcal{P}$ a *nested structure*. Otherwise, it is called a *crossing or pseudoknot structure*. Nested models enable a unique decomposition of $\mathcal{P}$ into secondary structure elements, which facilitates efficient RNA structure and interaction prediction methods. Therefore, we will focus on nested models only in the following.

A nested secondary structure $\mathcal{P}$ can be uniquely decomposed into structural elements. These elements are called *loops*. Each loop is defined by an enclosing base pair $(i, j) \in \mathcal{P}$. The loop type is determined by the directly enclosed base pairs $(k, l) \in \mathcal{P}$ with $i < k < l < j$. A base pair (k, l) is directly enclosed by (i, j), if there is no other (k, l)-enclosing base pair $(i', j') \in \mathcal{P}$ that is enclosed by (i, j) too. We distinguish the following loop types that are depicted in Fig. 1.

hairpin loop:	no enclosed base pair
stacking:	adjacent enclosed base pair, i.e. $i + 1 = k$ and $j - 1 = l$
bulge loop:	only one side adjacent, i.e. $(i{+}1{=}k \wedge j{-}1{>}l)$ or $(i{+}1{<}k \wedge j{-}1{=}l)$
interior loop:	non-stacked enclosed base pair, i.e. $(i + 1 < k \wedge j - 1 < l)$
multi-loop:	more than one directly enclosed base pair

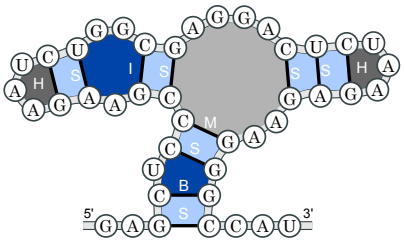

Figure 1: Loop decomposition of a nested RNA structure into **H**airpin loops (dark gray), **M**ulti-loops (light gray), **S**tackings (light blue), **B**ulges and **I**nterior loops (dark blue). Initials of the loop types are placed in white next to the enclosing base pair (black).

2.1 Individual structure prediction

One of the first and most fundamental algorithms for the prediction of nested RNA secondary structures was introduced by Ruth Nussinov and co-workers [8]. It applies dynamic programming techniques to efficiently identify a stable structure $\mathcal{P}$ for an RNA molecule with sequence $\mathcal{R}$ by maximizing the number of base pairs $|\mathcal{P}|$. To this end, the Nussinov algorithm recursively computes the maximum number of base pairs for each subsequence $\mathcal{R}_i..\mathcal{R}_j$ and stores this information in the dynamic programming matrix $N_{i,j}$. It can be filled by employing Eq. 1 (depicted in Fig. 2), which is a variant of the original recursions formulated in [8] to enable the relation to other approaches in the following. The maximum number of base pairs is found in the matrix entry $N_{1,n}$. The corresponding structure can be identified via traceback.

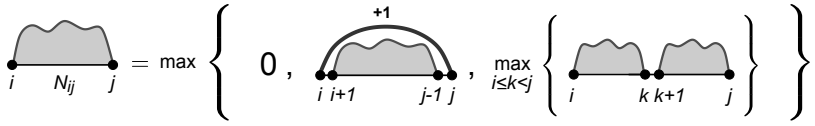

Figure 2: Graphical depiction of the Nussinov-like recursion from Eq. 1.

$$N_{i,j} = \max \begin{cases} 0 & : \text{if } (i + s_l) \geq j \\ N_{i+1,j-1} + 1 & : \text{if } R_i, R_j \text{ can form base pair} \\ \max_{i \leq k < j} \{N_{i,k} + N_{k+1,j}\} & : \text{decomposition} \end{cases} \tag{1}$$

While efficient and in theory applicable to the folding of RNAs, this simple optimization scheme shows poor prediction accuracy for several reasons. Firstly, it does not account for differences in base pairing strengths. In general, base pair complementarity is at the core of every RNA interaction prediction but a distinction needs to be made between stronger and weaker base pairs. While a `G-C` pair incorporates three hydrogen bonds, `G-U` and `A-U` base pairs only form two hydrogen bonds and are less stable when compared to a `G-C` pair. Furthermore, the stability influence of loop sizes, base pair stackings, loop context, multi-loop formations, etc. is not considered. The stacking of base pairs, for instance, is central to RNA helix stability [9]. Nevertheless, the algorithmic idea was transferred into more sophisticated optimization schemes that are discussed in the following.

Current RNA secondary structure prediction algorithms are usually energy minimization methods. Typically, they use the aforementioned loop decomposition of a structure (see Fig. 1) in combination with loop-specific energy contributions. This enables an incorporation of empirically determined loop type- and context-specific contributions [10, 11]. Thus, this so called *nearest-neighbor model* [12, 13] considers the directly neighboring bases and base pairs for each interaction. The overall energy $E(\mathcal{P})$ of a nested structure $\mathcal{P}$ can therefore be computed by the summation over energies for each loop enclosed by $(i, j) \in \mathcal{P}$ (Eq. 2).

$$E(\mathcal{P}) = \sum_{(i,j) \in \mathcal{P}} \begin{cases} E^{\text{loop}}(i, j, i, j) & : \text{if hairpin loop} \\ E^{\text{loop}}(i, j, k, l) & : \text{if stacking, bulge, or interior loop} \\ E^{\text{loop}}_{\text{multi}}(i, j, h, u) & : \text{if multi-loop} \end{cases} \tag{2}$$

where E^{loop} provides the loop's energy contribution for stackings, bulges, interior and hairpin

loops and $E^{\text{loop}}_{\text{multi}}$ gives the energy for a multi-loop element. To reduce complexity, multi-loop energies are typically estimated by $E^{\text{loop}}_{\text{multi}}(i,j,h,u) = E^m_c + hE^m_h + uE^m_u$. This uses empirically identified constants E^m_c, E^m_h and E^m_u, with E^m_c penalizing the multi-loop closure by (i,j), E^m_h accounting for the $h \geq 2$ directly enclosed base pairs (i.e. branching helices) and E^m_u weighting the u directly enclosed unpaired bases. For instance, the multi-loop in Fig. 1 results in $h = 2$ and $u = 6$. To respect the outer context of non-enclosed loops, so called *dangling end contributions* have to be added as well, which is neglected in Eq. 2 and in the following presentations for simplicity. Throughout the last decades, several parameter sets for the nearest neighbor model have been derived from experimental data [10, 14, 15].

Given such an energy decomposition, a dynamic programming scheme to compute the *minimal free energy (mfe) structure* $\mathcal{P}^{mfe}$ for an RNA $\mathcal{R}$ was introduced by Michael Zuker and Patrick Stiegler [16]. It uses three matrices to store results for distinct subproblems: $V_{i,j}$ provides the mfe for all possible structures that can be formed by the subsequence $\mathcal{R}_i..\mathcal{R}_j$ under the assumption that $\mathcal{R}_i$ and $\mathcal{R}_j$ form a base pair; $W^M_{i,j}$ handles multi-loop decompositions, where $\mathcal{R}_i..\mathcal{R}_j$ is enclosed in the multi-loop; and W_i encodes the mfe for the prefix $\mathcal{R}_1..\mathcal{R}_i$. Given the following recursions (Eq. 3-6), the mfe of the whole RNA can be found in W_n. Note, the multi-loop decomposition in the $W^M_{i,j}$ recursion is not unique, which makes it unsuitable for suboptimal structure prediction as addressed in [17]. When restricting the maximally allowed interior loop size in Eq. 5, this algorithm runs in $O(n^3)$ time and $O(n^2)$ space. Note, since this transfer of a base pair maximizing recursion (Nussinov algorithm) to energy minimization using the nearest neighbor model (Zuker algorithm) is generic, we will restrict where appropriate the algorithm presentations to a Nussinov-like form.

$$W_0 = 0 \tag{3}$$

$$W_j = \min \begin{cases} W_{j-1} & : j \text{ is unpaired} \\ \min\limits_{1 \le k < j} \{W_{k-1} + V_{k,j}\} & : \text{base pair } (k,j) \end{cases} \tag{4}$$

$$V_{i,j} = \min \begin{cases} \infty & : \mathcal{R}_i, \mathcal{R}_j \text{ cannot pair or } (i + s_l) \ge j \\ E^{\text{loop}}(i,j,i,j) & : (i,j) \text{ closes hairpin} \\ \min\limits_{i<k<l<j} \left\{E^{\text{loop}}(i,j,k,l) + V_{k,l}\right\} & : (i,j),(k,l) \text{ stacking, bulge, } \ldots \\ \min\limits_{i<k<j} \left\{E_c^m + W_{i+1,k}^M + W_{k+1,j-1}^M\right\} & : (i,j) \text{ closes multi-loop} \end{cases} \tag{5}$$

$$W_{i,j}^M = \min \begin{cases} \infty & : (i + s_l) \ge j \\ E_u^m + W_{i+1,j}^M & : i \text{ unpaired} \\ E_u^m + W_{i,j-1}^M & : j \text{ unpaired} \\ E_h^m + V_{i,j} & : (i,j) \text{ directly enclosed} \\ \min\limits_{i<k<j} \left\{W_{i,k-1}^M + V_{k,j} + E_h^m\right\} & : \text{decompose and } (k,j) \text{ directly enclosed} \end{cases} \tag{6}$$

The advanced versions of the Zuker algorithm are implemented in the standard folding programs UNAFOLD [18] (former MFOLD [19]) and RNAFOLD [20, 21]. Both implementations are being successfully used by the research community and show good prediction accuracy for RNAs such as Spot42 [22] and FnrS [23] to name just two.

The time complexity of the Nussinov and Zuker algorithm is $O(n^3)$, at least when restricting the interior loop size in Zuker. This is, however, still high for long RNA sequences as for instance large mRNAs, long non-coding RNAs and viral RNAs. For that reason, attempts have been made to reduce the overall time complexity on average to $O(n^2)$. When revisiting the recursion in Eq. 1, the split in the final decomposition case is the one causing the high complexity. Many of these splits will not lead to the optimal solution. This observation sparked the idea of sparsification techniques, first introduced by Ydo Wexler and colleagues [24], which is discussed in the supplementary material.

Finally, more and more tools now allow for the inclusion of structure probing data [25, 26], which can guide a more sophisticated structure prediction process based on experimentally established constraints [27, 28]. Examples are RNASTRUCTURE [29], RNASC [30] and RNAFOLD [31, 32].

2.2 Comparative structure prediction

In order to increase the prediction quality, it is often useful to take not only one but a set of evolutionarily related molecules into account. That is, one wants to compute a common structure for a set of sequences, which requires an alignment that takes both sequence and structure features into account. Paul R. Gardner and Robert Giegerich classified such approaches according to the applied plan [33]: Plan A) "first align then fold", Plan B) "simultaneously align and fold", Plan C) "first fold then align".

The class of methods, which is referred to as Plan A, is basically an extension of the individual structure prediction to alignments. That is, first all sequences are aligned based on their sequence similarity. This multiple sequence alignment can be successively folded into a consensus structure that is compatible with all sequences. Common approaches are RNAALIFOLD [34], PFOLD [35] or PETFOLD [36]. Such approaches are as efficient as individual structure prediction and work well for data sets with high sequence similarity. RNAALIFOLD for instance has been used to determine the conservation of the CyaR sRNA [37]. If sequence identity within the data drops below 60%, Plan A approaches have been shown to fail [38]. Here, the other two plans are more promising, since they utilize the observation that an RNA's structure is often more strongly conserved than its sequence on an evolutionary scale [38].

