Computational analysis and prediction of RNA–RNA interactions

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.)

vorgelegt dem Rat
der Technischen Fakultät
der Albert-Ludwigs-Universität Freiburg

von Diplom-Bioinformatiker
Andreas S. Richter

Dekan:
Prof. Dr. Bernd Becker

Gutachter:
Prof. Dr. Rolf Backofen
Prof. Dr. Wolfgang R. Hess

Datum der Promotion:
11. Juli 2012

Abstract

In the past decade, there has been an explosion in the number of identified regulatory non-protein-coding RNAs (ncRNAs). Famous representatives are microRNAs in eukaryotes and small RNAs (sRNAs) in bacteria, which both act as post-transcriptional regulators by base pairing with messenger RNAs (mRNAs). For the functional characterisation of these base-pairing ncRNAs, an understanding of the principles that govern the RNA–RNA interaction formation is as essential as is the availability of large-scale target identification approaches. Therefore, I addressed these aspects by means of computational methods and analyses.

In the first part of my thesis, I present `IntaRNA`, a new fast and accurate method for the prediction of interactions between two RNA molecules. The underlying dynamic programming algorithm incorporates the structural accessibility of interaction sites and the existence of an interaction seed region. Both features have a significant impact on the strength of RNA–RNA interactions and thus on the functionality of mRNA target sites. `IntaRNA` was evaluated on a dataset of experimentally verified sRNA–mRNA interactions and achieved the highest accuracy, of all compared methods, in terms of sensitivity and positive predictive value. In a genome-wide target search, `IntaRNA` performed as well as the best existing method, but with considerably lower computing time and memory requirement. For several sRNAs, I was able to predict whether the target is positively or negatively regulated. `IntaRNA` is integrated in the Freiburg RNA Tools web server to offer RNA–RNA interaction prediction for specific RNAs as well as genome-wide target searches via an easy to use web-based interface.

In the second part, I present the findings of a systematic analysis of RNA–RNA interactions in which I determined features that discriminate functional from non-functional interactions and assessed the influence of these features on genome-wide target predictions. For this purpose, I compiled a set of 74 experimentally verified sRNA–target interactions and collected genome-wide full-length 5' untranslated regions. Surprisingly, I found that only interaction sites in sRNAs, but not in targets, displayed significant sequence conservation. The base pairing complementarity between sRNAs and their targets was not conserved in general across more distantly related species. In contrast to conservation, structural accessibility of functional interaction sites was significantly higher in both sRNAs and targets in comparison to non-functional sites. Based on these observations, I successfully improved the specificity of genome-wide target predictions by constraining interaction seeds to highly accessible regions in both RNAs or unstructured conserved sRNA regions. The findings on interaction site accessibility and single-stranded conserved seeds were additionally confirmed in a case study in the cyanobacterium *Prochlorococcus* MED4. By using an ultraconserved sequence motif of the sRNA Yfr1 as seed, two novel Yfr1 mRNA targets were predicted and subsequently experimentally confirmed.

In the third part, I present `PETcofold`, a comparative method for the prediction of interactions and secondary structures of two multiple alignments of RNA sequences. On the premise that each of the two RNAs is structurally conserved and that conservation of their RNA–RNA interaction implies conserved function, `PETcofold` accounts for covariance information arising from compensatory exchanges in intra- and intermolecular base pairs. Furthermore, the method can predict pseudoknots between intra- and intermolecular base pairs by employing a hierarchical folding strategy. The ability of `PETcofold` to predict RNA–RNA interactions was demonstrated on a carefully curated dataset of sRNAs and their target mRNAs. On phylogenetically simulated sequences enriched for covariance patterns at the interaction sites, `PETcofold` performed better with increasing amounts of covariance. For evaluation of both RNA–RNA interaction and structure prediction, I exemplified that the prediction is improved by the comparative approach in comparison to existing single sequence-based methods.

In summary, my thesis presents a new approach for fast and accurate RNA–RNA interaction prediction, one of the first systematic studies on accessibility and conservation of interacting sRNAs and mRNAs, the identification of two novel targets of the cyanobacterial sRNA Yfr1, and the first comparative method to predict joint secondary structures of two interacting RNAs.

Zusammenfassung

Im letzten Jahrzehnt stieg die Anzahl der bekannten nicht-protein-kodierenden RNAs (ncRNAs) mit regulatorischer Funktion explosionsartig an. Bedeutende Vertreter sind microRNAs in Eukaryoten und kleine RNAs (sRNAs) in Bakterien. RNAs beider Klassen fungieren als post-transkriptionale Regulatoren durch Basenpaarung mit Boten-RNAs (mRNAs). Das Verstehen der Prinzipien, die der Ausbildung solcher RNA–RNA-Interaktionen zugrunde liegen, und die Verfügbarkeit von Methoden, die die Identifikation von mRNA-Zielgenen im großen Maßstab erlauben, haben grundlegende Bedeutung für die funktionale Charakterisierung dieser ncRNAs. Aus diesem Grund widme ich mich in dieser Arbeit obigen Fragestellungen mit Hilfe von computergestützten Methoden und Analysen.

Im ersten Teil dieser Arbeit stelle ich `IntaRNA`, eine neue schnelle und genaue Methode zur Vorhersage von Interaktionen zwischen zwei RNA-Molekülen, vor. Der zugrunde liegende Algorithmus nutzt dynamische Programmierung und berücksichtigt die strukturelle Zugänglichkeit der Interaktionsstellen sowie das Vorhandensein einer Interaktions-Startregion. Beide Merkmale haben einen wesentlichen Einfluss auf die Stärke von RNA–RNA-Interaktionen und die Auswahl von Bindestellen in den Zielgenen. `IntaRNA` wurde anhand eines Datensatzes mit experimentell bestätigten sRNA–mRNA-Interaktionen evaluiert und erreichte die höchste Vorhersage-Genauigkeit unter allen verglichenen Methoden hinsichtlich Sensitivität und positiven Vorhersagewert (PPV). In einer genomweiten Suche nach Zielgenen arbeitete `IntaRNA` genauso gut wie die beste bisher verfügbare Methode, benötigte aber deutlich weniger Rechenzeit und Speicher. Für mehrere sRNAs war es möglich, vorherzusagen, ob das Zielgen positiv oder negativ reguliert wird. `IntaRNA` ist in einem Web-Server mit weiteren RNA-Strukturvorhersage-Programmen integriert und ermöglicht damit sowohl die Vorhersage von RNA–RNA-Interaktionen für einzelne RNAs als auch die genomweite Suche von Zielgenen über eine einfach zu bedienende web-basierte Benutzeroberfläche.

Im zweiten Teil habe ich charakteristische Merkmale, welche funktionale von nicht-funktionalen Interaktionen unterscheiden, bestimmt und habe den Einfluss dieser Merkmale auf genomweite Vorhersagen von Zielgenen bewertet. Die Analyse basierte auf einem umfassenden Datensatz mit 74 experimentell verifizierten sRNA–mRNA-Interaktionen und genomweiten 5' untranslatierten Bereichen. Überraschenderweise stellte sich heraus, dass nur die Interaktionsstellen in den sRNAs, aber nicht in den mRNAs, signifikant sequenzkonserviert sind. Die Basenpaar-Komplementarität zwischen sRNAs und ihre Zielgenen ist in der Regel nicht über weiter entfernt verwandte Arten erhalten. Im Gegensatz zur Konservierung war die strukturelle Zugänglichkeit der funktionalen Interaktionsstellen sowohl in den sRNAs als auch in den mRNAs signifikant höher als in den nicht-funktionalen Stellen. Auf der Grundlage dieser Beobachtungen konnte die Spezifität genomweiter Vorhersagen von Zielgenen verbessert werden, wenn die Interaktions-Startregionen auf sehr zugängliche Regionen in beiden RNAs oder unstrukturierte sequenzkonservierte Regionen in den sRNAs eingeschränkt wurden. Obige Erkenntnisse über strukturelle Zugänglichkeit und einzelsträngige konservierte Interaktions-Startregionen wurden zusätzlich in einer Fallstudie in dem Cyanobakterium *Prochlorococcus* MED4 bestätigt. Dabei wurde ein hoch konserviertes Sequenzmotiv in der sRNA Yfr1 als Interaktions-Startregion bei einer Vorhersage von mRNA-Zielgenen verwendet und zwei der vorhergesagten Zielgene wurden anschließend experimentell bestätigt.

Im dritten Teil stelle ich `PETcofold`, eine komparative Methode für die Vorhersage von Interaktionen und Sekundärstrukturen für zwei multiple Alignments von RNA-Sequenzen, vor. Unter der Voraussetzung, dass jede der beiden RNAs strukturell konserviert ist und dass die Konservierung ihrer RNA–RNA-Interaktion eine Konservierung der Funktion impliziert, berücksichtigt `PETcofold` die Kovarianz-Information aus kompensatorischen Austauschen in intra- und intermolekularen Basenpaaren. Ferner kann die Methode Pseudoknoten zwischen intra- und intermolekularen Basenpaaren durch Verwendung einer hierarchischen Faltungsstrategie vorhersagen. `PETcofold`s Eignung zur Vorhersage von RNA–RNA-Interaktionen wurde auf einem sorgfältig zusammengestellten Datensatz von sRNAs und ihren Zielgenen gezeigt. Für phylogenetisch simulierte Sequenzen, bei welchen die Kovarianz in den Interaktionsstellen angereichert wurde, konnte mit steigender Kovarianz auch das Vorhersageergebnis verbessert werden. Für die gemeinsame Vorhersage von Interaktionen und Strukturen wurde anhand einiger Beispiele gezeigt, dass die Vorhersagegenauigkeit durch den komparativen Ansatz im Vergleich zu bisherigen Verfahren basierend auf Einzelsequenzen verbessert wird.

Zusammenfassend gesagt wurde in meiner Arbeit ein neuer Ansatz zur schnellen und genauen Vorhersage von RNA–RNA-Interaktionen entwickelt, eine der ersten systematischen Studien über die Konservierung und strukturelle Zugänglichkeit von interagierenden sRNAs und mRNAs durchgeführt, die erste komparative Methode zur gemeinsamen Vorhersage von Sekundärstrukturen und Interaktionen zweier RNAs entwickelt, und es wurden zwei neue Zielgene, die von der cyanobakteriellen sRNA Yfr1 reguliert werden, identifiziert.

Mein besonderer Dank gilt ...

... meinem Doktorvater Rolf Backofen für die Möglichkeit, bei ihm über ein spannendes und ergiebiges Thema promovieren zu können, für seine Betreuung und die exzellenten Arbeitsbedingungen in seiner Arbeitsgruppe mit einer gelungenen Mischung aus zu übernehmender Verantwortung und wissenschaftlicher Freiheit, und für seine inspirierende und motivierende Art, mit der er mich für die Wissenschaft begeistert hat,

... meinem Zweitgutachter Wolfgang R. Hess für das Interesse an den Fragestellungen meiner Doktorarbeit und für die Bereitschaft, die Begutachtung zu übernehmen,

... meinen Kollegen für die Abwechslung beim Doktorandendasein mit Dart (und sonstigem „Sport“), Bier oder einfach nur Kaffee, für Antworten auf unzählige Fragen, die die Wissenschaft und das Leben so aufwirft, für Unterstützung und Tipps (und dabei insbesondere Christina, Daniel und Sita für das auszugsweise und Steffen für das vollständige Korrekturlesen dieser Arbeit), und für eine erlebnisreiche gemeinsame Zeit bei Reisen im In- und Ausland und in Freiburg,

... meinen Kooperationspartnern in nah und fern, insbesondere Claudia, Stefan, Tanja und Jan, für die Zusammenarbeit an spannenden Projekten, für den sich daraus ergebenden inspirierenden Blick über den eigenen Tellerrand und für die erzielten gemeinsamen Publikationen,

... allen Studenten, mit denen ich in den letzten Jahren zusammengearbeitet habe und die zum Gelingen dieser Arbeit beigetragen haben, insbesondere Dragoş, Mohamed und Cameron,

... all meinen Freunden, die mit mir in den letzten Jahren durch dick und dünn gegangen sind, sodass (leider) die Zeit gefühlt wie im Fluge vergangen ist, für die gemeinsamen Erlebnisse und den abwechslungsreichen Ausgleich von der Arbeit, und dafür, dass dadurch Freiburg innerhalb kurzer Zeit mein Zuhause geworden ist,

... meiner Familie für so Vieles, aber insbesondere für ihr Vorbild, ihre Liebe und für ihre Unterstützung mit Rat und Tat zu jeder Zeit!

List of own publications

This thesis is based on the following publications:

1. Richter, A. S. and Backofen, R. Accessibility and conservation: General features of bacterial small RNA–mRNA interactions? *RNA Biology*, 2012. In press.

2. Richter, A. S. and Backofen, R. Accessibility and conservation in bacterial small RNA–mRNA interactions and implications for genome-wide target predictions. In *Proceedings of the German Conference on Bioinformatics (GCB 2011)*, 2011.

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

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

5. Smith, C.[*], Heyne, S.[*], Richter, A. S.[*], Will, S.[*], and Backofen, R. Freiburg RNA Tools: a web server integrating `IntaRNA`, `ExpaRNA` and `LocARNA`. *Nucleic Acids Research*, 38(Web Server issue):W373–7, 2010.

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

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

Further publications:

8. Jäger, D., Pernitzsch, S. R., Richter, A. S., Backofen, R., Sharma, C. M., and Schmitz, R. A. An archaeal sRNA targeting *cis*- and *trans*-encoded mRNAs via two

[*] Joint first authors

distinct domains. Submitted.

9. Sonnleitner, E., Gonzalez, N., Sorger-Domenigg, T., Heeb, S., Richter, A. S., Backofen, R., Williams, P., Hüttenhofer, A., Haas, D., and Bläsi, U. The small RNA PhrS stimulates synthesis of the *Pseudomonas aeruginosa* quinolone signal. *Molecular Microbiology*, 80(4):868–85, 2011.

10. Schilling, D., Findeiß, S., Richter, A. S., Taylor, J. A., and Gerischer, U. The small RNA Aar in *Acinetobacter baylyi*: a putative regulator of amino acid metabolism. *Archives of Microbiology*, 192(9):691–702, 2010.

11. Richter, A. S., Will, S., and Backofen, R. A sampling approach for the exploration of biopolymer energy landscapes. In Kyriakopoulos, A., Michalke, B., Graebert, A., and Beschnidt, G., editors, *Proceedings of the European Conference on Metallobiolomics (HMI Berlin, Germany, 2007)*, pages 27–38. Herbert Utz Verlag, München, 2008.

Contents

Chapter 1

Introduction

Towards a modern RNA world

In 1952, James D. Watson sketched his hypothesis of a two-stage scheme for protein synthesis [61]:

$$\circlearrowright \text{DNA} \rightarrow \text{RNA} \rightarrow \text{protein.}$$

The hypothesis states that the genetic information of DNA serves as a template for the synthesis of RNA, which in turn serves as a template for protein synthesis.[1] For the following many years, research in molecular biology was predominantly focused on only three different types of RNA that all are involved in protein synthesis: (i) messenger RNA (mRNA), which carries the genetic information of the DNA to the protein synthesis machinery of the ribosome, (ii) ribosomal RNA (rRNA), which forms together with proteins the ribosome and (iii) transfer RNA (tRNA), which transfers the amino acids to the ribosome [45]. In the late 1960s and early 1970s, the first bacterial RNAs not directly involved in protein synthesis were found in the enterobacterium *Escherichia coli* (*E. coli*) [67, 74, 194]. The functional roles of these RNA molecules named 4.5S RNA and 6S RNA were, however, unveiled much later [141, 152, 202]. In the 1970s, another type of non-(protein-)coding RNA (ncRNA) was discovered in eukaryotes, but these small nuclear RNAs (snRNAs) initially drew the interest of only a few researchers [215]. Later, snRNAs turned out to be part of the spliceosome, which carries out mRNA splicing [103].

In 1981, many years before the discovery of microRNAs (miRNAs) and small interfering RNAs (siRNAs) in eukaryotes, regulatory RNAs that act by base pairing with other RNAs have already been studied in bacteria [205]. The about 108 nucleotide (nt) long RNA I binds to precursors of replication primers, which results in inhibition of primer formation

[1] Watson's information flow scheme should not be confused with the central dogma of molecular biology formulated by Francis H. C. Crick, which states that "... [sequential] information cannot be transferred from protein to either protein or nucleic acid" [36]. While the former is a positive statement saying that an information transfer from DNA to protein exists, the latter is a negative statement saying there is no transfer from protein.

for replication of plasmid ColE1 [184]. The about 90 nt long CopA RNA was found to be also involved in plasmid copy number control, but it binds to the leader region of *repA* mRNA and inhibits synthesis of the replication initiator protein RepA of plasmid R1 [177]. Both RNA I and CopA RNA were some of the first representatives of the class of *cis*-encoded antisense RNAs. In 1984, the chromosomally encoded RNA MicF was discovered in *E. coli* [125]. MicF binds to its *trans*-encoded target *ompF* and causes translational repression and degradation of the *ompF* mRNA.

These three examples of base-pairing RNAs were some of the first RNAs that were found to be directly involved in gene regulation instead of making proteins. In other words, they were pioneering for a modern RNA world full of regulatory RNAs.

Problem statement and contribution

Starting with four systematic computational screens for small RNAs (sRNAs) in *E. coli* in the years 2001 and 2002 [7, 27, 154, 203], there has been an explosion in the number of identified sRNAs throughout the bacterial domain of life. Recent studies based on high-throughput RNA sequencing (RNA-seq) reported about 140 sRNAs in the proteobacteria *E. coli* and *Salmonella enterica* serovar Typhimurium (*Salmonella*) [98, 147], and 314 candidate sRNAs (of which 166 were verified as significantly expressed) in the cyanobacterium *Synechocystis* PCC6803 [124]. It seems safe to predict that the broad availability of RNA-seq will continue to unveil a large number of yet unknown sRNAs in a wide range of bacterial species. Upon identification of novel sRNAs, the **key questions** are: what are the functions of these novel sRNAs, and which targets do they regulate?

As many sRNAs regulate mRNA targets by base pair interactions, the identification and validation of the interaction partners and the precise interaction sites are important tasks in the functional characterisation of sRNAs. Experimental transcriptomics and proteomics approaches for the analysis of target regulation are, however, costly and time-consuming. Therefore, they are often complemented by computational approaches to cope with the steadily increasing number of discovered but uncharacterised sRNAs [162].

In this thesis, I discuss RNA–RNA interactions with a focus on bacterial sRNAs that base pair with *trans*-encoded target mRNAs. The presented methods can, however, also be applied to other ncRNAs like eukaryotic miRNAs and siRNAs. I present

- two novel computational approaches for the identification of targets of base-pairing RNAs and for the prediction of the precise RNA–RNA interaction pattern,

- one of the first systematic analyses of the properties of functional sRNA–mRNA interactions to improve the specificity of genome-wide sRNA target predictions, and

- a case study in which two novel targets of the sRNA Yfr1 have been identified in the ecologically important cyanobacterium *Prochlorococcus* MED4.

Overview of this thesis

This introductory chapter starts with an overview on bacterial translation initiation, as most sRNAs characterised to date are involved in regulation of this process. Then, I give an introduction to bacterial regulatory RNAs followed by a review on the regulatory mechanisms employed by *trans*-encoded sRNAs. In the second chapter of this thesis, I present `IntaRNA`, an efficient method for the prediction of interactions between two RNAs. The third chapter presents my findings on different features of sRNA–target interactions and how genome-wide target predictions can be improved by integrating these features. In the fourth chapter, I present `PETcofold`, a method for comparative interaction prediction integrating evolutionary information from multiple alignments. The final chapter summarises this thesis and gives an outlook for future work.

Scientific research nowadays mostly involves collaborations, which is reflected in the general use of ”we” in scientific writing. As most of the work presented in this thesis was also carried out in collaboration with other scientists who contributed with discussions, ideas and wet-lab experiments, first person plural instead of first person singular is used in the further course of this thesis.

1.1 Initiation of bacterial protein synthesis

Proteins are synthesised during the process of translation by large ribonucleoprotein complexes, the ribosomes. The initiation of translation is a key process in the cell as it determines the synthesis rate of proteins from the mRNA. The regulation of translation, which often takes place at the level of initiation, therefore provides an important layer of gene expression control besides transcriptional regulation. While regulation on transcriptional level allows to efficiently control the expression of large sets of genes by a single transcription factor, translational regulation allows for a sensitive and graded (but not necessarily much faster) response to changing environmental conditions [82]. In the following, we will give an overview on bacterial translation initiation and its regulation (based on reviews in references [96, 101, 111, 168]).

In bacteria, mRNA transcription and translation are directly coupled. The translation can be initiated on the mRNA while being synthesised by the RNA polymerase as, in contrast to eukaryotes, no mRNA splicing occurs and both transcription and translation take place in the same cellular compartment. Additionally, the translating ribosomes protect the nascent mRNA from degradation by nucleases.

Ribosomes are composed of a small and a large subunit. The small ribosomal subunit has a sedimentation rate of 30S (S denotes Svedberg units, which measure the sedimentation rate in a centrifuge) and is formed by the approximately 1500 nt 16S rRNA and 21 ribosomal proteins. The large 50S ribosomal subunit is composed of the 120 nt 5S rRNA, the approximately 2900 nt 23S rRNA and about 34 ribosomal proteins. Both subunits are assembled to the complete 70S ribosome during translation initiation.

The initiation process is mediated by the three initiation protein factors IF1, IF2 and IF3. Initiation factor IF3 prevents association of the 30S and 50S ribosomal subunit during initiation. The interaction between the 30S subunit and the mRNA is promoted by the Shine–Dalgarno (SD) sequence [167], a sequence motif typically 4–5 nt in length and located around 7 ± 2 nt upstream of the start codon. The SD sequence is bound by a complementary motif at the 3' tail of the 16S rRNA called anti-SD sequence. The base pairing between the conserved anti-SD sequence and the complementary SD sequence helps to position the ribosome over the translation start codon, which is crucial for selecting the correct translational reading frame. The term ribosome binding site (RBS) denotes the mRNA region that is covered by the ribosome during translation initiation and, thereby, protected against RNase digestion. Footprinting experiments identified the mRNA region -35 to $+19$ relative to the start codon as the maximal region that is covered by the 30S ribosomal subunit [83]. During formation of the translation initiation process, initiation factor IF1 promotes the binding of IF2 and IF3 to the 30S subunit and ensures together with IF2 that the initiator tRNA carrying formylmethionine (fMet) is positioned correctly in the 30S subunit. After formation of an interaction between the start codon of the mRNA and the anticodon of the initiator fMet-tRNA, the stable 30S initiation complex associates with the 50S ribosomal subunit while GTP bound to IF2 is hydrolysed and the IFs are ejected. Subsequently, the translation elongation phase starts and the polypeptide chain is synthesised according to the codon sequence of the mRNA till the stop codon is reached and translation terminates. More than 80 percent of all genes in *E. coli* K-12 use `AUG` as start codon [17]. The remaining genes use the weaker start codons `GUG` or `UUG`, which result in less efficient translation due to weaker pairing between start codon and fMet-tRNA.

The protein coding region of an mRNA enclosed by start and stop codon is denoted open reading frame (ORF). The translation rate of an ORF depends on the strength of its SD sequence in terms of similarity to the consensus motif `5'-AGGAGG-3'`, the spacing between SD sequence and start codon and the strength of the start codon. Additionally, a pyrimidine-rich region upstream of the SD can compensate for absent or weak SD sequences by acting as recognition motif for ribosomal protein S1. There are also several examples of leaderless transcripts that have no or just a very short 5' untranslated region (UTR) and, therefore, lack ribosome recruitment motifs apart from the start codon.

Bacterial mRNAs are often polycistronic, i.e. they contain multiple ORFs and, thus, encode multiple proteins. These polycistronic mRNAs allow to regulate the translation of a downstream cistron by coupling to the translation of a preceding cistron. This translational coupling can be achieved by base pairing between the coding sequence (CDS) of the upstream cistron with the RBS of the downstream cistron. When the ribosome moves along the upstream cistron, the pairing is disrupted and the downstream cistron can be translated. Translational coupling in combination with an overlap between stop codon or 3' end of an ORF and start codon of the next ORF allows for a mechanism called reinitiation: after terminating translation of the upstream cistron, the ribosome is reused for translation of the downstream cistron and does not have to be recruited again. This mechanism also enables the expression of cistrons without SD sequence at a low level.

Regulation of bacterial translation often occurs at the level of initiation, which can be realised via mRNA primary sequence (e.g. start codon, SD sequence and pyrimidine-rich S1 binding sites) or by structurally blocking ribosome access to the RBS. The former control mechanism has already been discussed above. Changes in the accessibility of the RBS can involve either *cis*-acting elements in the mRNA like RNA thermosensors and riboswitches or *trans*-acting elements like proteins and sRNAs. Regulatory RNA elements and regulatory sRNAs are presented in detail in the next section. Translational repressor proteins typically compete with ribosomes for mRNA binding or bind the 5' UTR of mRNAs and induce mRNA secondary structures that sequester the RBS. These mechanisms are, for example, used by ribosomal proteins to autoregulate their own expression. When the available rRNA is saturated, some ribosomal proteins inhibit protein production in excess by binding to their own mRNA to repress translation. The polycistronic gene arrangement of many ribosomal proteins thereby permits that several downstream cistrons are coordinately regulated through binding of a single repressor protein.

1.2 Classes of bacterial regulatory RNAs

Bacteria utilise a multitude of diverse regulatory RNA molecules that are involved in various mechanisms to modulate a large variety of physiological responses (reviewed in references [66, 176, 205]). These bacterial regulatory RNAs can be divided into *cis*- and *trans*-acting elements. The former class encompasses regulatory elements like riboswitches and RNA thermosensors, which both are located in the 5' UTR of the mRNA that they regulate. RNAs acting in *trans*, i.e. on a different molecule, encompass sRNAs that bind to target mRNAs or to proteins, and CRISPR (clustered regularly interspaced short palindromic repeats) RNAs. The base pairing sRNAs can be further divided into antisense RNAs, which are encoded in *cis*, i.e. on the DNA strand opposite of their target gene, and sRNAs that bind *trans*-encoded target genes. The *cis*-encoded antisense RNAs (asRNAs) show extensive complementarity with their target mRNAs, whereas *trans*-encoded sRNAs show only limited and imperfect complementarity with their targets.

Riboswitches are regulatory mRNA elements that selectively sense changes in the concentrations of metabolites and inorganic ions, and respond by regulation of gene expression (reviewed in references [21, 22]). A typical riboswitch consists of a well-conserved aptamer domain that binds the ligand and of a regulatory expression platform that controls expression of the downstream-located gene. Ligand binding to the aptamer region induces the formation of an alternative conformation in the expression platform, which commonly modulates (i) transcription termination by formation or disruption of a transcription terminator stem, or (ii) translation initiation by changing ribosome access to the SD sequence or RBS. The most frequent riboswitch class binds the coenzyme thiamine pyrophosphate (TPP) and was shown to reduce expression of genes involved in thiamine biosynthesis by both inhibiting translation and interrupting transcription [123, 212]. In eukaryotes, TPP riboswitches are known to control alternative splicing [26]. Another, only recently discovered, class are fluoride riboswitches, which allow many bacterial and archaeal species to respond to toxic levels of fluoride anions by activating the expression of toxicity resistance proteins [10]. RNA thermosensors are similar to riboswitches as they also regulate genes by a structural rearrangement at the RBS, but in response to changing temperatures instead of ligand binding. Johansson et al. [86], for example, discovered an RNA thermosensor in *Listeria monocytogenes* that controls expression of virulence genes. At 30 °C, the 5' UTR of the virulence-activating transcription factor PrfA forms a secondary structure that masks the RBS. When temperature is increased to the host temperature of 37 °C, this inhibitory structure melts and translation can be initiated.

Bacterial asRNAs are generally transcribed from the antisense strand of an annotated transcriptional unit (reviewed in references [58, 66]). They overlap either the 5' or 3' end of the mRNA, or are located internally of the mRNA. Two examples of plasmid-encoded asRNAs that control plasmid copy number have already been presented in the beginning of this chapter. Several other characterised asRNAs act as antitoxins and repress the translation of mRNAs that encode toxic proteins. A classical example of such a toxin-antitoxin systems is the *hok/sok* locus of plasmid R1 in *E. coli* [59]. Another regulatory mechanism of asRNAs is directed cleavage of the target mRNA. The asRNA and its complementary target form a duplex, which can result in degradation of both RNAs or specific processing of the mRNA. A well-characterised example of co-degradation is provided by *isiA*/IsrR in the cyanobacterium *Synechocystis* PCC6803 [43]. The asRNA IsrR is encoded opposite of the central part of the photosynthesis gene *isiA*. While IsrR is constitutively transcribed, transcription of *isiA* is induced by iron, redox or light stress. Once *isiA* transcription becomes activated, the mRNA forms a duplex with its asRNA. As the duplex is immediately degraded, *isiA* mRNA can only accumulate when *isiA* titrates IsrR away.

The most extensively studied class of sRNAs base pair *trans*-encoded target mRNAs to post-transcriptionally regulate their translation or alter their stability. As the analyses and evaluations presented in this thesis are predominantly based upon these sRNAs, *trans*-encoded sRNAs are discussed in detail in the next section.

In the first part of this chapter, we introduced the small 4.5S and 6S RNAs. The 4.5S RNA is the RNA component of the signal recognition particle and is thus involved in housekeeping functions [141, 152]. In contrast, 6S RNA acts as a regulator and modulates the activity of the σ^{70}-containing RNA polymerase in *E. coli* [202]. The 6S RNA secondary structure mimics an open promotor that is bound by RNA polymerase, resulting in downregulation of transcription from a subset of σ^{70}-dependent promotors [176]. Another example of a protein-binding sRNA is CsrB, which modulates the activity of the RNA-binding protein CsrA in *E. coli* [205]. The CsrA protein binds to `GGA` motifs in stem–loops of its target mRNAs and regulates translation and stability of these targets. As CsrB sRNAs have multiple CsrA binding sites, they can sequester CsrA from its target mRNAs by direct competition for binding.

CRISPR RNAs are essential components of an RNA-based adaptive immune system in bacteria and archaea (reviewed in references [65, 209]). The CRISPR system uses short RNAs to detect and destruct foreign nucleic acids from invading viruses (bacteriophages in the case of bacteria) and plasmids. A CRISPR locus in a bacterial chromosome is defined by a cluster of 20–50 nt palindromic repeats separated by similarly sized unique spacers. The spacer sequences are often identical to bacteriophage or plasmid DNA and provide a genetic memory of previous invasions. New spacers are acquired by integration of invading DNA into the CRISPR locus. CRISPR loci are transcribed as long single RNAs and cleaved into short guide RNAs by CRISPR-associated (Cas) proteins or RNase III. These guide RNAs (denoted crRNAs for CRISPR-derived RNAs) then recognise invading DNA or RNA by base pairing, followed by destruction of the foreign nucleic acids.

