Accessibility and conservation in bacterial small RNA-mRNA interactions and implications for genome-wide target predictions

Andreas S. Richter
Bioinformatics Group, University of Freiburg
Georges-Köhler-Allee 106, 79110 Freiburg
Germany
arichter@informatik.uni-freiburg.de

Rolf Backofen
Bioinformatics Group, University of Freiburg
Georges-Köhler-Allee 106, 79110 Freiburg
Germany
backofen@informatik.uni-freiburg.de

Abstract

Bacterial small RNAs (sRNAs) are a class of structural RNAs that often regulate mRNA targets via post-transcriptional base pair interactions. In this study, we assessed the accessibility and conservation of the interaction sites and the influence of these features on genome-wide target predictions. For this purpose, we compiled a set of 71 experimentally verified sRNA-target pairs from *Escherichia coli* and *Salmonella*, and collected genome-wide information on 5' untranslated regions of annotated genes. Then, features of the confirmed interactions were compared to a set of non-interactions.

We found that the interaction sites both in sRNA and target are more accessible than in the negative data. Furthermore, interaction sites in the sRNAs, but not in the targets, show high sequence conservation. The base pairing between sRNA and target was not found to be generally conserved across more distantly related species. We then present two approaches to constrain the region of interaction initiation to (1) well-accessible regions in both interaction partners or (2) unstructured conserved sRNA regions derived from reliability profiles of multiple sRNA alignments. Using these constraints, genome-wide target predictions were improved in terms of specificity.

1 Introduction

In recent years small regulatory RNAs (sRNAs) in bacteria have received growing attention. The largest and most extensively studied group of sRNAs base pair *trans*-encoded target mRNAs to post-transcriptionally regulate their translation or alter their stability [36]. Therefore, these sRNAs can be considered as functional analogues to eukaryotic microRNAs (miRNAs). In contrast to miRNAs, sRNAs are heterogeneous in size (typically 50-250 nt) and structure, but there are examples where a conserved 5' sRNA domain is used to regulate multiple targets in analogy to miRNA seed pairing [24].

Several computational approaches have been developed to date for the prediction of sRNA targets and sRNA-target interactions. The methods employed range from alignment-like scoring [35] and machine learning [5] over energy-based models incorporating interaction site accessibility [3, 23, 34] to complex joint secondary structure prediction models [6, 19, 30, 32].

The pairing between sRNA and target usually involves a core interaction of six to eight contiguous base pairs, which we denote seed region in the following [13]. Our method `IntaRNA` employs this interaction feature and predicts RNA-RNA interactions starting from such a seed region. Furthermore, its scoring accounts for the structure of the two interaction partners via interaction site accessibility. In previous work, we predicted targets of the sRNA Yfr1 using an ultraconserved unpaired sequence motif as putative interaction seed, which reduced the number of target candidates remarkably. Two Yfr1 targets could be confirmed experimentally [29].

In this study, we explore to which extent accessibility and sequence conservation are general features of interaction sites in sRNA-mRNA interactions. A recent study by Peer and Margalit [26] showed for *Escherichia coli* (*E. coli*) that interaction sites in sRNAs are marked by accessibility and conservation. Here, we analyse a more comprehensive set of 74 interactions from *E. coli* and *Salmonella typhimurium* (*Salmonella*) to study accessibility and conservation both in sRNAs and their targets. Furthermore, we

investigate the covariance in the sRNA-mRNA interaction. Finally, we combine our findings with our target prediction method `IntaRNA` to improve the specificity of genome-wide sRNA target predictions.

2 Material and Methods

2.1 Data set of sRNA-mRNA interactions

This analysis uses a data set of 71 sRNA-target pairs involving 19 distinct sRNAs from the two bacterial model organisms *E. coli* and *Salmonella* (see Supplementary Tables S1 and S2). Three sRNA-mRNA pairs include two separate interaction sites, such that the total number of interactions sums up to 74. All interactions were verified experimentally by *in vitro* (structural) probing or mutational studies at the interaction site. The seed lengths of these interactions range from 5 to 19 nt. The interaction sites in the targets are located between positions -131 and +78 relative to the translation start.

Homologous sRNA and mRNA sequences were searched in 21 enterobacterial species (see Supplementary Table S3). The complete genome of each species was retrieved from NCBI RefSeq database [27]. Groups of orthologous genes were identified with `OrthoMCL` [21].

The semi-global alignment tool `GotohScan` [15] was used to search homologous sequences of each *E. coli* and *Salmonella* sRNA (E-value < 0.01). Because sequence-based methods perform best for a pairwise sequence identity of at least 50-60% [11], sequences with < 60% sequence identity to the query sequence were rejected. Each set of homologous sRNA sequences was structurally aligned with `LocARNA-P` applying probabilistic consistency transformation (parameters `-probabilistic -consistency-transformation -max-diff=60`) [37, 38].