The first practical implementations of Plan C were RNAFORESTER [39] and MARNA [40]. They generate a multiple structure alignment for a set of given input structures. The latter are either known or stem from individual structure prediction. Thus, Plan C approaches depend on the accuracy of the input structures. Since there is only a limited number of known RNA structures, the overall alignment quality is often flawed by the individual structure predictions used instead.

Thus, the current state-of-the-art approaches are applying Plan B that was first introduced by David Sankoff [41]. Here, sequences are simultaneously folded *and* aligned leading to an algorithm with $O(n^6)$ time and $O(n^4)$ space complexity. The key idea of Sankoff is to simultaneously find two equivalent structures $\mathcal{P}^1$ and $\mathcal{P}^2$ for the given sequences $\mathcal{R}^1$ and $\mathcal{R}^2$, respectively, in combination with a compatible sequence alignment of $\mathcal{R}^1$ and $\mathcal{R}^2$. That is, we have to optimize the combination: $E(\mathcal{P}^1) + E(\mathcal{P}^2) + S$, where $S_{i,j,\,i',j'}$ provides the alignment score for the respective subsequences. The two structures are *equivalent* if they are of the same size ($|\mathcal{P}^1| = |\mathcal{P}^2|$) and show the same nesting. The sequence alignment is *compatible* with both structures if equivalent base pairs $(i,j) \in \mathcal{P}^1$ and $(i',j') \in \mathcal{P}^2$ are aligned to each other.

In the following, we present a reduced Nussinov-like version of the algorithm using a base pair maximization scheme. Here, the Nussinov-matrix N (Eq. 1) for one sequence is extended to a four-dimensional matrix F (Eq. 7) that encodes the optimal Sankoff-like score for two aligned subsequences $\mathcal{R}^1_{i..j}$ and $\mathcal{R}^2_{i'..j'}$.

$$F_{i,j,\,i',j'} = \max \begin{cases} S_{i,j,\,i',j'} & : \text{alignment of no structure} \\ F_{i+1,j-1,\,i'+1,j'-1} + 2 & : \text{if } (\mathcal{R}^1_i, \mathcal{R}^1_j) \text{ and } (\mathcal{R}^2_{i'}, \mathcal{R}^2_{j'}) \text{ are compl.,} \\ \quad +S_{i,i,\,i',i'} + S_{j,j,\,j',j'} & \quad \text{alignment of base pairs } (i,j),(i',j') \\ \max\limits_{k,k'} \left\{ F_{i,k,\,i',k'} + F_{k+1,k'+1,\,j,j'} \right\} & : \text{decomposition} \end{cases} \tag{7}$$

where the sequence alignment contributions are computed by

$$S_{i,j,\,i',j'} = \max \begin{cases} S_{i+1,j,\,i'+1,j'} & : \text{align } i \text{ and } i' \text{ if } \mathcal{R}^1_i,\, \mathcal{R}^2_{i'} \text{ match} \\ S_{i+1,j,\,i'+1,j'} - s_m & : \text{align } i \text{ and } i' \text{ if } \mathcal{R}^1_i,\, \mathcal{R}^2_{i'} \text{ mismatch} \\ S_{i+1,j,\,i',j'} - s_g & : \text{align } \mathcal{R}^1_i \text{ with gap} \\ S_{i,j,\,i'+1,j'} - s_g & : \text{align } \mathcal{R}^2_{i'} \text{ with gap} \end{cases} \tag{8}$$

using the penalties s_m and s_g (both ≥ 0) for mismatch and gap alignments, respectively. The introduction of base pairs into both structures at once ensures their equivalence and the direct inclusion of respective alignment scores ensures the compatibility of the alignment with the structures. As for the Nussinov-like recursion in Eq. 1, the Sankoff-like recursion from Eq. 7 can be extended to the nearest-neighbor energy model just like the Zuker algorithm.

To reduce the computational complexity, several simplifications have been introduced. One class of variants uses sequence-based heuristics to restrict the search space. Programs of this class are for instance FOLDALIGN [42, 43], DYNALIGN [44], and STEMLOC [45]. Another class of approaches, e.g. implemented in PMCOMP [46], LOCARNA [47], and FOLDALIGNM [48], do not restrict the alignment search space but use a simplified energy model based on base pair probabilities to reduce the considered structural search space and such computational complexity and runtime. Here, instead of directly considering loop energies, as done by Sankoff, energy terms are indirectly encoded within base pair probabilities that are efficiently precomputed using the algorithm by John S. McCaskill [49]. Both classes of simplifications are combined by RAF [50], which fuses heuristics based on subsequence alignment quality with the simplified

energy model of PMcomp, an approach first introduced in Stemloc.

3 RNA-RNA interaction prediction

3.1 Key components in RNA-RNA interactions

Recalling the commonly known double helical structure of DNA, one may naively assume that base pair complementarity is the sole component needed to form a stable interaction between two RNAs. However, further factors influence RNA-RNA interactions. One of these factors are the previously discussed intramolecular structures (see Sec. 2) that can be formed by each individual copy of the interacting RNAs before they *meet*. Given the hypothetical RNA sequence, where 5' denotes the five prime phosphate group and 3' denotes the three prime hydroxyl group, 5'-`GGG-GGGGGGGCCCCCCCCCC`-3'; if no intramolecular base pairs are assumed, one may be tempted to see a perfect duplex forming between two individual RNAs of this type (Fig. 3a). However, each individual RNA of this type is also capable of forming a hairpin structure incorporating a strong `G-C` stem enclosing a small loop. Hence, a more realistic assumption is that only the unpaired hairpin loop regions will interact and the duplex is not as long as naively expected (Fig. 3b). This theoretical scenario is also reflected by *cis*-antisense RNAs *in vivo* [51]. Neglecting or considering the component of intramolecular structures splits the prediction approaches that will be discussed in this review into two major groups. They will be reviewed in the sections 3.2 and 3.3 respectively.

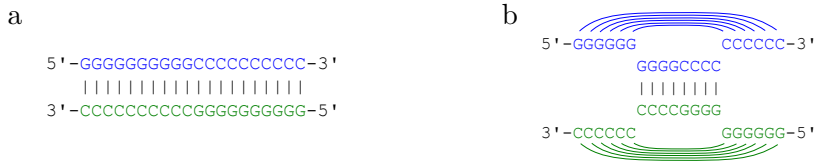

Figure 3: Potential interactions for two identical RNA molecules (blue and green) as predicted by the a) RNAhybrid webserver [52] and b) IntaRNA webserver [53] (intramolecular base pairs subsequently added, which form a kissing hairpin interaction). Inter- and intramolecular base pairs are indicated by vertical pipe symbols and arches respectively.

3.2 RNA-RNA interaction prediction not considering intramolecular base pairing

As previously mentioned (see Sec. 3.1), approaches to predict RNA-RNA interactions can be split into two major groups. The first group, that will be discussed here, neglects the impact

of intramolecular base pairing within the interaction partners. The algorithmic solutions are either purely sequence- or structure-based approaches. While the sequence-based models solely search for stretches of extensive base pair complementarity, a physical energy model is employed by structure-based approaches.

To find base pair complementarity, the basic local alignment search tool (BLAST) algorithm [54] is appropriate. Yet, next to the canonical Watson-Crick base pairs `G-C` and `A-U`, also the non-Watson-Crick base pair `G-U` can form within RNA-RNA interactions and must thus be considered. GUUGle [55] is an approach that incorporates the `G-U` wobble base pairs, and was mainly developed as filtering scheme to reduce the search space for more complex algorithms. The advantage of these sequence-based approaches is that they immediately inherit a method to calculate p-values from local alignment approaches [56].

TargetRNA [57], which was developed to predict the targets of bacterial sRNAs, approaches the problem from different angles. Two scoring schemes for an interaction of RNAs are proposed. Firstly, TargetRNA allows for a purely sequence-based solution using a variant of the Smith-Waterman alignment algorithm [58]. Therein, it searches for base pair complementarity rather than sequence similarity. Furthermore, loops within the interaction are penalized while `G-C` and `A-U` base pairs are favored over `G-U` base pairs. The second solution in TargetRNA uses an energy model. Here, the free energy of an RNA duplex is considered when scoring an interaction. A lower energy represents a more stable interaction. This second type of scoring resembles the energy model used for individual RNA structure prediction (see Sec. 2). RNAhybrid [59], which was developed prior to TargetRNA, also uses minimal free energy scoring primarily in order to predict eukaryotic miRNA targets. Yet, it has also been frequently applied in the prediction of prokaryotic sRNA target interactions. [37, 60, 61]. The scoring in these approaches strongly depends on the energies of stacked back-to-back base pairs, interior loops and bulges. The stacking energies, which were originally derived for intramolecular structures, are available from experimental testing (see Sec. 2). Energies for small interior loops and bulges are also available from experimental data.

Both RNAhybrid and TargetRNA restrict the length for long interior loops (i.e. considered values for p, q in Eq. 9/10), as these structures increase the computational complexity. The rationale behind this is that long interior loops do not represent structures that are favorable, and thus may be disregarded in the interaction prediction due to their limited *real world* relevance. RNAplex [62] on the other hand deviates from RNAhybrid's and TargetRNA's

type of treatment for long interior loops by using an affine function for scoring long interior loops and bulges within the interaction, while its energy model is similar to those applied in the previously mentioned approaches. Furthermore, RNAplex can avoid disproportionately long duplex predictions, which often occur in RNAhybrid, by imposing a penalty for every nucleotide in the interaction. Thereby, RNAplex provides a more realistic estimation of the potential *in vivo* duplex, which is especially helpful when predicting the targets of prokaryotic sRNAs. The constraint on duplex lengths can be regarded as a step towards consideration of intramolecular structures without specifically addressing them. A pre-filtering algorithm for RNA-RNA interaction prediction on the genomic scale employing a simplified Turner energy model is RIsearch [63].

The above mentioned energy-based approaches predict RNA-RNA interactions by minimizing the free energy of the resulting duplex. Specifically, this can be solved in polynomial time using dynamic programming. In principle, one can consider all possible interaction sites $i..k$ on the first sequence $\mathcal{R}^1$ together with all possible interaction sites $j..l$ on the second RNA $\mathcal{R}^2$, and store the minimal duplex energy in a matrix $D_{j..l}^{i..k}$. Note, that we number the first sequence in $5' \to 3'$, and the second sequence in reverse orientation ($3' \to 5'$) since we consider only sense-antisense interactions. To guarantee that these interaction sites are actually covered by a duplex, we have to enforce that i, j, k, l are occupied by intermolecular base pairs. By the non-crossing condition for the intermolecular base pairs, this is only possible if i is paired to k and j to l. Then, we can simply calculate all possible interactions by the following recursion:

$$D_{j..l}^{i..k} = \min \begin{cases} +\infty & \text{if } \mathcal{R}_i^1, \mathcal{R}_j^2 \text{ can } \textit{not} \text{ pair} \\ E^{\text{Init}}(i,j) & \text{if } i = k \text{ and } j = l \\ \min\limits_{p,q} \left\{ E^{\text{loop}}(i,j,p,q) + D_{q..l}^{p..k} \right\} & \text{if } i < k \text{ and } j < l \end{cases} \tag{9}$$

Here, $E^{\text{Init}}(i,j)$ is the free energy for the first intermolecular base pair in a duplex. Following [64], this comprises the dangling end contributions for the initial base pair (i,j) and the intermolecular initiation free energy (usually 4.10 kcal/mol). $E^{\text{loop}}(i,j,p,q)$ is the energy contribution for the loop enclosed by the base pairs (i,j) and (p,q), where $i < p$ and $j < q$. This can either be a stacking ($i = p-1$ and $j = q-1$), bulge ($i = p-1$ or $j = q-1$, but not both) or interior loop ($i < p-1$ and $j < q-1$) (see Sec. 2). When restricting the maximal size of interior loops (typically not more than 30 bases), this gives rise to an algorithm with

complexity $O(n^4)$ time and space. The best duplex interaction of $\mathcal{R}^1$ and $\mathcal{R}^2$ can be found by starting a traceback from the minimal $D_{j..l}^{i..k}$ entry, where (i, j) and (k, l) define the duplex's left- and rightmost intermolecular base pairs. The recursion is depicted in Fig. 4a.