1.3 Regulation by *trans*-encoded small RNAs

The most well-studied class of sRNAs base pair with *trans*-encoded target mRNAs (reviewed in references [66, 176, 205]). The interaction between sRNA and mRNA typically modulates the translation and the stability of the mRNA. For this reason, these sRNAs can be considered as functional analogues to eukaryotic miRNAs [205]. They differ, however, in their synthesis, their specific target regulation and other features. While miRNAs are, for example, only 21–24 nt in length [136], sRNAs are heterogeneous in size (typically 50–300 nt) as well as structure [135]. The interaction between the sRNA and its target mRNA is generally short and often imperfect, i.e. short stretches of complementarity are disrupted by mismatches. It has been shown for some examples that only a limited number of base pairs is required for effective target regulation: a core interaction of six consecutive base pairs (bp) has been reported as important for silencing of *ptsG* by SgrS [90], and two stretches of six and three consecutive bp are sufficient for regulation of *cirA* mRNA by OmrA/B sRNAs [70]. These short contiguous core interactions are often denoted as seed regions in analogy to miRNAs [66]. Interestingly, it has been shown that the sRNAs OmrA/B and RybB regulate multiple targets by a conserved 5’ sRNA domain, which is reminiscent of seed pairing and multiple target regulation of animal miRNAs [70, 135]. The interaction between sRNA and target mRNA is, in several bacteria, facilitated by

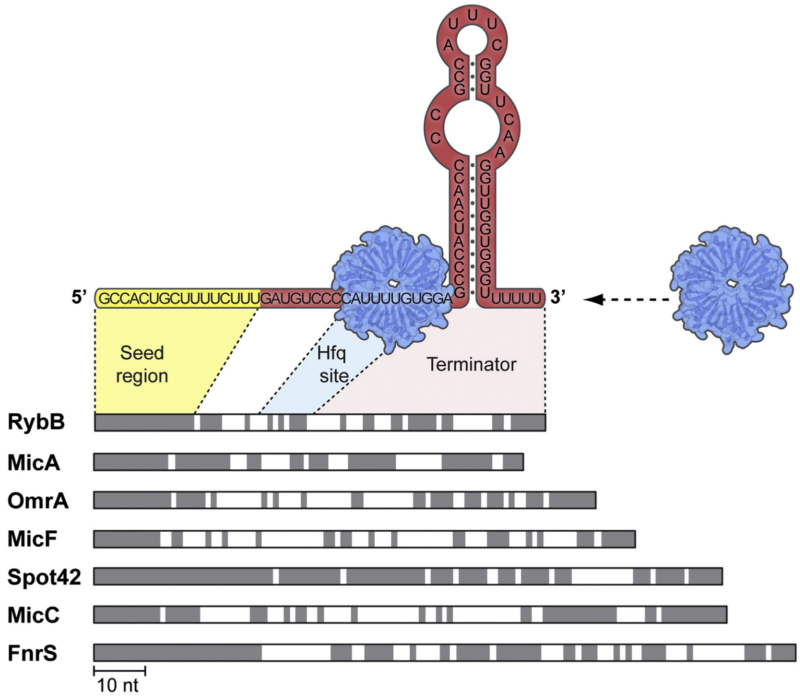


Figure 1.1. Typical structure of an Hfq-binding *trans*-encoded sRNA. The sRNA features a seed region that is involved in pairing to the target mRNA, a binding site for the Hfq protein and a Rho-independent transcription terminator stem followed by a poly(U) tail at the 3' end. In the seven enterobacterial sRNAs shown on the bottom, the regions with highest sequence conservation are shaded in grey. The figure is adapted from reference [176] with permission from Elsevier.

the RNA chaperone Hfq [87, 193]. It has been suggested that Hfq increases annealing rates, stabilises sRNA–mRNA duplexes and remodels RNA structure, but its specific contribution to interaction formation is still unknown. The general features of Hfq-dependent *trans*-encoded sRNAs are illustrated in Figure 1.1 [176]. They typically include one or more target-binding sites, a binding site for the RNA-binding protein Hfq and a stem–loop structure followed by a poly(U) tail at the sRNA 3' end to facilitate Rho-independent transcription termination.

The classical mechanism by which sRNAs regulate their target mRNAs is inhibition of translation. Often, the sRNA binds directly to the RBS, which prevents binding of the 30S ribosomal subunit to the mRNA and, thus, blocks translation initiation. After interaction formation, the sRNA–mRNA duplex is frequently degraded. The initiating 30S subunit was observed to cover maximally an mRNA region from positions −35 in the 5' UTR to +19 in the CDS [83]. The regulation of RybB sRNA, which represses its target *ompN* by pairing down to the fifth codon, can therefore still be explained by sequestration of the RBS [20]. A study by Holmqvist et al. [77] showed that the two sRNAs OmrA and OmrB translationally downregulate *csgD* by binding its 5' UTR at positions −79 to −61. The distance between interaction site and RBS precludes competition between sRNA and

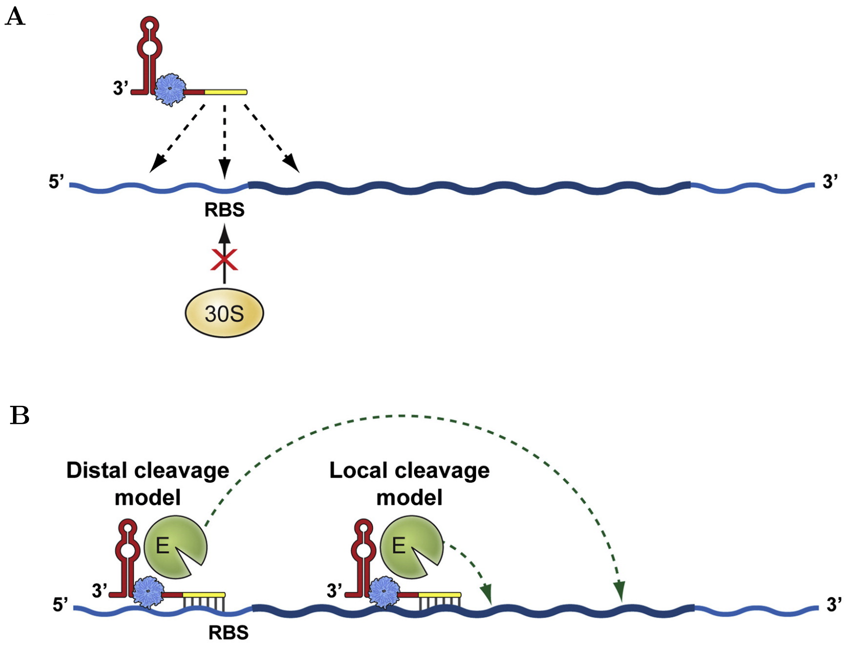


Figure 1.2. Regulatory mechanisms of *trans*-encoded sRNAs. **(A)** sRNA binding to a wide variety of mRNA positions can prevent binding of 30S ribosomal subunit to the mRNA, leading to translation inhibition. **(B)** sRNA binding can direct local or distal mRNA cleavage by RNase E. The figure is adapted from reference [176] with permission from Elsevier.

ribosome for mRNA binding. Although it was shown that the sRNA binding opens a local stem–loop structure in the 5' UTR of *csgD*, the precise mechanism of translation inhibition remains elusive for this example. The finding that sRNAs are able to block ribosome access by binding to diverse mRNA positions is illustrated in Figure 1.2A. Other sRNAs activate translation of their target mRNAs. In this case, the sRNA binds the 5' UTR and induces a structural change that opens an inhibitory structure at the RBS. Examples include the sRNA RyhB, which activates translation of *shiA* [142], and the sRNAs ArcZ, DsrA and RprA, which all activate the translation of *rpoS* [109, 110, 113].

More recently, it was discovered that sRNAs can also directly regulate the stability of their target genes instead of merely causing degradation of naked mRNA as secondary effect of repressed translation [198]. Pfeiffer et al. [140] showed that the sRNA MicC silences *ompD* via an interaction located around 70 nt downstream of the start codon in the CDS. The sRNA binding far downstream of the RBS excludes interference with translation initiation as regulatory mechanism. Instead, it was found that MicC promotes RNase E-dependent decay of *ompD* mRNA. Prévost et al. [143] showed that RyhB sRNA, which pairs its target *sodB* at the RBS, directs RNase E-dependent cleavage at a distal site that is located about 350 nt downstream from the interaction site. In summary, sRNA-induced cleavage can occur at sites adjacent to the sRNA–mRNA interaction as observed for *ompD* or at sites with large distance to the interaction as observed for *sodB*

(Figure 1.2B, [176]).

Chapter 2

IntaRNA: efficient prediction of RNA–RNA interactions

In this chapter, we start with an overview on important existing approaches for the prediction of RNA–RNA interactions. We concentrated on methods that are either general or specifically designed for bacterial sRNAs. Then, we present IntaRNA, a new fast and accurate method for the prediction of RNA–RNA interactions. In contrast to existing methods, IntaRNA incorporates both the existence of interaction seeds and the accessibility of interaction sites. Our tool was evaluated on a set of bacterial sRNAs on which it achieved the highest accuracy of all compared methods. Furthermore, we introduce a method to predict the mechanism of target regulation by the sRNA. Finally, we present a web server that allows to conduct RNA–RNA interaction predictions and genome-wide target predictions with IntaRNA via an easy to use web-based interface.

2.1 State-of-the-art prediction approaches

During the last decade, a multitude of regulatory and catalytic RNA molecules has been discovered, leading to a high demand for large-scale approaches that allow to characterise these novel ncRNAs and to assign them putative functions [3, 162, 166]. As ncRNAs in general and sRNAs in particular frequently interact with other RNAs [34, 205], a recent interest in the prediction of RNA–RNA interactions emerged. Currently, there exist four main classes of computational methods for predicting these interactions.

The first class includes methods that evaluate the stability of the duplex formed between two RNA molecules. Only base pairs involved in duplex formation are evaluated; the intramolecular structure of both RNA molecules is ignored. Algorithms based on this idea typically find the energetically most favourable hybridisation of two RNAs. The most popular tools representing this class are RNAhybrid [150], RNAduplex and RNAplex [178], and DINAMelt [41, 116]. RNAhybrid is primarily tailored for predicting potential miRNA binding sites in large target mRNAs. This method uses a modified version of the classical secondary structure prediction algorithm of Zuker and Stiegler [217] that omits multiloops. Furthermore, the loop size is restricted to a fixed value to reduce complexity. In principle, RNAduplex and RNAplex incorporate the same ideas as RNAhybrid. RNAplex, however, uses a simplified loop energy model, which gives a 10–27-fold improvement in runtime over RNAhybrid [178]. This initial version of RNAplex additionally introduced a fixed per-nucleotide penalty to favour short stable interactions, which made the tool also suitable for longer queries like sRNAs. Recently, RNAplex was extended by position-specific per-nucleotide penalties ([179], see below). Tjaden et al. [182] developed a tool named TargetRNA for the prediction of bacterial sRNA targets. TargetRNA provides two scoring schemes: (i) individual base pairs are scored by an extended version of the alignment algorithm of Smith and Waterman [170] or (ii) the minimum free energy of the duplex between the two RNA sequences is calculated similar to RNAhybrid. All above-mentioned tools have a time and space complexity of $O(n \cdot m)$ when restricting the size of loops, where n and m are the lengths of the two input sequences ($n > m$).

The second class of RNA–RNA interaction prediction methods determines a joint secondary structure of two RNAs, which can include both intra- and intermolecular base pairs. The two input RNA sequences are concatenated and the concatenation point is memorised. The concatenated sequence is then folded by an RNA folding algorithm, e.g. the algorithm of Zuker and Stiegler [217], in which the set of secondary structure elements is extended by a special loop that contains the linkage location of the two sequences. This special loop spanning the concatenation point is handled energetically as an external loop and not as a usual hairpin, internal or multiloop, which would result in an incorrect folding energy. The most prominent tools implementing the concatenation idea are PairFold [5] and RNAcofold [15]. They have a time complexity of $O((n+m)^3)$ and a space complexity of $O((n+m)^2)$ when restricting the size of loops. The sRNATarget web server is based on a machine learning approach that employs the simpler idea of concatenating a sRNA and an mRNA sequence by a short linker sequence to a single sequence [24, 214]. The minimum free energy (mfe) structure predicted by RNAfold [75] is then used to derive ten features like the distribution of secondary structure elements, length-normalised free energy, seed match length and A/U-content in single-stranded regions. Based on these features, a naive Bayes classifier was constructed to discriminate sRNA–mRNA interactions from non-interacting sRNAs and mRNAs. The main disadvantage of all tools based on the concatenation approach is their restriction on the set of interaction types. The utilised RNA folding algorithm can only predict secondary structures that are pseudoknot-free. Many interaction sites are, however, located in loop regions [23] and represent a pseudoknot in the context of the concatenated sequences. Consequently, this frequently occurring interaction type cannot be predicted by the concatenation approaches.

Representatives of the third class of interaction prediction methods allow for the prediction of complex interactions with a single interaction site. For eukaryotic miRNAs, it has been known for several years that the free energy of the miRNA–target duplex is a poor predictor for potential target sites [148] and several authors showed that the secondary structure of the target mRNA has a strong effect on target recognition [4, 91, 97, 106, 107, 161]. While the short miRNAs and siRNAs typically exhibit a high or even full complementarity to their targets, longer ncRNAs like bacterial sRNAs bind their target RNAs only partially. Therefore, the structure of the ncRNA should also be taken into account. The tool `RNAup` [127] calculates the thermodynamics of RNA–RNA interactions as the sum of two energy contributions: (i) the energy required to open the interaction sites and make them accessible, which is calculated from the partition function of the structural ensemble, and (ii) the hybridisation energy of the two interacting subsequences. The initial version of `RNAup` incorporated only the accessibility of the target site [127], but a more recent version also includes the accessibility of the interaction site in the binding RNA [128]. `RNAup` has a time complexity of $O(n^3 + n \cdot w^5)$ and a space complexity of $O(n^2 + n \cdot w^3)$ when restricting the size of loops to a fixed value and limiting the interaction length to w. In Tafer et al. [179], `RNAplex` was extended by a position-specific per-nucleotide penalty to approximate the competition between intra- and intermolecular base pairs. The position-dependent penalties are derived from accessibility profiles computed by `RNAplfold` [16] or `RNAup` [128]. This version of `RNAplex` has a time complexity of $O(n \cdot m + n^3)$ and a space complexity of $O(n \cdot m + n^2)$, where the second term refers to the computation of the accessibility profiles. The methods of this class are able to predict complex interactions like loop-loop interactions, but the interaction has to be restricted to one region. If an RNA–RNA interaction involves two or more sites as, e.g. OxyS–*fhlA* [6] and RNAIII–*rot* [18], which both form two kissing hairpin interactions, then only one of the sites can be predicted. It is, however, not clear whether the formation of multiple interactions is a common principle and whether two or more simultaneous interactions are frequently required for the regulatory function of the sRNA *in vivo*. For example, the sRNA RNAIII can form an imperfect duplex and a loop-loop interaction with its target mRNA *coa* in *Staphylococcus aureus* (*S. aureus*), but the kissing hairpin interaction is not essential for *in vivo* repression and contributes only moderately to the stability of the complex [29].

The final class of methods handles more complex joint secondary structures and allows for multiple interaction sites. The `IRIS` tool [138] predicts joint secondary structures with multiple interacting regions by maximising the number of base pairs. Alkan et al. [1] presented a more realistic energy model and showed the NP-completeness of the general RNA–RNA interaction prediction problem. Both methods have a time complexity of $O(n^3 \cdot m^3)$ and a space complexity of $O(n^2 \cdot m^2)$. Based on the type of joint structures considered by Alkan et al. [1], approaches were presented to compute the partition function of joint secondary structures [30, 79], to sample joint secondary structures [80] and to predict mfe structures [156]. All these approaches have a high time complexity of $O(n^6)$

and a space complexity of $O(n^4)$, which makes them not applicable for genome-wide scans. By sparsification of the dynamic programming matrices, a linear improvement in time and space complexity can be achieved on average for predicting mfe structures [157]. Although the methods of this class are based on a rather general interaction model, all of them except for `IRIS` still do not allow pseudoknotted structures or crossing interactions. They can, thus, not predict instances as the two kissing hairpin interaction between the sRNA RNAIII and its target *rot* in *S. aureus*, as the two loop-loop interactions constitute a crossing interaction [18].

In the following, we present `IntaRNA`, a new general approach for the prediction of **int**er**a**cting **RNA**s that includes interaction site accessibility and user-definable seed regions. `IntaRNA` calculates a combined interaction energy score as the sum of the free energy of hybridisation and the free energy required for making the interaction sites accessible. We present two variants: (i) a complete approach with a time complexity of $O(n^2 \cdot m^2 + n^3)$ and a space complexity of $O(n \cdot m + n^2)$ when restricting the size of loops, and (ii) a heuristic simplification of the complete approach with a time complexity of $O(n \cdot m + n^3)$ and a space complexity of $O(n \cdot m + n^2)$, where the second term in the complexity always refers to the time and space required for accessibility computation. We successfully applied `IntaRNA` to the genome-wide prediction of sRNA targets and accurately predicted the precise interaction between sRNA and mRNA.

2.2 Ensemble-based model for interaction prediction including accessibility

2.2.1 Combining hybridisation energy and interaction site accessibility

Let s^1 and s^2 be two potentially interacting RNA sequences of lengths n and m, respectively. We use the convention to number the first RNA sequence in 5' → 3' direction and the second RNA sequence in the reverse direction. The first component that determines the quality of an RNA–RNA interaction between the subsequence $s^1_i \ldots s^1_k$ of s^1 and the subsequence $s^2_j \ldots s^2_l$ of s^2 is the **hybridisation energy** $H(i, j, k, l)$. Its calculation is based on the algorithm of `RNAhybrid` [150] using the nearest neighbour energy model with the energy parameters of Mathews et al. [119]. $H(i, j)$ denotes the hybridisation energy of the best interaction, i.e. the hybridisation minimum free energy, of subsequences $s^1_i \ldots s^1_n$ and $s^2_j \ldots s^2_m$, where the leftmost positions of both subsequences, i and j, form a base pair (i, j). $H(i, j)$ can be calculated using a modified variant of the algorithm of Zuker and Stiegler [217] discarding multiloop structures. The algorithm has a time and space complexity of $O(n \cdot m)$ when restricting the length of loops. An entry $H(i, j)$ of the matrix

H is computed by the following recursion:

$$H(i,j) = \begin{cases} \min \left\{ \begin{array}{l} \min\limits_{p,q} \left\{ E^{\text{loop}}(i,j,p,q) + H(p,q) \right\} \\ E^{\text{dangle}}_{3'}(i,j,i+1) \\ E^{\text{dangle}}_{5'}(i,j,j+1) \\ E^{\text{term}}_{\text{mm}}(i,j,i+1,j+1) \\ 0 \end{array} \right\} & \text{if } (s^1_i, s^2_j) \text{ can pair,} \\ \infty & \text{otherwise,} \end{cases} \tag{2.1}$$

where the terms $E^{\text{dangle}}_{3'}(i,j,d_3)$ and $E^{\text{dangle}}_{5'}(i,j,d_5)$ denote the dangling end energy contributions of the unpaired nucleotides at sequence positions d_3 and d_5, which are 3' and 5' adjacent to base pair (i,j), respectively. The term $E^{\text{term}}_{\text{mm}}(i,j,d_3,d_5)$ denotes the energy contribution of the terminal mismatch (d_3,d_5), i.e. the non-canonical pair, adjacent to base pair (i,j). $E^{\text{loop}}(i,j,k,l)$ denotes the free energy of the loop enclosed by the left base pair (i,j) and the right base pair (k,l).

A loop enclosed by base pairs (i,j) and (k,l) can be any of the secondary structure elements stacked pair, bulge or internal loop. Its free energy is given by

$$E^{\text{loop}}(i,j,k,l) = \begin{cases} \mathit{stacked\_pair}(i,j,k,l) & \text{if } k-i=1 \text{ and } l-j=1, \\ \mathit{bulge}(i,j,k,l) & \text{if } 1 \leq k-i-1 \leq 16 \text{ and } l-j=1, \\ \mathit{bulge}(i,j,k,l) & \text{if } k-i=1 \text{ and } 1 \leq l-j-1 \leq 16, \\ \mathit{internal\_loop}(i,j,k,l) & \text{if } 1 \leq k-i-1 \leq 16 \text{ and } 1 \leq l-j-1 \leq 16, \\ \infty & \text{otherwise,} \end{cases}$$

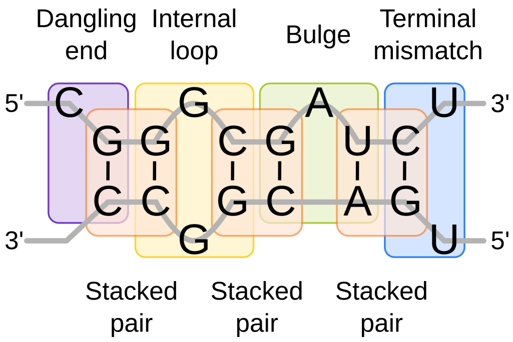
where $\mathit{stacked\_pair}(i,j,k,l)$, $\mathit{bulge}(i,j,k,l)$ and $\mathit{internal\_loop}(i,j,k,l)$ are the sequence-dependent free energy parameters of the respective secondary structure elements. Figure 2.1 gives an example interaction and illustrates the calculation of the hybridisation free energy from the free energy parameters of the different secondary structure elements.

The hybridisation minimum free energy of the full sequences s^1 and s^2 is calculated from

$$\min_{1 \leq i < n, 1 \leq j < m} \left\{ \begin{array}{l} H(i,j) \\ H(i,j) + E^{\text{dangle}}_{5'}(i,j,i-1) \\ H(i,j) + E^{\text{dangle}}_{3'}(i,j,j-1) \\ H(i,j) + E^{\text{term}}_{\text{mm}}(i,j,j-1,i-1) \\ 0 \end{array} \right\}.$$

The actual hybridisation pattern, i.e. the nucleotides involved in base pairing, is calculated by traceback in the dynamic programming matrix.

The second component contributing to the quality of an RNA–RNA interaction is the



Structural element	Nucleotides	Free energy [kcal/mol]
Dangling end	(C,GC)	-0.3
Stacked pair	(GC,GC)	-3.3
Internal loop	(GC,G|G,CG)	-2.1
Stacked pair	(CG,GC)	-2.4
Bulge	(GC,A,UA)	$+1.6$
Stacked pair	(UA,CG)	-2.4
Terminal mismatch	(CG,UU)	-1.2
Intermolecular initiation		$+4.1$
Overall free energy	$\sum$	-6.0

Figure 2.1. Example of an interaction formed by two short RNA sequences to illustrate the nearest neighbour (energy) model. The table lists the sequence-dependent energy parameters of the individual secondary structure elements that are highlighted in the RNA–RNA interaction. The overall folding free energy of the depicted interaction is the sum of the energy contributions of all secondary structure elements plus the intermolecular initiation energy. The energy parameters are given according to Mathews et al. [119] and were retrieved from the Nearest Neighbor Database (NNDB) [186].

accessibility of the interaction site, i.e. the subsequence participating in the hybridisation, in each sequence. The interaction site accessibility corresponds to the energy that is required to make the site single-stranded. For a given RNA sequence s, it is defined as the difference between the energy of the ensemble of all structures that can be formed by s and the energy of the ensemble of structures, in which the interaction site $s_i \dots s_k$ is single-stranded. This energy difference is denoted by $ED(i,k)$ and can be calculated using a partition function approach [121]. Let $\mathcal{S}$ be the set of all structures (called **ensemble**) that can be formed by a sequence s. The partition function of the ensemble $\mathcal{S}$ is defined by

$$Z = \sum_{Q \in \mathcal{S}} e^{-\frac{E(Q)}{RT}},$$

where $E(Q)$ is the free energy of a particular secondary structure Q formed by sequence s, R is the gas constant and T is the temperature. The free energy of the ensemble $\mathcal{S}$ is

$$E^{\text{ens}}(\mathcal{S}) = -RT \ln Z.$$

Let $\mathcal{S}^{\text{unpaired}}_{i,k}$ be the set of all structures of s that have nucleotides $s_i, s_{i+1}, \dots, s_k$ unpaired. Then,

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

which is greater than or equal to zero by definition. The probability $PU(i,k)$ that the complete region between positions i and k is unpaired can be calculated by

$$\begin{aligned} PU(i,k) &= \frac{Z_{i,k}^{\text{unpaired}}}{Z} \\ &= \frac{\sum\limits_{Q \in \mathcal{S}_{i,k}^{\text{unpaired}}} e^{-\frac{E(Q)}{RT}}}{\sum\limits_{Q \in \mathcal{S}} e^{-\frac{E(Q)}{RT}}} \\ &= \frac{e^{-\frac{E^{\text{ens}}(\mathcal{S}_{i,k}^{\text{unpaired}})}{RT}}}{e^{-\frac{E^{\text{ens}}(\mathcal{S})}{RT}}} \quad = e^{-\frac{E^{\text{ens}}(\mathcal{S}_{i,k}^{\text{unpaired}}) - E^{\text{ens}}(\mathcal{S})}{RT}} \\ &= e^{-\frac{ED(i,k)}{RT}}, \end{aligned} \tag{2.2}$$

where $Z_{i,k}^{\text{unpaired}}$ is the partition function of the ensemble $\mathcal{S}_{i,k}^{\text{unpaired}}$ [127].

The accessibilities of all sequence intervals of a given RNA sequence can be obtained by `RNAplfold` [16] or `RNAup` [128], each with parameter `-u`, in $O(n^3)$ time and $O(n^2)$ space. `RNAplfold` folds RNA sequences locally using a sliding window approach and allows only structures with a given maximal base pair span; this approach was shown to be appropriate for mRNA sequences [100]. `RNAplfold` returns the mean probability $\overline{PU}(i,k)$ that the region between i and k is unpaired. Their logarithm $-RT \ln(\overline{PU}(i,k))$ is not exactly equivalent to the mean energy difference $\overline{ED}(i,k)$, but can be used as an approximation. `RNAup` predicts, in contrast to `RNAplfold`, global RNA structures and returns accessibilities in terms of ED values.

Next, the two energy contributions presented above are combined. The **extended hybridisation energy** of two interacting subsequences is defined as the sum of their hybridisation energy and ED values. For calculation of the ED values, the first and the last interacting position in both sequences must be known. Hence, the basic recursion for calculating the extended hybridisation energy requires a four-dimensional matrix C. The extended hybridisation energy $C(i,j,k,l)$ of a specific hybridisation between the subsequences $s_i^1 \dots s_k^1$ and $s_j^2 \dots s_l^2$ is defined by

$$C(i,j,k,l) = \begin{cases} \min \left\{ \begin{array}{l} H(i,j,k,l) + ED_1(i,k^*) + ED_2(j,l^*) \\ H(i,j,k,l) + E_{5'}^{\text{dangle}}(i,j,i-1) \\ +ED_1(i-1,k^*) + ED_2(j,l^*) \\ H(i,j,k,l) + E_{3'}^{\text{dangle}}(i,j,j-1) \\ +ED_1(i,k^*) + ED_2(j-1,l^*) \\ H(i,j,k,l) + E_{\text{mm}}^{\text{term}}(i,j,j-1,i-1) \\ +ED_1(i-1,k^*) + ED_2(j-1,l^*) \end{array} \right\} & \begin{array}{l}\text{if } (s_i^1, s_j^2) \text{ and } (s_k^1, s_l^2) \\ \text{can pair,}\end{array} \\ \infty & \text{otherwise,} \end{cases} \tag{2.3}$$

$$C^{k,l}(i,j) = \min E$$

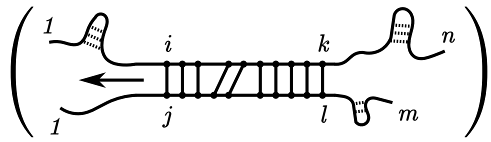


Figure 2.2. Interpretation of the matrix entry $C^{k,l}(i,j)$. The upper RNA sequence is of length n and the lower RNA sequence is of length m. The hybridisation starts with the rightmost base pair (k,l), is extended to the left and ends with the leftmost base pair (i,j).

where ED_1 and ED_2 denote the ED values of the sequences s^1 and s^2, respectively. The position k^* refers to sequence positions k or $k+1$, depending on whether the dangling end or terminal mismatch energy of the nucleotide adjacent to the rightmost base pair contributes to the optimal hybridisation energy $H(i,j,k,l)$ in Equation (2.4) or not. In the former case, nucleotide $k+1$ is required to be unpaired. Analogously, l^* refers to l or $l+1$. $H(i,j,k,l)$ is the four-dimensional variant of the matrix defined in Equation (2.1):

$$H(i,j,k,l) = \begin{cases} \min\limits_{p,q} \left\{ E^{\text{loop}}(i,j,p,q) + H(p,q,k,l) \right\} & \text{if } (s^1_i, s^2_j) \text{ and } (s^1_k, s^2_l) \text{ can pair,} \\ & i \neq k \text{ and } j \neq l, \\ \min \left\{ \begin{array}{l} E^{\text{dangle}}_{3'}(i,j,i+1) \\ E^{\text{dangle}}_{5'}(i,j,j+1) \\ E^{\text{term}}_{\text{mm}}(i,j,i+1,j+1) \\ 0 \end{array} \right\} & \text{if } (s^1_i, s^2_j) \text{ can pair, } i = k \text{ and } j = l, \\ \infty & \text{otherwise.} \end{cases} \tag{2.4}$$

This approach has a complexity of $O(n^2 \cdot m^2)$ time and $O(n^2 \cdot m^2)$ space when limiting the size of the loops. Limiting the interaction length to w (as done in RNAup) reduces the complexity to $O(n{\cdot}m{\cdot}w^2)$ time and $O(n{\cdot}m{\cdot}w^2)$ space. While such a restriction is reasonable for short RNAs as miRNAs (by just setting w to the miRNA length), it is problematic for long sRNAs as their expected interaction length is not known in advance. In the following, we show how to improve both time and space complexity without restricting the interaction length and additionally integrating the requirement for a seed region. The energy contributions of dangling ends and terminal mismatches will be disregarded for the purpose of simplification.