To compile a set of all mRNAs with accurate 5' regions, the lengths of 5' untranslated regions (5'UTRs) were obtained from two genome-wide studies that characterised the transcription unit architecture of the *E. coli* genome by high-throughput sequencing or directed mapping of transcription start sites (TSSs) [7, 22]. Since both data sets missed the TSS of two *E. coli* and six *Salmonella* genes of our interaction data set, we determined their 5'UTR lengths from literature. In total, the 5'UTR lengths of 2313 different *E. coli* genes, which is about 56% of all annotated genes, were obtained. 5'UTR lengths of *Salmonella* genes were derived from the length of the corresponding *E. coli* orthologs. In case of ambiguities, the 5'UTR length was set to the maximal value found. For each annotated mRNA, the 5'UTR sequence and 150 nt of the coding sequence (CDS) were extracted from the genomic sequence. These sequences are denoted target regions in the following. If the TSS position was unknown, e.g., because the gene is part of an operon, 200 nt upstream of the start codon were used instead of the 5'UTR. A sequence length of 200 nt covers the majority of *E. coli* 5'UTRs, which mostly vary from 20 to 40 nt in length [22]. Target regions of orthologous genes were then locally aligned with `MAFFT` (method E-INS-i for generalized affine gap costs) [20].

2.2 Negative data

Selected features of the experimentally validated sRNA-mRNA interactions were evaluated by comparison to a negative data set. To study the interaction site features accessibility and sequence conservation independent of the hybridization between the two RNAs, each non-interaction has to closely resemble the intermolecular base pairing and free energy of the respective proven interaction. Furthermore, for each interaction in the negative data, the mRNA and the interaction site in the sRNA have to be distinct from the validated interaction.

To this end, we first predicted putative hybridisations between all *E. coli* and *Salmonella* sRNAs and all target regions with at least one identified ortholog. The hybridisations were predicted with `IntaRNA` neglecting accessibility (parameters `-a 0 -b 0 -p 6 -s 2`) [3]. Then, for each known interaction,

a set comprising all sub-interactions of these predicted hybridisations satisfying the following properties was determined: mRNA different from true target, sRNA interaction site non-overlapping with true sRNA interaction site, equal interaction pattern or number of base pairs (and optionally equal interaction length) with true interaction, and mRNA interaction site located in CDS if the same applies to the true target since protein-coding and non-coding regions are subject to different evolutionary constraints. Finally, the hybridisation free energies of each true interaction and all sub-interactions complying with the aforementioned requirements were calculated with `RNAeval` [18]. For each validated interaction, the non-interaction having the closest energy to the validated interaction was selected as negative interaction.

The sRNA GcvB is known to directly regulate 21 mRNAs, which is the largest number of validated targets for a single sRNA [33]. In total, GcvB alters mRNA expression levels of ~1% of all protein-coding genes. Assuming that each sRNA has a similar number of targets, it is very unlikely that a random mRNA selected as non-target is actually a true target of the sRNA, which is additionally paired at exactly the region chosen in the negative data set.

Negative data could have also been obtained from the database sRNATarBase, which contains experimentally proven non-interactions [4]. However, it was not used in this study as it does not contain enough entries to obtain a non-interaction for each true interaction.

2.3 Accessibility

The accessibility of interacting regions was assessed by the free energy needed to open the region, which is equivalent to its probability PU of being unpaired. This measure has the advantage to account for all possible secondary structures.

The unpaired probabilities of sRNA sequences were computed by global folding with `RNAup` (parameters `-d2 -u 1-10`) [23]. mRNA sequences were locally folded with a sliding window approach by `RNAplfold` allowing a maximal base pair span of 70 in a folding window of 140 nt (parameters `-d2 -W 140 -L 70 -u 10`) [1].

The length of interaction sites varies for each sRNA-target pair. Since unpaired probabilities can only be compared for regions of equal length, we instead computed the expected fraction EF of unpaired bases at the interaction site [16]. This measure is length-independent.

The $EF_{a,b}$ of a subsequence $S_a \dots S_b$ of RNA sequence S is defined by

$$EF_{a,b} = \frac{\sum_{i=a}^{b} PU_{i,i}}{b-a+1},$$

where $PU_{i,i}$ is the probability that nucleotide S_i is unpaired.

Furthermore, let S^1 and S^2 be two RNA sequences where the subsequences $S^1_i \dots S^1_j$ and $S^2_k \dots S^2_l$ form an interaction enclosed by base pairs (i,k) and (j,l). We then define the joint probability $PU^*_{i,j,k,l}$ that the interacting subsequences $S^1_i \dots S^1_j$ and $S^2_k \dots S^2_l$ are unpaired by

$$PU^*_{i,j,k,l} = PU_{i,j} \cdot PU_{k,l},$$

where $PU_{i,j}$ and $PU_{k,l}$ are the probabilities that the respective subsequences are unpaired. This definition is based on the assumption that both sequences fold independently, i.e., $PU_{i,j}$ and $PU_{k,l}$ are stochastically independent.