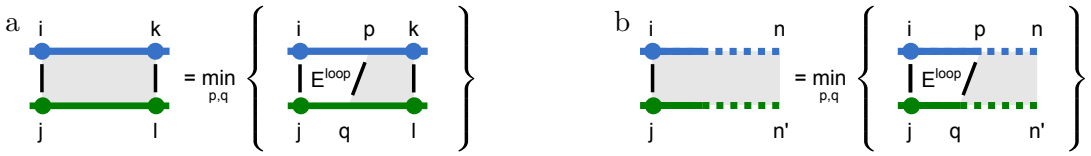

Figure 4: Recursion depiction of interaction prediction via a) $D_{j..l}^{i..k}$ and b) C_j^i.

When neglecting the intramolecular base pairing, then information about the interacting sites as used in $D_{j..l}^{i..k}$ is actually not required to determine the best duplex. Instead, one can use a matrix C_j^i that provides the minimal duplex energies for the suffixes $\mathcal{R}_i^1 \dots \mathcal{R}_n^1$ and $\mathcal{R}_j^2 \dots \mathcal{R}_{n'}^2$ (see Fig.4b), where n and n' are the respective sequence lengths. The simplified recursion is then given by

$$C_j^i = \min \begin{cases} +\infty & \text{if } \mathcal{R}_i^1,\ \mathcal{R}_j^2 \text{ can } \textit{not} \text{ pair} \\ E^{\text{Init}}(i, j) & \\ \min\limits_{p,q} \left\{ E^{\text{loop}}(i, j, p, q) + C_q^p \right\} & \end{cases} \tag{10}$$

The actual interaction sites can be subsequently generated via traceback starting at the minimal C_j^i entry, where (i, j) marks the leftmost intermolecular base pair. Assuming $n' \in O(n)$, this simplified recursion has an $O(n^2)$ time and space complexity and according variants are used in RNAhybrid, RNAduplex [21] and TargetRNA.

Evidently, the approaches described in the current section can be split into purely sequence-based methods and methods that incorporate an energy model for RNA-RNA interaction prediction. While the methods that solely rely on sequence complementarity are a useful initial approximation for potential interactions of RNA molecules, their disregard of the energetic properties of RNA duplexes represents a major pitfall. Hence, the minimum-free-energy-based algorithms have several advantages. Firstly, they have the same runtime complexity as the sequence-based methods and their use of experimentally derived energy values for specific structure elements allows a more realistic approximation of RNA duplexes. Furthermore, the incorporation of a thermodynamic context allows the consideration of temperature, which is a key factor when defining the structural states of molecules such as RNA. The temperature is a dynamic pa-

rameter and can thus be adjusted to accommodate for the investigated system's temperature properties. The latter may be especially helpful for predictions on systems that are not assumed to be at 37°C (human body temperature), like the native environment of organisms such as the thermophile archaeons belonging to the *Sulfolobus* genus [65].

The inclusion of thermodynamic parameters into the prediction of RNA-RNA interactions represents an advance in this field of research when compared to the purely sequence based methods. Yet, by not directly addressing the influence of intramolecular structures within the interaction partners, a major *in vivo* factor of RNA-RNA interactions is neglected and can cause duplex predictions that may never occur. Thus, a common artifact of the methods that disregard intramolecular structures can be unproportionately long duplex predictions [62, 66] as these are generally favored by the energy model. To counter such effects, recent versions of RNAplex [67] (and its webserver RNApredator [68]), TargetRNA2 [69] and RIsearch2 [70] incorporate the accessibility of the interacting sites for intermolecular base pairing prediction. This concept is more closely discussed in the following section.

3.3 RNA-RNA interaction prediction accounting for intramolecular base pairing

Intramolecular base pairing plays a key role for the *in vivo* interplay of distinct RNA molecules. Hence, *in silico* predictions taking intramolecular base pairs into account currently belong to the most sophisticated and successful approaches in this field of biocomputational research. The algorithms can be split into concatenation-based, accessibility-based and general joint structure approaches, which is also the order that the algorithms are going to be presented in.

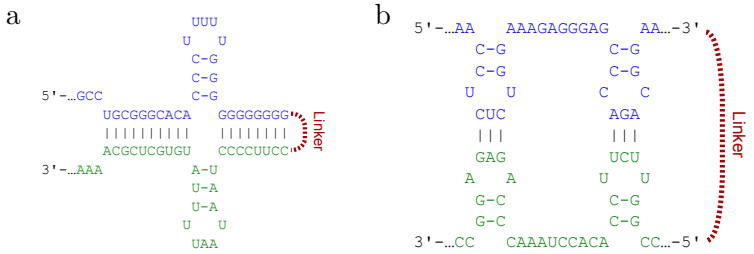

Figure 5: a) Intramolecular structure enclosing interaction predictable by concatenation-based approaches. b) Double kissing hairpin interaction that can *not* be predicted since it forms a pseudoknot when linked; red dotted line denotes the linker, blue/green denotes the first/second sequence, respectively. Base pairs are indicated by pipe or dash symbols.

3.3.1 Concatenation-based approaches

In general, the concatenation-based approaches make extensive use of the predefined algorithms for single RNA secondary structure prediction (see Sec. 2) by concatenating the putatively interacting RNAs, usually interspaced with a so called linker sequence [20, 64]. Tools that allow RNA-RNA interaction prediction in this manner are e.g. MFOLD/UNAFOLD [71], PAIRFOLD [72], RNACOFOLD [73] and one of NUPACK's utilities [74]. The approaches record the position of the linker and fold the concatenated sequences, thus returning the joint minimum free energy structure for the two sequences and the linker. The main difference of the concatenation-based approaches to general RNA folding is that they need a special handling of loops containing the linker sequence, as the linker is an artificially introduced entity. Hence, the energy contributions added by the structures including the linker need to be adjusted. Figure 5a shows a 'hairpin context'. Instead of treating the structure formed by the linker sequence as a bulge or hairpin, respectively, one rather treats them as structure ends. As a consequence, the high energy penalties for the embedding of unpaired regions can be appropriately reduced. Technically, this is solved by an extension of the Zuker recursions from Sec. 2. Here, for every matrix (see Eq. 3-6), one has to treat the case that the linker position is covered by the current loop in addition to the normal case (see [72] for details). Still, the computational complexity of Zuker's algorithm is retained.

Conveniently, the nature of the concatenation-based approaches also allows for the calculation of the partition function and base pair probabilities for the joint structures under application of the principles described by John S. McCaskill [49]. Furthermore, interactions that form a multi-loop (Fig. 5a) can also be considered. A downside, however, is the inability of these approaches to detect interactions that represent pseudoknots in the concatenated model, namely kissing hairpin interactions [75, 76], shown in Fig. 3b and 5b. Here, complementary nucleotides within the hairpin loops, which are not entangled in intramolecular base pairs, form an intermolecular duplex between sRNA and mRNA thereby tying the two RNA molecules together. This common kind of RNA-RNA interaction represents an unpredictable pseudoknot structure in the nested concatenation-based approach, and thus is a central limitation of these approaches. An experimentally verified example of an interaction involving a pseudoknot in the concatenation context is the interplay of the *Escherichia coli* sRNA RyhB and its target mRNA encoded by the *sodB* gene. Here, the second loop of RyhB interacts with the translation initiation region of its target mRNA and thereby represses its translation into the superoxide

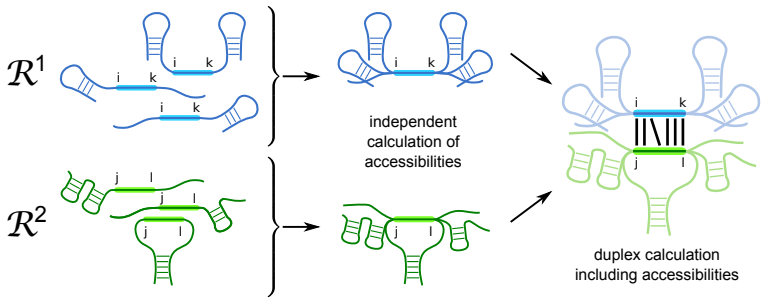

Figure 6: The ensemble-based approach for interaction prediction. Instead of considering only a single individual structure for the RNAs $\mathcal{R}^1$ and $\mathcal{R}^2$, RNAUP and INTARNA introduce a sequence-specific accessibility term, which represents all structures with an accessible (i.e. not covered by intramolecular base pairs) interaction site $i..k$ and $j..l$. These are incorporated into a modified duplex calculation.

dismutase protein [77].

3.3.2 Accessibility-based approaches

Accessibility-based approaches like INTARNA [66] and RNAUP [78] can predict kissing hairpin interactions while still considering the contribution of intramolecular structures. These approaches specifically evaluate the structuredness of putative RNA-RNA interaction sites within the interaction partners and penalize intermolecular duplexes that require the breakup of intramolecular base pairs. Hence, interactions between commonly unstructured or accessible regions, like hairpin loops, are favored.

For this, both RNAUP and INTARNA incorporate the hybridization energy (E^{hybrid}) for the interacting RNAs $\mathcal{R}^1, \mathcal{R}^2$ and the unfolding energies ($ED^{\mathcal{R}^1}$, $ED^{\mathcal{R}^2}$) required to make the interacting regions in both RNAs accessible. The general strategy of these ensemble-based algorithms is given in Fig. 6. E^{hybrid} is calculated by employing the energy model from RNAHYBRID and the energy parameters from Mathews et al. [10] (see $D^{i..k}_{j..l}$ in Eq. 9), while the unfolding energy is derived under application of a partition function approach. The partition function $Z_{\mathfrak{P}}$ for all possible structures $\mathfrak{P}$ of a given RNA is defined as:

$$Z_{\mathfrak{P}} = \sum_{\mathcal{P} \in \mathfrak{P}} e^{-\frac{E(\mathcal{P})}{RT}} \tag{11}$$

where $E(\mathcal{P})$ is the free energy of a specific structure $\mathcal{P} \in \mathfrak{P}$ that the given RNA can form (see Eq. 2), R is the gas constant and T is the temperature of the system. John S. McCaskill introduced an efficient algorithm for the partition function computation [49] that adapts the structure prediction approach of Zuker (see Sec. 2) with equal complexity. Given the loop

decomposition of the energy function, i.e. $E(\mathcal{P}) = \sum_{(i,j)\in\mathcal{P}} E\left(\text{loop } (i,j)\right)$ (see Eq. 2), the Boltzmann weight $e^{-\frac{E(\mathcal{P})}{RT}}$ in Eq. 11 can be replaced by a loop-based product:

$$e^{-\frac{E(\mathcal{P})}{RT}} = e^{-\frac{1}{RT}\sum\limits_{(i,j)\in\mathcal{P}} E(\text{loop}(i,j))} = \prod_{(i,j)\in\mathcal{P}} e^{-\frac{E(\text{loop}(i,j))}{RT}}. \tag{12}$$

Based on that, one can replace energy summations within the Zuker recursions with Boltzmann weight multiplication (following Eq. 12) and change the minimization strategy to a weight summation of structural alternatives (Eq. 11), which results in the computation of $Z_{\mathfrak{P}}$. Once $Z_{\mathfrak{P}}$ is identified, the free energy of the ensemble (E^{ens}) of all possible structures $\mathfrak{P}$ of the given RNA can be calculated by$E^{\text{ens}}(\mathfrak{P}) = -RT\ ln(\ Z_{\mathfrak{P}}\ )$.