2.2.2 Reducing the space complexity

The space complexity of the recursion derived from the combination of Equations (2.3) and (2.4) can be improved by calculating all interactions for a common interaction start in one step. This leads to a two-dimensional matrix $C^{k,l}$, for which an entry $C^{k,l}(i,j)$ is basically a slice of $C(i,j,k,l)$ for fixed k and l. The hybridisation that starts with base pair (k,l) is elongated to the left and ends with the leftmost base pair (i,j) (Figure 2.2).

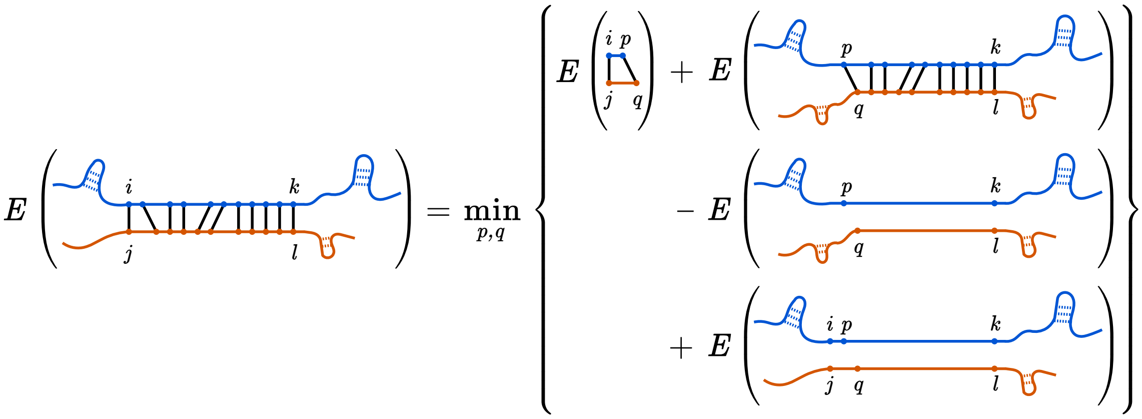


Figure 2.3. Visualisation of the recursion for calculating the matrix entry $C^{k,l}(i,j)$. The hybridisation between the mRNA and the sRNA is shown in black, while the energy required to make the mRNA and the sRNA interaction site accessible (ED_1 and ED_2) is indicated in blue and orange, respectively. Since ED values are not additive, i.e. $ED_1(i,k) \neq ED_1(i,p) + ED_1(p,k)$, we need to subtract $ED_1(p,k)$ and $ED_2(q,l)$, and add $ED_1(i,k)$ and $ED_2(j,l)$ to get the final value for $C^{k,l}(i,j)$.

The matrix $C^{k,l}$ can be filled according to the following recursion:

$$C^{k,l}(i,j) = \begin{cases} \min\limits_{p,q} \left\{ \begin{array}{l} E^{\text{loop}}(i,j,p,q) + C^{k,l}(p,q) \\ -ED_1(p,k) - ED_2(q,l) \\ +ED_1(i,k) + ED_2(j,l) \end{array} \right\} & \begin{array}{l}\text{if } (s^1_i, s^2_j) \text{ and } (s^1_k, s^2_l) \text{ can pair,} \\ i \neq k \text{ and } j \neq l,\end{array} \\ \infty & \text{otherwise.} \end{cases} \tag{2.5}$$

For the initial case,

$$C^{k,l}(k,l) = \begin{cases} ED_1(k,k) + ED_2(l,l) & \text{if } (s^1_k, s^2_l) \text{ can pair,} \\ \infty & \text{otherwise.} \end{cases} \tag{2.6}$$

The recursion for $C^{k,l}(i,j)$ as given in Equation (2.5) includes accessibility contributions in the form of ED values. Since ED values are not additive, previous ED values have to be subtracted and new values have to be added when extending the hybridisation. The idea of the recursion for calculating $C^{k,l}(i,j)$ is illustrated in Figure 2.3.

Finally, only a two-dimensional matrix C is required to store for all left-end base pairs (i,j) the best energy score found so far for all rightmost base pairs (k,l) with $i \leq k$ and $j \leq l$. This matrix is defined by

$$C(i,j) = \min_{k,l} \left\{ C^{k,l}(i,j) \right\}. \tag{2.7}$$

Computing $C^{k,l}(i,j)$ for all (k,l) first and finding the minimal value $C(i,j)$ afterwards gives still a time and space complexity of $O(n^2 \cdot m^2)$ for filling the full matrices. To reduce the space complexity, $C(i,j)$ can instead be updated successively after evaluation of each right-end base pair (k,l) and the matrix $C^{k,l}$ can be reused:

$$C(i,j) = \min\left\{C(i,j),\ C^{k,l}(i,j)\right\}. \tag{2.8}$$

This approach results in an $O(n^2 \cdot m^2)$ time algorithm that requires only $O(n \cdot m)$ space for computing the matrix C.

2.2.3 Incorporation of seed features

A feature that is commonly observed in RNA–RNA interactions is the presence of a seed region, i.e. an initial interaction region of (nearly) perfect sequence complementarity. We introduce the following **seed features** that define the properties of a seed region:

- P: number of perfectly paired bases in the seed region,
- $b^{\max}$, $b_m^{\max}$ and $b_s^{\max}$: maximal number of unpaired nucleotides in the seed region in both RNAs, in the mRNA and in the sRNA, respectively,
- $[a_s, b_s]$: optional constraint of the seed location to sRNA region $s^2_{a_s} \ldots s^2_{b_s}$.

The seed features are a variable part of our algorithm and can be specified by the user. By default, we require only the existence of a single seed region at any position in the two interacting sequences. Certain ncRNAs, however, have a preferred seed position, e.g. the 5' end in the case of miRNAs and some sRNAs as RybB or OmrA/B [12, 70, 135]. Therefore, the user can optionally constrain the seed location in the ncRNA sequence to a specific region.

Seed regions are integrated in the IntaRNA algorithm by an additional matrix *seed*. An entry *seed*$(i,j,k,l;P')$ contains the minimal free energy of a hybridisation between the subsequences $s^1_i \ldots s^1_k$ and $s^2_j \ldots s^2_l$ that includes exactly P' base pairs. For given i, j, k, l and P', the numbers of unpaired nucleotides in the mRNA and the sRNA are fixed to $k - i + 1 - P'$ and $l - j + 1 - P'$, respectively. The matrix *seed* is defined by

$$seed(i,j,k,l;P') = \begin{cases} \min\limits_{\substack{p,q \text{ with} \\ k-p+1 \geq P'-1 \\ l-q+1 \geq P'-1}} \left\{ \begin{array}{l} E^{\text{loop}}(i,j,p,q) \\ +seed(p,q,k,l;P'-1) \end{array} \right\} & \begin{array}{l} \text{if } (s^1_i, s^2_j) \text{ and } (s^1_k, s^2_l) \\ \text{can pair and} \\ 2 < P' \leq P, \end{array} \\ E^{\text{loop}}(i,j,k,l) & \begin{array}{l} \text{if } (s^1_i, s^2_j) \text{ and } (s^1_k, s^2_l) \\ \text{can pair and } P' = 2, \end{array} \\ \infty & \text{otherwise.} \end{cases}$$

The conditions $k - p + 1 \geq P' - 1$ and $l - q + 1 \geq P' - 1$ ensure that $P' - 1$ base pairs can be formed between $s^1_p \ldots s^1_k$ and $s^2_q \ldots s^2_l$, respectively. Let $l_m = k - i + 1$ and $l_s = l - j + 1$

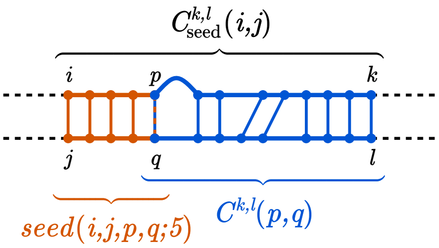


Figure 2.4. The matrix $C^{k,l}_{\text{seed}}$ and its relation to the other matrices. $seed(i,j,p,q;5)$ contains the minimal hybridisation energy of a seed region with five base pairs that is enclosed by base pairs (i,j) and (p,q). $C^{k,l}(p,q)$ contains the minimal extended hybridisation energy of the subsequences $s^1_p \dots s^1_k$ and $s^2_q \dots s^2_l$. Note that $seed(i,j,p,q;5)$, in contrast to $C^{k,l}(p,q)$, does not include accessibilities in terms of ED values.

be the lengths of intervals $[i,k]$ and $[j,l]$, respectively. Then, $seed(i,j,k,l;P')$ is only valid (i.e. different from ∞) if $l_m - P' \le b^{\max}_m$, $l_s - P' \le b^{\max}_s$ and $l_m + l_s - 2P' \le b^{\max}$. These three conditions assure compliance with the user-defined seed features. If the position of the seed region in the sRNA is constrained to the interval $[a_s, b_s]$, then the condition $a_s \le i \le b_s - P' + 1$ has to be met additionally.

While $seed(i,j,k,l;P')$ contains the minimal hybridisation free energy for two fixed intervals $[i,k]$ and $[j,l]$, all valid intervals have to be considered to find the optimal seed region. This is realised during the calculation of a second two-dimensional matrix $C^{k,l}_{\text{seed}}$, which stores the minimal extended hybridisation free energy of interactions that include a seed region (Figure 2.4). The matrix $C^{k,l}_{\text{seed}}$ is filled by the following recursion:

$$C^{k,l}_{\text{seed}}(i,j) = \begin{cases} \min \left\{ \begin{array}{l} \min\limits_{p,q} \left\{ \begin{array}{l} E^{\text{loop}}(i,j,p,q) + C^{k,l}_{\text{seed}}(p,q) \\ -ED_1(p,k) - ED_2(q,l) \\ +ED_1(i,k) + ED_2(j,l) \end{array} \right\} \\ \min\limits_{\substack{p,q \text{ with} \\ P \le l_m \le b^{\max}_m + P \\ P \le l_s \le b^{\max}_s + P \\ l_m + l_s \le b^{\max} + 2P}} \left\{ \begin{array}{l} seed(i,j,p,q;P) + C^{k,l}(p,q) \\ -ED_1(p,k) - ED_2(q,l) \\ +ED_1(i,k) + ED_2(j,l) \end{array} \right\} \end{array} \right\} & \text{if } (s^1_i, s^2_j) \text{ and } (s^1_k, s^2_l) \text{ can pair, } i \neq k \text{ and } j \neq l, \\ \infty & \text{otherwise,} \end{cases} \tag{2.9}$$

where $l_m = p - i + 1$ and $l_s = q - j + 1$ are now the lengths of intervals $[i,p]$ and $[j,q]$, respectively. The first of the two inner minima in the recursion addresses the case that a seed region was already included in the interaction right of base pair (p,q). The second inner minimum refers to the case that no seed region was included right of base pair (p,q), but a seed region is enclosed by base pairs (i,j) and (p,q).

Extending the algorithm by seed regions does not increase its complexity. The final energy score values are stored in matrix C by replacing $C^{k,l}(i,j)$ with $C^{k,l}_{\text{seed}}(i,j)$ in

Equation (2.8).

2.2.4 Reducing the time complexity while preserving quadratic space complexity

Although the IntaRNA algorithm incorporating seeds as presented above has a space complexity of $O(n \cdot m)$, its time complexity of $O(n^2 \cdot m^2)$ is still impractical for genome-wide searches. Therefore, we describe in the following how the time complexity can be reduced.

Before introducing a version of IntaRNA with reduced time complexity while preserving the quadratic space complexity, we summarise the full version of the algorithm by combining Equations (2.5), (2.6) and (2.7) into a single equation:

$$C(i,j) = \begin{cases} \min \left\{ \begin{array}{l} \min\limits_{p,q,k,l} \left\{ \begin{array}{l} E^{\text{loop}}(i,j,p,q) + C^{k,l}(p,q) \\ -ED_1(p,k) - ED_2(q,l) \\ +ED_1(i,k) + ED_2(j,l) \end{array} \right\} \\ ED_1(i,i) + ED_2(j,j) \end{array} \right\} & \text{if } (s^1_i, s^2_j) \text{ can pair,} \\ \infty & \text{otherwise.} \end{cases} \tag{2.10}$$

To reduce the time complexity while preserving the quadratic space complexity, we use a heuristic simplification that is inspired by the sparsification technique. The basic idea is that the four-dimensional matrix C as defined in Equation (2.3) is sparse in such a way that many entries have the same values. This is due to the fact that many right hybridisation ends will not be used in the computation of subsequent matrix entries. Therefore, we simplify the algorithm by storing the values $C^{k,l}(i,j)$ for only one interaction start (k,l). Consequently, for finding the best hybridisation with leftmost base pair (i,j), only one right hybridisation end (k,l) is considered instead of all possible right ends. We use a matrix *che* (C hybridisation end) in which an entry $che(i,j)$ stores the right hybridisation end (k,l) of the hybridisation with left end (i,j). It should be noted that this heuristic simplification does not guarantee to yield the optimal interaction in terms of extended hybridisation energy as only one fixed right end is considered during the extension of interactions to the left. Since $che(i,j)$ stores a base pair, i.e. $che(i,j) = (k,l)$, $che_1(i,j)$ denotes its first component k and $che_2(i,j)$ denotes its second component l. After introducing the matrix *che*, we can use a two-dimensional matrix C' instead of matrices $C^{k,l}$ and C to compute interactions by optimising their extended hybridisation energy according to the following recursion:

$$C'(i,j) = \begin{cases} \min \left\{ \begin{array}{ll} \min\limits_{p,q} \left\{ \begin{array}{l} E^{\text{loop}}(i,j,p,q) + C'(p,q) \\ -ED_1(p, che_1(p,q)) - ED_2(q, che_2(p,q)) \\ +ED_1(i, che_1(p,q)) + ED_2(j, che_2(p,q)) \end{array} \right\} & \text{(A)} \\ ED_1(i,i) + ED_2(j,j) & \text{(B)} \end{array} \right\} & \begin{array}{l}\text{if } (s^1_i, s^2_j) \\ \text{can pair,}\end{array} \\ \infty & \text{otherwise.} \end{cases} \tag{2.11}$$

After the calculation of each entry $C'(i,j)$, the corresponding value in $che(i,j)$ has to be updated with

$$che(i,j) = \begin{cases} che(p,q) & \text{if (A) is the minimum in Equation (2.11),} \\ (i,j) & \text{if (B) is the minimum in Equation (2.11).} \end{cases}$$

The recursion for computing interactions that incorporate a seed region as given by Equation (2.9) is simplified in the same way:

$$C'_{\text{seed}}(i,j) = \begin{cases} \min \left\{ \begin{array}{ll} \min\limits_{p,q} \left\{ \begin{array}{l} E^{\text{loop}}(i,j,p,q) + C'_{\text{seed}}(p,q) \\ -ED_1(p, che^{\text{seed}}_1(p,q)) \\ -ED_2(q, che^{\text{seed}}_2(p,q)) \\ +ED_1(i, che^{\text{seed}}_1(p,q)) \\ +ED_2(j, che^{\text{seed}}_2(p,q)) \end{array} \right\} & \text{(A)} \\ \min\limits_{\substack{p,q \text{ with} \\ P \leq l_m \leq b^{\max}_m + P \\ P \leq l_s \leq b^{\max}_s + P \\ l_m + l_s \leq b^{\max} + 2P}} \left\{ \begin{array}{l} seed(i,j,p,q;P) + C'(p,q) \\ -ED_1(p, che_1(p,q)) \\ -ED_2(q, che_2(p,q)) \\ +ED_1(i, che_1(p,q)) \\ +ED_2(j, che_2(p,q)) \end{array} \right\} & \text{(B)} \end{array} \right\} & \begin{array}{l}\text{if } (s^1_i, s^2_j) \\ \text{can pair,}\end{array} \\ \infty & \text{otherwise.} \end{cases} \tag{2.12}$$

The matrix che^{seed} is defined by

$$che^{\text{seed}}(i,j) = \begin{cases} che^{\text{seed}}(p,q) & \text{if (A) is the minimum in Equation (2.12),} \\ che(p,q) & \text{if (B) is the minimum in Equation (2.12).} \end{cases}$$

The best hybridisation score of interactions that include a seed is obtained by computing

$$\min_{1 \leq i < n, 1 \leq j < m} \left\{ \; 0, \; C'_{\text{seed}}(i,j) \; \right\}.$$

When comparing the full and the heuristic version of the IntaRNA algorithm as presented in Equations (2.10) and (2.11), then the heuristic version computes the minimum over only two instead of four variables. The range of these two variables p and q is restricted by the maximal loop size, which is set to a default value of 16. The matrices C' and C'_{seed} can be filled in $O(n \cdot m)$ time and they can be stored in $O(n \cdot m)$ space. The computation of accessibilities in terms of ED values is realised by RNAplfold and RNAup [16, 128], which are directly integrated into IntaRNA via the Vienna RNA library [75]. The accessibilities are computed from global sRNA structures and local mRNA structures, which both requires $O(n^3)$ time and $O(n^2)$ space. Overall, IntaRNA has a time complexity of $O(n \cdot m + n^3 + m^3)$ and a space complexity of $O(n \cdot m + n^2 + m^2)$. The complexity is reduced to $O(n \cdot m)$ time and space when precomputed accessibilities are available.

2.2.5 Suboptimal hybridisations

IntaRNA can predict multiple potential interactions for each sRNA. Suboptimal hybridisations are obtained by computing suboptimal tracebacks in the dynamic programming matrices. Since interactions at different mRNA locations are especially of interest, an interaction is accepted as suboptimal if the target site does not overlap with any other previously reported target site in interactions with better energy score. Whenever a (sub-)optimal interaction is found, the corresponding interaction site is masked in the mRNA. Subsequently, another suboptimal hybridisation that does not overlap with the masked mRNA subsequence(s) is searched in the matrix C'_{seed} and its hybridisation pattern is computed by traceback. This procedure is iteratively repeated until the user-specified maximal number of reported suboptimal predictions or energy threshold is reached, or no further suboptimal hybridisations are available.

2.3 Performance on prediction of sRNA targets

2.3.1 Evaluation dataset and compared methods

In order to assess the performance of the heuristic IntaRNA algorithm as presented in Section 2.2, we used the program to predict targets of bacterial regulatory sRNAs. The test set consisted of ten experimentally verified sRNA–mRNA interactions from *E. coli* and eight interactions from *Salmonella* that were previously published (Table 2.1). For each sRNA, we predicted interactions for all genes of the respective genome. The genome sequences were downloaded from the RefSeq database of the National Center for Biotechnology Information (NCBI) [144]. Since the majority of sRNAs from our test set bind their target gene in close proximity to the RBS, we defined a subsequence of 150 nt up- and 50 nt downstream of the first base of the start codon as the (putative) target region. We obtained 4294 target regions from the *E. coli* genome (RefSeq accession number

NC_000913) and 4425 target regions from the *Salmonella* genome (RefSeq accession number NC_003197).

The seed features and other parameters that were used for the target prediction with IntaRNA were chosen according to known interactions. They included a seed of at least eight consecutive bp and no restriction on the interaction length. All interactions except OxyS–*fhlA* have a continuous hybridisation pattern. Among all examples, the Spot42–*galK* interaction is the longest one with a length of 75 nt. We compared the results with several state-of-the-art methods for the prediction of RNA–RNA interactions, namely TargetRNA, RNAhybrid, RNAplex, and RNAup. Although RNAhybrid is primarily designed for the prediction of miRNA target sites, it has been used occasionally for predictions related to sRNAs [e.g. 163, 189]. Therefore, it has been included in our comparison for the sake of completeness using the default parameters. For TargetRNA, we used the web application [181] with default parameters, except that the search was focused on our target regions and that the p-value threshold was increased to obtain the best 100 target predictions per sRNA. RNAplex was used with a penalty of 0.3 kcal/mol per nucleotide as suggested by Tafer and Hofacker [178]. We used RNAup including the accessibility of both RNAs [128] and set the maximal interaction length to 80, which is slightly longer than the maximal length that was found in our dataset.

2.3.2 Accuracy of predicted sRNA–mRNA interactions

In a first experiment, we assessed whether IntaRNA is able to predict precisely the interaction between each sRNA and its mRNA target. Therefore, we computed the sensitivity (SENS) and the positive predictive value (PPV) for each sRNA–target pair:

$$\text{SENS} = \frac{\text{number of correctly predicted base pairings}}{\text{number of true base pairings}} \quad \text{and}$$
$$\text{PPV} = \frac{\text{number of correctly predicted base pairings}}{\text{number of predicted base pairings}}.$$

These measures have already been used in the past to compare different RNA secondary structure prediction methods [e.g. 42]. As shown in Table 2.1, IntaRNA outperformed the compared methods in the average accuracy of the predicted interactions. TargetRNA achieved the second best average sensitivity and the third best average PPV, but reported only 12 out of 18 interactions due to its cut-off. The program RNAhybrid tended to maximise the length of hybridisation, which led to a high average sensitivity, but very low average PPV. Thus, the program is more appropriate to predict interactions between short RNAs (like miRNAs) and long RNAs. To overcome this problem, RNAplex introduced a length penalty, which significantly increased its average PPV compared with RNAhybrid. RNAup achieved the third best average sensitivity and the second best average PPV. Among all programs compared, it had an overall accuracy closest to IntaRNA.

Table 2.1. Prediction accuracy of `IntaRNA` compared with other leading RNA–RNA interaction prediction methods on a set of experimentally verified sRNA–mRNA interactions.

sRNA–mRNA	Ref.	Sensitivity					PPV				
		`IntaRNA`	`TargetRNA`	`RNAhybrid`	`RNAplex`	`RNAup`	`IntaRNA`	`TargetRNA`	`RNAhybrid`	`RNAplex`	`RNAup`
DsrA–*rpoS*	[151]	0.808	0.808	0.000	0.808	0.808	0.778	0.778	0.000	0.778	0.778
GcvB–*argT*	[163]	0.950	1.000	1.000	0.000	0.900	0.950	0.625	0.160	0.000	0.947
GcvB–*dppA*	[163]	1.000	0.941	0.941	0.765	1.000	0.586	0.421	0.132	0.448	0.459
GcvB–*gltI*	[163]	0.000	-	0.875	1.000	0.000	0.000	-	0.210	0.857	0.000
GcvB–*livJ*	[163]	0.955	-	1.000	0.955	0.955	0.955	-	0.180	0.955	0.955
GcvB–*livK*	[163]	0.542	-	0.542	0.542	0.542	0.565	-	0.108	0.565	0.565
GcvB–*oppA*	[163]	1.000	1.000	1.000	1.000	1.000	0.957	0.957	0.200	0.957	0.957
GcvB–STM4351	[163]	0.760	0.000	0.000	0.000	0.880	0.905	0.000	0.000	0.000	0.957
IstR–*tisB*	[195]	0.879	0.939	0.939	0.750	0.667	0.690	0.775	0.403	1.000	1.000
MicA–*ompA*	[187]	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.302	1.000	1.000
MicA–*lamB*	[19]	1.000	-	0.609	1.000	0.826	0.821	-	0.318	1.000	0.704
MicC–*ompC*	[28]	1.000	0.636	1.000	0.000	0.727	0.537	0.286	0.333	0.000	0.410
MicF–*ompF*	[159]	0.960	0.560	0.960	0.920	0.800	0.960	0.636	0.545	0.958	0.952
OxyS–*fhlA*	[6]	0.500	-	0.938	0.563	0.375	1.000	-	0.288	0.750	1.000
RyhB–*sdhD*	[118]	0.588	0.882	0.794	0.824	0.794	1.000	0.909	0.403	1.000	0.794
RyhB–*sodB*	[57]	1.000	1.000	1.000	1.000	1.000	0.818	0.375	0.167	0.818	0.900
SgrS–*ptsG*	[90]	0.739	-	0.000	0.739	0.739	1.000	-	0.000	1.000	1.000
Spot42–*galK*	[126]	0.409	0.545	0.523	0.432	0.523	0.643	0.558	0.280	0.655	0.523
Average		**0.783**	0.776	0.729	0.683	0.752	**0.787**	0.610	0.224	0.708	0.772

For each sRNA–target pair, sensitivity and PPV were calculated for the highest scoring interaction predicted. '-' means that no interaction was predicted. The best average result for each measure is highlighted in bold.

Table 2.2. Accuracy of suboptimal interaction predictions of the tools `IntaRNA`, `RNAhybrid` and `RNAplex` for sRNA–mRNA examples with incorrect energetically optimal prediction (compare with Table 2.1).

sRNA–mRNA	Tool	Suboptimal energy	Mfe	Sensitivity	PPV
DsrA–*rpoS*	`RNAhybrid`	-46.500	-47.700	0.808	0.488
GcvB–*argT*	`RNAplex`	-26.300	-30.400	0.950	0.950
GcvB–*gltI*	`IntaRNA`	-18.284	-18.356	0.375	0.500
GcvB–STM4351	`RNAhybrid`	-15.700	-87.800	0.000	0.000
GcvB–STM4351	`RNAplex`	-9.700	-33.600	0.000	0.000
MicC–*ompC*	`RNAplex`	-16.400	-28.000	0.000	0.000
SgrS–*ptsG*	`RNAhybrid`	-13.100	-92.100	0.000	0.000

For each sRNA–target pair, all suboptimal predictions within an energy range of $0.5 * \text{mfe}$ kcal/mol were analysed until a prediction with sensitivity and PPV different from zero was found. If no such prediction within that range existed, the suboptimal solution with the best energy was used.

The programs `IntaRNA` and `RNAup` were the only ones whose optimal solution located every sRNA target site correctly, except for the interaction GcvB–*gltI*. For this interaction, both `IntaRNA` and `RNAup` predicted optimal hybridisations that did not share a single base pair with the experimentally verified interaction. As `IntaRNA`, `RNAhybrid` and `RNAplex` can compute suboptimal solutions, we used these to compute sensitivity and PPV for all interactions for which the optimal solution was incorrect (see Table 2.2 and compare with Table 2.1). When the suboptimal predictions were taken into account, the average sensitivity/PPV on the whole test set improved to 0.774/0.251 and 0.736/0.761 for `RNAhybrid` and `RNAplex`, respectively. `IntaRNA` predicted a suboptimal interaction for GcvB–*gltI* that overlapped with the correct target site; its energy score of -18.28 kcal/mol was only marginally inferior to the optimal energy score of -18.36 kcal/mol. When including this suboptimal prediction, `IntaRNA` achieved on the whole test set an average sensitivity and PPV greater than 0.8.

To study the influence of seed features on `IntaRNA`'s prediction quality, we repeated the experiments without requiring an interaction seed (Table 2.3). In this case, the average sensitivity and PPV of `IntaRNA` were 0.699 and 0.728, respectively, which is below the accuracy of `IntaRNA` with seed features and `RNAup`. The difference to the latter, which uses a similar scoring scheme, can be explained by the heuristic behaviour of `IntaRNA`.

Altogether, the results demonstrate that `RNAup` and `IntaRNA`, which both incorporate the accessibility of interaction sites, perform better in the prediction of sRNA–mRNA interactions than programs that neglect the accessibility. Furthermore, the quality of `IntaRNA`'s predictions is substantially improved when seed features are additionally taken into account.

Table 2.3. Prediction accuracy of IntaRNA without requiring an interaction seed. The prediction results of IntaRNA with seed features as reported in Table 2.1 are included for comparison.

sRNA–mRNA	IntaRNA without seed		IntaRNA with seed	
	Sensitivity	PPV	Sensitivity	PPV
DsrA–*rpoS*	0.462	0.667	0.808	0.778
GcvB–*argT*	0.950	0.950	0.950	0.950
GcvB–*dppA*	1.000	0.586	1.000	0.586
GcvB–*gltI*	0.000	0.000	0.000	0.000
GcvB–*livJ*	0.000	0.000	0.955	0.955
GcvB–*livK*	0.542	0.565	0.542	0.565
GcvB–*oppA*	1.000	0.957	1.000	0.957
GcvB–STM4351	0.760	0.905	0.760	0.905
IstR–*tisB*	0.806	0.690	0.879	0.690
MicA–*ompA*	1.000	0.821	1.000	1.000
MicA–*lamB*	1.000	1.000	1.000	0.821
MicC–*ompC*	1.000	0.537	1.000	0.537
MicF–*ompF*	0.960	0.960	0.960	0.960
OxyS–*fhlA*	0.375	1.000	0.500	1.000
RyhB–*sdhD*	0.588	1.000	0.588	1.000
RyhB–*sodB*	1.000	0.818	1.000	0.818
SgrS–*ptsG*	0.739	1.000	0.739	1.000
Spot42–*galK*	0.409	0.643	0.409	0.643
Average	0.699	0.728	0.783	0.787

For each sRNA–target pair, sensitivity and PPV were calculated for the highest scoring interaction prediction.

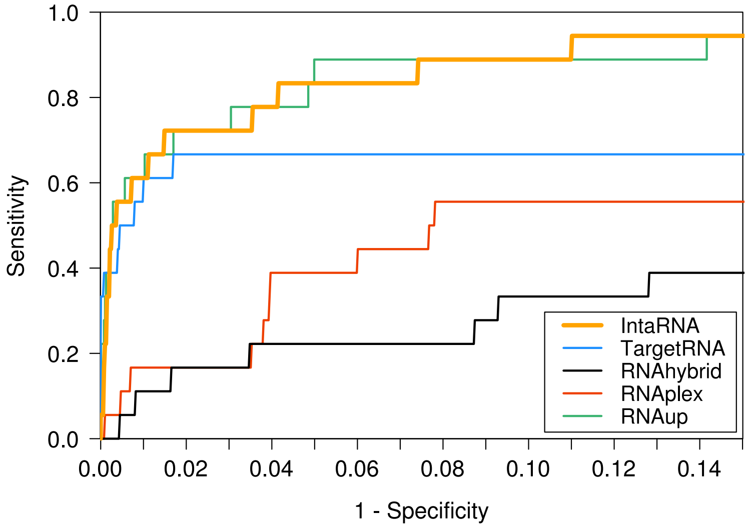


Figure 2.5. Performance of `IntaRNA` and other methods in genome-wide target predictions for our test set of 10 sRNAs. The sensitivity (true positive rate) is shown as a function of the false positive rate (1 - specificity). For each prediction method, the target candidates for each sRNA were sorted by their energy score. Each ROC curve was generated from the rate of true and false predictions, while varying the number of predicted targets per sRNA.