2.4 Sequence conservation

The sequence conservation of interacting regions was assessed by their information content. This measure allows to compare the values of alignments that differ in the number of included species. We used an extended expression of the information content that incorporates scoring of gaps in the alignment [12].

The information content I_i of an alignment position $\mathcal{A}_i$ is defined by

$$I_i = \sum_{k \in A} I_{ik} = \sum_{k \in A} q_{ik} \log_2 \frac{q_{ik}}{p_k},$$

where $A = \{\texttt{A},\texttt{C},\texttt{G},\texttt{U},\texttt{-}\}$ is the set of nucleotides including gaps, q_{ik} is the observed frequency of symbol $k \in A$ at position i, and p_k is the background symbol distribution [12]. We set $p_- = 1$ and assume uniform background nucleotide distribution, i.e., $p_k = 0.25$.

We then define the sequence conservation $C_{a,b}$ of consecutive alignment columns $\mathcal{A}_a$ to $\mathcal{A}_b$ by

$$C_{a,b} = \frac{\sum_{i=a}^{b} I_i}{b-a+1}.$$

When calculating the sequence conservation of a particular sRNA and mRNA, we included only sequences of species where homologs of both the sRNA and its target were found.

2.5 Determining conserved and accessible sRNA regions

Probabilistic alignment with `LocARNA-P` [37] gives column-wise reliabilities for sequence and base pair matches. These reliability profiles can be used to determine regions that are weakly structured and well-conserved in sequence. Given a multiple sRNA alignment $\mathcal{A}$, we first determined structure and sequence reliability background signals $\text{strrel}_{\mathcal{A}}^{bg}$ and $\text{seqrel}_{\mathcal{A}}^{bg}$ as average of the respective reliabilities over all alignment columns. Then, we identified windows of a fixed length n with an average sequence reliability $\text{seqrel}_{\mathcal{A}}^{win} \geq \gamma\ \text{seqrel}_{\mathcal{A}}^{bg}$ and an average structure reliability $\text{strrel}_{\mathcal{A}}^{win} \leq \delta\ \text{strrel}_{\mathcal{A}}^{bg}$. In this study, we used $\gamma = 1.0$ and $\delta = 0.8$. Each window satisfying these two conditions was considered as conserved accessible region.

2.6 Genome-wide target prediction

Genome-wide sRNA target predictions with `IntaRNA` [3] were performed with the following settings: minimal seed length of seven consecutive base pairs and local mRNA folding with a maximal base pair span of 70 in a folding window of 140 nt (parameters `-p 7 -w 140 -L 70`). Optionally, seed constraints were used as introduced below. All target regions (full 5'UTR and first 150 nt of CDS) with at least one identified orthologous gene were considered for the target search. Since the target sites of all experimentally confirmed interactions are located between positions -131 to +78 relative to the start codon, we filtered all predictions to be in range -150 to +100. The list of putative targets was then sorted by the `IntaRNA` energy score, which is the sum of hybridisation energy and opening energy of both interaction sites.

Genome-wide target predictions with `TargetRNA` [35] were performed with its default settings, but the search was restricted to the region -150 to +100 relative to the start codon. Furthermore, the p-value threshold was increased to obtain 100 target predictions per sRNA, which is the maximal number supported by the web server.

3 Results

3.1 Accessibility of interaction sites

The accessibility of the interaction sites in sRNAs and mRNAs was compared between the experimentally verified interactions and the non-interactions. For this purpose, we computed individually for each