Finally, the $ED^{\mathcal{R}}$ value, to make a stretch of bases $i..k$ within an RNA $\mathcal{R}$ unpaired, can be computed by subtracting the ensemble energy E^{ens} of all possible structures $\mathfrak{P}$ from the ensemble energy of all structures with $i..k$ unpaired $\mathfrak{P}_{i..k}^{unpaired}$, which can also be calculated using an extension of McCaskill's algorithm [78, 79]. Note from the following, ED values can either be derived from the ensemble energies or via the unpaired probabilities $Pr^{unpaired}(i..k)$ of subregions.

$$\begin{aligned} ED^{\mathcal{R}}(i,k) &= E^{\text{ens}}\left(\mathfrak{P}_{i..k}^{unpaired}\right) - E^{\text{ens}}(\mathfrak{P}) \\ &= -RT\ ln\left(\ Z_{\mathfrak{P}_{i..k}^{unpaired}}\ /\ Z_{\mathfrak{P}}\ \right) \\ &= -RT\ ln\left(Pr^{unpaired}(i..k)\right) \end{aligned} \tag{13}$$

The $ED^{\mathcal{R}}$ term from Eq. 13 is positive by definition (since $\mathfrak{P}_{(i,k)}^{unpaired} \subseteq \mathfrak{P}$) and thus represents a penalty in the RNA-RNA interaction context, as smaller energies are considered to be favorable. Given an RNA-RNA interaction between the closing base pairs (i,j) and (k,l), where i and k denote the outermost bases of a stretch of RNA $\mathcal{R}^1$ that is written in $5' \rightarrow 3'$ direction, and j and l denote the outermost bases of a stretch of an RNA $\mathcal{R}^2$ that is written in $3' \rightarrow 5'$ direction, the extended hybridization energy computed by RNAUP [78] is given as follows:

$$E^{\text{RNAup}}(i,j,k,l) = D_{j..l}^{i..k} + ED^{\mathcal{R}^1}(i,k) + ED^{\mathcal{R}^2}(j,l), \tag{14}$$

where $D_{j..l}^{i..k}$ is the duplex energy as calculated by Eq. 9. To get the interaction details for the E^{RNAup} entry that minimizes Eq. 14, one has to traceback the according hybridization entry $D_{j..l}^{i..k}$, which directly defines the duplex's boundaries (i,j) and (k,l). This gives rise to an

$O(n^4)$ time and space complexity, which can be reduced to $O(n^2w^2)$ when using a maximal interaction length w. Nevertheless, this complexity is too high for genomic screens. For that reason, IntaRNA [66, 80] was introduced to replace the exhaustive recursions from the presented approach with a heuristic. Therein, for each possible leftmost interaction base pair only respective best interaction (and its right end-point) are stored and considered. Thus, the recursion gets similar to the hybrid-only recursion in Eq. 10 but additionally considers ED values. This reduces the complexity to $O(n^2)$ but keeps the predictive power of the accessibility-based model.

IntaRNA also enforces a seed region as a necessary constraint for two RNAs to be able to interact. This means, that a stretch of perfectly complementary base pairs, the seed, needs to be present within the potentially interacting RNAs in order to make a prediction. This constraint is biologically warranted both for sRNAs [81] and miRNAs [82]. The seed length is usually assumed to be between six and eight nucleotides. To speed up genome-wide target prediction, tools like RIsearch2 [70] or RIblast [83] apply suffix-array-based screens to identify seed regions that are subsequently extended in both directions using an accessibility-based prediction approach.

IntaRNA (version ¿ 2.0) [80] can emulate most hybrid-only and accessibility-based approaches that have been previously discussed. While IntaRNA and similar approaches can predict interactions between single hairpin loops (Fig. 3b), they can not be used to predict interaction patterns forming multiple kissing hairpins (Fig. 5b). An *in vivo* verified example of double kissing hairpins is the interaction of the enterobacterial sRNA OxyS pairing with its target mRNA encoded by the *fhlA* gene [84]. Furthermore, interactions that would represent a multi-loop structure within the interacting region (Fig. 5a) cannot be predicted. Such interactions are conceivable for RNAs such as the glucose activated enterobacterial sRNA Spot42 [22], that could potentially interact with targets using its highly accessible unstructured regions I and III simultaneously. These aspects highlight that both concatenation-based approaches and accessibility-based approaches have intrinsic limitations that restrict the extent of interactions that can be predicted. General joint structure prediction algorithms, which will be discussed next, attempt to tackle these limitations.

3.3.3 General joint structure approaches

Can Alkan and co-workers have shown that unrestricted prediction of RNA-RNA interactions is an NP-hard problem [85]. Nevertheless, they were able to identify topological constraints that

enable efficient prediction schemes that are also satisfied by the following approaches.

Dmitri D. Pervouchine introduced and applied the first intermolecular RNA interaction search (IRIS) method [86] that can predict *general duplex structures* incorporating the structural context of both interacting RNAs. The supplementary material provides further details on the underlying recursions. IRIS was applied to computationally reconstruct certain known interactions of prokaryotic RNAs such as OxyS with the *fhlA* mRNA and several further examples [86], which form a double kissing hairpin interaction as shown in Fig. 5b. While the joint structure prediction enables a wide spectrum of predictable interaction patterns, its extreme runtime complexity of $O(n^6)$ renders it inapplicable to genome-wide screens or similar problems.

The previously mentioned approaches predict a single optimal interaction. However, as in the case of the folding of a single RNA-sequence, the *mathematically* optimal structure is not necessarily the biologically functional one. For that reason, the partition function version of RNA-RNA interaction was independently introduced in [87] and [88], leading to an $O(n^6)$ time and $O(n^4)$ space algorithm. This partition function version of RNA-RNA interaction prediction allows not only to predict sub-optimal interactions and their probabilities, but also allows the computation of probabilities for intermolecular interactions and melting curves, which can be used to assess the stability of the interaction.

Due to the high complexity of these methods, several approaches for reducing the requirements of these algorithms have been introduced. In [89], the idea of sparsification (see supplementary material) was applied to this problem. In [90] and [91], an approach was introduced that extends the accessibility-based approaches to multiple binding sites.

As already discussed for intramolecular RNA structure prediction in Sec. 2, one can increase the prediction quality when considering sets of evolutionarily related sequences instead of solely considering individual examples. The following section discusses such comparative approaches for RNA-RNA interaction prediction.

3.4 Comparative RNA-RNA interaction prediction

Currently, one of the standard applications for RNA-RNA interaction prediction algorithms is whole genome target prediction. This is usually the first step towards characterizing the function of an RNA regulator that exerts its function by directly base pairing with its target RNA. Unfortunately, the pool of potential targets can be huge. The bacterium *Escherichia coli* for instance has over 4,000 protein coding genes, which must all be considered as putative

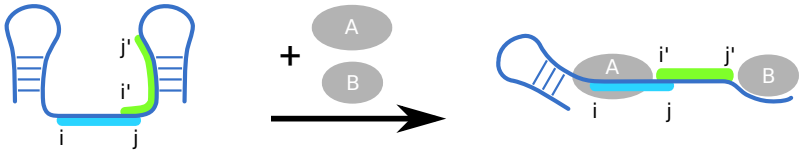

Figure 7: (left) *In silico* accessibility scenario without consideration of RNA-binding factors (A, B). Here, region $i..j$ is accessible while $i'..j'$ is blocked by intramolecular base pairs. (right) Putative *in vivo* situation with bound factors A and B. The *in silico* accessible site $i..j$ is blocked by A while $i'..j'$ becomes accessible due to structural reconfiguration upon binding of A and B.

targets. This number is significantly higher in eukaryotes. Due to the fact that an RNA-RNA duplex can be predicted between most RNA molecules, the magnitude of potential interaction partners often leads to many false positive predictions and thus represents the central limitation of RNA target prediction algorithms. Specifically, this means that predictions for real targets may be lost on the genomic scale due to the noise created by the high abundance of false positives. Simply put, the lists of putative targets are very long and oftentimes the real targets are not on the top ranks.

An explanation for false positive predictions is that the fundamental principles used by most RNA-RNA interaction algorithms generally neglect external factors such as proteins or other RNAs. Therefore, the system that most algorithms imply is an *in vitro* system in which only the two potentially interacting RNAs are present. Clearly this is far from a realistic *in vivo* setting in which RNAs are usually densely covered, for example by factors like proteins, RNAs or small ligands. Hence, regions of the interacting RNAs, may appear accessible within the thermodynamic model even though they are blocked due to additional factor binding. This means that sites considered accessible *in silico* may be inaccessible *in vivo* or vice versa (see Fig. 7).

To at least partially resolve this issue, data from more than one source or organism can be used, which can greatly aid in the reduction of false positive predictions. In fact, it has been stressed that RNA target predictions should be carried out in a comparative manner if possible [59]. A selection of homologous input sequences for comparative predictions can be obtained by using tools such as GLASSGO [92], RNALIEN [93] or GOTOHSCAN [94]. There are two major approaches for comparative RNA-RNA interaction prediction. The first one, as implemented in PETCOFOLD [95, 96] and RIPALIGN [97], uses the same ideas as applied for comparative structure prediction (see Sec. 2.2). That is, instead of predicting the interaction of two single sequences, one predicts the interaction for two alignments. This not only assumes a

conserved interaction site, but also a conserved interaction structure, which is a strong signal. Furthermore, TargetRNA2 also optionally incorporates a phylogenetic target prediction based on an assessment of conservation within the regions of the sRNA input [69]. However, as shown in [98], interaction sites are not necessarily conserved. This implies that the potential of two RNAs to interact might be conserved without a strictly conserved interaction site, which is not in agreement with most alignment based assumptions.

For that reason, the second major approach to comparative RNA-RNA interaction prediction combines individual RNA-RNA interaction predictions without enforcing a strict consensus in order to obtain a more reliable result, given that the regulation is also conserved throughout the considered systems. In principle this is like asking several people a question and making a joined conclusion or decision without closely investigating how each individual reached his or her answer. Such a joined conclusion is most likely better when compared to the conclusion derived from the answer of a single person. In the original RNAhybrid publication [59], Rehmsmeier et al. present a scheme that uses distinct RNAhybrid predictions for homologous miRNAs from different organisms on orthologous targets of said organisms in order to achieve superior predictions.

Duplex energies can not be directly used to make a combined prediction, because they are strongly influenced by the GC-content and dinucleotide frequency of the organism they are made for. For instance, organisms with higher GC-content will generally produce duplex predictions with lower energies. Hence, the duplex energies predicted by RNAhybrid need to be transformed to p-values, which are then comparable. In the following, we will introduce how p-values can be derived and how p-values from different organisms can be combined to enable comparative RNA-RNA interaction prediction.