2.3.3 Performance in genome-wide target predictions

In a second experiment, we compared `IntaRNA` and the existing methods with respect to their ability of finding sRNA targets on a genome-wide scale. We applied every program to our test set and, for each sRNA, searched all target regions for potential target sites. The resulting list of target candidates for each sRNA was sorted by the computed energy score. All programs except `TargetRNA` and `IntaRNA` give an interaction for each putative target region. `TargetRNA` reports at most 100 putative interaction sites per sRNA. `IntaRNA` returns interactions that have both a seed with specified features and an energy score below 0.0 kcal/mol. For each method, we calculated the sensitivity $(\frac{\text{True Positives}}{\text{True Positives + False Negatives}})$ and specificity $(\frac{\text{True Negatives}}{\text{True Negatives + False Positives}})$. Our test set contains 18 targets that can be predicted as true positive. Each correctly predicted target was counted as true positive regardless of whether the interaction site was predicted correctly or not. Each of the nine sRNAs in *E. coli* may interact with any of the 4294 genes, and each sRNA in *Salmonella* may interact with any of the 4425 genes. Consequently, there are 47496 potential interactions, of which 47478 are considered as non-interactions, i.e. negatives. A similar approach to evaluate the performance on prediction of sRNA targets has been used previously by Tjaden et al. [182].

The ROC curves in Figure 2.5 illustrate the performance of different target prediction methods on our test set. We generated each ROC curve by calculating sensitivity and specificity while varying the number of predicted targets that were taken into account for each sRNA. The plot shows that `IntaRNA` and `RNAup` are the methods performing best

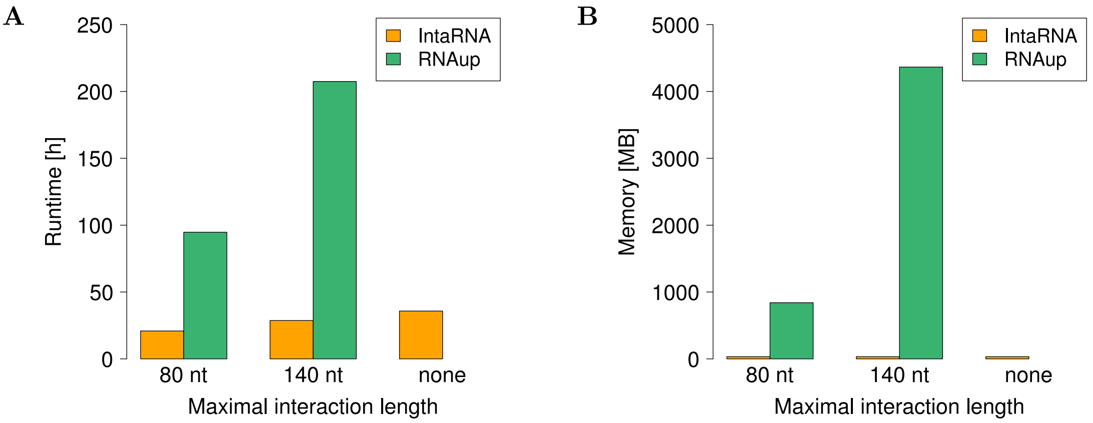


Figure 2.6. Comparison of **(A)** runtime and **(B)** memory requirements of `IntaRNA` (including computation of accessibilities) and `RNAup` for a GcvB target search in *Salmonella*. Without restricting the interaction length, `RNAup` used up the complete available memory and, as a consequence, crashed.

on prediction of sRNA targets. Both `RNAhybrid` and `RNAplex` achieved a low sensitivity suggesting that these programs are suitable only to a limited degree for genome-wide sRNA target searches. Instead, especially the latter method could potentially be used as a prefilter to reduce the number of target candidates before analysis with more sensitive methods. `TargetRNA` reports at most 100 putative interaction sites per sRNA. Taking this in consideration, it achieved a fairly high sensitivity at a low false positive rate, although only an alignment-like algorithm based on base pairing potential is used. However, it can be assumed that the program will perform worse on interactions that show lower sequence complementarity, but underlie more complex duplex formation rules. The curves show that `IntaRNA` and `RNAup` have a similar performance on predicting sRNA targets and perform best among all studied programs. There is, however, a clear difference in the practical applicability of both programs (Figure 2.6). On an Intel Xeon 5160 (3.0 GHz) with 7.8 GB available RAM, a GcvB target search in all *Salmonella* target regions allowing a maximal interaction length of 80 nt took 21 h and required 33 MB RAM with `IntaRNA`. The same search with `RNAup` needed 95 h and 840 MB RAM. An increase of the maximal interaction length to 140 nt raised `IntaRNA`'s runtime to 29 h with unchanged memory usage, whereas `RNAup` then required 207 h and 4.3 GB RAM. Without a restriction on the interaction length, `IntaRNA` took 36 h and required again 33 MB RAM. Since `RNAup` requires a restriction on the maximal interaction length, we limited it to the length of the sRNA. This caused exhaustion of the complete memory and, as a consequence, a crash of `RNAup`. The dramatic increase of `RNAup`'s resource requirements results from its higher asymptotic complexity and impairs its applicability on normal work stations with limited available memory. Note that this benchmark was based on an `IntaRNA` version that used a previous implementation of `RNAplfold` [14]. The computation of all *ED* values therefore

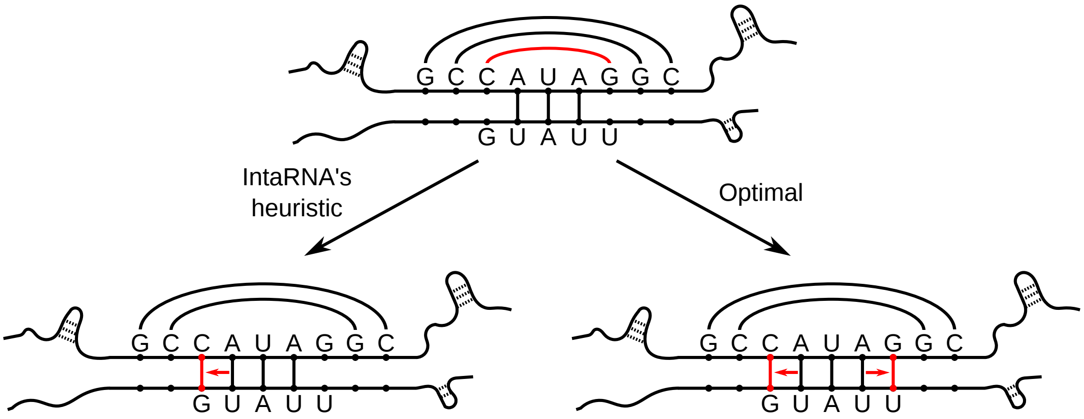


Figure 2.7. Schematic illustration of the difference between an uni- and a bidirectional RNA–RNA interaction extension strategy. The three interacting nucleotides `AUA` in the upper RNA sequence are located in a hairpin loop such that the interaction site is spanned by one intramolecular `G-C` and two intramolecular `C-G` base pairs. When the innermost intramolecular `C-G` base pair (highlighted in red in the interaction on top) is broken to extend the interaction by the intermolecular base pair `C-G` to the left, it would be energetically favourable to extend the interaction by the `G-U` base pair to the right as well. `IntaRNA`, however, extends interactions only unidirectional due to its heuristic behaviour. The illustration is adapted from reference [48].

required $O(n \cdot L^3)$ time, where L is the size of the sequence window in which the sequences are folded (here: $L = n$). The same task requires only $O(n^3)$ time in the recent `IntaRNA` version, which further reduces the runtime of genome-wide target predictions.

It should be noted that the calculation of sensitivity and specificity is rather conservative, since we relied only on experimentally verified interactions that were published by the end of 2007. Several top-ranking target predictions of `IntaRNA` that were verified as true targets in subsequent studies [e.g. 35, 165] have therefore not been taken into consideration in this performance evaluation.

2.4 Improving the heuristic of `IntaRNA`

The heuristic version of `IntaRNA` presented in Section 2.2 predicts interactions by minimising an extended hybridisation energy score. When computing the optimal score according to Equation (2.12), the resulting interaction might occasionally be just an approximation of the mfe interaction as not all possible rightmost hybridisation ends are considered. Instead, for each left hybridisation end (i, j), either an interaction with fixed rightmost base pair $che(i, j)$ or $che^{\text{seed}}(i, j)$ is extended unidirectional to the left or a new interaction is started with base pair (i, j). It might, however, be energetically favourable to extend an interaction bidirectional such that a hybridisation extension by further base pairs to the left is also followed by an extension to the right beyond the fixed rightmost end. Figure 2.7 illustrates the difference between an uni- and a bidirectional interaction extension strategy with an example. There, the interaction site is located in a hairpin loop in one of the two sequences. To extend the interaction to the left by base pair C-G, the intramolecular base pair C-G spanning the interaction site is broken. As a consequence of opening this intramolecular base pair, the base pair G-U could be added to the rightmost hybridisation end without requiring further free energy to make nucleotide G single-stranded in the upper sequence. Although an interaction extension in both directions would improve the overall energy score in this example, the heuristic IntaRNA algorithm does not accommodate this. Therefore, we present an improved heuristic version of the IntaRNA algorithm in the following (see reference [48] for more details).

To tackle the above-mentioned problem, a compromise between considering only one and all rightmost hybridisation ends can be achieved by storing multiple but not all right ends for a given leftmost hybridisation end. Let V be the number of different right interaction ends that are considered. For each leftmost base pair (i,j), the entry $C(i,j)$ of matrix C stores an array containing the V best hybridisation energies, for which all energy scores correspond to hybridisations with mutually distinct right ends. An entry $C(i,j,v)$ contains the v^{th} best energy of all interactions with distinct right ends that start with left base pair (i,j). The corresponding right hybridisation end is stored in $che(i,j,v)$. When extending Equation (2.11) to incorporate multiple distinct rightmost hybridisation ends, we get

$$C(i,j,v) = \begin{cases} v^{\text{th}}\text{-}\min\limits_{\substack{p,q,s \text{ with} \\ 1\leq s\leq V}}^{\text{dre}} \left\{ \begin{array}{ll} E^{\text{loop}}(i,j,p,q) + C(p,q,s) & \\ -ED_1(p, che_1(p,q,s)) & \\ -ED_2(q, che_2(p,q,s)) & \text{(A)} \\ +ED_1(i, che_1(p,q,s)) & \\ +ED_2(j, che_2(p,q,s)) & \\ & \\ ED_1(i,i) + ED_2(j,j) & \text{(B)} \end{array} \right\} & \text{if } (s_i^1, s_j^2) \text{ can pair,} \\ \infty & \text{otherwise,} \end{cases} \tag{2.13}$$

where the operator v^{th}-min$^{\text{dre}}$ gives the v^{th} minimum of the energy scores of all interactions with distinct right ends (dre). For any two hybridisations with leftmost base pair (i,j) that have the same rightmost base pair, the hybridisation with lower energy score is taken. After calculation of each entry $C(i,j,v)$, $che(i,j,v)$ has to be updated with

$$che(i,j,v) = \begin{cases} che(p,q,s) & \text{if (A) is the } v^{\text{th}} \text{ minimum in Equation (2.13),} \\ (i,j) & \text{if (B) is the } v^{\text{th}} \text{ minimum in Equation (2.13).} \end{cases}$$

Hybridisations that include a seed region are computed analogously to Equations (2.12) and (2.13) by the following recursion:

$$C_{\text{seed}}(i,j,v) = \begin{cases} v^{\text{th}}\text{-min}^{\text{dre}} \left\{ \begin{array}{l} \underset{\substack{p,q,s \text{ with} \\ 1\leq s\leq V}}{v\text{-min}^{\text{dre}}} \left\{ \begin{array}{l} E^{\text{loop}}(i,j,p,q) \\ +C_{\text{seed}}(p,q,s) \\ -ED_1(p, che_1^{\text{seed}}(p,q,s)) \\ -ED_2(q, che_2^{\text{seed}}(p,q,s)) \\ +ED_1(i, che_1^{\text{seed}}(p,q,s)) \\ +ED_2(j, che_2^{\text{seed}}(p,q,s)) \end{array} \right\} \text{(A)} \\ \underset{\substack{p,q \text{ with} \\ P\leq l_m\leq b_m^{\max}+P \\ P\leq l_s\leq b_s^{\max}+P \\ l_m+l_s\leq b^{\max}+2P \\ 1\leq s\leq V}}{v\text{-min}^{\text{dre}}} \left\{ \begin{array}{l} seed(i,j,p,q;P) \\ +C(p,q,s) \\ -ED_1(p, che_1(p,q,s)) \\ -ED_2(q, che_2(p,q,s)) \\ +ED_1(i, che_1(p,q,s)) \\ +ED_2(j, che_2(p,q,s)) \end{array} \right\} \text{(B)} \end{array} \right\} & \text{if } (s_i^1, s_j^2) \text{ can pair,} \\ \infty & \text{otherwise,} \end{cases} \tag{2.14}$$

where the operator v-min$^{\text{dre}}$ gives the set of v minimal values of the energy scores of all interactions with distinct right ends (dre) (in contrast to v^{th}-min$^{\text{dre}}$, which gives the v^{th} minimal value). The value $C_{\text{seed}}(i,j,v)$ is the v^{th} minimum of the union of the two sets obtained from cases (A) and (B), with each of the two sets having v elements. The variables $l_m = p - i + 1$ and $l_s = q - j + 1$ are the lengths of intervals $[i,p]$ and $[j,q]$, respectively. $che^{\text{seed}}(i,j,v)$ is defined by

$$che^{\text{seed}}(i,j,v) = \begin{cases} che^{\text{seed}}(p,q,s) & \text{if (A) is the } v^{\text{th}} \text{ minimum in Equation (2.14),} \\ che(p,q,s) & \text{if (B) is the } v^{\text{th}} \text{ minimum in Equation (2.14).} \end{cases}$$

The best energy score of a hybridisation including a seed is then

$$\min_{\substack{1\leq i<n \\ 1\leq j<m \\ 1\leq v\leq V}} \Big\{ \; 0, \; C_{\text{seed}}(i,j,v) \; \Big\}.$$

The algorithm presented above has a time complexity of $O(n \cdot m \cdot V^2)$. Each entry of the array $C(i,j)$ contains V energy scores; for each of these V scores has to be ensured that it belongs to a hybridisation with a rightmost end that is distinct from all other $V-1$ right hybridisation ends. In total, these operations require $O(V^2)$ time. The space complexity of the algorithm is $O(n \cdot m \cdot V)$.

The performance of the improved heuristic IntaRNA algorithm was evaluated with a prototype implementation using a test set of 36 experimentally validated sRNA–mRNA interactions from *E. coli* and *Salmonella* [48]. Each predicted interaction was evaluated in terms of F-measure, which is the harmonic mean of sensitivity (SENS) and PPV, $F = \frac{2 \times \text{SENS} \times \text{PPV}}{\text{SENS}+\text{PPV}}$. When the value of parameter V, i.e. the number of considered right hybridisation ends, was increased, both the mfe and the F-measure of the predicted interactions improved on average. Depending on the parameters for the seed region, it was sufficient to set V to values of three to five to yield interactions with optimal energy, i.e. the energy score was equal to the mfe computed by the non-heuristic algorithm. Optimal F-measures were already achieved for V equal to three. The overall improvement by considering multiple hybridisation ends in the new heuristic algorithm is, however, marginal; only two percent improvement could be achieved for the mean mfe and the accuracy of the predicted interactions in terms of mean F-measure. The runtime of the program increased by approximately factor two when three right hybridisation ends instead of one were used.

In conclusion, the heuristic of `IntaRNA` as presented in Section 2.2 gives already optimal or near-optimal solutions. The energy score and F-measure of the predicted interactions can be improved by considering more than one rightmost hybridisation end for each leftmost hybridisation end. The gain in prediction performance is, however, too marginal in light of the resulting doubled runtime.

2.5 Functional analysis of predicted target sites

In bacteria, the initiation of translation is typically stimulated by an interaction between the 30S ribosomal subunit and the mRNA. The 3' tail of the 16S rRNA binds the mRNA at the SD sequence, a 4–5 nt sequence motif located around 7 ± 2 nt upstream of the start codon [96, 101]. Many sRNAs regulate the translation of their target mRNA by changing the accessibility of the SD sequence for ribosome binding. Base pairing of the sRNA at or in the vicinity of the RBS typically results in translation inhibition. In contrast, some sRNAs activate translation of their target mRNAs by opening an inhibitory structure that sequesters the RBS [53]. In the following, we study the consequences of sRNA binding to the target mRNA regarding the accessibility of the SD sequence, and, thus, the mode of translational regulation by the sRNA–mRNA interaction. To this end, SD sequence locations were predicted for all given mRNAs. Then, the change in the accessibility of the predicted SD sequence as a result of the predicted sRNA–mRNA interaction was computed separately for each interaction.

The SD sequence location of a specific mRNA was predicted by simulating the hybridisation between the mRNA and the 16S rRNA following the approach of Starmer et al. [173]. In brief, hybridisations were computed between the single-stranded 16S rRNA 3' tail and the mRNA region covering 35 nt up- and downstream of the first base of the start codon. For the 16S rRNA 3' tail, we used the sequences `5'-GAUCACCUCCUUA-3'` for

E. coli and `5'-GAUCACCUCCUUACC-3'` for *Salmonella*. The SD sequence was then located by the position of the optimal 16S rRNA–mRNA hybridisation if the hybridisation free energy was below a significance threshold derived from core SD sequences. We successfully predicted SD sequence locations for about 74 percent of all *E. coli* genes and about 75 percent of all *Salmonella* genes.

The influence of the sRNA–mRNA interaction on the accessibility of the predicted SD sequence was studied by the change in the probability that the SD sequence is unpaired before and after sRNA binding. Let s be a given mRNA sequence, in which the SD sequence is located at positions $s_i \dots s_j$ and the sRNA binds the mRNA at positions $s_k \dots s_l$. The probabilities that the SD sequence is unpaired before ($PU_{SD}^{\text{nohybrid}}$) and after ($PU_{SD}^{\text{hybrid}}$) the hybridisation of the sRNA and the mRNA are defined by

$$\begin{aligned} PU_{SD}^{\text{nohybrid}} &= e^{\frac{E^{\text{ens}}(\mathcal{S}) - E^{\text{ens}}(\mathcal{S}_{i,j}^{\text{unpaired}})}{RT}} \quad \text{and} \\ PU_{SD}^{\text{hybrid}} &= e^{\frac{E^{\text{ens}}(\mathcal{S}_{k,l}^{\text{unpaired}}) - E^{\text{ens}}(\mathcal{S}_{i,j,k,l}^{\text{unpaired}})}{RT}}, \end{aligned}$$

respectively, where E^{ens} is ensemble free energy, $\mathcal{S}$ is the ensemble of all structures formed by s, $\mathcal{S}_{i,j}^{\text{unpaired}}$ is the ensemble of all structures of s with nucleotides $s_i \dots s_j$ unpaired, $\mathcal{S}_{i,j,k,l}^{\text{unpaired}}$ is the ensemble of all structures with nucleotides $s_i \dots s_j$ and $s_k \dots s_l$ unpaired, R is the gas constant, and T is the temperature.

ΔPU_{SD}, which is the change in the probability that the SD sequence is unpaired due to the sRNA–mRNA hybridisation, is defined by

$$\Delta PU_{SD} = PU_{SD}^{\text{hybrid}} - PU_{SD}^{\text{nohybrid}}.$$

$\Delta PU_{SD} > 0$ suggests that the sRNA–mRNA interaction results in structural rearrangements that increase the accessibility of the SD sequence to activate translation. Conversely, $\Delta PU_{SD} < 0$ suggests translational repression. The higher the absolute value, the higher is the expected regulatory outcome. However, a special case arises if the mRNA target site overlaps with or is in close vicinity to the SD sequence. Then, the RBS is inaccessible for ribosome binding and translation repression is expected.

Using this approach, we analysed the regulatory outcome of the predicted sRNA–mRNA binding for all interactions of the test set introduced in Section 2.3. We determined for each mRNA whether the sRNA binding site predicted by `IntaRNA` is at or close to the SD sequence such that the SD sequence is already blocked for ribosome binding. Otherwise, we calculated ΔPU_{SD} for the mRNA. The results are shown in Table 2.4. In 11 out of 18 examples, our approach successfully predicted the type of translational regulation by the sRNA. For three of the remaining seven interactions, the SD sequence could either not be located (GcvB–STM4351) or was located at an incorrect position (MicF–*ompF* and OxyS–*fhlA*). Another sRNA, IstR, blocks translation of its mRNA target

Table 2.4. Predicted changes in the accessibility of the SD sequence as a result of the interaction between the sRNA and its target mRNA.

sRNA–mRNA	Regulation	Target site	SD	ΔPU_{SD}	Predicted regulation
DsrA–*rpoS*	Activation	-126 – -97	-12 – -7	**0.07**	**Activation**
GcvB–*argT*	Repression	-57 – -37	-10 – -5	0.01	Activation
GcvB–*dppA*	Repression	-43 – -11	-12 – -7	**[SD]**	**Repression**
GcvB–*gltI*	Repression	-62 – -45	-10 – -5	0.00	n/a
GcvB–*livJ*	Repression	-51 – -28	-18 – -13	0.06	Activation
GcvB–*livK*	Repression	-44 – -17	-13 – -8	**-0.46**	**Repression**
GcvB–*oppA*	Repression	-8 – 16	-15 – -10	**[SD]**	**Repression**
GcvB–STM4351	Repression	-45 – -19	n/a	n/a	n/a
IstR–*tisB*	Repression	-145 – -102	-11 – -6	0.00	n/a
MicA–*ompA*	Repression	-21 – -6	-14 – -9	**[SD]**	**Repression**
MicA–*lamB*	Repression	-17 – 18	-10 – -5	**[SD]**	**Repression**
MicC–*ompC*	Repression	-62 – -15	-15 – -10	**[SD]**	**Repression**
MicF–*ompF*	Repression	-16 – 10	20 – 25	0.00	n/a
OxyS–*fhlA*	Repression	34 – 41	15 – 20	0.00	n/a
RyhB–*sdhD*	Repression	-33 – -13	-12 – -7	**[SD]**	**Repression**
RyhB–*sodB*	Repression	-6 – 5	-12 – -7	**[SD]**	**Repression**
SgrS–*ptsG*	Repression	-28 – -9	-16 – -11	**[SD]**	**Repression**
Spot42–*galK*	Repression	-18 – 14	-13 – -8	**[SD]**	**Repression**

The target site positions as predicted by IntaRNA and the predicted SD sequence positions are given as distance to the annotated translation start site. GcvB–*gltI* shows the first suboptimal target prediction. 'n/a' indicates that no significant SD sequence location was found or that our method could not predict the regulatory effect of the sRNA–mRNA interaction. ΔPU_{SD} represents the change, due to sRNA binding, in the probability that the SD sequence is unpaired. '[SD]' indicates that either ΔPU_{SD} cannot be calculated because the sRNA binds at the predicted SD sequence or the target site is in the immediate vicinity of the predicted SD sequence location (distance at most 2 nt). If predicted and observed regulatory effect agreed, then the sRNA–mRNA interaction was marked in bold in the last two columns.

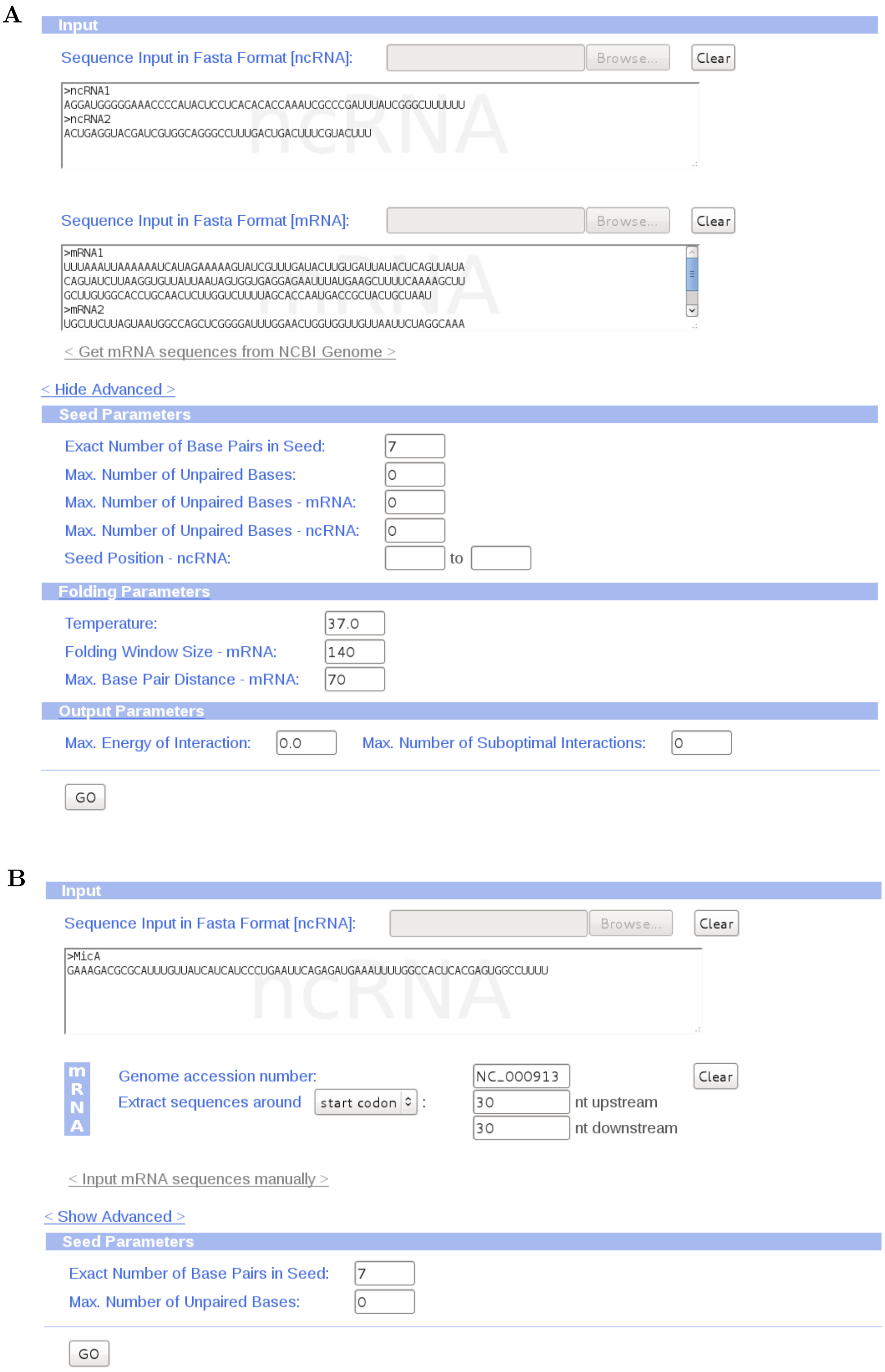
tisB by binding 100 nt upstream of the start codon without inducing structural changes at the RBS. Instead, IstR blocks a ribosome standby site that is essential for translation initiation [37]. The remaining three interactions all involve the sRNA GcvB. Its targets *argT*, *livJ* and *gltI* are bound upstream of the ribosome binding site, and the inhibitory activity cannot be directly explained by competition with ribosome binding. At least for the last example, translational repression by a simple interference model or by masking a ribosome standby site is unlikely [163]. Consequently, the regulation cannot be predicted by our model.

2.6 A web interface for genome-wide target predictions

The RNA–RNA interaction prediction tool `IntaRNA` is integrated in the Freiburg RNA tools web server, which gives access to several tools for different RNA analysis tasks via a common web-based user interface.

The input of the `IntaRNA` web server consists of a set of ncRNA sequences and a set of mRNA sequences, both in FASTA format. The sequences can be either entered directly or uploaded from files. Instead of manually providing mRNA sequences, the user can also request that mRNA sequences are automatically extracted from a genomic molecule. For this purpose, an NCBI RefSeq genome accession number [145] has to be specified; any complete genomic molecule including genomes, chromosomes, organelles and plasmids is allowed. Subsequences of all genes that are annotated in this genomic sequence are then extracted using BioPerl [171]. Each subsequence contains an user-specified number of nucleotides up- and downstream of the start or stop codon. Furthermore, the input page provides parameters for `IntaRNA` with reasonable default settings. For user convenience, the server distinguishes between basic parameters as the number of paired and unpaired bases in the seed region and advanced parameters as the folding temperature, window size and base pair span for local mRNA folding, and some more sophisticated seed parameters. The advanced parameters are hidden by default and can be unfolded on demand. In this way the server provides broad flexibility without confusing the less experienced user. The input is validated and the user is informed of inconsistencies as early as possible. Furthermore, we provide example input for demonstration purposes, a video tutorial and online help that describes `IntaRNA`, its input, available parameters and output. Figure 2.8 shows two example input pages of the `IntaRNA` web server.

Each query is processed following a general scheme: jobs are scheduled to a computing cluster managed by Sun Grid Engine such that jobs can be computed in parallel and resources are flexibly adapted to the server load. After submission, the current status of the job is reported and the user receives a URL allowing access to the job status or output. The user can also provide an email address to receive a notification when the job has finished. Upon job completion the result page is displayed online in the web browser.



Input

Sequence Input in Fasta Format [ncRNA]:

```
>ncRNA1
AGGAUGGGGGAAACCCCAUACUCCUCACACACCAAAUCGCCCGAUUUAUCGGGCUUUUUU
>ncRNA2
ACUGAGGUACGAUCGUGGCAGGGCCUUUGACUGACUUUCGUACUUU
```

Sequence Input in Fasta Format [mRNA]:

```
>mRNA1
UUUAAAUUAAAAAAUCAUAGAAAAAGUAUCGUUUGAUACUUGUGAUUAUACUCAGUUAUA
CAGUAUCUUAAGGUGUUAUUAAUAGUGGUGAGGAGAAUUUAUGAAGCUUUUCAAAAGCUU
GCUUGUGGCACCUGCAACUCUUGGUCUUUUAGCACCAAUGACCGCUACUGCUAAU
>mRNA2
UGCUUCUUAGUAAUGGCCAGCUCGGGGAUUUGGAACUGGUGGUUGUUAAUUCUAGGCAAA
```

Seed Parameters

Exact Number of Base Pairs in Seed: 7

Max. Number of Unpaired Bases: 0

Max. Number of Unpaired Bases - mRNA: 0

Max. Number of Unpaired Bases - ncRNA: 0

Seed Position - ncRNA: to

Folding Parameters

Temperature: 37.0

Folding Window Size - mRNA: 140

Max. Base Pair Distance - mRNA: 70

Output Parameters

Max. Energy of Interaction: 0.0 Max. Number of Suboptimal Interactions: 0

Input

Sequence Input in Fasta Format [ncRNA]:

```
>MicA
GAAAGACGCGCAUUUGUUAUCAUCAUCCCUGAAUUCAGAGAUGAAAUUUUGGCCACUCACGAGUGGCCUUUU
```

Genome accession number: NC_000913

Extract sequences around start codon : 30 nt upstream 30 nt downstream

Seed Parameters

Exact Number of Base Pairs in Seed: 7

Max. Number of Unpaired Bases: 0

Figure 2.8. Screenshots of the IntaRNA web server input page. **(A)** Input forms for direct submission of ncRNA and mRNA sequences and the available parameters for IntaRNA. **(B)** The mRNA sequences will be automatically extracted from the *E. coli* K-12 genome (RefSeq accession number NC_000913). Target sites of the sRNA MicA will be searched in regions spanning 30 nt of the 5' UTR and 30 nt of the CDS of each annotated gene.