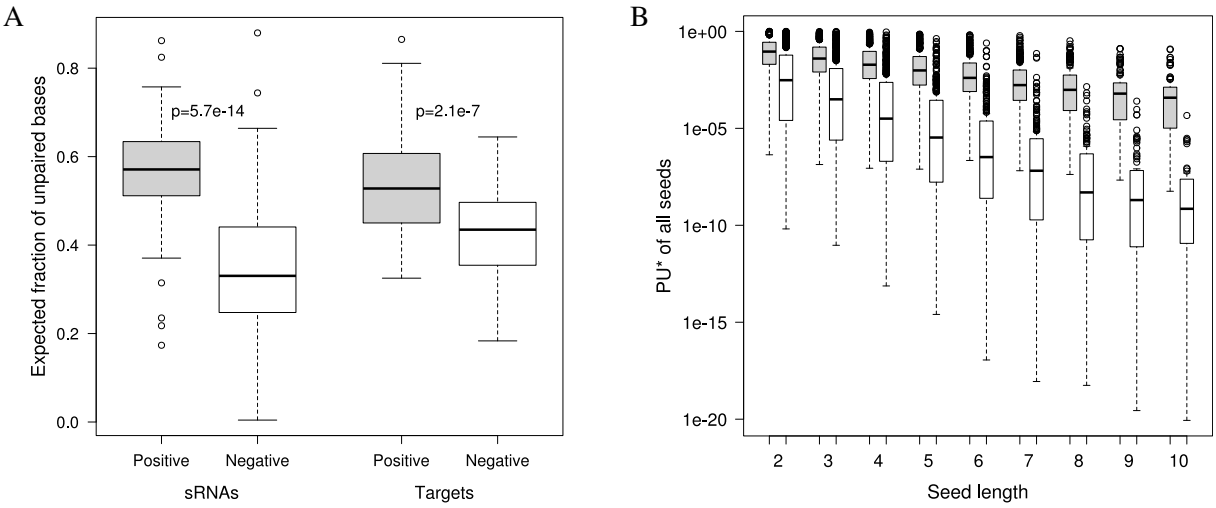

Figure 1: Comparison of the interaction site accessibility between interactions (grey) and non-interactions (white). The plots show the accessibility of (A) the whole interaction site and (B) the joint probability of being unpaired (PU^*) of all seeds of given length. Interaction sites in experimentally verified interactions are significantly more accessible than random regions. p-values were calculated by Wilcoxon rank sum test.

sRNA and its target mRNA(s) the expected fraction EF of bases unpaired at the interaction site (see Section 2.3).

As shown in Fig. 1A, the interaction sites of the experimentally verified interactions are more accessible than the random interaction regions of the negative data. This difference in accessibility is statistically significant both for sRNAs and targets (p-value of 5.7×10^{-14} and 2.1×10^{-7} for sRNAs and targets, respectively, calculated by Wilcoxon rank sum test).

These results motivated us to explore how the accessibility information of the interacting RNAs can be combined into a single feature that discriminates interactions from non-interactions. As elucidated in Section 2.3, the degree of single-strandedness of a specific RNA region can be evaluated by its PU value, which is the probability that the region is unpaired. To assess the accessibility of two interacting regions, we computed the joint unpaired probability PU^* of all perfectly matching sub-interactions of length two to ten. We then compared the joint unpaired probabilities of these interaction seeds between true interactions and non-interactions. For all analysed seed lengths, the PU^* of the positive data were significantly higher (Fig. 1B, $p < 4.8 \times 10^{-19}$ by Wilcoxon rank sum test).

Our target prediction tool `IntaRNA` predicts RNA-RNA interactions starting from an interaction seed, i.e., a region of (nearly) perfect sequence complementarity to facilitate interaction initiation. This region of initial pairing is often located at well-accessible structures such as hairpin loops [2]. Therefore, it seems reasonable to require for genome-wide sRNA target searches that the interaction seed is located at highly accessible regions in both RNAs, which can be ensured by only allowing seeds with a high PU^*. Since the background accessibility signal depends on sequence composition, e.g., `GC`-content, and folding parameters such as temperature and folding windows, we wanted to avoid fixed PU^* thresholds that define a valid seed. Instead, the PU^* cut-off is computed individually for each pair of RNA sequences as q-quantile of the sequences' background PU^* for a user-defined q.

We integrated into `IntaRNA` the feature of allowing only interaction seeds with a PU^* greater than the q-quantile of the background PU^* (which is computed from all subsequences of length equal to the seed). An evaluation of this feature is presented in Section 3.3.