A p-value represents a statistical measure for the quality of a given prediction and, if correctly estimated, also enables comparability. Following the conclusions from extreme value theory [99], it is appropriate to regard the results of RNA-RNA interaction predictions as extreme value distributed. The density function f of the generalized extreme value (GEV) distribution is introduced and discussed in the supplementary material.

A p-value is the probability that a certain event x or something more extreme ($\geq x$) is observed for a specific background model. Given the density function f of the events, a p-value can be computed by the integral $\int_x^\infty f(x)\mathrm{d}x$. The cumulative distribution F for the GEV distribution (see supplementary material) provides the integral $F(x) = \int_{-\infty}^x f(x)\mathrm{d}x$ for

events $\leq x$, such that we can compute the p-value by $1 - F(x)$.

Since the significant p-values depend on the right tail of the distribution, its correct estimation via the parameters μ, σ, and ε is central to their quality. An appropriate volume of background predictions for estimating the parameters of the GEV can be obtained by dinucleotide shuffling sequences that are actually present in the real search space (e.g. putative target sequences that are present in the investigated genome) and predicting RNA-RNA interactions for these shuffled sequences. It is important to retain the dinucleotide frequency because the duplex prediction depends on base pair stacking and mononucleotide shuffling would thus no longer yield random sequences that still appropriately represent the properties of the non-random system. If target sequences of differing lengths are used, the energy scores need to be normalized with $E_n = -E/\ln(mn)$ to prevent inappropriately high abundance of better predictions for longer sequences.

The negative normalized energy is denoted as E_n while the unnormalized energy is denoted as E with m and n being the lengths of the target and the binding RNA respectively [59]. The parameters for the GEV can then be derived by fitting a GEV to the empiric background's energy scores after duplex prediction and if necessary length normalization. A p-value for a given energy score within the search space can then be inferred from the GEV's cumulative distribution function. For whole genome target predictions, it has been shown that the GEV's parameters can also be estimated by using all the predictions on real putative target sequences and fitting the GEV to these [100]. These predictions are clearly not all completely random due to functionally correct predictions presumably being present. Yet, the majority of predicted duplexes can be assumed as not functionally relevant *in vivo*. While the p-value quality might be inferior compared to a shuffled background model, this strategy leads to strongly reduced runtimes, which is important when performing predictions on a genomic scale.

The individual p-values (p_i) for an orthologous putative target can then be combined to a joint p-value (p_{joint}) by selecting the biggest individual p-value and raising it to the power of the amount of participating organisms (k) (Eq. 15).

$$p_{joint} = (\max\{p_1, ..., p_k\})^k \tag{15}$$

This, however, assumes complete statistical autonomy of the individual p-values, which is not the case for the given biological scenario. Here, the investigated species are assumed to be

descended from a common ancestor, which intuitively implies a certain degree of mutual dependence. Consequently, smaller evolutionary distance between organisms leads to higher statistical dependence of individual results. Hence, the effective number of organisms or sequences (k_{eff}) needs to be assessed. k_{eff} lies between 1 and the amount of participating organisms (k).

k_{eff} can be estimated by shuffling the homologous sRNA or miRNA sequences while retaining the original dinucleotide frequencies followed by duplex predictions for homologous targets. This supplies a background set of optimal duplexes for each participating homologous target from which extreme value distribution parameters can be assessed. With these parameters, the duplex energies can be transformed into p-values. These p-values are then joined following Eq. 15 using several consecutive k' values instead of one single k. k' can lie between 1 and k. Finally, the k' yielding the most uniform distribution of joint p-values is set as k_{eff}. A small k_{eff} implies high dependence between the putative target sequences. Specifically, k_{eff}=1 would mean that no additional information could be gained by incorporation of predictions for homologous sequences. The method for joining the p-values (Eq. 15) can be regarded as very conservative due the fact that it always selects the highest/worst individual p-value for p-value combination. In other words, this means that a putative target needs to return a good prediction for each organism participating in the comparative analysis to be considered as true target. This most likely leads to a lowered number of false positives but may also cause many false negatives depending on the set of organisms that is used.

The comparative prediction algorithm for sRNA targets CopraRNA [100, 101], which was developed to make whole genome target predictions, also uses the concept of transforming energy scores to p-values. CopraRNA, unlike other comparative methods that enforce a consensus interaction site within the homologous putative targets [95], is very unrestrictive. It solely enforces an interaction (as predicted by IntaRNA) to be present anywhere in the putative target sequence without demanding a consistent duplex pattern throughout the homologs. Also, homologs of a putative target need to be present in at least half of the participating organisms to ensure a sensible degree of conservation. Missing single organism p-values are sampled from a multivariate normal distribution, which is based on clusters of homologous genes that contain a homolog from every organism participating in the comparative analysis.

The first step in CopraRNA is to compute individual whole genome target predictions with the homologous sRNAs for each organism participating in the analysis. The energy scores for each putative target are computed by IntaRNA. Following the logic that most duplexes pre-

dicted by IntaRNA are likely to be non-functional (i.e. they represent a random background), the whole genome prediction can be used to estimate extreme value distribution parameters for each of the homologous sRNAs. Based on these parameters, the energy scores can be transformed to p-values. The next step, like in RNAhybrid, is the combination of p-values for homologous putative targets. CopraRNA employs Hartung's method for the combination of dependent p-values [102] to calculate a joined p-value for a cluster of homologous genes with size K. For this, the initial p-values need to be transformed to probits (t_i), which is done based on the inverse of the cumulative distribution function. The combination of the probits t_{joint} is computed following Eq. 16.

$$t_{\text{joint}} = \frac{\sum_{i=1}^{K} \lambda_i t_i}{\sqrt{(1-\rho)\sum_{i=1}^{K} \lambda_i^2 + \rho(\sum_{i=1}^{K} \lambda_i)^2}} \tag{16}$$

The result of Eq. 16, t_{joint}, can then be transformed back to a p-value. Hartung's method includes both a correction for the dependence in the data (ρ) and a weighting (λ_i) for each individual single p-value. The rationale for the dependence correction is the same as described previously for the comparative approach in RNAhybrid and ρ is assessed in a similar manner as k_{eff}. ρ can adopt values between 0 and 1 and higher values for ρ indicate higher dependence in the data. The optimal ρ is the ρ that yields the most uniform distribution of joint CopraRNA p-values. CopraRNA uses the organisms' 16S rDNA to construct a phylogenetic tree. Organisms that are very similar need to have a lower individual weight λ_i when compared to an organism that is evolutionarily very distant. The weights are calculated by a recursive scheme that computes the relative weights of an organism in all subtrees and then multiplies these. The weights are subsequently subjected to a root function to reduce overly strong effects of outlier organisms.

CopraRNA was originally benchmarked on a set of 101 experimentally verified enterobacterial sRNA-target interactions and significantly outperformed its competitor algorithms while also rivaling experimental target predictions derived from microarray experiments. An independent comprehensive benchmark has since confirmed this finding [103]. Furthermore, CopraRNA has been successfully applied in non-enterobacterial systems [104, 105, 106, 107, 108] and it has been shown that it can benefit from the incorporation of Hfq binding data [109]. As a concept, CopraRNA may also be promising in a eukaryotic setting but has not yet been implemented to accommodate for such a context. Due to CopraRNA using IntaRNA as a

background RNA-RNA interaction model, it partly inherits IntaRNA's limitations. Yet, the general concept of phylogeny-guided p-value combination is detached from IntaRNA and allows the application of other interaction prediction algorithms. For the future, this means that CopraRNA can benefit from advances in single organism target prediction.

4 Outlook

The RNA structure and RNA-RNA interaction prediction approaches that have been discussed here are typically used either via according webservers or command-line interfaces of local installations. In order to enhance reproducibility and to accommodate for large-scale application, pipeline and workflow systems like Galaxy [110, 111] and bioconda [112] have been developed. Recently, the 'RNA workbench' extension of Galaxy was published [113], which features many of the approaches outlined here. This enables their high-throughput application in sophisticated (partially pre-defined) workflows for non-expert scientists [114].

All algorithms discussed here are tailored for linear RNA molecules. However, circular non-coding RNAs also exists and have been reported in e.g. eukaryotic cells [115, 116] and archaea [117]. To tackle this new class of ncRNAs, some of the sketched approaches have been appropriately adapted [118, 119].

Overall, a lot of prokaryotic RNA research has been intensely focusing on the one-by-one functional classification of newly identified RNA based regulators and many such projects are still ongoing. One of the central pillars in most of these studies are RNA structure and RNA-RNA interaction elucidation aided by the computational tools that have been mentioned in this review. However, more recently transcriptome wide RNA interactomics data has been produced [120, 121, 122, 123, 25, 26], which can be expected to strongly shift interest towards large scale projects. The opportunities and information within the newly acquired data might be able to answer long standing questions in the predictive community and allow for the development of more sophisticated data driven algorithms.

References

[1] F. H. Crick. Codon–anticodon pairing: the wobble hypothesis. *J Mol Biol*, 19(2):548–55, 1966. [PubMed:5969078] [doi:10.1016/S0022-2836(66)80022-0].

[2] A. P. Gerber and W. Keller. An adenosine deaminase that generates inosine at the wobble position of tRNAs. *Science*, 286(5442):1146–9, 1999. [PubMed:10550050] [doi:10.1126/science.286.5442.1146].

[3] Frank V. 4th Murphy and V. Ramakrishnan. Structure of a purine-purine wobble base pair in the decoding center of the ribosome. *Nat Struct Mol Biol*, 11(12):1251–2, 2004. [PubMed:15558050] [doi:10.1038/nsmb866].

[4] J. D. Watson and F. H. Crick. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature*, 171(4356):737–8, 1953. [doi:10.1038/171737a0].

[5] E. Gerhart H. Wagner and Pascale Romby. Small RNAs in bacteria and archaea: who they are, what they do, and how they do it. *Adv Genet*, 90:133–208, 2015. [PubMed:26296935] [doi:10.1016/bs.adgen.2015.05.001].

[6] Stefan L. Ameres and Phillip D. Zamore. Diversifying microRNA sequence and function. *Nat Rev Mol Cell Biol*, 14(8):475–88, 2013. [PubMed:23800994] [doi:10.1038/nrm3611].

[7] Lauren S. Waters and Gisela Storz. Regulatory RNAs in bacteria. *Cell*, 136(4):615–28, 2009. [PubMed:19239884] [PubMed Central:PMC3132550] [doi:10.1016/j.cell.2009.01.043].

[8] Ruth Nussinov, George Pieczenik, Jerrold R. Griggs, and Daniel J. Kleitman. Algorithms for loop matchings. *SIAM J Appl Math*, 35(1):68–82, July 1978. [doi:10.1137/0135006].

[9] Howard DeVoe and Ignacio Tinoco. The stability of helical polynucleotides: Base contributions. *Journal of Molecular Biology*, 4(6):500–517, 1962. [PubMed:13885894] [doi:10.1016/S0022-2836(62)80105-3].

[10] DH Mathews, J Sabina, M Zuker, and DH Turner. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, 288(5):911–40, 1999. [PubMed:10329189] [doi:10.1006/jmbi.1999.2700].

[11] Douglas H. Turner and David H. Mathews. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res*, 38(Database issue):D280–2, 2010. [PubMed:19880381] [PubMed Central:PMC2808915] [doi:10.1093/nar/gkp892].

[12] Ignacio Tinoco Jr, PN Borer, B Dengler, MD Levin, OC Uhlenbeck, DM Crothers, and J Bralla. Improved estimation of secondary structure in ribonucleic acids. *Nature New Biology*, 246(150):40–41, 1973. [PubMed:4519026] [doi:10.1038/newbio246040a0].