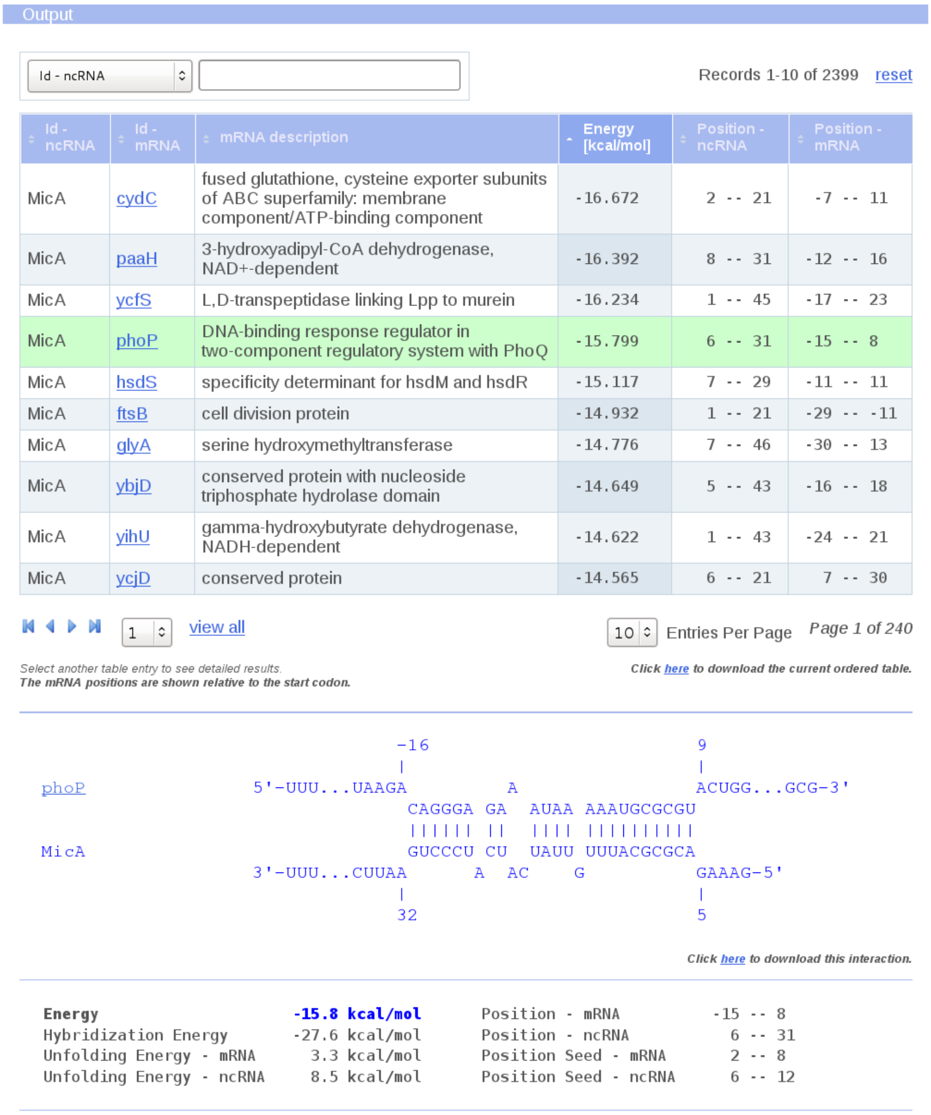
The output of the IntaRNA web server consists of a table that summarises the prediction results and links to all predicted interactions between the ncRNAs and mRNAs (Figure 2.9). The output table can be sorted by columns to allow selection of interactions by sequence identifier, annotated mRNA description, interaction energy score or interaction position in each sequence. The table can also be searched by sequence identifier or mRNA description. All mRNA identifier link to the corresponding entry in NCBI's Entrez Gene database [108], which provides the user with a lot of additional information on the putative target genes. The IntaRNA raw results, the output table in current sorting and the selected interaction can be downloaded as a text file. Finally, the server provides a link to the source code of IntaRNA as the stand-alone command-line version is more convenient and appropriate for large-scale studies.

The Freiburg RNA tools web server can be accessed at http://rna.informatik.uni-freiburg.de. The web server is based on a general framework developed for the CPSP web tools server [114] and has been continuously improved. Performing complex RNA analysis tasks, our server complements available web servers such as the Vienna RNA web suite [68] and the UNAFold web server [116, 216]. Other web servers specifically designed for the prediction of sRNA targets are, for example, provided by TargetRNA [181] and RNApredator [46]. The latter is based on the tool RNAplex and allows to analyse the predicted targets for enrichment of Gene Ontology (GO) and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway terms.

2.7 Discussion

In this chapter, we presented IntaRNA, a new method for the prediction of interactions between two RNAs based on minimisation of an extended hybridisation energy score via dynamic programming. IntaRNA accounts for two important features that influence the strength of RNA–RNA interactions and the selection of target sites: (i) the accessibility of the interaction sites in both RNAs and (ii) the existence of an interaction seed region. In contrast to previous methods for the prediction of RNA–RNA interactions, both features are integrated in a general approach for arbitrary RNAs. Although IntaRNA was applied to predict bacterial sRNA targets in this work, the program can readily be used to find other RNA–RNA interactions as well. For example, it was recently applied to miRNA–mRNA interactions by Marín and Vaníček [115] and Starczynowski et al. [172].

IntaRNA's prediction performance was evaluated by predicting targets for bacterial sRNAs and comparing the results to several state-of-the-art methods for the prediction of RNA–RNA interactions. IntaRNA outperformed the other methods in the accuracy of the predicted interaction. It performed as well as the best other program, RNAup, on a genome-scale search for putative sRNA targets, while the required CPU time and memory decreased drastically. Overall, the results showed that our method is well suited



Output

Id - ncRNA ⇕ □

Records 1-10 of 2399 reset

Id - ncRNA	Id - mRNA	mRNA description	Energy [kcal/mol]	Position - ncRNA	Position - mRNA
MicA	cydC	fused glutathione, cysteine exporter subunits of ABC superfamily: membrane component/ATP-binding component	-16.672	2 -- 21	-7 -- 11
MicA	paaH	3-hydroxyadipyl-CoA dehydrogenase, NAD+-dependent	-16.392	8 -- 31	-12 -- 16
MicA	ycfS	L,D-transpeptidase linking Lpp to murein	-16.234	1 -- 45	-17 -- 23
MicA	phoP	DNA-binding response regulator in two-component regulatory system with PhoQ	-15.799	6 -- 31	-15 -- 8
MicA	hsdS	specificity determinant for hsdM and hsdR	-15.117	7 -- 29	-11 -- 11
MicA	ftsB	cell division protein	-14.932	1 -- 21	-29 -- -11
MicA	glyA	serine hydroxymethyltransferase	-14.776	7 -- 46	-30 -- 13
MicA	ybjD	conserved protein with nucleoside triphosphate hydrolase domain	-14.649	5 -- 43	-16 -- 18
MicA	yihU	gamma-hydroxybutyrate dehydrogenase, NADH-dependent	-14.622	1 -- 43	-24 -- 21
MicA	ycjD	conserved protein	-14.565	6 -- 21	7 -- 30

1 ⇕ view all 10 ⇕ Entries Per Page *Page 1 of 240*

Select another table entry to see detailed results.
The mRNA positions are shown relative to the start codon.

Click here to download the current ordered table.

```
                             -16                      9
                             |                        |
phoP            5'-UUU...UAAGA         A                ACUGG...GCG-3'
                              CAGGGA GA  AUAA AAAUGCGCGU
                              |||||| ||  |||| ||||||||||
MicA                          GUCCCU CU  UAUU UUUACGCGCA
                3'-UUU...CUUAA      A  AC    G          GAAAG-5'
                             |                          |
                             32                         5
```

Click here to download this interaction.

```
Energy                      -15.8 kcal/mol    Position - mRNA         -15 -- 8
Hybridization Energy        -27.6 kcal/mol    Position - ncRNA          6 -- 31
Unfolding Energy - mRNA       3.3 kcal/mol    Position Seed - mRNA      2 -- 8
Unfolding Energy - ncRNA      8.5 kcal/mol    Position Seed - ncRNA     6 -- 12
```

Figure 2.9. Screenshot of the IntaRNA web server result page for a target prediction with input as given in Figure 2.8B. The table summarises all predicted interactions. It can be sorted by clicking on the header of a column. The interaction shown below the table is highlighted in green. For this prediction, additional information such as the interaction positions and the different contributions of the interaction energy score are given.

both for general searches for putative sRNA target sites and the prediction of accurate sRNA–mRNA interactions. The comparison with RNAhybrid, whose hybridisation energy model is the basis of our more sophisticated extended hybridisation energy, shows that the incorporation of interaction site accessibility and the existence of an interaction seed improve the prediction quality. In addition, we determined SD sequence locations for the sRNA targets and analysed the change in the probability that the SD sequence is unpaired as consequence of the predicted sRNA–mRNA interaction. This approach allowed us to successfully predict the regulatory effect on translation initiation for a number of sRNAs from our test set.

Only a fraction of IntaRNA's runtime is spent for the calculation of the actual hybridisation, whereas the calculation of interaction site accessibilities in terms of ED values is computationally much more intensive. Although IntaRNA is already considerably faster than the best other method RNAup, it could be further sped up by precomputing the ED values of all target regions only once for a given genome. Then, only the ED values of the ncRNA and the hybridisation energy have to be calculated for a target prediction in this genome. When all accessibilities are precomputed, the IntaRNA algorithm runs in only $O(n \cdot m)$ time and space.

The fast runtime of IntaRNA is achieved by a simplification: for each leftmost interaction base pair, only one right interaction end is stored and evaluated when searching for the optimal interaction. Despite this heuristic simplification, the prediction accuracy is high and the interaction energy predicted by IntaRNA is close to or equal to the mfe, which can be computed by the non-heuristic version of the algorithm. The heuristic of IntaRNA could be further improved by considering multiple different right interaction ends instead of just one. A study on 36 interactions showed that three to five different right ends have to be stored to achieve the same performance as the non-heuristic version. The performance gain is, however, marginal and achieved at the expense of a doubled runtime.

The interaction between the sRNA OxyS and its target mRNA *fhlA* is the only one in our test set with a discontinuous hybridisation pattern. In fact, the two RNAs form kissing hairpins at two sites [6]. The complex joint secondary structure prediction approaches that form the fourth class of methods presented in Section 2.1 can be used to predict interactions that consist of such independent substructures. These algorithms are, however, rather expensive with a time complexity of $O(n^3 \cdot m^3)$. Further extensions of IntaRNA could incorporate some basic ideas of those approaches to allow for prediction of more complex interactions with two or multiple interaction sites.

Chapter 3

Evaluation of bacterial RNA–RNA interaction features

In the following chapter, we present an analysis on how specific interaction features can be used to improve genome-wide target predictions for bacterial sRNAs. This is especially important as there exists a high demand for accurate large-scale sRNA target identification approaches to functionally characterise novel sRNAs. First, we determined features that displayed significant differences between functional and non-functional interactions. Second, we evaluated how genome-wide target searches improve by using these features as additional criteria. We focused especially on structural accessibility, sequence and interaction conservation, and interaction seed regions. Third, we assessed the importance of interaction site accessibility and conserved seed regions in a case study in the cyanobacterium *Prochlorococcus* MED4 and identified two novel mRNA targets of the sRNA Yfr1.

3.1 Understanding interaction formation principles to improve target predictions

A key task in the functional characterisation of base-pairing sRNAs is the identification of their interaction partners. To tackle this problem, a growing number of studies complements experimental approaches for target identification [162] by computational approaches for the prediction of sRNA targets and sRNA–target interactions [9]. All current *in silico* methods, however, still suffer from a high false positive rate, which becomes obvious when applying the methods to genome-wide target searches (Section 2.3). Therefore, an in-depth understanding of the mechanisms and principles that govern sRNA–target interactions is required to improve the identification of target genes. The pairing between sRNAs and their targets usually involves a seed region of six to eight contiguous bp [66]. The seed feature is employed by `IntaRNA` and other sRNA target prediction methods [182]. The sRNAs typically utilise well-accessible regions, i.e. hairpin loops or extended single-stranded sequence stretches, to recognise their targets [176]. Therefore, `IntaRNA` and some other RNA–RNA interaction prediction approaches account for the structure of the interaction partners via interaction site accessibility [127, 156, 179]. A recent study by Peer and Margalit [137] showed for a set of sRNAs from *E. coli* that the target-binding regions in the sRNAs exhibit characteristically high accessibility and conservation. Little systematic investigation, however, has been made on features associated with the binding sites of mRNAs that are targeted by sRNAs. Therefore, we explored to which extent accessibility and sequence conservation are general features of interaction sites in sRNAs and their target mRNAs and whether these features can be used to discriminate functional from non-functional interactions. To this end, we compiled a comprehensive set of 74 experimentally verified sRNA–target interactions from the enterobacteria *E. coli* and *Salmonella*, and generated an appropriate dataset of non-functional interactions. By comparison between functional and non-functional interactions, we found that true interaction sites are significantly more accessible in both sRNAs and targets. In contrast to accessibility, only interaction sites in sRNAs show high sequence conservation, while the conservation of target sites and target regulation is rather limited. Comparative sRNA target prediction approaches using target conservation information can therefore only predict a minor subclass of interactions that exhibit broad evolutionary conservation. When a specific target gene is, however, broadly conserved and evolves slowly, then this gene might be of high functional importance. An analysis of the nucleotide composition of interaction sites and flanking regions showed an enrichment of putative binding sites of the RNA-binding protein Hfq, which facilitates base pairing between sRNAs and their targets [193]. Finally, we combined our findings with the target prediction tool `IntaRNA` to improve the specificity of genome-wide sRNA target searches.

Apart from our enterobacterial dataset of sRNA–target interactions, we additionally investigated the importance of accessibility and of a seed region in a case study in the cyanobacterium *Prochlorococcus*. We predicted putative interaction partners of the sRNA Yfr1 and experimentally validated these candidates by a reporter system based on green fluorescent protein (GFP). These experiments confirmed that Yfr1 regulates two mRNAs at the post-transcriptional level. Furthermore, this case study showed that the incorporation of a seed region and a scoring of interaction site accessibility can successfully reduce the number of predicted sRNA target candidates for subsequent experimental validation.

3.2 Features of functional interaction sites

3.2.1 Dataset of experimentally verified sRNA–mRNA interactions

This analysis used a dataset of 71 sRNA–target pairs involving 19 distinct sRNAs from the two bacterial model organisms *E. coli* and *Salmonella* (Figure 3.1 and Tables A.1

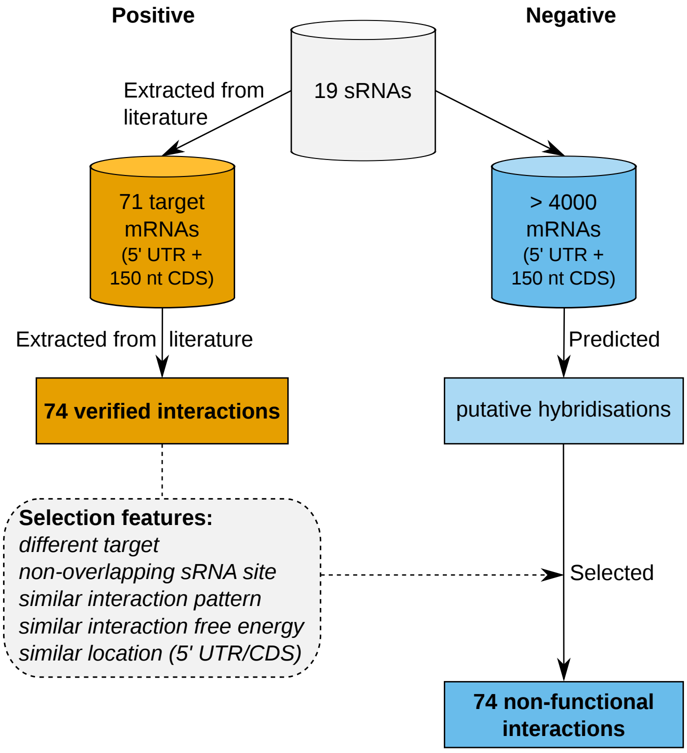

Positive
Negative
19 sRNAs
Extracted from literature
71 target mRNAs (5' UTR + 150 nt CDS)
> 4000 mRNAs (5' UTR + 150 nt CDS)
Extracted from literature
Predicted
74 verified interactions
putative hybridisations
Selection features:
different target
non-overlapping sRNA site
similar interaction pattern
similar interaction free energy
similar location (5' UTR/CDS)
Selected
74 non-functional interactions

Figure 3.1. Overview of data used in this analysis. The positive dataset (orange) consists of experimentally validated sRNA–mRNA interactions from literature and the negative dataset (blue) consists of predicted non-functional interactions that closely resemble the positive data.

and A.2). These two species were selected due to the availability of a high number of validated interactions. Three of the 71 sRNA–mRNA pairs interact at two separate sites: GcvB sRNA uses redundant regions to pair its target *cycA* [165], RybB sRNA can pair two alternative sites within its target *ompD* [11], whereas OxyS sRNA forms two kissing hairpin interactions with *fhlA* mRNA [6]. Thus, there is a total of 74 interactions in our dataset. All interactions were experimentally verified by *in vitro* (structural) probing or mutational studies at the interaction sites (see Tables A.1 and A.2 for references). The interaction seed lengths, which are defined by the length of the longest continuously paired region, range from 5 to 19 bp. The interaction sites in the targets are located between positions -131 and $+78$, relative to the translation start.

For the analysis of conservation, 21 enterobacterial species were included (Figure 3.2 and Table A.3). To search for homologous sRNA and mRNA sequences, the complete genomes of these species were retrieved from NCBI RefSeq database [145]. Homologs of each *E. coli* and *Salmonella* sRNA were identified in these 21 genome sequences using the semi-global alignment tool `GotohScan` (E-value cut-off of 0.01) [72]. Sequence-based alignment methods like `GotohScan` are appropriate for structural RNAs when the pairwise sequence identity is at least 50–60 percent [54]. Therefore, to reduce the number of false positives, sequences identified as homologs were rejected when the pairwise sequence

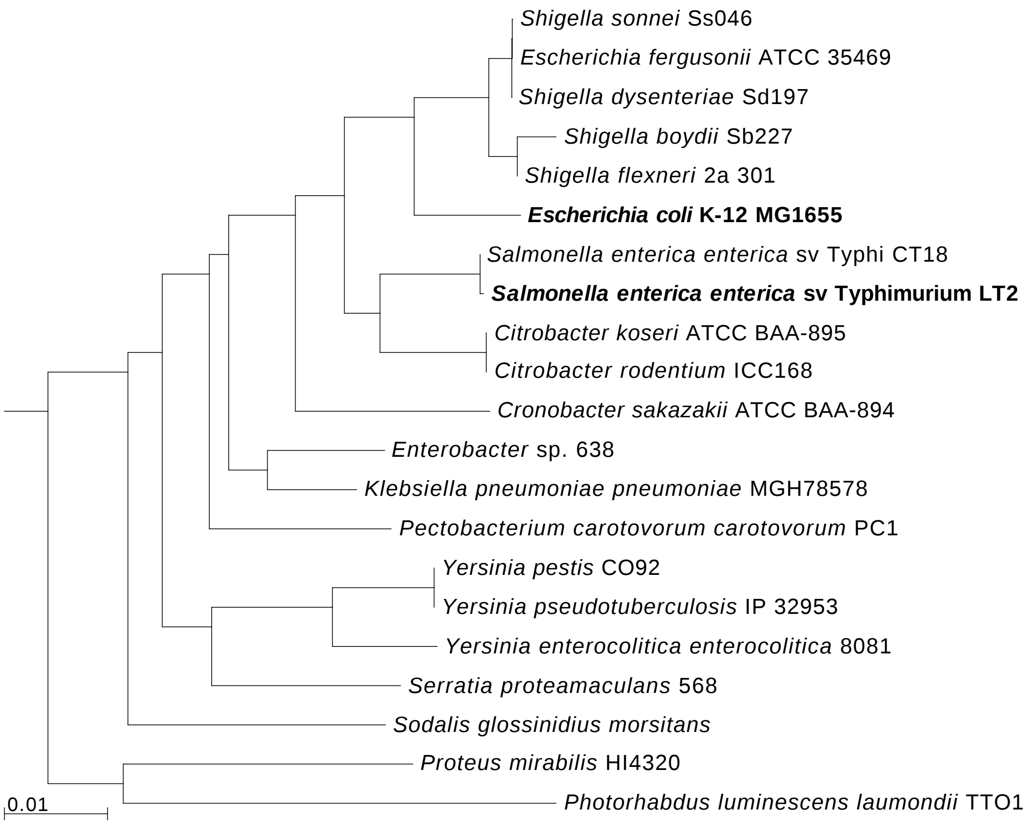

Shigella sonnei Ss046
Escherichia fergusonii ATCC 35469
Shigella dysenteriae Sd197
Shigella boydii Sb227
Shigella flexneri 2a 301
***Escherichia coli* K-12 MG1655**
Salmonella enterica enterica sv Typhi CT18
***Salmonella enterica enterica* sv Typhimurium LT2**
Citrobacter koseri ATCC BAA-895
Citrobacter rodentium ICC168
Cronobacter sakazakii ATCC BAA-894
Enterobacter sp. 638
Klebsiella pneumoniae pneumoniae MGH78578
Pectobacterium carotovorum carotovorum PC1
Yersinia pestis CO92
Yersinia pseudotuberculosis IP 32953
Yersinia enterocolitica enterocolitica 8081
Serratia proteamaculans 568
Sodalis glossinidius morsitans
Proteus mirabilis HI4320
Photorhabdus luminescens laumondii TTO1
0.01

Figure 3.2. Phylogenetic tree of the 21 enterobacterial species used for conservation analysis. Distances are based on 16S rRNA genes. Positive data was verified experimentally in *E. coli* and *Salmonella*, which are highlighted in bold. The tree was generated using the integrated microbial genomes system (IMG) [117].

identity to the query sequence was less than 60 percent. Each set of homologous sRNA sequences was then structurally aligned with `LocARNA-P`, applying probabilistic consistency transformation [211].

Groups of homologous (specifically orthologous) mRNAs were identified with the tool `OrthoMCL` [105] using all annotated mRNAs except pseudo genes as input. The accurate calculation of structural RNA properties such as thermodynamic stability or accessibility requires the precise definition of transcripts, but transcription start sites (TSSs) are currently not part of the gene annotation in genome databases. Therefore, we compiled a set of all mRNAs with accurate 5' UTRs. The 5' UTR lengths were obtained from two genome-wide studies that experimentally determined TSSs in *E. coli* by high-throughput sequencing and directed mapping [31, 122]. Since both datasets missed the TSS of two *E. coli* genes of our interaction dataset (*dpiB* and *nanC*) and of six further genes of which the *Salmonella* ortholog is included in our interaction dataset (*ompD*, *ompF*, *ompN*, *ompS*, STM3216 and STM4351), we determined the 5' UTR lengths of these genes from the literature that reports the corresponding interaction. In total, 5' UTR lengths were obtained

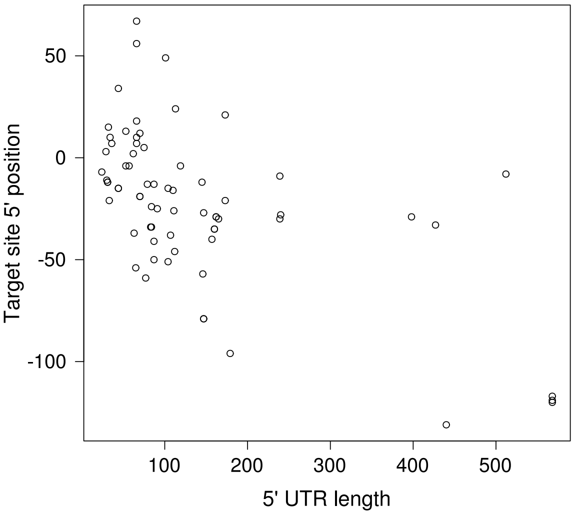


Figure 3.3. Positions of the 5' end of all target sites from *E. coli* and *Salmonella* vs. the lengths of the 5' UTR of the target genes. Target site positions are given as distance to the annotated translation start site. All genes that are located within an operon are excluded. Target site location and 5' UTR length show modest negative correlation (Spearman's correlation coefficient $\rho = -0.54$).

for 2313 different *E. coli* genes, which is about 56 percent of all annotated genes. 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 of the mRNA was set to the maximal 5' UTR length of all orthologs.

The lengths of the 5' UTR of the target genes and the positions of the target interaction sites relative to the translation start site show a modest negative correlation (Spearman's correlation coefficient $\rho = -0.54$ with p-value of 1.3×10^{-6}, Figure 3.3). Constrained by the transcript length, an interaction site can of course only be located far upstream of the start codon if the 5' UTR is sufficiently long (compare, e.g. DsrA–*rpoS* and Spot42–*gltA*). A long 5' UTR, however, does not necessarily imply that the interaction site is located upstream and in large distance to the translation start (compare, e.g. Spot42–*nanC* and RybB–*ompA*).

For each annotated mRNA, the 5' UTR sequence and the first 150 nt CDS were extracted from the genomic sequence. If the TSS position was unknown or if the gene was encoded within 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 [122]. The sequences of orthologous genes were then aligned with `MAFFT`, more precisely with method E-INS-i from the `MAFFT` package that uses a generalised affine gap cost model [89].

Alignments of homologous sRNA sequences were generated incorporating structural information, which is advisable for structural RNAs to obtain high-quality alignments. In contrast, homologous mRNA sequences can contain large unalignable regions, especially in the 5' UTRs, and mRNAs are not expected to fold into a common global structure. Therefore, we resorted to a pure sequence-based alignment method for the mRNA sequences. Note that we only compared functional and non-functional sites in either sRNAs or mRNAs. The use of two different tools could lead to a bias in results on conservation when comparing sRNAs with mRNAs.

3.2.2 Dataset of non-functional interactions

Interaction site features of the experimentally validated sRNA–mRNA interactions were evaluated by comparison to a negative dataset that contained one non-functional interaction per validated interaction. Since we wanted to investigate interaction site features independent of specific RNA–RNA hybridisation patterns, we took great care to generate a negative dataset in which each non-functional interaction closely resembles the intermolecular base pairing and hybridisation free energy of the respective functional interaction. Ideally, the precise form of the hybridisation duplex is maintained and only the associated sequences are exchanged, which was possible for about half of the true interactions. In the other cases, we resorted to the next best option, namely preserving the number of interaction base pairs. Furthermore, each non-functional interaction was required to involve another mRNA gene and another region in the sRNA than the respective validated interaction.

To this end, we first predicted putative hybridisations between each *E. coli* and *Salmonella* sRNA and the full 5' UTR and 150 nt CDS of all genes for which orthologous genes were identified. The hybridisations were predicted with `IntaRNA` neglecting accessibility, which typically results in extended stretches of complementary sequences. We then extracted all sub-hybridisations of these predicted hybridisations for which the hybridisation pattern was equal to the verified interaction. When such a sub-hybridisation did not exist, we searched for a sub-hybridisation where the number of base pairs (and optionally the interaction length) was equal to the verified interaction. Additionally, the sub-hybridisations had to satisfy the following properties: the mRNA is not the true target and the sRNA interaction site does not overlap the true sRNA interaction site. The last condition that the mRNA interaction site is in the CDS if and only if the verified interaction site is in the CDS is motivated by the fact that protein-coding and non-coding regions are subject to different evolutionary constraints. Finally, the sub-hybridisation with the closest hybridisation free energy to the validated interaction was selected as the corresponding non-functional interaction. By selecting only one non-functional interaction for every validated interaction, we gained a balanced set of functional and non-functional instances. Alignments of non-functional targets with their homologous genes predicted by `OrthoMCL` were generated as described above. An overview on the construction of the negative dataset is presented in Figure 3.1. The set of non-functional interactions is given in Table A.4. In addition, a second negative dataset was created using the aforementioned approach except that, in the final step, non-targets were not selected based on the free energy of the sub-hybridisation. Instead, the overall accessibility of each non-functional target had to be as close as possible to the overall accessibility of the corresponding true target. To evaluate the overall accessibility, we computed the expected fraction of unpaired bases in the mRNA sequence from the average probability that a single nucleotide is unpaired over all positions in the sequence.

The sRNA GcvB is known to directly regulate 21 mRNAs, which is the largest number of validated targets for a single sRNA [165]. In total, GcvB alters mRNA expression levels of about one percent of all protein-coding genes in *Salmonella*. Assuming that each sRNA has a similar number of targets, it is very unlikely that an mRNA randomly selected as a non-target is actually a true target of the sRNA.

Negative data could have also been obtained from the database sRNATarBase, which contains experimentally proven non-functional interactions [25]. However, it was not used in this study as it does not contain enough entries to obtain a non-functional interaction for each verified interaction. Furthermore, by constraining the predicted hybridisations to be as close as possible to the verified interactions, it was possible to focus on interaction site features independent of the actual hybridisation pattern.