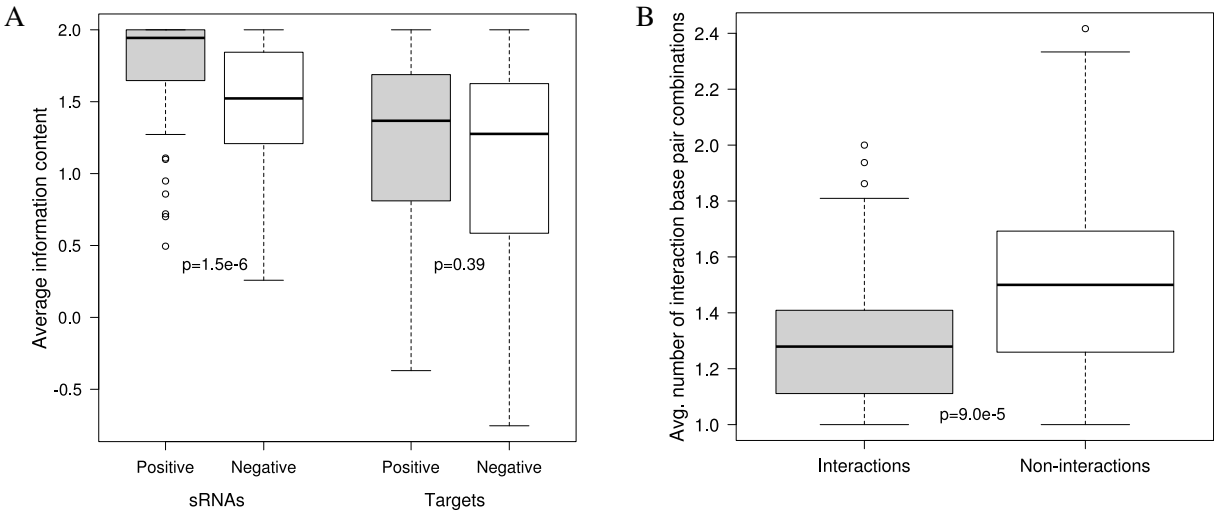

Figure 2: Comparison of (A) interaction site sequence conservation and (B) average number of different interaction base pairings between interactions (grey) and non-interactions (white). True interaction sites in sRNAs, but not in mRNAs, are significantly more conserved than random regions. Furthermore, the average number of intermolecular base pair combinations is significantly lower in the true interactions. p-values were calculated by Wilcoxon rank sum test.

3.2 Conservation of interaction sites

The sequence conservation of the sRNA and mRNA interaction sites was also compared between the positive and negative interaction data set. We assessed the conservation of each sRNA and mRNA interaction site by its average information content (see Section 2.4).

Fig. 2A shows that true sRNA interaction sites are significantly more conserved than random regions ($p = 1.5 \times 10^{-6}$ by Wilcoxon rank sum test). Intriguingly, the target sites exhibit no significant difference in sequence conservation ($p = 0.39$ by Wilcoxon rank sum test).

The missing sequence conservation in the targets, in contrast to the sRNAs, gives rise to the suspicion that conservation of sRNA-mRNA interactions among related bacterial species is not a general feature. However, despite lack of target sequence conservation, the intermolecular base pairings could still be preserved by consistent mutations in the target. In consistent mutations, only one of the two pairing bases changes, e.g., `A-U` mutates to `G-U` [17]. We checked this possibility by comparing the number of base pair types (out of the possible combinations `C-G`, `G-C`, `A-U`, `U-A`, `G-U` and `U-G`) per interaction position between positive and negative data to examine to which extent consistent or compensatory mutations occurred. The results in Fig. 2B show that there are even less different base pair types in the confirmed interactions than in the random data ($p = 9.0 \times 10^{-5}$ by Wilcoxon rank sum test). Hence, we can conclude that interactions between sRNAs and their targets are not structurally conserved in general.

3.3 Seed constraints in genome-wide target prediction

In the previous two sections, we showed that high interaction site accessibility and strong sRNA interaction site sequence conservation seem to be common features of bacterial sRNA-mRNA interactions. These observations suggest the following strategy to improve the false positive rate of sRNA target predictions: (i) identify conserved and weakly structured regions in the sRNA that might serve as target-binding region, and (ii) focus the target search to interactions that are located at such a region. A possibility to focus a target search to a specific region is to constrain the position of the interaction

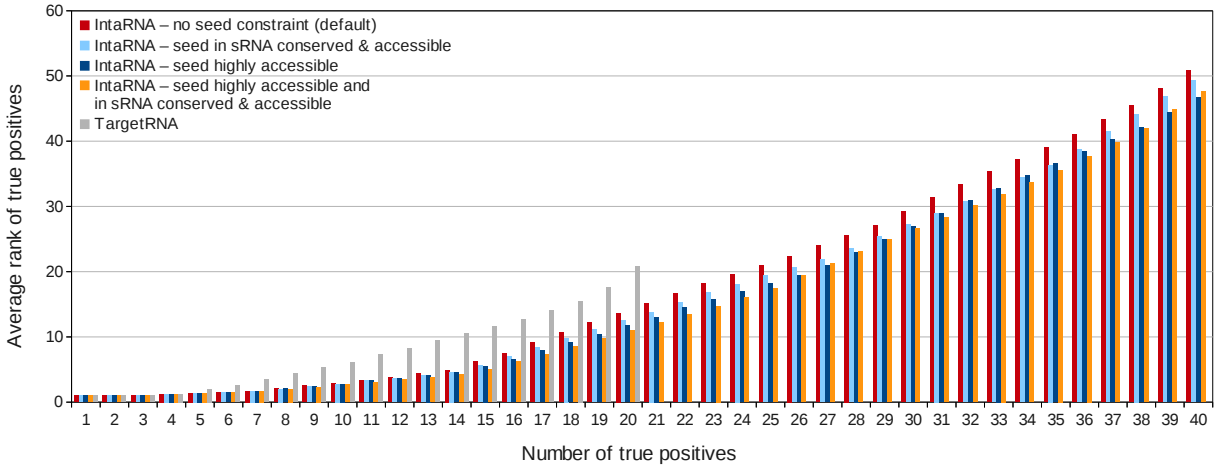

Figure 3: Performance of `IntaRNA` and `TargetRNA` in genome-wide target predictions for 25 sRNAs. `IntaRNA` was used with four different parameter settings. For each sRNA, the list of putative targets was sorted by (energy) score and the ranks of true positive predictions were determined. Subsequently, the ranks were sorted. The plot shows the average rank (y-axis) for the n best true positives (x-axis). The prediction performance of `IntaRNA` was best when requiring a seed region that is highly accessible in both interaction partners (dark blue and orange bars). Independent of the used parameter setting, `IntaRNA` clearly outperformed `TargetRNA` (grey bars).