[13] Philip N. Borer, Barbara Dengler, Ignacio Tinoco, and Olke C. Uhlenbeck. Stability of ribonucleic acid double-stranded helices. *Journal of Molecular Biology*, 86(4):843–853, 1974. [PubMed:4427357] [doi:10.1016/0022-2836(74)90357-X].

[14] Mirela Andronescu, Anne Condon, Holger H. Hoos, David H. Mathews, and Kevin P. Murphy. Efficient parameter estimation for RNA secondary structure prediction. *Bioinformatics*, 23(13):i19–28, 2007. [PubMed:17646296] [doi:10.1093/bioinformatics/btm223].

[15] D. H. Turner, N. Sugimoto, J. A. Jaeger, C. E. Longfellow, S. M. Freier, and R. Kierzek. Improved parameters for prediction of RNA structure. *Cold Spring Harb Symp Quant Biol*, 52:123–33, 1987. [PubMed:2456874] [doi:10.1101/SQB.1987.052.01.017].

[16] M. Zuker and P. Stiegler. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–48, 1981. [PubMed:6163133] [PubMed Central:PMC326673] [doi:10.1093/nar/9.1.133].

[17] S. Wuchty, W. Fontana, I. L. Hofacker, and P. Schuster. Complete suboptimal folding of RNA and the stability of secondary structures. *Biopolymers*, 49(2):145–65, 1999. [PubMed:10070264] [doi:10.1002/(SICI)1097-0282(199902)49:2<145::AID-BIP4>3.0.CO;2-G].

[18] Nicholas R. Markham and Michael Zuker. *UNAFold: software for nucleic acid folding and hybridization*, pages 3–31. Humana Press, Totowa, NJ, 2008. [PubMed:18712296] [doi:10.1007/978-1-60327-429-6_1].

[19] M. Zuker. Prediction of RNA secondary structure by energy minimization. *Methods in Molecular Biology*, 25:267–94, 1994. [PubMed:7516239] [doi:10.1385/0-89603-276-0:267].

[20] Ivo L. Hofacker, Walter Fontana, Peter F. Stadler, Sebastian Bonhoeffer, Manfred Tacker, and Peter Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte Chemie*, 125:167–188, 1994. [doi:10.1007/BF00818163].

[21] Ronny Lorenz, Stephan H. Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. ViennaRNA Package 2.0.

Algorithms Mol Biol, 6:26, 2011. [PubMed:22115189] [PubMed Central:PMC3319429] [doi:10.1186/1748-7188-6-26].

[22] Thorleif Møller, Thomas Franch, Christina Udesen, Kenn Gerdes, and Poul Valentin-Hansen. Spot 42 RNA mediates discoordinate expression of the *E. coli* galactose operon. *Genes Dev*, 16(13):1696–706, 2002. [PubMed Central:PMC186370] [doi:10.1101/gad.231702].

[23] Sylvain Durand and Gisela Storz. Reprogramming of anaerobic metabolism by the FnrS small RNA. *Mol Microbiol*, 75(5):1215–31, 2010. [PubMed:20070527] [PubMed Central:PMC2941437] [doi:10.1111/j.1365-2958.2010.07044.x].

[24] Ydo Wexler, Chaya Zilberstein, and Michal Ziv-Ukelson. A study of accessible motifs and RNA folding complexity. *J Comput Biol*, 14(6):856–72, 2007. [PubMed:17691898] [doi:10.1089/cmb.2007.R020].

[25] Pablo Cordero, Julius B. Lucks, and Rhiju Das. An RNA mapping database for curating RNA structure mapping experiments. *Bioinformatics*, 28(22):3006, 2012. [PubMed:22976082] [PubMed Central:PMC3496344] [doi:10.1093/bioinformatics/bts554].

[26] Matthew Norris, Chun Kit Kwok, Jitender Cheema, Matthew Hartley, Richard J. Morris, Sharon Aviran, and Yiliang Ding. FoldAtlas: a repository for genome-wide RNA structure probing data. *Bioinformatics*, 33(2):306, 2017. [PubMed:27663500] [PubMed Central:PMC5254078] [doi:10.1093/bioinformatics/btw611].

[27] Justin T. Low and Kevin M. Weeks. SHAPE-directed RNA secondary structure prediction. *Methods*, 52(2):150–158, 2010. [PubMed:20554050] [PubMed Central:PMC2941709] [doi:10.1016/j.ymeth.2010.06.007].

[28] Stefan Washietl, Ivo L. Hofacker, Peter F. Stadler, and Manolis Kellis. RNA folding with soft constraints: reconciliation of probing data and thermodynamic secondary structure prediction. *Nucleic Acids Research*, 40(10):4261, 2012. [PubMed:22287623] [PubMed Central:PMC3378861] [doi:10.1093/nar/gks009].

[29] Katherine E. Deigan, Tian W. Li, David H. Mathews, and Kevin M. Weeks. Accurate SHAPE-directed RNA structure determination. *Proceedings of the National Academy*

of Sciences, 106(1):97–102, 2009. [PubMed:19109441] [PubMed Central:PMC2629221] [doi:10.1073/pnas.0806929106].

[30] Kourosh Zarringhalam, Michelle M. Meyer, Ivan Dotu, Jeffrey H. Chuang, and Peter Clote. Integrating chemical footprinting data into RNA secondary structure prediction. *PLOS ONE*, 7(10):1–13, 10 2012. [PubMed:23091593] [PubMed Central:PMC3473038] [doi:10.1371/journal.pone.0045160].

[31] Ronny Lorenz, Ivo L. Hofacker, and Peter F. Stadler. RNA folding with hard and soft constraints. *Algorithms for Molecular Biology*, 11(1):8, 2016. [PubMed:27110276] [PubMed Central:PMC4842303] [doi:10.1186/s13015-016-0070-z].

[32] Ronny Lorenz, Dominik Luntzer, Ivo L. Hofacker, Peter F. Stadler, and Michael T. Wolfinger. SHAPE directed RNA folding. *Bioinformatics*, 32(1):145, 2016. [PubMed:26353838] [PubMed Central:PMC4681990] [doi:10.1093/bioinformatics/btv523].

[33] Paul P. Gardner and Robert Giegerich. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics*, 5:140, 2004. [PubMed:15458580] [PubMed Central:PMC526219] [doi:10.1186/1471-2105-5-140].

[34] Ivo L. Hofacker, Martin Fekete, and Peter F. Stadler. Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, 319(5):1059–66, 2002. [doi:10.1016/S0022-2836(02)00308-X].

[35] Bjarne Knudsen and Jotun Hein. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, 31(13):3423–8, 2003. [PubMed:12824339] [PubMed Central:PMC169020] [doi:10.1093/nar/gkg614].

[36] Stefan E. Seemann, Jan Gorodkin, and Rolf Backofen. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res*, 36(20):6355–62, 2008. [PubMed:18836192] [PubMed Central:PMC2582601] [doi:10.1093/nar/gkn544].

[37] Kai Papenfort, Verena Pfeiffer, Sacha Lucchini, Avinash Sonawane, Jay C. D. Hinton, and Jorg Vogel. Systematic deletion of *Salmonella* small RNA genes identifies CyaR, a conserved CRP-dependent riboregulator of OmpX synthesis. *Mol Microbiol*, 68(4):890–906, 2008. [PubMed:18399940] [doi:10.1111/j.1365-2958.2008.06189.x].

[38] Paul P. Gardner, Andreas Wilm, and Stefan Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res*, 33(8):2433–9, 2005. [PubMed:15860779] [PubMed Central:PMC1087786] [doi:10.1093/nar/gki541].

[39] Matthias Höchsmann, Thomas Töller, Robert Giegerich, and Stefan Kurtz. Local similarity in RNA secondary structures. In *Proceedings of Computational Systems Bioinformatics (CSB 2003)*, volume 2, pages 159–168. IEEE Computer Society, 2003. [PubMed:16452790] [doi:10.1109/CSB.2003.1227315].

[40] Sven Siebert and Rolf Backofen. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. *Bioinformatics*, 21(16):3352–9, 2005. [PubMed:15972285] [doi:10.1093/bioinformatics/bti550].

[41] David Sankoff. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM J. Appl. Math.*, 45(5):810–825, 1985. [doi:10.1137/0145048].

[42] J Gorodkin, LJ Heyer, and GD Stormo. Finding the most significant common sequence and structure motifs in a set of RNA sequences. *Nucleic Acids Res*, 25(18):3724–32, 1997. [PubMed:9278497] [PubMed Central:PMC146942] [doi:10.1093/nar/25.18.3724].

[43] Jakob Hull Havgaard, Rune B. Lyngso, Gary D. Stormo, and Jan Gorodkin. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, 21(9):1815–24, 2005. [PubMed:15657094] [doi:10.1093/bioinformatics/bti279].

[44] David H. Mathews and Douglas H. Turner. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol*, 317(2):191–203, 2002. [PubMed:11902836] [doi:10.1006/jmbi.2001.5351].

[45] Robert K. Bradley, Lior Pachter, and Ian Holmes. Specific alignment of structured RNA: stochastic grammars and sequence annealing. *Bioinformatics*, 24(23):2677–83, 2008. [PubMed:18796475] [PubMed Central:PMC2732270] [doi:10.1093/bioinformatics/btn495].

[46] I. L. Hofacker, S. H. Bernhart, and P. F. Stadler. Alignment of RNA base pairing probability matrices. *Bioinformatics*, 20(14):2222–7, 2004. [PubMed:15073017] [doi:10.1093/bioinformatics/bth229].

[47] Sebastian Will, Kristin Reiche, Ivo L. Hofacker, Peter F. Stadler, and Rolf Backofen. Inferring non-coding RNA families and classes by means of genome-scale

structure-based clustering. *PLoS Comput Biol*, 3(4):e65, 2007. [PubMed:17432929] [doi:10.1371/journal.pcbi.0030065].

[48] Elfar Torarinsson, Jakob H. Havgaard, and Jan Gorodkin. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, 23(8):926–32, 2007. [PubMed:17324941] [doi:10.1093/bioinformatics/btm049].

[49] J. S. McCaskill. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–19, 1990. [PubMed:1695107] [doi:10.1002/bip.360290621].

[50] Chuong B. Do, Chuan-Sheng Foo, and Serafim Batzoglou. A max-margin model for efficient simultaneous alignment and folding of RNA sequences. *Bioinformatics*, 24(13):i68–76, 2008. [PubMed:18586747] [PubMed Central:PMC2718655] [doi:10.1093/bioinformatics/btn177].

[51] Jens Georg and Wolfgang R. Hess. *cis*-antisense RNA, another level of gene regulation in bacteria. *Microbiol Mol Biol Rev*, 75(2):286–300, 2011. [doi:10.1128/MMBR.00032-10].

[52] Jan Kruger and Marc Rehmsmeier. RNAhybrid: microRNA target prediction easy, fast and flexible. *Nucleic Acids Res*, 34(Web Server issue):W451–4, 2006. [PubMed:16845047] [PubMed Central:PMC1538877] [doi:10.1093/nar/gkl243].

[53] Patrick R. Wright, Jens Georg, Martin Mann, Dragos A. Sorescu, Andreas S. Richter, Steffen Lott, Robert Kleinkauf, Wolfgang R. Hess, and Rolf Backofen. CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains. *Nucleic Acids Res*, 42(Web Server issue):W119–23, 2014. [PubMed:24838564] [PubMed Central:PMC4086077] [doi:10.1093/nar/gku359].