3.2.3 Interaction sites are significantly accessible

The accessibility of an interaction site can be assessed by its probability of being unpaired (denoted by PU), which can be calculated from the ensemble free energy needed to open the region. This measure has the advantage to account for all secondary structures that can be formed by a particular RNA sequence, i.e. the whole thermodynamic ensemble of structures is considered instead of a specific mfe structure. A formal definition of PU values is given in Equation (2.2) in Section 2.2. Since the length of interaction sites varies for each sRNA–target pair and the expected PU values decrease with length, PU values can only be compared for regions of equal length. Therefore, we used the PU values to compute the expected fraction of unpaired bases at each interaction site (denoted by EF), which is a length-independent measure [73]. The expected fraction of unpaired bases of a subsequence $s_a \dots s_b$ of an RNA sequence s is defined by

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

The PU values, which are required for the calculation of EF values, were computed for sRNA sequences by global folding with `RNAup` [127]. As mRNA sequences should be folded locally in contrast to sRNA sequences [100], PU values of mRNA sequences were computed with `RNAplfold` [16] using a sliding window approach with a 140 nt folding window and a maximal base pair span of 70.

The accessibilities, i.e. EF values, of the interacting regions in sRNAs and mRNAs were

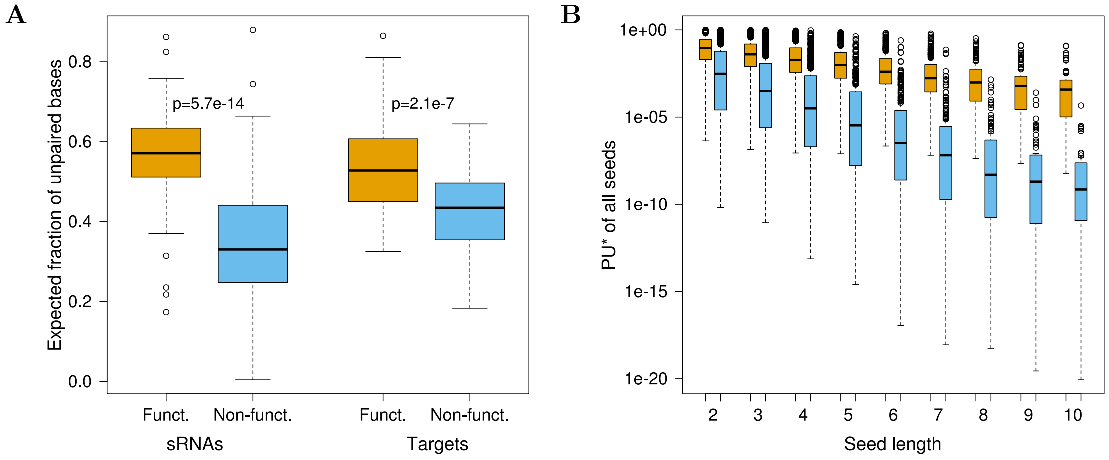


Figure 3.4. Comparison of the interaction site accessibility between functional (orange) and non-functional (blue) interactions. The plots show the **(A)** accessibility of the whole interaction site and **(B)** joint probability of being unpaired (PU^*) of all interaction seeds, i.e. all perfectly matching sub-interactions, of length two to ten. Interaction sites both in sRNAs and targets are significantly more accessible in functional than in non-functional interactions (p-values calculated by Wilcoxon rank sum test). In addition, interaction seeds in the functional interactions are significantly more accessible than in the non-functional interactions ($p < 4.8 \times 10^{-19}$ calculated by Wilcoxon rank sum test).

then compared between the experimentally verified interactions and the non-functional interactions. As shown in Figure 3.4A, the interaction sites of the experimentally verified interactions are more accessible than the corresponding sites in the non-functional set. 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).

To ensure that the observed high target site accessibility is not just an artefact from negative data construction, we compared the overall structuredness of functional and non-functional targets in terms of expected fraction of unpaired bases in the full 5' UTR and 150 nt CDS. We observed that the functional targets are slightly more accessible over the whole sequence (Figure 3.5A). For the set of non-functional interactions, we aimed for finding non-functional targets that share as many features as possible with the functional targets. The functional targets showed, however, a slightly larger overall accessibility than the non-functional ones. Therefore, we additionally created a second set of non-functional interactions, in which we selected each non-target to have an overall accessibility as close as possible to the corresponding true target. The overall accessibilities of these non-functional targets do not differ significantly from the functional targets (Figure 3.5A, $p = 0.389$ by Wilcoxon rank sum test). For both negative datasets, the differences in median and mean accessibility between functional and non-functional targets are much larger for the interaction sites only than for the whole sequence (more than 5-fold and 10-fold increase for original and second negative dataset, respectively, see Figure 3.5). In summary, the

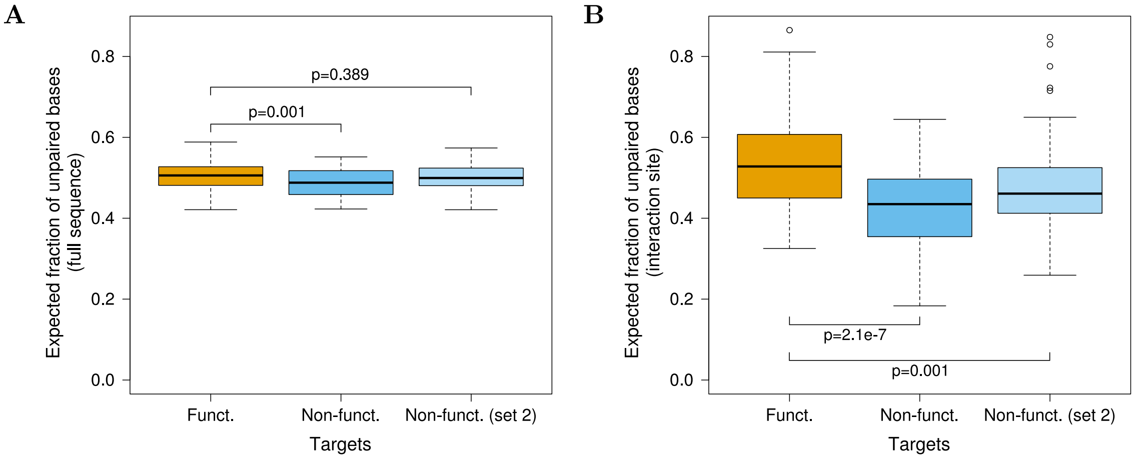


Figure 3.5. Comparison of the target gene accessibility between a functional (orange) and two non-functional datasets (blue and light blue). The plots show the accessibility of **(A)** the full 5' UTR and 150 nt CDS, and **(B)** the interaction site only, respectively. In non-functional dataset 2 (light blue), each non-functional target was not selected by the interaction free energy (as done for the first non-functional dataset), but by the overall accessibility, which had to be as similar as possible to the overall accessibility of the corresponding true target. For both negative datasets, the difference in mean accessibility between functional and non-functional targets is larger and much more significant for the interaction sites only than for the whole sequence. All p-values were calculated by Wilcoxon rank sum test.

higher accessibility of the functional target sites cannot be explained by differences in the structuredness of the compared mRNA datasets alone.

The results on interaction site accessibility in both sRNA and target motivated us to explore the accessibility information of the interacting RNAs in greater detail. Based on the observation that two short well-accessible regions often form the initial interaction [23], we examined the accessibility of all putative seed regions defined by perfectly matching sub-interactions (allowing Watson–Crick and `G-U` wobble base pairs) of length two to ten. We assessed whether the accessibility information of two interacting RNAs can be combined into a single feature.

For this purpose, 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 a (sub-)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.

We then compared the joint probability of being unpaired (PU^*) for all seed regions of length two to ten between true interactions and non-functional interactions. Figure 3.4B shows that PU^* of the functional interactions is significantly higher for all analysed seed lengths ($p < 4.8 \times 10^{-19}$ by Wilcoxon rank sum test). Consequently, the accessibility of interaction seed regions, which is represented by the single feature PU^*, can be used to discriminate functional from non-functional interactions.

3.2.4 Interaction sites are only significantly conserved in sRNAs

An analysis of evolutionary conservation was performed on alignments of homologous sRNAs and mRNAs per interaction site. The sequence conservation of each interacting region was assessed by the average information content of the alignment columns corresponding to the known interaction site in *E. coli* or *Salmonella*, respectively. The information content allows a comparison between alignments that differ in the number of included species. We used an extended expression of this measure that also incorporates scoring of gaps in the alignment [63].

The information content I_i of an alignment column $\mathcal{A}_i$ 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 $A = \{\texttt{A,C,G,U,-}\}$ is the set of nucleotides including gaps, q_{ik} is the observed frequency of the symbol $k \in A$ in alignment column $\mathcal{A}_i$, and p_k is the background symbol distribution [63]. We set $p_{\texttt{-}} = 1$ 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 from $\mathcal{A}_a$ 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.

The sequence conservation of all functional sRNA and mRNA interaction sites was then evaluated by comparison to the sequence conservation of the sites involved in the non-functional interactions. Figure 3.6A shows that true sRNA interaction sites are significantly more conserved than non-functional interaction sites ($p = 1.5 \times 10^{-6}$ 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, indicates that conservation of sRNA–mRNA interactions among related bacterial species might not be a general feature. Despite lack of target sequence conservation, it may, however, be that intermolecular base pairings are still 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` [76]. To examine to which extent consistent or compensatory mutations occurred, we counted the number of base pair types (out of the possible combinations `C-G`, `G-C`, `A-U`, `U-A`,

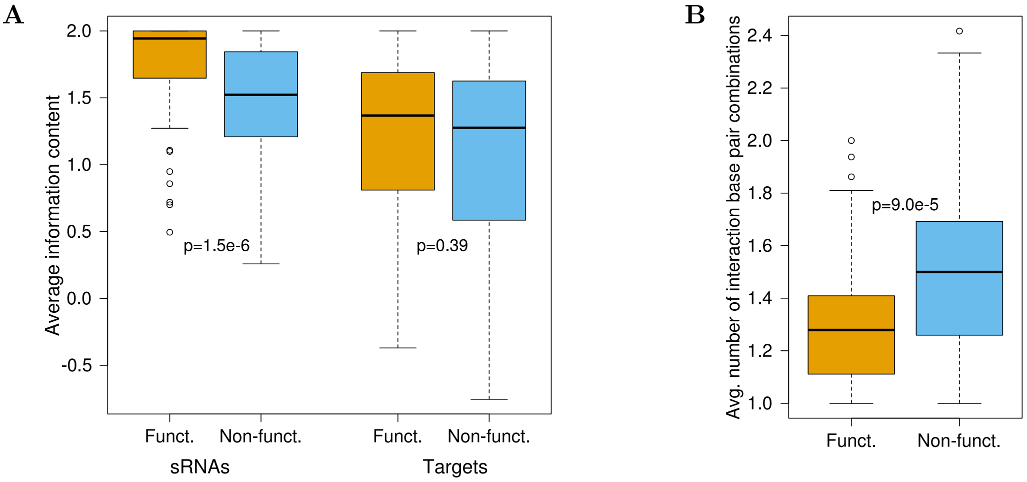


Figure 3.6. Comparison of interaction site features between functional (orange) and non-functional (blue) interactions. The plots show the **(A)** interaction site sequence conservation and **(B)** average number of different interaction base pairings. Only interaction sites in sRNAs, but not in targets, show significant evolutionary conservation. The average number of intermolecular base pair combinations is significantly smaller in the functional interactions. All p-values were calculated by Wilcoxon rank sum test.

G-U and U-G) per interaction position in the alignments. The functional set utilised the interactions experimentally validated in *E. coli* or *Salmonella* and the non-functional set utilised the hybridisations predicted for the *E. coli* or *Salmonella* sequences. An example is given in Figure 3.7. The results in Figure 3.6B show that the number of different base pair types is smaller in the confirmed interactions than in the non-functional interactions ($p = 9.0 \times 10^{-5}$ by Wilcoxon rank sum test). Hence, we can conclude that interactions between sRNAs and their targets are neither sequentially nor structurally conserved in general.

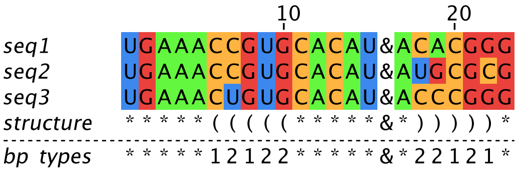

```
                    10            20
seq1       UGAAACCGUGCACAU&ACACGGG
seq2       UGAAACCGUGCACAU&AUGCGCG
seq3       UGAAACUGUGCACAU&ACCCGGG
structure  *****(((((*****&*)))))*
------------------------------------
bp types   *****12122*****&*22121*
```

Figure 3.7. Schematic illustration of different interaction base pairings between two interacting RNAs. The sequence alignment shows two multiple RNA alignments concatenated by the linker symbol '&'. Round brackets in the structure indicate intermolecular base pairs between the two RNAs and stars indicate positions that do not participate in the interaction. The number of different interaction base pairings is given for each interaction position. For example, alignment columns 9 and 19 support the base pair U-A in *seq1* and U-G in *seq2*, whereas *seq3* contains a mismatch. Consequently, alignment columns 9 and 19 contain two different base pair combinations. The average number of different interaction base pair combinations in this example is 1.6.

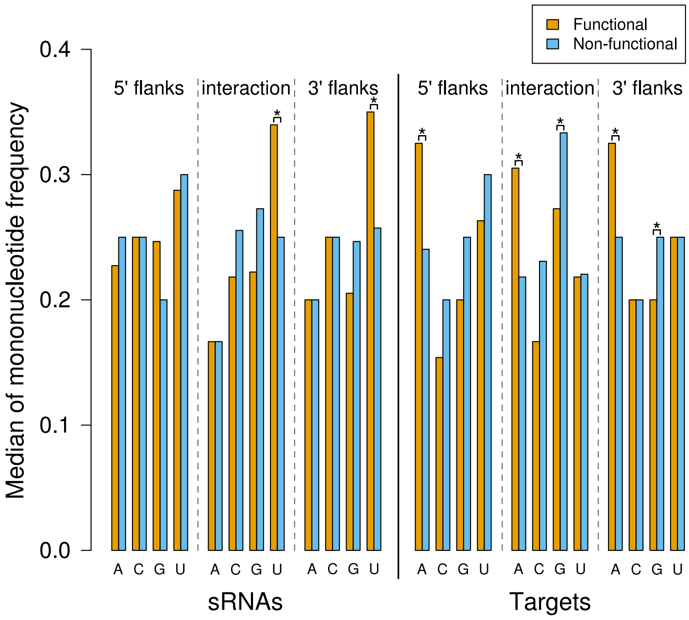


Figure 3.8. Comparison of the mononucleotide frequencies at interaction sites and 20 nt flanking regions between functional (orange) and non-functional (blue) interactions. The short poly(U) tails at the sRNA 3' ends were excluded from the analysis. p-values were calculated by Wilcoxon rank sum test; bars are marked by '*' if the differences in the mononucleotide frequencies are significant at the 0.01 level.

3.2.5 Sequence composition of interaction sites and flanking regions

To analyse whether functional interaction sites are characterised by specific sequence compositions, mononucleotide frequencies were determined at interaction sites and their flanking regions of at most 20 nt for both sRNAs and mRNAs. Bacterial sRNAs commonly possess a short poly(U) tail at their 3' end, which forms, together with the preceding stem–loop structure, the Rho-independent transcription terminator. In the following analysis these poly(U) tails were disregarded to avoid a bias in the sequence composition.

Figure 3.8 shows a comparison of mononucleotide frequencies of interaction sites and flanks between the functional and non-functional interactions. We found that the true interaction sites in sRNAs contain significantly more `U` nucleotides than the corresponding regions in the non-functional data, while the target sites contain significantly more `A` and less `G` (p-values of 0.0002, 3.4×10^{-6} and 0.001, respectively, by Wilcoxon rank sum test). The mutual enrichment of `U`s and `A`s in sRNAs and targets, respectively, ensures base pair complementarity between the two interacting RNAs. As both `A` and `G` are complementary to `U`, but alleviated `G` frequency was observed at target sites, `A-U` interaction base pairs might be favoured over less stable non-Watson–Crick `G-U` base pairs in sRNA–mRNA duplexes. Moreover, not only the true sRNA interaction sites, but also their 3' flanking regions have a significantly higher frequency of `U` than the corresponding regions in the non-functional data ($p = 0.0003$ by Wilcoxon rank sum test). Likewise, the regions flanking the target sites in both directions also show significantly higher A frequencies (p-values of 3.6×10^{-5} and 0.001 for 5' and 3' flanks, respectively, by Wilcoxon rank sum test). Low G content was found for the 3' flanks of the target sites ($p = 3.7 \times 10^{-5}$ by Wilcoxon rank sum test). Noteworthy, the nucleotide pattern observed here is consistent with the binding preference of the RNA chaperone Hfq toward A/U-rich regions [193].

3.2.6 Conservation of sRNA–target complementarity is limited

Many sRNAs of our dataset directly regulate multiple targets by binding via a single interaction site (although some sRNAs possess more than one target-binding site, e.g. FnrS, GcvB and Spot42). To gain further insight into the relationship between interaction site conservation in sRNAs and their targets, we selected two sRNAs with multiple targets, RyhB and RybB, and investigated the conservation of their target regulation in detail. For each target mRNA of these sRNAs, we analysed to which degree the base pairing between the two RNAs is conserved in related species by manual inspection of the multiple sRNA and mRNA sequence alignments. We distinguished between preserved complementarity of the full interaction and of a core interaction of at least six consecutive bp. Both consistent and compensatory mutations in the intermolecular pairing were considered.

The sequence of the first analysed sRNA, RyhB, was found to be conserved in 19 out of the 21 enterobacterial species considered here. Five RyhB targets have been experimentally verified in *E. coli* to date, of which *shiA* is translationally activated and the other four are subject to translational repression [40, 57, 142, 158, 191]. The target-binding site of RyhB is located between sequence positions 34 to 76. All interaction seeds are located in the highly conserved RyhB region between positions 34 to 55. Among the RyhB targets, the interaction with *cysE* is conserved in 17 out of 19 species when requiring a core interaction of at least six consecutive bp (Table 3.1). The full interaction is preserved in six species. For the target *sodB*, the interaction site is fully conserved in 12 out of the 16 species, in which an ortholog of *sodB* was identified. A conserved core interaction was additionally found in one species. The remaining three species with *sodB* ortholog carry a single mismatch within the nine bp interaction. In both *cysE* and *sodB*, the RyhB target site is located around the start codon. The lowest interaction conservation was found for the targets *fur* and *shiA*, each with a preserved complementarity in only six species.

The sequence of the second analysed sRNA, RybB, is conserved in all 21 species. Its 5' end sequence is fully conserved up to position 19. In *Salmonella*, it was shown that this 5' RybB domain base pairs ten mRNAs, which results in translational repression and mRNA destabilisation [11, 20, 135]. The base pairing between RybB and its target *ompA* is fully conserved in all analysed species except *Shigella dysenteriae*, where the target site includes a single mismatch (Table 3.2). Among the other nine RybB targets, the lowest degree of interaction conservation was found for *ompD* and *ompS* with conserved base pairing in only four and five species, respectively.

Table 3.1. Conservation of interactions between RyhB sRNA and its target mRNAs in 19 enterobacterial species.

Organism	Interaction conservation				
	cysE	*fur*	*iscS*	*shiA*	*sodB*
Escherichia coli K-12	X	X	X	X	X
Shigella dysenteriae	X	X	X	X	X
Escherichia fergusonii	X	X	X	x	X
Shigella sonnei	X	X	X	–	X
Shigella flexneri	X	x	X	–	X
Shigella boydii	X$^{\#}$	X	X	–	X
Salmonella Typhi	–	–	x	n/a	X
Salmonella Typhimurium	–	–	X	–	X
Citrobacter koseri	x	–	x	x$^{\#}$	X
Citrobacter rodentium	x	–	x	x$^{\#}$	X
Klebsiella pneumoniae	x	–	X	–	X
Enterobacter sp. 638	x	–	x	x	X
Pectobacterium carotovorum	x	–	X$^{\#}$	-	n/a
Yersinia pestis	x$^{\#}$	–	–	n/a	-
Yersinia pseudotuberculosis	x$^{\#}$	–	–	n/a	-
Yersinia enterocolitica	x	–	–	n/a	-
Sodalis glossinidius	x$^{\#}$	–	–	n/a	n/a
Proteus mirabilis	x*	–	x$^{\#}$	n/a	x
Photorhabdus luminescens	x	–	–	n/a	n/a
Conserved interactions	0.89	0.32	0.74	0.32	0.68

Orthologs of target genes were identified with `OrthoMCL` and by gene annotations. The last row gives the fraction of species in which interaction conservation was found. 'X' indicates full interaction conservation, 'x' indicates conservation of a core interaction (i.e. at least six consecutive bp), and '-' indicates no interaction conservation. 'n/a' indicates that no target ortholog was found. '#' and '*' mark interactions that contain consistent and compensatory mutations, respectively. Organisms are sorted by evolutionary distance to *E. coli* based on 16S rRNA genes.

Table 3.2. Conservation of interactions between RybB sRNA and its target mRNAs in 21 enterobacterial species.

Organism	Interaction conservation									
	chiP	*fadL*	*ompA*	*ompC*	*ompD*	*ompF*	*ompN*	*ompS*	*ompW*	*tsx*
Salmonella Typhimurium	X	X	X	X	X	X	X	X	X	X
Salmonella Typhi	X	X	X	X	n/a	X	X	X	X	X
Citrobacter koseri	X	X	X	x	x	X	X	X	X	x$^{\#}$
Citrobacter rodentium	X$^{\#}$	X	X	x	x	X	X	X	X	x
Shigella dysenteriae	-	X	x	x	n/a	-	n/a	n/a	X	X
Escherichia fergusonii	X	X	X	X	-	X	X	n/a	X	X
Shigella sonnei	X	X	X	x	-	X	x	n/a	X	X
Shigella flexneri	X	X	X	x	-	X	X	n/a	X	X
Escherichia coli K-12	X	X	X	x	n/a	X	x	n/a	X	X
Shigella boydii	n/a	X$^{\#}$	X	x	n/a	X	x	n/a	X	X
Klebsiella pneumoniae	X	-	X	X	n/a	-	X	X	-	x$^{\#}$
Cronobacter sakazakii	-	X	X	X$^{\#}$	n/a	X	X	n/a	-	x$^{\#}$
Enterobacter sp. 638	X$^{\#}$	X	X	X	x	X	X$^{\#}$	n/a	X	X
Pectobacterium carotovorum	n/a	X$^{\#}$	X$^{\#}$	n/a	n/a	X	n/a	n/a	-	-
Serratia proteamaculans	-	X	X$^{\#}$	X$^{\#}$	n/a	n/a	-	n/a	-	-
Yersinia pestis	-	-	X$^{\#}$	X$^{\#}$	n/a	X	n/a	n/a	-	n/a
Yersinia pseudotuberculosis	-	-	X$^{\#}$	X$^{\#}$	n/a	-	-	n/a	-	n/a
Yersinia enterocolitica	-	X	X$^{\#}$	X$^{\#}$	n/a	X	n/a	n/a	-	n/a
Sodalis glossinidius	n/a	n/a	X$^{\#}$	x$^{\#}$	n/a	x$^{\#}$	n/a	n/a	n/a	n/a
Proteus mirabilis	-	-	X$^{\#}$	n/a	n/a	X	n/a	n/a	-	n/a
Photorhabdus luminescens	n/a	X	X	n/a	n/a	n/a	-	n/a	-	n/a
Conserved interactions	0.48	0.76	1.00	0.86	0.19	0.76	0.57	0.24	0.52	0.62

Orthologs of target genes were identified with `OrthoMCL` and by gene annotations. The ortholog clusters of the *ompF* and *ompN* genes were hand-curated due to an incorrect assignment of evolutionary relationship (as already observed for bacterial porin genes in previous studies [39, 130]). The *ompD* gene contains two RybB interaction sites, but only the site at positions 18 to 26 is conserved, and thus only this site is included above. Organisms are sorted by evolutionary distance to *Salmonella* based on 16S rRNA genes. See Table 3.1 for details on the symbols.

3.3 Interaction seed constraints in genome-wide target predictions

In Section 3.2, we showed that high interaction site accessibility and strong sRNA interaction site sequence conservation are common features of bacterial sRNA–mRNA interactions. These observations suggest the following strategy to improve the false positive rate of genome-wide sRNA target predictions: (i) identify complementary regions in sRNA and putative target that are highly accessible, or (ii) identify conserved and weakly structured, i.e. accessible, regions in the sRNA that might serve as target-binding region. Subsequently, focus the target search to interactions that include these regions, which can be achieved by, e.g. constraining the position of the interaction seed region.

Interaction seeds were restricted to highly accessible regions in both RNAs by only allowing seeds with a high joint probability of being unpaired (PU^*). The background accessibility signal of a particular RNA sequence depends on sequence composition, e.g. `GC`-content, and folding parameters such as temperature and folding windows. Therefore, to define valid (i.e. accessible) seeds, the PU^* cut-off is computed individually for each pair of RNA sequences as the q-quantile of the sequences' background PU^* for a user-defined q. The target prediction tool `IntaRNA` already predicts RNA–RNA interactions starting from an interaction seed. We extended `IntaRNA` by optionally allowing only interaction seeds with a PU^* greater than the q-quantile of the background PU^* (which is computed as the average PU^* of all subsequences of length equal to the seed). Previously, `IntaRNA`'s interaction scoring already included an overall accessibility term, but did not allow to specifically restrict interaction seeds to highly accessible regions.

Candidate sRNA seeds in weakly structured and conserved regions were obtained from reliability profiles computed with the sequence-structure alignment tool `LocARNA-P` [211]. Positions in the input sequences are matched structurally by `LocARNA-P` if they are part of conserved base pairs, otherwise positions are matched non-structurally. The former case contributes to the structural reliability. In the latter case, the sequence positions are matched based on their sequence similarity, which contributes to the sequence reliability. Probabilistic alignment with `LocARNA-P` gives the reliabilities for sequence and base pair matches in each alignment column, which can be visualised in a reliability plot. Figure 3.9 gives an example of such a plot for RyhB sRNA. A stretch of alignment columns with high sequence but low structure reliability indicates a region with trustworthy alignment but without conserved base pairs, i.e. with conserved unstructuredness. Note that high sequence reliability is an indication for high sequence similarity, i.e. sequence conservation, but the two measures do not represent the same. We therefore computed the sequence identity of regions with high sequence but low structure reliability to identify regions that are conserved on sequence level, but without conserved secondary structure. We then extended `IntaRNA` by an optional constraint that allows to restrict the position of the

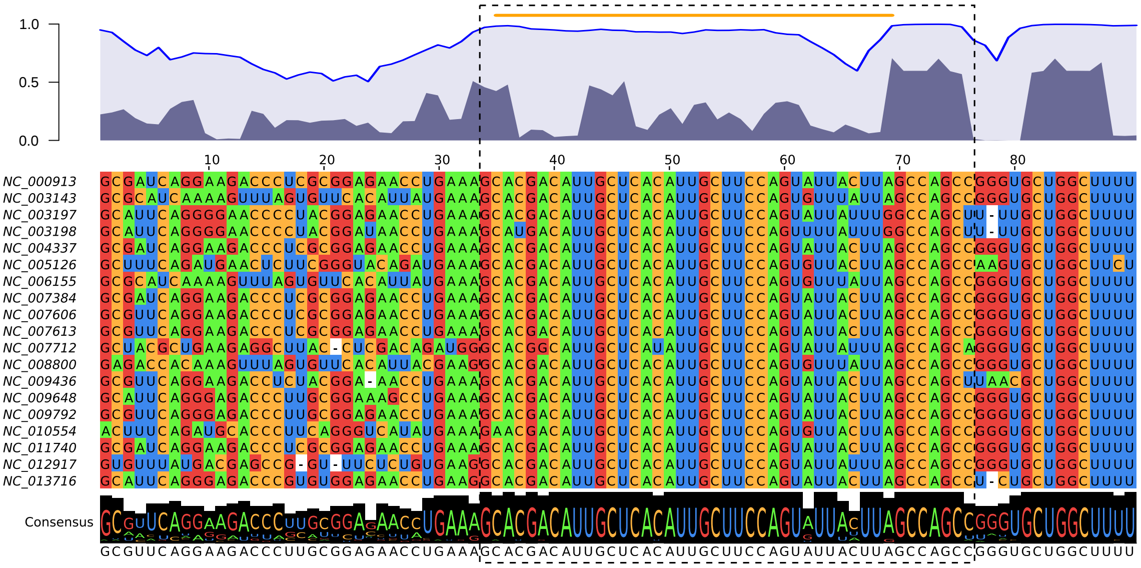


Figure 3.9. Alignment and reliability profile plot of RyhB sRNA homologs and the conserved and accessible RyhB region derived from them. In the reliability plot on top, the dark and light blue regions represent alignment column-wise structure and sequence reliabilities, respectively, and the blue line shows the combined column reliabilities. Below the alignment, the consensus sequence and the sequence conservation are shown. The RyhB target-binding region is boxed with a black dashed line. The region identified as conserved and accessible by comparison to background signals is indicated by the orange line; this region was used as seed constraint in the genome-wide prediction of RyhB targets with `IntaRNA`. Sequences in the alignment are labelled by the RefSeq genome accession number of each organism. The plots are projected to the *E. coli* sequence, i.e. columns with gaps in the *E. coli* sequence are excluded.

interaction seed to these sRNA regions.

In detail, reliability profiles together with the corresponding alignment were used as follows to determine well-conserved regions without conserved secondary structure: given a multiple sRNA alignment $\mathcal{A}$, we first determined the background signals of sequence identity, structure and sequence reliability, which are denoted by $\text{seqid}_{\mathcal{A}}^{bg}$, $\text{strrel}_{\mathcal{A}}^{bg}$ and $\text{seqrel}_{\mathcal{A}}^{bg}$, respectively. The background signal is defined as the average sequence identity or reliability over all alignment columns. Then, we identified windows of a fixed length n with an average sequence identity $\text{seqid}_{\mathcal{A}}^{win} \geq \gamma\ \text{seqid}_{\mathcal{A}}^{bg}$, an average structure reliability $\text{strrel}_{\mathcal{A}}^{win} \leq \delta\ \text{strrel}_{\mathcal{A}}^{bg}$ and an average sequence reliability $\text{seqrel}_{\mathcal{A}}^{win} \geq \varepsilon\ \text{seqrel}_{\mathcal{A}}^{bg}$. In this study, we used $\gamma = 1.0$, $\delta = 0.9$, $\varepsilon = 1.0$ and window length equal to the seed length. The windows satisfying the three conditions were considered as accessible conserved regions.