seed region. Therefore, we extended the `IntaRNA` program by an option that allows to constrain the seed search to several user-definable regions. These candidate sRNA seed regions were derived from `LocARNA-P` reliability profiles using the approach described in Section 2.5 with window length equal to seed length followed by projecting the alignment window to the sRNA sequence of interest.

To evaluate the seed constraints introduced above and in Section 3.1, we conducted genome-wide target predictions in *E. coli* and *Salmonella* for every sRNA of our data set. Four different `IntaRNA` settings were used: (1) seed without accessibility and conservation constraints (default), (2) seed constraints derived from `LocARNA-P` reliability profile, (3) seed with PU^* in upper quartile (0.75-quantile) of background distribution, and (4) combining seed constraints of settings (2) and (3). Additionally, we present the prediction results of the widely used sRNA target prediction tool `TargetRNA` for comparison.

The bar chart in Fig. 3 shows the average rank of the true positive predictions vs. the number of true positives for all four `IntaRNA` settings and for `TargetRNA`. It can be seen that the correctly predicted targets rank best, i.e., lowest, in the two `IntaRNA` settings that required a seed region that is highly accessible in both interacting RNAs (dark blue and orange bars). Restricting the interaction seeds to conserved and weakly structured sRNA regions alone (light blue bars) gave a smaller improvement over the default method without constraining the seed region (red bars). For all parameter settings, `IntaRNA` clearly outperformed `TargetRNA` (grey bars), which returned only 20 true positive predictions in total.

It is expected that many more true targets can be found among the predictions that we considered as "false positives". For example, a very recent study by Sharma *et al.* identified thirteen additional targets of the GcvB sRNA that were not yet considered in our analysis [33]. Thus, both `IntaRNA` and `TargetRNA` should indeed perform better in predicting novel targets than observed in our experiments.

4 Discussion

In this study, we compiled a set of 71 sRNA-target pairs including 74 interaction sites. By comparing these interactions to a negative data set, we found that both sRNA and target interaction sites are highly accessible, and that the interaction sites in the sRNAs are additionally highly conserved. The overall interaction site accessibility in the targets was lower than in the sRNAs and the difference to the non-interactions was also less pronounced (even though still highly significant). This observation can be explained by the finding that structural RNAs, but not mRNAs, have in general lower folding energies than random RNAs of the same dinucleotide frequency [8, 9, 39]. Furthermore, prediction of local mRNA structures with a sliding window approach might be less reliable than prediction of global sRNA structures.

Our results on sRNA interaction site accessibility and conservation are in line with the results of a recent study by Peer and Margalit [26]. However, we could not observe a considerable reduction of false positives by narrowing down the search space to interactions in conserved accessible sRNA regions as suggested in their study. Instead, we found that the number of putative targets can be reduced more successfully by restricting the target search to interactions that contain a seed region that is highly accessible in both interaction partners. This finding supports the idea that target recognition is mediated by initial annealing of two well-accessible RNA regions, which form a strong duplex due to high sequence complementarity. Restricting the interactions seeds to highly accessible regions instead of conserved unstructured sRNA regions has also the advantage that it does not require the identification of sRNA candidate seed sites, e.g., by a probabilistic classifier or `LocARNA-P` reliability plots. Thus, our approach solely based on seed accessibility does not employ machine learning and does not depend on additional parameters apart from a cut-off relative to the background signal.

The comparison between verified interactions and non-interactions did not provide evidence that interaction sites in targets are generally conserved. This observation is somewhat contrary to miRNAs, where target site conservation is an often used feature for target prediction, e.g., [10]. Furthermore, our results also did not show an enrichment for compensating or consistent mutations in the interactions. Taken together, these observations indicate that the base pairing between sRNAs and their targets is not generally conserved across related species. Furthermore, our results suggest that comparative methods using covariance scoring does not appear to be very promising for the prediction of bacterial sRNA-mRNA interactions. The paucity of sequence covariation between sRNA and target (which is consistent with our recent findings in [31]) can be explained by the high evolutionary conservation of the sRNA interaction site and missing consistent mutations in the target.