[54] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–10, 1990. [doi:10.1016/S0022-2836(05)80360-2].

[55] Wolfgang Gerlach and Robert Giegerich. GUUGle: a utility for fast exact matching under RNA complementary rules including G-U base pairing. *Bioinformatics*, 22(6):762–4, 2006. [PubMed:16403789] [doi:10.1093/bioinformatics/btk041].

[56] S. Karlin and S. F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci USA*, 87(6):2264–8, 1990. [PubMed:2315319] [PubMed Central:PMC53667].

[57] Brian Tjaden, Sarah S. Goodwin, Jason A. Opdyke, Maude Guillier, Daniel X. Fu, Susan Gottesman, and Gisela Storz. Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Res*, 34(9):2791–802, 2006. [PubMed:16717284] [PubMed Central:PMC1464411] [doi:10.1093/nar/gkl356].

[58] T. F. Smith and M. S. Waterman. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–7, 1981. [PubMed:7265238] [doi:10.1016/0022-2836(81)90087-5].

[59] Marc Rehmsmeier, Peter Steffen, Matthias Höchsmann, and Robert Giegerich. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–17, 2004. [PubMed:15383676] [PubMed Central:PMC1370637] [doi:10.1261/rna.5248604].

[60] Hao Gong, Gia-Phong Vu, Yong Bai, Elton Chan, Ruobin Wu, Edward Yang, Fenyong Liu, and Sangwei Lu. A salmonella small non-coding RNA facilitates bacterial invasion and intracellular replication by modulating the expression of virulence factors. *PLoS Pathog*, 7(9):e1002120, 2011. [PubMed:21949647] [PubMed Central:PMC3174252] [doi:10.1371/journal.ppat.1002120].

[61] Kai Papenfort, Yan Sun, Masatoshi Miyakoshi, Carin K. Vanderpool, and Jorg Vogel. Small RNA-mediated activation of sugar phosphatase mRNA regulates glucose homeostasis. *Cell*, 153(2):426–37, 2013. [PubMed:23582330] [PubMed Central:PMC4151517] [doi:10.1016/j.cell.2013.03.003].

[62] Hakim Tafer and Ivo L. Hofacker. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, 24(22):2657–63, 2008. [PubMed:18434344] [doi:10.1093/bioinformatics/btn193].

[63] Anne Wenzel, Erdinc Akbasli, and Jan Gorodkin. RIsearch: fast RNA-RNA interaction search using a simplified nearest-neighbor energy model. *Bioinformatics*, 28(21):2738–46, 2012. [PubMed:22923300] [PubMed Central:PMC3476332] [doi:10.1093/bioinformatics/bts519].

[64] D. H. Mathews, M. E. Burkard, S. M. Freier, J. R. Wyatt, and D. H. Turner. Predicting oligonucleotide affinity to nucleic acid targets. *RNA*, 5(11):1458–69, 1999. [PubMed:10580474] [PubMed Central:PMC1369867].

[65] T. D. Brock, K. M. Brock, R. T. Belly, and R. L. Weiss. Sulfolobus: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature. *Arch Mikrobiol*, 84(1):54–68, 1972. [PubMed:4559703] [doi:10.1007/BF00408082].

[66] Anke Busch, Andreas S. Richter, and Rolf Backofen. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–56, 2008. [PubMed:18940824] [PubMed Central:PMC2639303] [doi:10.1093/bioinformatics/btn544].

[67] Hakim Tafer, Fabian Amman, Florian Eggenhofer, Peter F. Stadler, and Ivo L. Hofacker. Fast accessibility-based prediction of RNA-RNA interactions. *Bioinformatics*, 27(14):1934, 2011. [PubMed:21593134] [doi:10.1093/bioinformatics/btr281].

[68] Florian Eggenhofer, Hakim Tafer, Peter F. Stadler, and Ivo L. Hofacker. RNApredator: fast accessibility-based prediction of sRNA targets. *Nucleic Acids Research*, 39:W149, 2011. [PubMed:21672960] [PubMed Central:PMC3125805] [doi:10.1093/nar/gkr467].

[69] Mary Beth Kery, Monica Feldman, Jonathan Livny, and Brian Tjaden. TargetRNA2: identifying targets of small regulatory RNAs in bacteria. *Nucleic Acids Res*, 42(Web Server issue):W124–9, 2014. [PubMed:24753424] [PubMed Central:PMC4086111] [doi:10.1093/nar/gku317].

[70] Ferhat Alkan, Anne Wenzel, Oana Palasca, Peter Kerpedjiev, AndersFrost Rudebeck, Peter F. Stadler, Ivo L. Hofacker, and Jan Gorodkin. RIsearch2: suffix array-based large-scale prediction of RNA-RNA interactions and siRNA off-targets. *Nucleic Acids Research*, 45(8):e60, 2017. [PubMed:28108657] [PubMed Central:PMC5416843] [doi:10.1093/nar/gkw1325].

[71] Michael Zuker. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research*, 31(13):3406, 2003. [PubMed:12824337] [PubMed Central:PMC169194] [doi:10.1093/nar/gkg595].

[72] Mirela Andronescu, Zhi Chuan Zhang, and Anne Condon. Secondary structure prediction of interacting RNA molecules. *J Mol Biol*, 345(5):987–1001, 2005. [PubMed:15644199] [doi:10.1016/j.jmb.2004.10.082].

[73] Stephan H. Bernhart, Hakim Tafer, Ulrike Muckstein, Christoph Flamm, Peter F. Stadler, and Ivo L. Hofacker. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol*, 1(1):3, 2006. [PubMed:16722605] [PubMed Central:PMC1459172] [doi:10.1186/1748-7188-1-3].

[74] Robert M. Dirks, Justin S. Bois, Joseph M. Schaeffer, Erik Winfree, and Niles A. Pierce. Thermodynamic analysis of interacting nucleic acid strands. *SIAM Review*, 49(1):65–88, 2007. [doi:10.1137/060651100].

[75] Kung-Yao Chang and Ignacio Tinoco. The structure of an RNA "kissing" hairpin complex of the HIV TAR hairpin loop and its complement. *Journal of Molecular Biology*, 269(1):52 – 66, 1997. [PubMed:9193000] [doi:10.1006/jmbi.1997.1021].

[76] Nilshad Salim, Rajan Lamichhane, Rui Zhao, Tuhina Banerjee, Jane Philip, David Rueda, and AndrewL. Feig. Thermodynamic and kinetic analysis of an RNA kissing interaction and its resolution into an extended duplex. *Biophysical Journal*, 102(5):1097 – 1107, 2012. [PubMed:22404932] [doi:10.1016/j.bpj.2011.12.052].

[77] Branislav Vecerek, Isabella Moll, Taras Afonyushkin, Vladimir Kaberdin, and Udo Blasi. Interaction of the RNA chaperone Hfq with mRNAs: direct and indirect roles of Hfq in iron metabolism of Escherichia coli. *Mol Microbiol*, 50(3):897–909, 2003. [PubMed:14617150] [doi:10.1046/j.1365-2958.2003.03727.x].

[78] Ulrike Mückstein, Hakim Tafer, Jörg Hackermüller, Stephan H. Bernhart, Peter F. Stadler, and Ivo L. Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10):1177–82, 2006. [doi:10.1093/bioinformatics/btl024].

[79] Stephan H. Bernhart, Ulrike Mückstein, and Ivo L. Hofacker. RNA accessibility in cubic time. *Algorithms for Molecular Biology*, 6(1):3, 2011. [PubMed:21388531] [PubMed Central:PMC3063221] [doi:10.1186/1748-7188-6-3].

[80] Martin Mann, Patrick R. Wright, and Rolf Backofen. IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Res*, 45(W1):W435–W439, 2017. [PubMed:28472523] [doi:10.1093/nar/gkx279].

[81] Roberto Balbontín, Francesca Fiorini, Nara Figueroa-Bossi, Josep Casadesús, and Lionello Bossi. Recognition of heptameric seed sequence underlies multi-target regulation by RybB small RNA in *Salmonella enterica*. *Mol Microbiol*, 78(2):380–94, 2010. [doi:10.1111/j.1365-2958.2010.07342.x].

[82] Julius Brennecke, Alexander Stark, Robert B. Russell, and Stephen M. Cohen. Principles of microRNA-target recognition. *PLoS Biol*, 3(3):e85, 2005. [doi:10.1371/journal.pbio.0030085].

[83] Tsukasa Fukunaga and Michiaki Hamada. RIblast: an ultrafast RNA-RNA interaction prediction system based on a seed-and-extension approach. *Bioinformatics*, 33(17):2666–2674, 2017. [PubMed:28459942] [doi:10.1093/bioinformatics/btx287].

[84] L. Argaman and S. Altuvia. *fhlA* repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. *J Mol Biol*, 300(5):1101–12, 2000. [doi:10.1006/jmbi.2000.3942].

[85] Can Alkan, Emre Karakoç, Joseph H. Nadeau, S. Cenk Sahinalp, and Kaizhong Zhang. RNA-RNA interaction prediction and antisense RNA target search. *J Comput Biol*, 13(2):267–82, 2006. [PubMed:16597239] [doi:10.1089/cmb.2006.13.267].

[86] Dmitri D. Pervouchine. IRIS: intermolecular RNA interaction search. *Genome Inform*, 15(2):92–101, 2004. [PubMed:15706495].

[87] Hamidreza Chitsaz, Raheleh Salari, S. Cenk Sahinalp, and Rolf Backofen. A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, 25(12):i365–73, 2009. [PubMed:19478011] [PubMed Central:PMC2687966] [doi:10.1093/bioinformatics/btp212].

[88] Fenix W. D. Huang, Jing Qin, Christian M. Reidys, and Peter F. Stadler. Partition function and base pairing probabilities for RNA-RNA interaction prediction. *Bioinformatics*, 25(20):2646–54, 2009. [PubMed:19671692] [doi:10.1093/bioinformatics/btp481].

[89] Raheleh Salari, Mathias Möhl, Sebastian Will, S. Cenk Sahinalp, and Rolf Backofen. Time and space efficient RNA-RNA interaction prediction via sparse folding. In Bonnie Berger, editor, *Proc. of RECOMB 2010*, volume 6044 of *Lecture Notes in Computer Science*, pages 473–490. Springer-Verlag Berlin Heidelberg, 2010. [doi:10.1007/978-3-642-12683-3_31].

[90] Hamidreza Chitsaz, Rolf Backofen, and S. Cenk Sahinalp. biRNA: Fast RNA-RNA binding sites prediction. In Steven Salzberg and Tandy Warnow, editors, *Proc. of the 9th Workshop on Algorithms in Bioinformatics (WABI)*, volume 5724 of *Lecture Notes in Computer Science*, pages 25–36. Springer Berlin / Heidelberg, 2009. [doi:10.1007/978-3-642-04241-6_3].

[91] Raheleh Salari, Rolf Backofen, and S. Cenk Sahinalp. Fast prediction of RNA-RNA interaction. *Algorithms Mol Biol*, 5:5, 2010. [PubMed:20047661] [doi:10.1186/1748-7188-5-5].