To evaluate the above seed constraints, we conducted genome-wide target predictions in *E. coli* and *Salmonella* for every sRNA in our dataset. Four different `IntaRNA` settings were used: (1) seed without accessibility and conservation constraints (default), (2) seed constraints derived from sRNA `LocARNA-P` reliability profile (e.g. orange line, Figure 3.9), (3) seed with PU^* in 0.8-quantile of background distribution, i.e. highly accessible in both RNAs, and (4) a combination of the seed constraints in (2) and (3). The other `IntaRNA` parameters were as follows: minimal seed length of seven consecutive base pairs and local mRNA structure folding with a maximal base pair span of 70 in a folding window of 140 nt. Putative interactions were searched in the full 5' UTR and 150 nt CDS of all genes, for which orthologous genes were identified. 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 the range −150 to +100. Additionally, we present the prediction results of the widely used sRNA target prediction tool `TargetRNA` [182] for comparison. `TargetRNA` was used 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 the best 100 target predictions per sRNA, which is the maximal number of targets that the web server returns for each target search.

The ROC-like plot in Figure 3.10 shows the total number of true positive predictions vs. the number of predicted targets per sRNA for all four `IntaRNA` settings and for `TargetRNA`. The best prediction performance was achieved when interaction seeds were restricted to conserved and weakly structured sRNA regions (orange line, setting [2]). Restricting the seeds to highly accessible regions in both target mRNA and sRNA (dark blue line, setting [3]) resulted in an almost similar performance. A combination of the two constraints did not further improve the results (light blue line, setting [4]). For all parameter settings including the default method without constraining the seed region (black line, setting [1]), `IntaRNA` clearly outperformed `TargetRNA` (grey line). The plot was restricted to the 100 best predictions per sRNA as this is the maximal number of targets reported by `TargetRNA`.

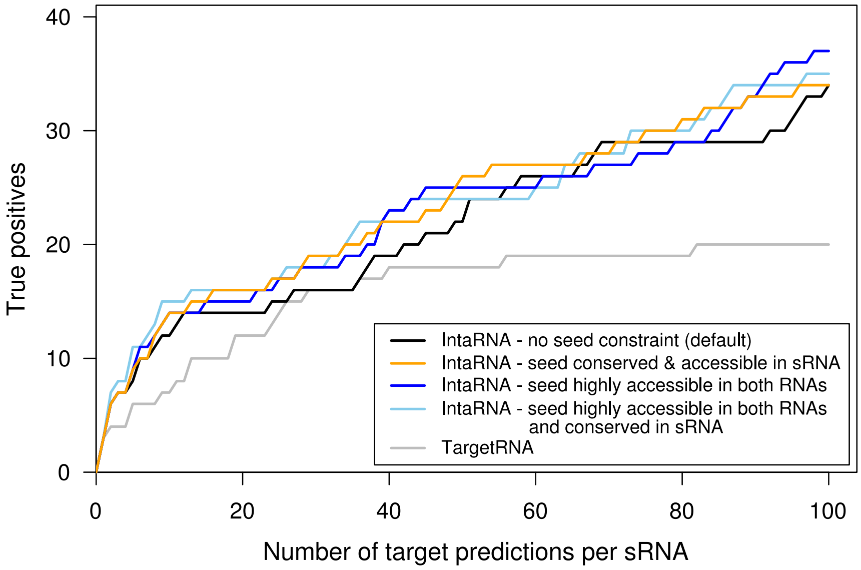


Figure 3.10. Genome-wide target predictions for 25 sRNAs to evaluate different constraints on the interaction seeds. The prediction performance of the tool `IntaRNA` using four different parameter settings is compared with the tool `TargetRNA`. The ROC-like plot shows the overall number of correctly predicted targets (y-axis) vs. the number of predictions per sRNA (x-axis) sorted by (energy) score. All `IntaRNA` predictions with constraints on the seed region (orange, light and dark blue lines) achieved a higher sensitivity (true positive rate) than `IntaRNA` without seed constraints (black line). Independent of the parameter setting used, `IntaRNA` always clearly outperformed `TargetRNA` (grey line).

3.4 Seed-based target identification for Yfr1 sRNA – a case study

3.4.1 Small RNAs in the cyanobacterium *Prochlorococcus*

For the two bacterial model organisms *E. coli* and *Salmonella*, it seems reasonable to assume that a great deal of the sRNAs expressed under standard experimental conditions have been determined, especially with the advent of RNA-seq [98, 147, 169]. These two organisms also account for the majority of experimentally characterised sRNA targets and were therefore selected for the analysis in Section 3.2. However, sRNA regulators are of course not restricted to model bacteria, but occur ubiquitously in bacteria. In the following case study, we investigate the ecologically important cyanobacterium *Prochlorococcus*. This photoautotrophically dwelling organism accounts for up to 50 percent of the organic biomass in the oligotrophic areas of the open oceans, and is thus a crucial component of the food web [62, 190]. A recent systematic survey of sRNAs in *Prochlorococcus* MED4 revealed a large number of potential regulatory RNAs comparable with those found in other bacteria [174]. This finding was very surprising, as *Prochlorococcus* has experienced an evolutionary streamlining of its genome, leading to very compact genomes between 1.64 and 2.68 Mb, which notably results in a small number of regulatory proteins [92].

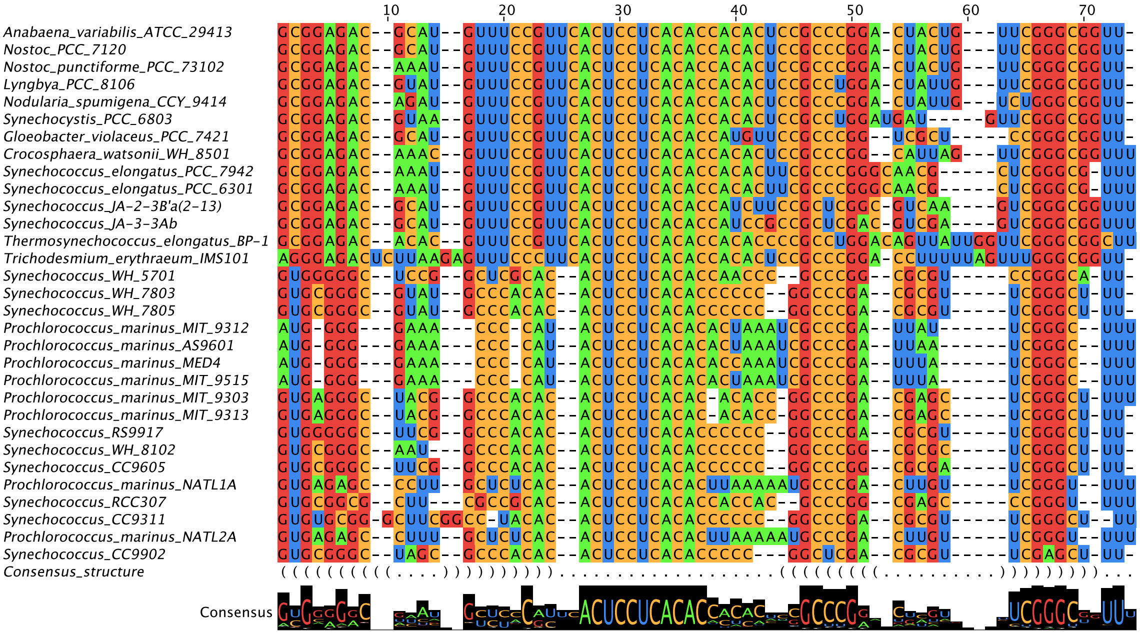

10
20
30
40
50
60
70
Anabaena_variabilis_ATCC_29413
Nostoc_PCC_7120
Nostoc_punctiforme_PCC_73102
Lyngbya_PCC_8106
Nodularia_spumigena_CCY_9414
Synechocystis_PCC_6803
Gloeobacter_violaceus_PCC_7421
Crocosphaera_watsonii_WH_8501
Synechococcus_elongatus_PCC_7942
Synechococcus_elongatus_PCC_6301
Synechococcus_JA-2-3B'a(2-13)
Synechococcus_JA-3-3Ab
Thermosynechococcus_elongatus_BP-1
Trichodesmium_erythraeum_IMS101
Synechococcus_WH_5701
Synechococcus_WH_7803
Synechococcus_WH_7805
Prochlorococcus_marinus_MIT_9312
Prochlorococcus_marinus_AS9601
Prochlorococcus_marinus_MED4
Prochlorococcus_marinus_MIT_9515
Prochlorococcus_marinus_MIT_9303
Prochlorococcus_marinus_MIT_9313
Synechococcus_RS9917
Synechococcus_WH_8102
Synechococcus_CC9605
Prochlorococcus_marinus_NATL1A
Synechococcus_RCC307
Synechococcus_CC9311
Prochlorococcus_marinus_NATL2A
Synechococcus_CC9902
Consensus_structure
Consensus

Figure 3.11. Multiple sequence alignment of Yfr1 homologs from 31 cyanobacteria. The sequence motif in alignment columns 27 to 37 is perfectly conserved in all species and predicted to be single-stranded in the consensus structure. The alignment and the consensus structure are based on Voß et al. [196]. In the structure, matching brackets indicate base pairs and dots indicate unpaired positions. The alignment was visualised with `Jalview` [204].

The identification of sRNA targets in *Prochlorococcus* constitutes a big challenge, since common experimental approaches such as knockouts of these sRNAs cannot be applied. Instead, the only possible approach is a combination of *in silico* target prediction, followed by *in vivo* experimental validation (in a heterologous expression system).

An interesting sRNA candidate to study is Yfr1, which is an abundant RNA with ubiquitous appearance in all lineages of cyanobacteria except for two *Prochlorococcus* strains [196]. Recent studies have shown that Yfr1 is constitutively expressed and accumulates up to 18000 copies per cell in *Synechococcus elangatus* PCC6301 [129]. The high copy numbers of Yfr1 raise the question of whether this RNA acts as a *trans*-encoded sRNA through base pairing with its mRNA targets, or whether it modulates protein activity as the 6S RNA, which downregulates mRNA transcription by mimicking an open promoter complex [201]. However, a prominent feature of Yfr1 is the ultraconserved 11 nt long sequence motif located in an unpaired sequence stretch flanked by two stem–loops (Figure 3.11). We showed in Section 3.2 that sequence conservation and single-strandedness are significant features that characterise target-binding sites in enterobacterial sRNAs. Therefore, we verify in the following whether the cyanobacterial sRNA Yfr1 regulates *trans*-encoded mRNAs via base pairing in analogy to *E. coli* and *Salmonella*.

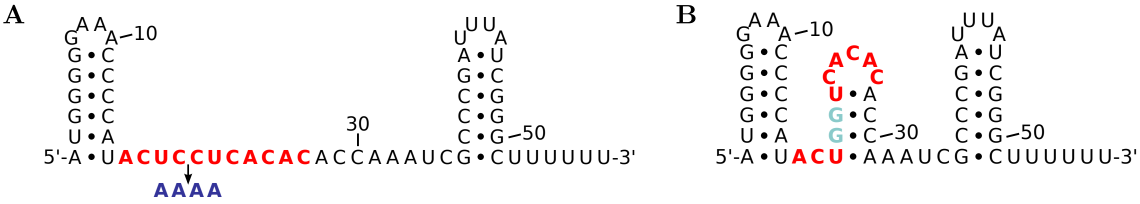

A
B
5'-A•UACUCCUCACACACCAAAUCG•CUUUUUU-3'
AAAA
-10
30
-50
5'-A•UACU•AAAUCG•CUUUUUU-3'
-30

Figure 3.12. **(A)** Secondary structure of *Prochlorococcus* MED4 Yfr1, as predicted by `RNAfold` [75]. The ultraconserved single-stranded region is highlighted in red. The arrow indicates the introduced mutation M2 (dark blue). **(B)** Secondary structure resulting from mutation M1 (substituted positions highlighted in light blue).

3.4.2 Computational prediction of Yfr1 targets

For the target prediction, a 400 nt subsequence including 250 nt upstream and 150 nt downstream of the start codon was extracted for all annotated genes of the *Prochlorococcus* MED4 genome (GenBank accession number BX548174 [155] using the updated annotation by Kettler et al. [92]). In total, we obtained 1964 sequences covering the full 5' UTR (if not longer than 250 nt) and the beginning of the CDS of each gene to search for interactions with Yfr1.

Putative interactions with Yfr1 were predicted with `IntaRNA` based on hybridisation energy and accessibility of the interaction sites. For the accessibility calculation, we assumed global folding of Yfr1. In contrast, mRNAs do not fold globally *in vivo* due to the helicase activity of the translating ribosome [180]. Hence, mRNA subsequences were folded locally in a 200 nt window with a maximal base pair distance of 100 nt. For each gene, the optimal interaction and up to five suboptimal interactions were computed.

In *Prochlorococcus* MED4, the ultraconserved motif 5'-`ACUCCUCACAC`-3' covers positions 17 to 27 of Yfr1 RNA (Figure 3.12A). This motif was predicted to be single-stranded in the consensus secondary structure of Yfr1 homologs from 31 cyanobacteria (Figure 3.11 and reference [196]). For the target search with `IntaRNA`, we required an interaction seed of eight paired bases and at most one unpaired base and constrained its position in the sRNA sequence to the aforementioned conserved Yfr1 motif. To investigate the influence of interaction seeds, another target prediction was conducted without requiring a seed region.

We also tested a modified energy score that weights the accessibility against the hybridisation energy with factor α:

$$E = H + \alpha \cdot ED_{\mathrm{mRNA}} + \alpha \cdot ED_{\mathrm{sRNA}},$$

where H denotes the hybridisation energy of the interaction, and ED_{mRNA} and ED_{sRNA} denote the energy required to make the interaction site accessible in the mRNA and the sRNA sequence, respectively. The original `IntaRNA` scoring does not weight the unfolding

Table 3.3. Highest scoring Yfr1 target candidates and their ranks under different IntaRNA parameter settings.

Target		Fixed seed			No seed		
	α	1	0.5	0	1	0.5	0
PMM1119 (*som*)		1	4	22	1	11	173
PMM0494 (*ppa*)		2	3	25	2	10	195
PMM1121 (*som*)		3	6	32	4	32	184
PMM1697		4	7	12	12	19	10
PMED4_09051		5	1	8	52	2	46
PMM0538		6	16	14	26	84	91
PMM0130		7	13	72	13	60	238
PMM1021		8	22	69	5	51	830
PMM1405		9	12	26	37	61	115
PMM0050 (*argJ*)		10	11	7	40	8	11

Only interactions at the RBS (−39 to +19 relative to the start codon, see Hüttenhofer and Noller [83]) were considered. All ranks are given according to IntaRNA energy score. α is a weighting factor for the accessibility in the extended hybridisation energy.

energy of the interaction sites, i.e. $\alpha = 1$.

In addition to the IntaRNA energy score, the location of the interaction in the mRNA is used as a further criterion to evaluate the quality of a predicted interaction. Many of the characterised *trans*-encoded sRNAs downregulate their targets by base pairing to the RBS. Therefore, the predicted target candidates were filtered for interactions that involve the mRNA region from −39 to +19 relative to the start codon, which is the maximal region covered by ribosomes [83].

Table 3.3 lists the ten highest scoring candidates of the Yfr1 target prediction. Out of these, we experimentally tested the six monocistronic target candidates with known transcriptional start sites and interaction sites predicted in the 5' UTR or at the start codon.

3.4.3 Yfr1 represses translation of two mRNAs

The predicted Yfr1 target candidates were experimentally validated using a two-plasmid reporter system based on green fluorescent protein (GFP) (described in detail by Urban and Vogel [188]). In brief, full-length 5' UTRs and the first coding residues of the targets of interest were fused to a *gfp* reporter gene. Then, the sRNA and the target-*gfp* fusion were co-expressed under control of constitutive promotors within the same *E. coli* cell. The level of post-transcriptional regulation was assessed by measuring the GFP fluorescence.

We tested potential interactions of Yfr1 sRNA with the 5' UTRs of the putative targets PMM0050 (*argJ*, bifunctional ornithine acetyltransferase/N-acetylglutamate synthase), PMM0494 (*ppa*, putative inorganic pyrophosphatase), PMM0538 (unknown function), PMM1119 (*som*, outer membrane protein), PMM1121 (*som*, outer membrane protein) or PMM1697 (type II alternative σ factor). Target-*gfp* fusions as well as control plasmids pXG-0 (negative control) and pXG-1 (positive control) were tested in the presence of a nonsense RNA and Yfr1 sRNA, respectively. Furthermore, two Yfr1 mutants were generated and tested (see Figure 3.12). In mutant Yfr1 M1, `CC` at positions 20 and 21 was substituted by `GG` leading to the formation of a stem–loop structure in the normally unpaired region. In mutant Yfr1 M2, `UCCU` at positions 19 to 22 was substituted by `AAAA` without changing the structure. The secondary structures of both mutants were predicted by `RNAfold` and verified by manual inspection of the ensemble-based base pairing probability matrix (`RNAfold` with option `-p`). All interaction studies were carried out in *E. coli* strain Top10. Single cell fluorescence was determined by flow cytometry. The mean fluorescence per plasmid combination was calculated from 10000 events (cells) of six individual clones.

The predicted interactions for targets with a GFP fluorescence signal above background (indicating measurable expression) are shown in Figure 3.13. Two of the six tested target candidates are translationally repressed by Yfr1, as shown by a reduced GFP fluorescence signal (Figure 3.14). The first clusters of the bar chart in Figure 3.14 constitute the negative controls (*E. coli* strain Top10 without plasmid or with plasmid pXG-0 devoid of *gfp*, respectively) and the positive control (*E. coli* strain Top10 with plasmid pXG-1 carrying *gfp*). The remaining clusters represent the 5' UTR-*gfp* fusions for the targets of interest. Each *gfp* fusion plasmid was tested in the presence of a second plasmid containing a nonsense RNA (white bars), Yfr1 sRNA (red bars) and the two mutated Yfr1 sRNAs M1 and M2 (light and dark blue bars) (Figure 3.14).

In the presence of the nonsense RNA, no regulation of the 5' UTR-*gfp* fusions by an interaction is expected (Figure 3.14, white bars), and the fluorescence measured here represents the 5' UTR-specific translation efficiency. The different GFP fluorescence intensities can be explained by differences in the affinities of the ribosomes for the translation initiation region. The strongest inhibition by Yfr1 was detected for the 5' UTRs of the two *som* genes PMM1119 and PMM1121 (3.0- and 2.7-fold reduced GFP signal, red bars in Figure 3.14). No change in GFP fluorescence was observed for PMM1697 and PMM0538 5' UTRs in the presence of Yfr1. For PMM0494 and PMM0050, no fluorescence above the background level (dashed line in Figure 3.14) could be detected for any tested plasmid combination.

Translation inhibition of the two *som*s was abolished by the introduction of a mutation in the conserved Yfr1 motif exchanging `CC` by `GG` (Yfr1 M1, light blue bars in Figure 3.14). These two substitutions involve the region predicted to base pair with the RBS of the two *som* mRNAs. Furthermore, mutation M1 led to a structural change by introducing a stem–loop in the single-stranded region of wild-type Yfr1 (Figure 3.12B). Thus, mutation M1 results in both a sequential and structural change at the interaction site. To

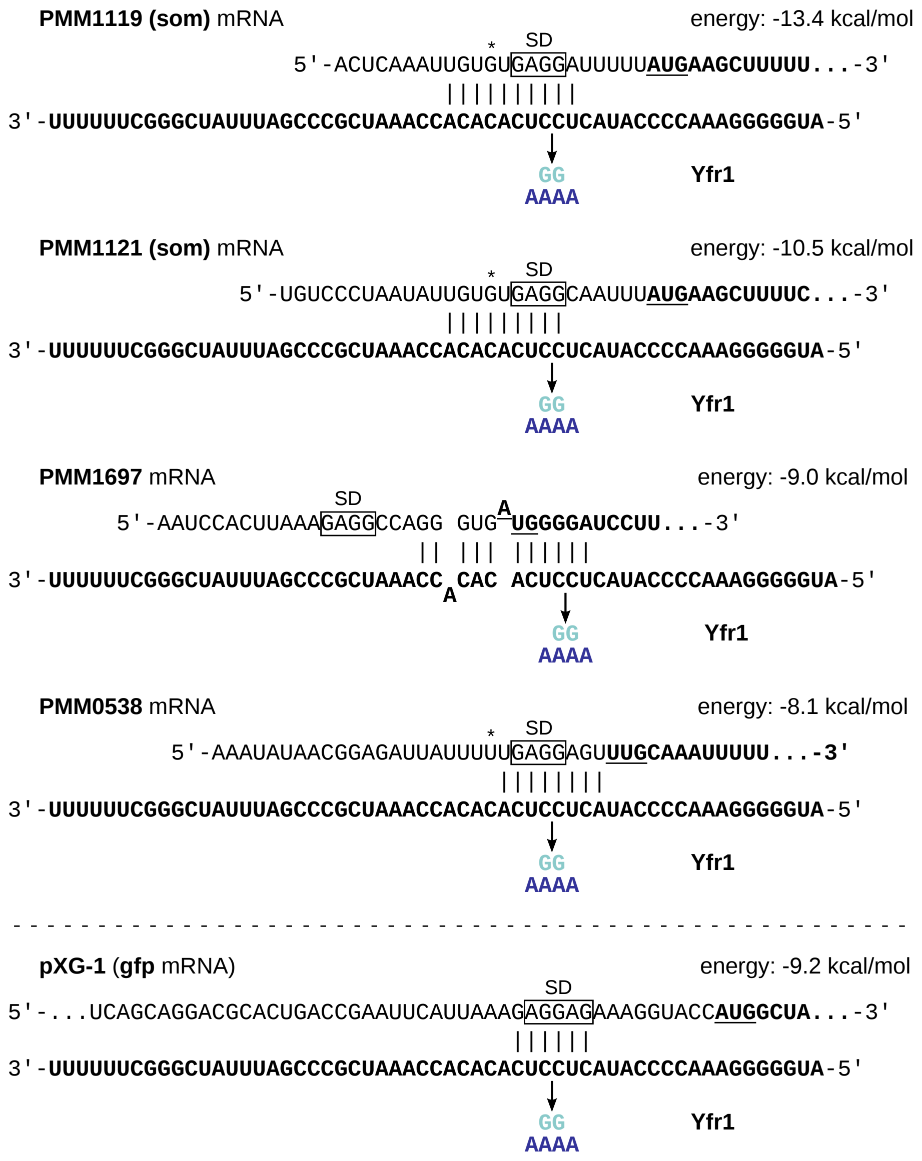

PMM1119 (som) mRNA
energy: -13.4 kcal/mol
*
SD
5'-ACUCAAAUUGUGUGAGGAUUUUUAUGAAGCUUUUU...-3'
3'-UUUUUUCGGGCUAUUUAGCCCGCUAAACCACACACUCCUCAUACCCCAAAGGGGGUA-5'
GG
AAAA
Yfr1
PMM1121 (som) mRNA
energy: -10.5 kcal/mol
*
SD
5'-UGUCCCUAAUAUUGUGUGAGGCAAUUUAUGAAGCUUUUC...-3'
3'-UUUUUUCGGGCUAUUUAGCCCGCUAAACCACACACUCCUCAUACCCCAAAGGGGGUA-5'
GG
AAAA
Yfr1
PMM1697 mRNA
energy: -9.0 kcal/mol
SD
A
5'-AAUCCACUUAAAGAGGCCAGG GUG UGGGGAUCCUU...-3'
3'-UUUUUUCGGGCUAUUUAGCCCGCUAAACC CAC ACUCCUCAUACCCCAAAGGGGGUA-5'
A
GG
AAAA
Yfr1
PMM0538 mRNA
energy: -8.1 kcal/mol
*
SD
5'-AAAUAUAACGGAGAUUAUUUUUGAGGAGUUUGCAAAUUUUU...-3'
3'-UUUUUUCGGGCUAUUUAGCCCGCUAAACCACACACUCCUCAUACCCCAAAGGGGGUA-5'
GG
AAAA
Yfr1
pXG-1 (gfp mRNA)
energy: -9.2 kcal/mol
SD
5'-...UCAGCAGGACGCACUGACCGAAUUCAUUAAAGAGGAGAAAGGUACCAUGGCUA...-3'
3'-UUUUUUCGGGCUAUUUAGCCCGCUAAACCACACACUCCUCAUACCCCAAAGGGGGUA-5'
GG
AAAA
Yfr1

Figure 3.13. Interactions between Yfr1 and target mRNA 5' UTRs predicted by `IntaRNA`. Additionally, a putative interaction between Yfr1 and the positive control pXG-1 is presented. The 5' ends of the mRNAs were experimentally mapped by deep sequencing (C. Steglich, unpublished data). Yfr1 RNA and coding sequences of the mRNAs are set in bold. Start codons are underlined. SD sequences are marked with a box. Asterisks denote start codons that are presumably misannotated in the *Prochlorococcus* MED4 genome sequence. The arrows indicate mutations M1 (light blue) and M2 (dark blue) introduced in Yfr1.

test whether the destruction of the antisense complementarity alone (without structural change) abolishes regulation by Yfr1, we constructed another Yfr1 mutant. In the Yfr1 mutant M2, nucleotides `UCCU` were substituted by `AAAA` without changing the secondary structure of wild-type Yfr1 (Figure 3.12A). Again, translation of PMM1119 and PMM1121 was restored (Figure 3.14, dark blue bars). These results indicate that Yfr1 inhibits translation of the two *som* mRNAs by direct base pairing at the RBS. Furthermore, the results

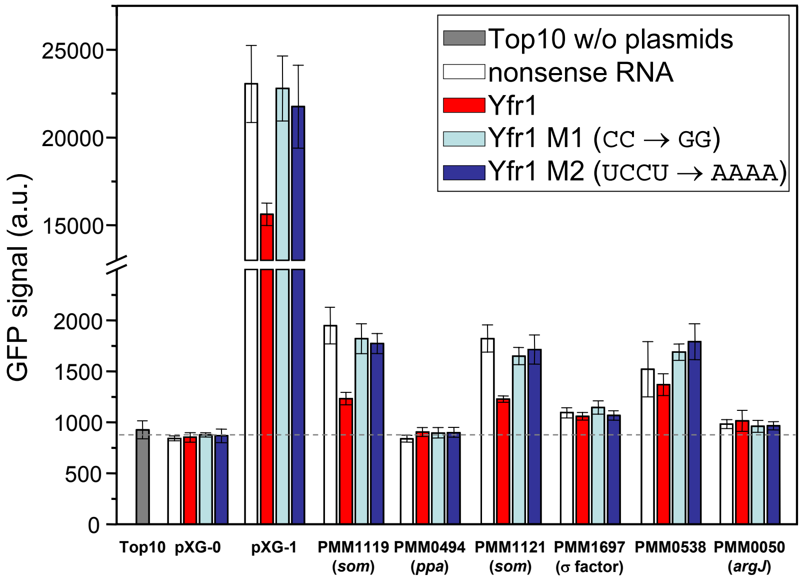


Figure 3.14. Experimental validation of Yfr1 target predictions. The relative decrease in GFP fluorescence as determined by flow cytometry indicates the strength of Yfr1-mediated regulation. The dashed line indicates background fluorescence (i.e. cellular autofluorescence), determined as the mean GFP signal of the negative controls. Fold changes of reduced GFP signal for PMM1119 (3.0-fold), PMM1121 (2.7-fold) and pXG1 (1.5-fold) were calculated after background subtraction from absolute fluorescence values (see Urban and Vogel [188]).

strongly indicate that both sequence and structure are important for Yfr1 regulation.

Surprisingly, we also observed a 1.5-fold reduction in GFP fluorescence for the positive control pXG-1 in the presence of Yfr1 and restored translation under the control of Yfr1 M1 and M2. However, the strong RBS in the 5' UTR of *gfp* in pXG-1 [188] shows a perfect complementarity to part of the conserved Yfr1 motif. Thus, Yfr1 can form a perfect 6 nt duplex with the 5' UTR (Figure 3.13), which can explain the observation of a reduction in translation in the presence of Yfr1 and restored translation in the presence of the two Yfr1 mutants.

3.4.4 Seed region and accessibility improve Yfr1 target prediction

Using the sRNA Yfr1 as example, we also investigated the importance of accessibility and of a seed region. Therefore, we computed lists of putative targets without enforcing a seed region and with enforcing a seed at the conserved Yfr1 motif. When requiring the fixed seed position, we obtained a short list of only 29 target candidates with the experimentally validated Yfr1 targets PMM1119 and PMM1121 ranked at positions 1 and 3, respectively (Table 3.3). Without the seed requirement, 1418 target candidates were obtained with the two true positives ranked at positions 1 and 4. Even without using a seed constraint, the interactions predicted for the true positives include the conserved single-stranded region of Yfr1. Thus, the combination of complementarity and accessibility alone resulted in interactions with an implicit seed.

In addition to the effect of a seed requirement, we studied the influence of accessibility on the Yfr1 target prediction. In the original IntaRNA scoring, hybridisation energy and interaction site accessibilities contribute equally to the energy score. Here, we tested a modified energy score, where the interaction site accessibility of both sequences was weighted by factor α with the values 0, 0.5 and 1. For both seed requirements studied, the true positives PMM1119 and PMM1121 were ranked best with the original scoring (Table 3.3). One interesting observation was that in the case of Yfr1, a full weighting of the interaction site accessibility, i.e. $\alpha = 1$, was required for a correct target site prediction. When both the seed region and accessibility were neglected, the two verified Yfr1 targets were not found within the top 150 predictions. When the seed position was fixed to the conserved region but accessibility was not included in the scoring, the validated targets were ranked at positions 22 and 32. However, in this case, predicted interactions involved almost the entire Yfr1 sequence. This observation is consistent with the findings of Tjaden et al. [182] and our findings presented in Section 2.3 that an energy model based solely on hybridisation energy tends to maximise the length of hybridisation, resulting in a small fraction of correctly predicted base pairs (i.e. low positive predictive value).

3.5 Discussion

3.5.1 Impact of accessibility and sequence conservation

We compiled a set of 71 sRNA–target pairs including 74 experimentally verified interaction sites to determine features that discriminate functional from non-functional interactions. We found that both sRNA and target interaction sites are highly accessible, and that the interaction sites in the sRNAs are additionally well conserved. The overall interaction site accessibility in the targets was lower than in the sRNAs and the difference to the non-functional interactions was also less pronounced (although still highly significant). There are two possible explanations for this observation: (i) Structural RNAs (e.g. sRNAs), but not mRNAs, generally have lower folding energies than random RNAs of the same dinucleotide frequency [32, 51, 213]. Consequently, the difference in accessibility between structured and unstructured regions might be higher for structural RNAs than for mRNAs. (ii) Although local folding of mRNAs is more accurate than global folding, a sliding window approach introduces a prediction bias by generating artificial sequence boundaries [100]. In contrast, the sRNAs are short and have well-defined sequence boundaries, making them suitable for global structure prediction. Therefore, the accessibilities for sRNAs might be more reliable than the accessibilities for mRNAs. We also observed that the overall accessibility is slightly higher in the target sequences than in the non-functional targets, so it might well be that unstructuredness is a selection criterion for sRNA target sites.