The question remains why sRNA interaction sites exhibit a very high sequence conservation when neither target sites are sequentially conserved nor interactions are structurally conserved. A possible explanation for missing target site conservation are technical issues as misalignment of target sites or false positive prediction of orthologous genes. However, we could not find any indication of these when inspecting a subset of the analysed targets. Another possible explanation is that, for particular sRNAs, regulation of the target could be conserved, but not the interaction itself. In this case, the target site has been shifted to another position (which is also known as target site mobility). Finally, for sRNAs with multiple targets, conservation of regulation also does not have to be present for all target genes. Many sRNAs of our data set directly regulate multiple targets by binding via a single interaction site (although some sRNAs also use more than one site, e.g., FnrS, GcvB and Spot42). Consequently, conservation of the sRNA in distant species does not necessarily imply full conservation of all of its target genes or of the base pairing even if the target gene is conserved [14, 25, 28]. Instead, for a particular sRNA, only one or some particular targets out of multiple targets might be critical for the evolution of this sRNA and, thus, be linked to the evolutionary conservation of the sRNA interaction site [13].

Acknowledgements

We thank Sita Lange for inspiring discussions, especially on generation of negative data, Jens Georg for discussions on multiple target regulation and conservation, Sebastian Will for providing the latest version of `LocARNA-P`, and Dominic Rose for his valuable comments on the manuscript. We are further thankful to the anonymous reviewer for his comments on conservation of interactions.

This work was supported by the German Research Foundation (DFG grants BA 2168/2-2 SPP 1258 and BA 2168/3-1 to R.B.).

References

[1] S. H. Bernhart, U. Mückstein, and I. L. Hofacker. RNA Accessibility in cubic time. *Algorithms Mol Biol*, 6(1):3, 2011.

[2] C. Brunel, R. Marquet, P. Romby, and C. Ehresmann. RNA loop-loop interactions as dynamic functional motifs. *Biochimie*, 84(9):925–44, 2002.

[3] A. Busch, A. S. Richter, and R. Backofen. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–56, 2008.

[4] Y. Cao, J. Wu, Q. Liu, Y. Zhao, X. Ying, L. Cha, L. Wang, and W. Li. sRNATarBase: a comprehensive database of bacterial sRNA targets verified by experiments. *RNA*, 16(11):2051–7, 2010.

[5] Y. Cao, Y. Zhao, L. Cha, X. Ying, L. Wang, N. Shao, and W. Li. sRNATarget: a web server for prediction of bacterial sRNA targets. *Bioinformation*, 3(8):364–6, 2009.

[6] H. Chitsaz, R. Salari, S. C. Sahinalp, and R. Backofen. A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, 25(12):i365–73, 2009.

[7] B.-K. Cho, K. Zengler, Y. Qiu, Y. S. Park, E. M. Knight, C. L. Barrett, Y. Gao, and B. O. Palsson. The transcription unit architecture of the *Escherichia coli* genome. *Nat Biotechnol*, 27(11):1043–9, 2009.

[8] P. Clote, F. Ferre, E. Kranakis, and D. Krizanc. Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, 11(5):578–91, 2005.

[9] E. Freyhult, P. P. Gardner, and V. Moulton. A comparison of RNA folding measures. *BMC Bioinformatics*, 6:241, 2005.

[10] R. C. Friedman, K. K.-H. Farh, C. B. Burge, and D. P. Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res*, 19(1):92–105, 2009.

[11] P. P. Gardner, A. Wilm, and S. Washietl. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Research*, 33(8):2433–9, 2005.

[12] J. Gorodkin, L. J. Heyer, S. Brunak, and G. D. Stormo. Displaying the information contents of structural RNA alignments: the structure logos. *Comput Appl Biosci*, 13(6):583–6, 1997.

[13] S. Gottesman and G. Storz. Bacterial Small RNA Regulators: Versatile Roles and Rapidly Evolving Variations. *Cold Spring Harb Perspect Biol*, 2010. doi:10.1101/cshperspect.a003798.

[14] M. Guillier and S. Gottesman. The 5' end of two redundant sRNAs is involved in the regulation of multiple targets, including their own regulator. *Nucleic Acids Research*, 36(21):6781–94, 2008.

[15] J. Hertel, D. de Jong, M. Marz, D. Rose, H. Tafer, A. Tanzer, B. Schierwater, and P. F. Stadler. Non-coding RNA annotation of the genome of *Trichoplax adhaerens*. *Nucleic Acids Research*, 37(5):1602–15, 2009.

[16] M. Hiller, R. Pudimat, A. Busch, and R. Backofen. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Research*, 34(17):e117, 2006.

[17] I. L. Hofacker, M. Fekete, and P. F. Stadler. Secondary structure prediction for aligned RNA sequences. *Journal of Molecular Biology*, 319(5):1059–66, 2002.