[92] Steffen Lott, Richard Schäfer, Martin Mann, Wolfgang Hess, Björn Voß, and Jens Georg. GLASSgo - automated and reliable detection of sRNA homologs from a single input sequence. }*, 2017. [submitted].*

[93] Florian Eggenhofer, Ivo L. Hofacker, and Christian Honer Zu Siederdissen. RNAlien - Unsupervised RNA family model construction. Nucleic Acids Res*, 44(17):8433–41, 2016. [PubMed:27330139] [PubMed Central:PMC5041467] [doi:10.1093/nar/gkw558].*

[94] Jana Hertel, Danielle de Jong, Manja Marz, Dominic Rose, Hakim Tafer, Andrea Tanzer, Bernd Schierwater, and Peter F. Stadler. Non-coding RNA annotation of the genome of Trichoplax adhaerens*.* Nucleic Acids Res*, 37(5):1602–15, 2009.*

[95] Stefan E. Seemann, Andreas S. Richter, Tanja Gesell, Rolf Backofen, and Jan Gorodkin. PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. Bioinformatics*, 27(2):211–219, 2011. [PubMed:21088024] [PubMed Central:PMC3018821] [doi:10.1093/bioinformatics/btq634].*

[96] S. E. Seemann, P. Menzel, R. Backofen, and J. Gorodkin. The PETfold and PETcofold web servers for intra- and intermolecular structures of multiple RNA sequences. Nucleic Acids Res*, 39:W107–11, 2011. [PubMed:21609960] [PubMed Central:PMC3125731] [doi:10.1093/nar/gkr248].*

[97] *Andrew X. Li, Manja Marz, Jing Qin, and Christian M. Reidys. RNA-RNA interaction prediction based on multiple sequence alignments.* Bioinformatics, *27(4):456–63, 2011. [PubMed:21134894] [doi:10.1093/bioinformatics/btq659].*

[98] *Andreas S. Richter and Rolf Backofen. Accessibility and conservation: General features of bacterial small RNA-mRNA interactions?* RNA Biol, *9(7):954–65, 2012. [PubMed:22767260] [PubMed Central:PMC3495738] [doi:10.4161/rna.20294].*

[99] *E.J. Gumbel.* Statistics of extremes. *Columbia University Press, 1958.*

[100] *Patrick R. Wright, Andreas S. Richter, Kai Papenfort, Martin Mann, Jorg Vogel, Wolfgang R. Hess, Rolf Backofen, and Jens Georg. Comparative genomics boosts target prediction for bacterial small RNAs.* Proc Natl Acad Sci USA, *110(37):E3487–96, 2013. [PubMed:23980183] [PubMed Central:PMC3773804] [doi:10.1073/pnas.1303248110].*

[101] *Patrick R. Wright.* Predicting small RNA targets in prokaryotes - a challenge beyond the barriers of thermodynamic models. *PhD thesis, Albert-Ludwigs-University Freiburg, December 2016.*

[102] *J Hartung. A note on combining dependent tests of significance.* Biom J, *41(7):849–55, 1999. [doi:10.1002/(SICI)1521-4036(199911)41:7¡849::AID-BIMJ849¿3.0.CO;2-T].*

[103] *Adrien Pain, Alban Ott, Hamza Amine, Tatiana Rochat, Philippe Bouloc, and Daniel Gautheret. An assessment of bacterial small RNA target prediction programs.* RNA Biol, *12(5):509–13, 2015. [PubMed:25760244] [PubMed Central:PMC4615726] [doi:10.1080/15476286.2015.1020269].*

[104] *Jens Georg, Dennis Dienst, Nils Schurgers, Thomas Wallner, Dominik Kopp, Damir Stazic, Ekaterina Kuchmina, Stephan Klahn, Heiko Lokstein, Wolfgang R. Hess, and Annegret Wilde. The Small Regulatory RNA SyR1/PsrR1 Controls Photosynthetic Functions in Cyanobacteria.* Plant Cell, *26(9):3661–79, 2014. [PubMed:25248550] [PubMed Central:PMC4213160] [doi:10.1105/tpc.114.129767].*

[105] *Aaron Overloper, Alexander Kraus, Rosemarie Gurski, Patrick R. Wright, Jens Georg, Wolfgang R. Hess, and Franz Narberhaus. Two separate modules of the conserved regulatory RNA AbcR1 address multiple target mRNAs in and outside of the translation initia-*

tion region. RNA Biol*, 11(5), 2014. [PubMed:24921646] [PubMed Central:PMC4152367] [doi:10.4161/rna.29145].*

[106] Marta Robledo, Benjamin Frage, Patrick R. Wright, and Anke Becker. A stress-induced small RNA modulates alpha-rhizobial cell cycle progression. PLoS Genet*, 11(4):e1005153, 2015. [PubMed:25923724] [PubMed Central:PMC4414408] [doi:10.1371/journal.pgen.1005153].*

[107] Stephan Klahn, Christoph Schaal, Jens Georg, Desiree Baumgartner, Gernot Knippen, Martin Hagemann, Alicia M. Muro-Pastor, and Wolfgang R. Hess. The sRNA NsiR4 is involved in nitrogen assimilation control in cyanobacteria by targeting glutamine synthetase inactivating factor IF7. Proc Natl Acad Sci USA*, 112(45):E6243–52, 2015. [PubMed:26494284] [PubMed Central:PMC4653137] [doi:10.1073/pnas.1508412112].*

[108] Sylvain Durand, Frederique Braun, Efthimia Lioliou, Cedric Romilly, Anne-Catherine Helfer, Laurianne Kuhn, Noe Quittot, Pierre Nicolas, Pascale Romby, and Ciaran Condon. A nitric oxide regulated small RNA controls expression of genes involved in redox homeostasis in Bacillus subtilis. PLoS Genet*, 11(2):e1004957, 2015. [PubMed:25643072] [PubMed Central:PMC4409812] [doi:10.1371/journal.pgen.1004957].*

[109] Erik Holmqvist, Patrick R. Wright, Lei Li, Thorsten Bischler, Lars Barquist, Richard Reinhardt, Rolf Backofen, and Jörg Vogel. Global RNA recognition patterns of post-transcriptional regulators Hfq and CsrA revealed by UV crosslinking in vivo. EMBO J*, 35(9):991–1011, 2016. [PubMed:27044921] [doi:10.15252/embj.201593360].*

[110] Enis Afgan, Jeremy Goecks, Dannon Baker, Nate Coraor, Anton Nekrutenko, and James Taylor. Galaxy: A Gateway to Tools in e-Science*, pages 145–177. Springer London, London, 2011. [doi:10.1007/978-0-85729-439-5_6].*

[111] Enis Afgan, Dannon Baker, Marius vandenBeek, Daniel Blankenberg, Dave Bouvier, Martin Cech, John Chilton, Dave Clements, Nate Coraor, Carl Eberhard, Bjrn Grning, Aysam Guerler, Jennifer Hillman-Jackson, Greg VonKuster, Eric Rasche, Nicola Soranzo, Nitesh Turaga, James Taylor, Anton Nekrutenko, and Jeremy Goecks. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Research*, 44(W1):W3, 2016. [PubMed:27137889] [PubMed Central:PMC4987906] [doi:10.1093/nar/gkw343].*

[112] Björn Grüning, Ryan Dale, Andreas Sjödin, Jillian Rowe, Brad A. Chapman, Christopher H. Tomkins-Tinch, Renan Valieris, Bioconda team, and Johannes Köster. Bioconda: A sustainable and comprehensive software distribution for the life sciences. bioRxiv*, 2017. [doi:10.1101/207092].*

[113] Björn A. Grüning, Jörg Fallmann, Dilmurat Yusuf, Sebastian Will, Anika Erxleben, Florian Eggenhofer, Torsten Houwaart, Berenice Batut, Pavankumar Videm, Andrea Bagnacani, Markus Wolfien, Steffen C. Lott, Youri Hoogstrate, Wolfgang R. Hess, Olaf Wolkenhauer, Steve Hoffmann, Altuna Akalin, Uwe Ohler, Peter F. Stadler, and Rolf Backofen. The RNA workbench: best practices for RNA and high-throughput sequencing bioinformatics in Galaxy. Nucleic Acids Research*, 35(W1):W560–W566, 2017. [PubMed:28582575] [doi:10.1093/nar/gkx409].*

[114] Björn A. Grüning, Eric Rasche, Boris Rebolledo-Jaramillo, Carl Eberhard, Torsten Houwaart, John Chilton, Nate Coraor, Rolf Backofen, James Taylor, and Anton Nekrutenko. Jupyter and Galaxy: Easing entry barriers into complex data analyses for biomedical researchers. PLOS Computational Biology*, 13(5):1–10, 05 2017. [PubMed:28542180] [PubMed Central:PMC5444614] [doi:10.1371/journal.pcbi.1005425].*

[115] Liang Chen, Chuan Huang, Xiaolin Wang, and Ge Shan. Circular RNAs in Eukaryotic cells. Current Genomics*, 16(5):312–318, 2015. [PubMed:27047251] [PubMed Central:PMC4763969] [doi:10.2174/1389202916666150707161554].*

[116] Linda Koch. RNA: Translated circular RNAs. Nat Rev Genet*, 18(5):272–273, 2017. [PubMed:28366936] [doi:10.1038/nrg.2017.27].*

[117] Miri Danan, Schraga Schwartz, Sarit Edelheit, and Rotem Sorek. Transcriptome-wide discovery of circular RNAs in Archaea. Nucleic Acids Research*, 40(7):3131, 2012. [PubMed:22140119] [PubMed Central:PMC3326292] [doi:10.1093/nar/gkr1009].*

[118] Ivo L. Hofacker and Peter F. Stadler. Memory efficient folding algorithms for circular RNA secondary structures. Bioinformatics*, 22(10):1172, 2006. [PubMed:16452114] [doi:10.1093/bioinformatics/btl023].*

[119] Ivo L. Hofacker, Christian M. Reidys, and Peter F. Stadler. Symmetric circular matchings and RNA folding. Discrete Mathematics*, 312(1):100–112, 2012. [doi:10.1016/j.disc.2011.06.004].*

[120] Sahar Melamed, Asaf Peer, Raya Faigenbaum-Romm, Yair E. Gatt, Niv Reiss, Amir Bar, Yael Altuvia, Liron Argaman, and Hanah Margalit. Global Mapping of Small RNA-Target Interactions in Bacteria. Mol Cell*, 63(5):884–97, 2016. [PubMed:27588604] [PubMed Central:PMC5145812] [doi:10.1016/j.molcel.2016.07.026].*

[121] Shafagh A. Waters, Sean P. McAteer, Grzegorz Kudla, Ignatius Pang, Nandan P. Deshpande, Timothy G. Amos, Kai Wen Leong, Marc R. Wilkins, Richard Strugnell, David L. Gally, David Tollervey, and Jai J. Tree. Small RNA interactome of pathogenic E. coli revealed through crosslinking of RNase E. EMBO J*, 2016. [PubMed:27836995] [PubMed Central:PMC5286369] [doi:10.15252/embj.201694639].*

[122] Eesha Sharma, Tim Sterne-Weiler, Dave O'Hanlon, and Benjamin J. Blencowe. Global Mapping of Human RNA-RNA Interactions. Mol Cell*, 62(4):618–26, 2016. [PubMed:27184080] [doi:10.1016/j.molcel.2016.04.030].*

[123] Zhipeng Lu, Qiangfeng Cliff Zhang, Byron Lee, Ryan A. Flynn, Martin A. Smith, James T. Robinson, Chen Davidovich, Anne R. Gooding, Karen J. Goodrich, John S. Mattick, Jill P. Mesirov, Thomas R. Cech, and Howard Y. Chang. RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. Cell*, 165(5):1267–79, 2016. [PubMed:27180905] [PubMed Central:PMC5029792] [doi:10.1016/j.cell.2016.04.028].*