[18] I. L. Hofacker, W. Fontana, P. F. Stadler, S. Bonhoeffer, M. Tacker, and P. Schuster. Fast folding and comparison of RNA secondary structures. *Monatshefte Chemie*, 125:167–188, 1994.

[19] F. W. D. Huang, J. Qin, C. M. Reidys, and P. F. Stadler. Partition function and base pairing probabilities for RNA-RNA interaction prediction. *Bioinformatics*, 25(20):2646–54, 2009.

[20] K. Katoh and H. Toh. Recent developments in the MAFFT multiple sequence alignment program. *Brief Bioinform*, 9(4):286–98, 2008.

[21] L. Li, C. J. J. Stoeckert, and D. S. Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9):2178–89, 2003.

[22] A. Mendoza-Vargas, L. Olvera, M. Olvera, R. Grande, L. Vega-Alvarado, B. Taboada, V. Jimenez-Jacinto, H. Salgado, K. Juárez, B. Contreras-Moreira, A. M. Huerta, J. Collado-Vides, and E. Morett. Genome-wide identification of transcription start sites, promoters and transcription factor binding sites in *E. coli*. *PLoS One*, 4(10):e7526, 2009.

[23] U. Mückstein, H. Tafer, J. Hackermüller, S. H. Bernhart, P. F. Stadler, and I. L. Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22(10):1177–82, 2006.

[24] K. Papenfort, M. Bouvier, F. Mika, C. M. Sharma, and J. Vogel. Evidence for an autonomous 5' target recognition domain in an Hfq-associated small RNA. *Proc. Natl. Acad. Sci. USA*, 107(47):20435–40, 2010.

[25] K. Papenfort, N. Said, T. Welsink, S. Lucchini, J. C. D. Hinton, and J. Vogel. Specific and pleiotropic patterns of mRNA regulation by ArcZ, a conserved, Hfq-dependent small RNA. *Mol Microbiol*, 74(1):139–58, 2009.

[26] A. Peer and H. Margalit. Accessibility and evolutionary conservation mark bacterial small-RNA target-binding regions. *J Bacteriol*, 193(7):1690–701, 2011.

[27] K. D. Pruitt, T. Tatusova, W. Klimke, and D. R. Maglott. NCBI Reference Sequences: current status, policy and new initiatives. *Nucleic Acids Research*, 37(Database issue):D32–6, 2009.

[28] J. B. Rice and C. K. Vanderpool. The small RNA SgrS controls sugar-phosphate accumulation by regulating multiple PTS genes. *Nucleic Acids Research*, 39(9):3806–19, 2011.

[29] A. S. Richter, C. Schleberger, R. Backofen, and C. Steglich. Seed-based IntaRNA prediction combined with GFP-reporter system identifies mRNA targets of the small RNA Yfr1. *Bioinformatics*, 26(1):1–5, 2010.

[30] R. Salari, M. Möhl, S. Will, S. Sahinalp, and R. Backofen. Time and space efficient RNA-RNA interaction prediction via sparse folding. In B. Berger, editor, *Proc. of RECOMB 2010*, volume 6044 of *Lecture Notes in Computer Science*, pages 473–490. Springer Berlin / Heidelberg, 2010.

[31] S. E. Seemann, A. S. Richter, T. Gesell, R. Backofen, and J. Gorodkin. PETcofold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. *Bioinformatics*, 27(2):211–219, 2011.

[32] S. E. Seemann, A. S. Richter, J. Gorodkin, and R. Backofen. Hierarchical folding of multiple sequence alignments for the prediction of structures and RNA-RNA interactions. *Algorithms Mol Biol*, 5:22, 2010.

[33] C. M. Sharma, K. Papenfort, S. R. Pernitzsch, H.-J. Mollenkopf, J. C. D. Hinton, and J. Vogel. Pervasive post-transcriptional control of genes involved in amino acid metabolism by the Hfq-dependent GcvB small RNA. *Mol Microbiol*, 2011. doi:10.1111/j.1365-2958.2011.07751.x.

[34] H. Tafer, F. Amman, F. Eggenhofer, P. F. Stadler, and I. L. Hofacker. Fast accessibility-based prediction of RNA-RNA interactions. *Bioinformatics*, 27(14):1934–40, 2011.

[35] B. Tjaden, S. S. Goodwin, J. A. Opdyke, M. Guillier, D. X. Fu, S. Gottesman, and G. Storz. Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Research*, 34(9):2791–802, 2006.

[36] L. S. Waters and G. Storz. Regulatory RNAs in bacteria. *Cell*, 136(4):615–28, 2009.

[37] S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, and R. Backofen. LocARNA-P: Accurate boundary prediction and improved detection of structured RNAs for genome-wide screens. 2011. Submitted.

[38] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen. Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Computational Biology*, 3(4):e65, 2007.

[39] C. Workman and A. Krogh. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Research*, 27(24):4816–22, 1999.