An exception would be sRNAs that act as direct translational activators by opening an inhibitory mRNA structure at the RBS [53].

One sRNA often targets multiple mRNAs via the same interaction site (e.g. CyaR, FnrS, GcvB, OmrA/B, RybB and RyhB). As a consequence, the target-binding region in the sRNA is likely to show high sequence conservation since base pairing with multiple targets is expected to constrain the evolution of the sRNA [66]. If there was only one target, sRNA and mRNA sequence would presumably have coevolved instead. Conversely, if only a single mRNA target is known for a sRNA with well-conserved interaction site, it seems very likely that there exist several other yet unknown targets. The idea that sRNAs typically target multiple mRNAs is additionally supported by the finding that the number of mRNAs bound by Hfq, which often facilitates RNA–RNA interaction formation, is considerably larger than the number of sRNAs associated with Hfq in *Salmonella* [169].

When comparing the nucleotide composition in the verified functional interactions to the non-functional interactions, we observed that true sRNA interaction sites and 3' flanks of 20 nt length are enriched for uridines. In accordance with sequence complementarity, we found an enrichment of adenosines in target interaction sites and 20 nt flanks on either side. Guanosine frequencies were reduced at target interaction sites and 3' flanks. The pairing of sRNAs with their target mRNAs is commonly facilitated by the RNA-binding protein Hfq, which has been recently reviewed by Vogel and Luisi [193]. Hfq has two binding surfaces, which preferentially bind single-stranded `U`-rich sequences and `ARN(N)` motifs, respectively. These sequence motifs match our observations, which suggests that the majority of the sRNA–target pairs analysed here can be bound simultaneously by Hfq. In addition, it was previously reported that sRNA target sites have a propensity for a flanking 3' adenosine [135], which accounts for about two third of the difference in 3' flank adenosine frequency between functional and non-functional sites.

The evaluation of the interaction site features accessibility and sequence conservation in genome-wide sRNA target predictions with `IntaRNA` includes a comparison to `TargetRNA`, which is a widely used tool and, thus, was included for sake of completeness. When comparing the overall prediction performance of `IntaRNA` and `TargetRNA`, the former ranked the true targets on average better than the latter. However, it is expected that many of the predicted “false positives” are actually true targets because our dataset is not an exhaustive set of interaction pairs. For example, Sharma et al. [165] recently identified 13 additional targets of the GcvB sRNA by *gfp* reporter gene fusions, but without a mapping of the exact interaction sites; thus they were not considered in our analysis. Consequently, both `IntaRNA` and `TargetRNA` are likely to perform better in predicting novel targets than in our experiments.

Our observation that sRNA interaction sites show characteristically high accessibility and sequence conservation is in line with a recent study by Peer and Margalit [137]. In their study, the authors also suggested for target predictions to narrow down the search space to interactions in conserved and accessible sRNA regions. Here, we required that the interaction seed only is located at an unstructured conserved sRNA region, which successfully increased the sensitivity of genome-wide target predictions with `IntaRNA`. We found that a similar improvement in sensitivity can be achieved 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. The overall quality of predictions does not further increase, but also does not decrease, when combining both restrictions. Restricting the interaction seeds to highly accessible regions, but not additionally to unstructured and conserved sRNA regions, has the advantage to require neither the availability of homologous sRNA sequences nor 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 structure prediction that is required to compute the accessibility of the interacting RNAs is already part of interaction prediction methods as `IntaRNA` and, thus, does not create any computational overhead.

The comparison between verified interactions and non-functional interactions provided no evidence that interaction sites in target mRNAs are generally conserved (in contrast to interaction sites in sRNAs). Consistently, a survey of the two sRNAs RybB and RyhB and their respective targets revealed that, although the sRNA interaction site is highly conserved, the actual seed base pair complementarity is maintained on average in only 60 percent of the species. For miRNAs, the functional analogues of sRNAs in eukaryotes, it was also found that a substantial fraction of experimentally verified target sites is non-conserved [47], albeit target site conservation being frequently used to increase the specificity of miRNA target prediction [52, 99]. Furthermore, our results did not show an enrichment for compensatory or consistent mutations in the interactions. Taken together, these observations suggest that the base pairing between sRNAs and their targets is not generally conserved across related species. The high evolutionary conservation of the sRNA interaction site and the missing consistent mutations in the target result in an overall paucity of sequence covariation between sRNA and target, which is consistent with our findings in Section 4.2. Consequently, our results further suggest that comparative methods will benefit from a covariance scoring only for a subclass of interactions.

The question remains why sRNA interaction sites exhibit a very high sequence conservation when neither interaction sites in the targets are sequentially conserved nor interactions are structurally conserved. A possible explanation is that, for particular sRNAs, regulation of the target could be conserved, but not the interaction site location. Instead, the interaction site has been shifted to another location in the target. As a result, this target site mobility could lead to an interaction site that is conserved in sequence, but found in a different sequence context. Another explanation for missing target site conservation results from the observation that many sRNAs regulate multiple targets. A specific gene that is a target in a particular organism does not have to be a target in each of the other organisms in which the sRNA is conserved. Often, multiple targets are regulated via the same binding site in the sRNA. But high conservation of this sRNA site does neither imply full conservation of all target genes nor conservation of the base pairing even if the genes are conserved [70, 134, 153]. Instead, regulation of individual targets might have been acquired or lost very recently in evolution. However, for a particular sRNA, 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 [66]. For example, the gene *ompA* is very broadly conserved and its base pairing potential with the 5' end of RybB sRNA is preserved in all 21 analysed species; thus, one could speculate that only *ompA* might have originally constrained the evolution of the RybB interaction site in these species. However, the RybB 5' end is recently involved in the regulation of several other broadly conserved targets (see Table 3.2). These other targets beside *ompA* now pose additional evolutionary constraints to the RybB interaction site, such that the 5' sequence will be preserved even if *ompA* is lost as a target.

3.5.2 Identification of two novel Yfr1 target genes

We showed that the cyanobacterial sRNA Yfr1 modulates the translation of two high-scoring predicted targets by an antisense interaction. This result proves that the combination of computational and experimental methods is a promising approach for the identification of sRNA targets in organisms where genetic manipulation constitutes a great challenge. Furthermore, our results showed for the first time that *trans*-encoded targets of a cyanobacterial sRNA are regulated in a mode of action similar to other studied bacteria. Both Yfr1 target mRNAs code for outer membrane proteins [71]. This class of proteins constitutes a major functional class that is regulated by bacterial sRNAs in *E. coli* and *Salmonella* [205]. The result was surprising as, until now, no highly abundant sRNAs have been shown to act via base pair interaction. However, both mRNA targets identified herein are also highly abundant (among the 10 most expressed mRNAs and with long half-lives of about 30 minutes [175]), which may require a high copy number of Yfr1 for efficient regulation. Furthermore, an mRNA with a long half-life can be regulated more efficiently by translational control than by transcriptional control.

Additionally, we assessed in this case study the influence of seed regions and interaction site accessibility on the prediction quality of Yfr1 targets. Analogously to many enterobacterial sRNAs, e.g. GcvB and RybB, Yfr1 contains a conserved single-stranded region, which seems to constitute a perfect interaction seed. When requiring this region as seed for the target prediction, the number of putative Yfr1 targets was remarkably smaller than without a seed requirement (29 versus 1418 candidates). The two true positives were, however, under the highest ranking candidates in both settings. This is due to the fact that, if the scoring includes accessibility, the interaction is implicitly formed in the seed region because of its single-strandedness in Yfr1. When neglecting both accessibility and a seed region, the true Yfr1 targets could not be found amongst the top 150 predictions.

Chapter 4

Comparative prediction of joint secondary structures

Many ncRNA candidates in bacteria and eukaryotes have emerged from alignment-based genomic screens that accounted for conserved RNA secondary structure. The availability of these structural alignments and the fact that many ncRNAs form RNA–RNA interactions motivated us to develop a comparative approach for the prediction of RNA–RNA interactions. In this chapter, we present `PETcofold`, the first approach for the prediction of conserved interactions and secondary structures of two multiple alignments of RNA sequences that takes compensatory exchanges in intra- and intermolecular base pairs into account. We showed in controlled tests on simulated data that covariance information improves the performance of RNA–RNA interaction prediction. Furthermore, we showed that our comparative method `PETcofold` outperforms single sequence-based methods in the prediction of joint secondary structures of sRNA–mRNA complexes.

4.1 Candidate ncRNAs are frequently available as multiple alignments

A substantial number of putative ncRNAs has emerged from genomic *in silico* screens for RNA structure taking compensatory base pair changes into account [e.g. 8, 154, 185, 197, 200, 207, 208]. Deep sequencing approaches and high-density tiling arrays are another growing source of novel ncRNAs [e.g. 69, 85, 164, 183]. One step towards assigning functions to these putative ncRNAs is to consider RNA–RNA interactions, as it has already been highlighted in previous chapters. Almost all interaction prediction methods that have been reviewed in Section 2.1 evaluate only interactions between a pair of single sequences. `RNAplex` and `RNArip` are an exception, as they were recently extended to work also on multiple sequence alignments [104, 179]. Many ncRNA candidates detected by genome-wide *in silico* screens are already available as multiple alignments for further analysis.

The existence of these alignments with *de novo* predicted RNA structures is therefore a key motivation for implementing a method that can predict RNA–RNA interactions between multiple alignments of RNA sequences, as the identification of potential interaction partners will help to elucidate the function of these candidate ncRNAs. For an approach that makes use of multiple alignments, our assumption is that a non-negligible amount of the existing RNA–RNA interaction contains compensatory changes across the binding sites. Whereas this is a motivating aspect, the general variation in the sequences with even a completely conserved interaction site might also yield improvements over single sequence-based prediction methods.

The literature contains only limited examples of conserved RNA–RNA interactions with likely compensating base pair changes. An example is the MicA–*ompA* interaction, where base pairing is preserved by compensatory changes in several enterobacterial species [187]. The authors of this study also developed an interaction prediction method that accounts for phylogenetic conservation, but the approach has not been published at the time of writing. In Section 3.2, we report some additional examples of interactions with consistent or compensatory base pair changes for the sRNAs RybB and RyhB. Compensatory (and not merely consistent) exchanges were, however, observed in only one of the examples. One explanation for the limited amount of data containing covariance information might well be that most existing data have been found using sequence similarity-based methods such as BLAST [2] to find homologs of the interacting RNAs. Such an approach is expected to lead to a collection of RNAs that are highly conserved in the primary sequence rather than structure. Although in these cases it is likely that the interaction pattern is conserved as well, only a small number of compensatory base pair changes is expected. Therefore, we also include simulated data based on substitution statistics of interacting base pairs to evaluate the performance of the method presented below. Simulated data has also been used in several previous studies to make controlled tests or to supplement existing data [e.g. 95].

In the following, we present PETcofold, a method for searching for RNA–RNA interactions between two multiple RNA sequence (or sequence-structure) alignments. PETcofold computes a semi-optimal combination of intra- and intermolecular base pairs to predict a joint secondary structure of two alignments, each representing its own evolutionary and structurally conserved RNA. Our method makes use of the idea of a linker from RNAcofold [15] by concatenating both RNA sequences, but employs this idea in the context of PETfold [160] along with a strategy for hierarchical folding [e.g. 56, 84]. PETfold is based on Pfold [94], which provides reliabilities for evolutionarily conserved base pairs, and unifies them with folding energies in one model. Hierarchical folding allows for the prediction of pseudoknots between intra- and intermolecular base pairs but is still fast enough for genome-scale applications in contrast to general pseudoknot search algorithms.

4.2 PETcofold: prediction of conserved joint structures

4.2.1 Algorithm and implementation

The method PETcofold, which is introduced in the following, uses a hierarchical folding approach to predict conserved RNA–RNA interactions between two multiple alignments of RNA sequences. The multi-step approach is motivated by the observation that interaction formation is often initiated at well-accessible intramolecular structures such as hairpin loops [23]. Furthermore, several existing methods for RNA–RNA interaction prediction are based on this observation and assume that the interaction sites are made accessible to allow for hybridisation of the two RNAs (e.g. our method IntaRNA as presented in Section 2.2 or RNAup [127]). The input of PETcofold consists of two RNA alignments $\mathcal{A}_1$ and $\mathcal{A}_2$ in which the first alignment represents a ncRNA and the second alignment represents an mRNA (or another ncRNA). In the first folding step, reliable base pairs in the two single RNAs are identified using the scoring of the PETfold approach [160]. In the second folding step, reliable base pairs in the concatenated sequences are predicted using a constrained version of the PETfold scoring scheme. In both steps, we use a combined scoring that evaluates consensus base pairs in the alignment based on thermodynamic stability and evolutionary conservation. The two-step hierarchical folding approach allows for the prediction of joint secondary structures that contain pseudoknots like kissing hairpins. The workflow of the PETcofold pipeline is shown in Figure 4.1.

For a given sequence s or alignment $\mathcal{A}$, let (i, j) denote a Watson–Crick or G-U wobble base pair between sequence positions i and j with $1 \leq i < j \leq |s|$ or between alignment columns i and j with $1 \leq i < j \leq |\mathcal{A}|$, respectively. A secondary structure σ is defined as a set of non-crossing base pairs. The set of single-stranded positions in the structure σ of a sequence s is defined as $\text{ss}(\sigma) = \{i \mid 1 \leq i \leq |s| \wedge \forall j = 1, \ldots, |s| : (i, j) \notin \sigma \wedge (j, i) \notin \sigma\}$. The Pfold model [93] allows to compute the probability $\Pr[\sigma \mid \mathcal{A}, T, M]$ of a consensus structure σ given an alignment $\mathcal{A}$, a phylogenetic tree T relating the sequences of the alignment and a general background model M for RNA secondary structures. The model M is based on a simple stochastic context-free grammar (SCFG) with the production rules

$$S \rightarrow LS \mid L \qquad F \rightarrow dFd \mid LS \qquad L \rightarrow s \mid dFd,$$

where s symbolises single-stranded bases and dFd symbolises a pairing between bases in a stem. As the phylogenetic tree T is calculated from the alignment $\mathcal{A}$ and the model M is constant, we abbreviate $\Pr[\sigma \mid \mathcal{A}, T, M]$ by $\Pr^{\text{evo}}[\sigma \mid \mathcal{A}]$ in the following. The evolutionary reliability $\mathcal{R}^{\text{evo}}_{\text{bp}}(i, j, \mathcal{A})$ of a base pair (i, j) is defined by

$$\mathcal{R}^{\text{evo}}_{\text{bp}}(i, j, \mathcal{A}) = \sum_{\substack{\sigma \text{ with} \\ (i,j) \in \sigma}} \Pr^{\text{evo}}[\sigma \mid \mathcal{A}].$$

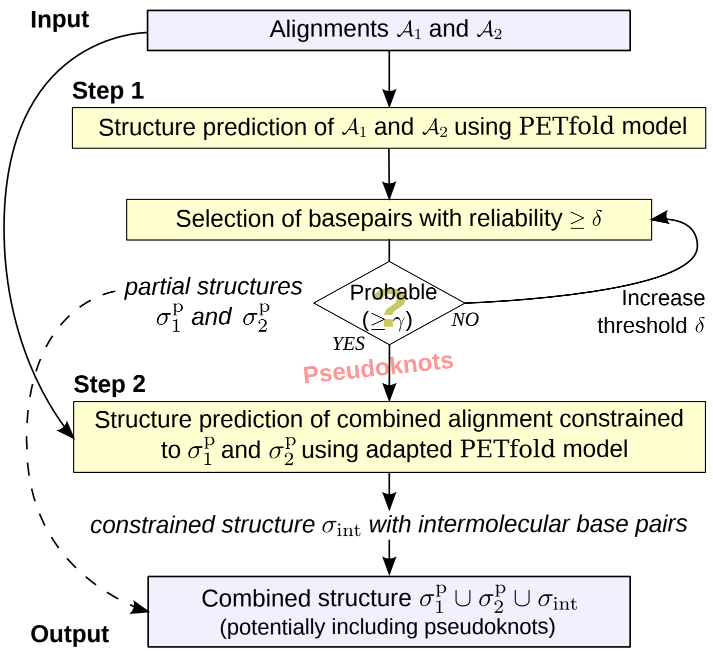

Input
Alignments $\mathcal{A}_1$ and $\mathcal{A}_2$
Step 1
Structure prediction of $\mathcal{A}_1$ and $\mathcal{A}_2$ using PETfold model
Selection of basepairs with reliability $\geq \delta$
partial structures σ_1^{p} and σ_2^{p}
Probable ($\geq\gamma$)
NO
YES
Increase threshold δ
Pseudoknots
Step 2
Structure prediction of combined alignment constrained to σ_1^{p} and σ_2^{p} using adapted PETfold model
constrained structure σ_{int} with intermolecular base pairs
Combined structure $\sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}} \cup \sigma_{\mathrm{int}}$ (potentially including pseudoknots)
Output

Figure 4.1. The `PETcofold` pipeline consists of two steps: (1) intramolecular folding by `PETfold` and selection of a set of highly reliable base pairs that decrease the probability of the ensemble in some pre-defined range only; (2) intermolecular folding by an adapted `PETfold` using the constraints from folding step 1. Finally, the partial structures and the constrained intermolecular structure are combined to the joint RNA secondary structure including pseudoknots.

The thermodynamic probability $\mathrm{Pr}^{\mathrm{th}}[\sigma \mid s]$ of the secondary structure σ for a sequence s can be computed from the partition function as defined by McCaskill [121]. The thermodynamic probability $\mathrm{Pr}^{\mathrm{th}}_{\mathrm{bp}}(k,l)$ that a base pair (k,l) is formed by the sequence s is given by

$$\mathrm{Pr}^{\mathrm{th}}_{\mathrm{bp}}(k,l) = \sum_{\substack{\sigma \text{ with} \\ (k,l)\in\sigma}} \mathrm{Pr}^{\mathrm{th}}[\sigma \mid s].$$

These base pair probabilities can by computed efficiently by, e.g. `RNAfold` [75] using option `-p`.

To combine the evolutionary reliabilities and the thermodynamic probabilities, we define $\sigma(s^u, \mathcal{A})$ as the structure of the u-th sequence s^u in the alignment $\mathcal{A}$, which is obtained by mapping the consensus structure σ of $\mathcal{A}$ to sequence s^u. The combined reliability $\mathcal{R}_{\mathrm{bp}}(i,j)$ of a base pair (i,j) is then given by

$$\begin{aligned}\mathcal{R}_{\mathrm{bp}}(i,j) &= \sum_{\substack{\sigma \text{ with} \\ (i,j)\in\sigma}} \mathrm{Pr}^{\mathrm{evo}}[\sigma \mid \mathcal{A}] + \frac{\beta}{n} \times \sum_{s\in\mathcal{A}} \sum_{\substack{\sigma \text{ with} \\ (i,j)\in\sigma}} \mathrm{Pr}^{\mathrm{th}}[\sigma(s,\mathcal{A}) \mid s] \\ &= \quad \mathcal{R}^{\mathrm{evo}}_{\mathrm{bp}}(i,j,\mathcal{A}) \quad + \frac{\beta}{n} \times \sum_{s\in\mathcal{A}} \mathrm{Pr}^{\mathrm{th}}_{\mathrm{bp}}(i,j,s), \end{aligned} \qquad (4.1)$$

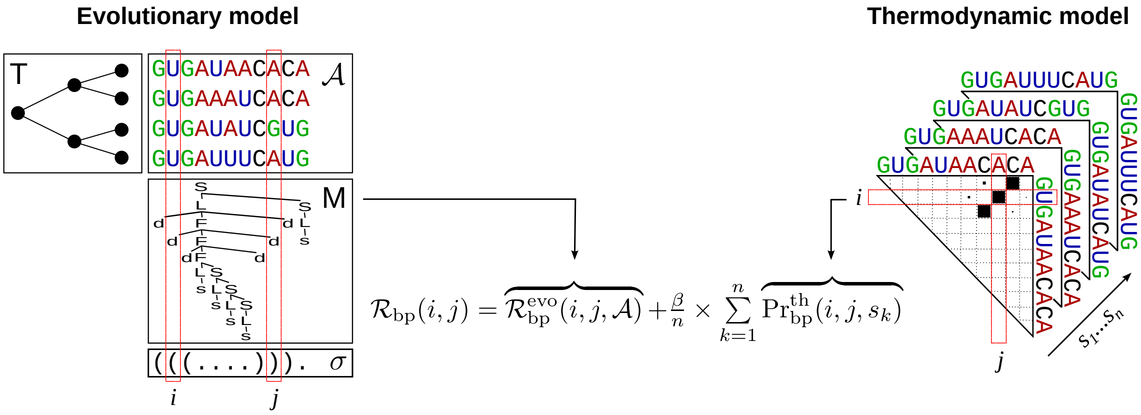

Evolutionary model
Thermodynamic model
T
$\mathcal{A}$
GUGAUAACACA
GUGAAAUCACA
GUGAUAUCGUG
GUGAUUUCAUG
M
(((. . . .))) . σ
i
j
$\mathcal{R}_{\text{bp}}(i,j) = \overbrace{\mathcal{R}^{\text{evo}}_{\text{bp}}(i,j,\mathcal{A})} + \frac{\beta}{n} \times \sum_{k=1}^{n} \overbrace{\text{Pr}^{\text{th}}_{\text{bp}}(i,j,s_k)}$
$s_1 \ldots s_n$

Figure 4.2. Reliability scoring of base pairs by PETfold. Evolutionary reliabilities computed by Pfold and thermodynamic probabilities computed by RNAfold are integrated into one model. Evolutionary base pair reliabilities are calculated from the probabilities of all consensus structures that include a specific base pair, given a sequence alignment $\mathcal{A}$, a phylogenetic tree T relating the sequences and an RNA secondary structure background model M. Thermodynamic base pair probabilities are calculated, for each sequence of the alignment, from the ensemble of structures that can be formed by the sequence. The illustration is based on a figure provided by Stefan E. Seemann, who is the developer of the PETfold program.

where β is a weighting factor for the thermodynamic score, n is the number of sequences in the alignment and $\text{Pr}^{\text{th}}_{\text{bp}}(i,j,s)$ is the probability of the base pair in sequence s that corresponds to alignment columns i and j. Figure 4.2 illustrates the idea of the combined reliability scoring.

In **step 1 of the PETcofold pipeline**, we search in the two input RNA alignments $\mathcal{A}_1$ and $\mathcal{A}_2$ for highly reliable base pairs which are interpreted as being not accessible for the RNA–RNA interaction. The base pair reliabilities are calculated separately for $\mathcal{A}_1$ and $\mathcal{A}_2$ according to Equation (4.1). The set of all base pairs with high base pair reliability $\mathcal{R}_{\text{bp}}$ in the folding of an individual alignment, i.e. $\mathcal{R}_{\text{bp}}$ is greater than or equal to a threshold δ, forms a partial structure σ^{p}. The partial structures for $\mathcal{A}_1$ and $\mathcal{A}_2$ are denoted by σ^{p}_1 and σ^{p}_2, respectively. The threshold δ should be set to at least 0.5 to avoid crossing structures. The ensemble of all specific structures σ' that are compatible with the partial structure σ^{p} is defined by $\mathcal{E}(\sigma^{\text{p}}) = \{\sigma' \mid \sigma' \supseteq \sigma^{\text{p}}\}$. To ensure that the highly reliable base pairs of a partial structure σ^{p} are also part of the final (consensus) structure, the resulting ensemble of structures $\mathcal{E}(\sigma^{\text{p}})$ should have a sufficiently high probability $\text{Pr}[\mathcal{E}(\sigma^{\text{p}})]$ in either the thermodynamic or the evolutionary model. A high value of $\text{Pr}[\mathcal{E}(\sigma^{\text{p}})]$ is guaranteed by the introduction of a second threshold γ. The threshold δ for highly reliable base pairs is now increased until the probability of $\mathcal{E}(\sigma^{\text{p}})$ exceeds γ in either the evolutionary or the thermodynamic model, i.e.

$$\text{Pr}^{\text{evo}}[\mathcal{E}(\sigma^{\text{p}}) \mid \mathcal{A}] \geq \gamma \quad \text{or} \quad \frac{1}{n} \sum_{s \in \mathcal{A}} \text{Pr}^{\text{th}}[\mathcal{E}(\sigma^{\text{p}}(s,\mathcal{A})) \mid s] \geq \gamma,$$

where n is the number of sequences in the alignment $\mathcal{A}$, and $\Pr^{\text{evo}}[\mathcal{E}(\sigma^{\text{p}}) \mid \mathcal{A}]$ ($= \Pr[\mathcal{E}(\sigma^{\text{p}}) \mid \mathcal{A}, T, M]$) is the probability of the partial structure σ^{p} given the alignment $\mathcal{A}$, the background model M and the tree T. $\Pr^{\text{evo}}[\mathcal{E}(\sigma^{\text{p}}) \mid \mathcal{A}]$ can be calculated by `Pfold` [94]. $\Pr^{\text{th}}[\mathcal{E}(\sigma^{\text{p}}(s,\mathcal{A})) \mid s]$ is the probability of the partial structure σ^{p} for sequence s in the thermodynamic model. This probability can be calculated from the partition function using constrained folding:

$$\Pr^{\text{th}}[\mathcal{E}(\sigma^{\text{p}}(s,\mathcal{A})) \mid s] = \frac{Z^{\mathcal{E}(\sigma^{\text{p}}(s,\mathcal{A}))}}{Z} = \frac{e^{-\frac{E^{\text{ens}}(\mathcal{E}(\sigma^{\text{p}}(s,\mathcal{A})))}{RT}}}{e^{-\frac{E^{\text{ens}}(\mathcal{S})}{RT}}} = e^{\frac{E^{\text{ens}}(\mathcal{S}) - E^{\text{ens}}(\mathcal{E}(\sigma^{\text{p}}(s,\mathcal{A})))}{RT}},$$

where R is the gas constant, T is the temperature, Z and $Z^{\mathcal{E}(\sigma^{\text{p}}(s,\mathcal{A}))}$ are the partition functions over all structures and over all structures compatible with $\sigma^{\text{p}}(s,\mathcal{A})$, respectively. $E^{\text{ens}}(\mathcal{S})$ and $E^{\text{ens}}(\mathcal{E}(\sigma^{\text{p}}(s,\mathcal{A})))$ are the free energies of the ensemble $\mathcal{S}$ of all structures and of the ensemble $\mathcal{E}(\sigma^{\text{p}}(s,\mathcal{A}))$ of structures with the base pairs of $\sigma^{\text{p}}(s,\mathcal{A})$ as constraints, respectively. The partition function and free energy of constrained structures can be calculated by `RNAfold` with options `-p -C`.

The selection of highly reliable intramolecular base pairs, which are constrained as single-stranded in the second folding step, can result in incomplete intramolecular stems. Therefore, constrained stems in the partial structure σ^{p} can be optionally extended by inner and outer base pairs. The base pairs are added to the constrained stems as long as the average reliability of the extended stem is greater than or equal to δ and the partial structure probability $\Pr[\mathcal{E}(\sigma^{\text{p}})]$ exceeds γ. This feature is enabled by the `PETcofold` option `-extstem`.

In **step 2 of the `PETcofold` pipeline**, we concatenate the sequences of the input alignments $\mathcal{A}_1$ and $\mathcal{A}_2$ with a linker symbol '&' to search for conserved interactions and structures of these sequences. On the concatenated alignment, we apply an adapted `PETfold` model that can handle fixed partial structures σ_1^{p} and σ_2^{p} from the first step by constrained expected accuracy scoring, which is an extension of `PETfold`'s maximum expected accuracy scoring for constrained folding. We search for a joint structure σ of the combined alignment that extends both σ_1^{p} and σ_2^{p}, i.e. $\sigma \supseteq \sigma_1^{\text{p}} \cup \sigma_2^{\text{p}}$. `PETfold` itself cannot handle pseudoknots and the linker in step 2 forbids pseudoknots in the concatenated sequences, i.e. the resulting structure has to be nested. The hierarchical folding strategy of `PETcofold`, however, allows for pseudoknots between intramolecular base pairs from step 1 and intermolecular (as well as intramolecular) base pairs from step 2. This is achieved by restricting the positions of the concatenated alignments that are covered by base pairs from σ_1^{p} and σ_2^{p} to be single-stranded. Under these constraints, thermodynamic probabilities $\Pr_{\text{raw}}^{2,\text{th}}$ are calculated with `RNAcofold` (using options `-p -C`) and evolutionary reliabilities $\mathcal{R}_{\text{raw}}^{2,\text{evo}}$ are calculated with a modified version of `Pfold` that incorporates constraints. This constrained folding results in raw probabilities and reliabilities, which are then weighted by the product of the partial structure probabilities $\Pr[\mathcal{E}(\sigma^{\mathrm{p}})]$ from step 1. However, to avoid underestimating the probabilities of step 2, we here replace the product with the geometric mean of partial structure probabilities. The probabilities and reliabilities of all intramolecular base pairs from σ_1^{p} and σ_2^{p} are set to the partial structure probabilities $\Pr[\mathcal{E}(\sigma_1^{\mathrm{p}})]$ and $\Pr[\mathcal{E}(\sigma_2^{\mathrm{p}})]$, respectively.

To summarise, in the thermodynamic model, the probability of a base pair (i, j) in the sequence $s_1\&s_2 \in \mathcal{A}_1\&\mathcal{A}_2$, given the partial consensus structures σ_1^{p} and σ_2^{p} for the alignment $\mathcal{A}_1$ and $\mathcal{A}_2$ as calculated in step 1, is defined by

$$\Pr_{\mathrm{bp}}^{2,\mathrm{th}}(i,j,s_1\&s_2) = \begin{cases} \Pr^{\mathrm{th}}[\mathcal{E}_1(\sigma_1^{\mathrm{p}}(s_1,\mathcal{A}_1) \mid s_1] & \text{if } (i,j) \in \sigma_1^{\mathrm{p}}, \\ \Pr^{\mathrm{th}}[\mathcal{E}_2(\sigma_2^{\mathrm{p}}(s_2,\mathcal{A}_2) \mid s_2] & \text{if } (i,j) \in \sigma_2^{\mathrm{p}}, \\ \Pr_{\mathrm{bp,raw}}^{2,\mathrm{th}}(i,j) \times \prod_{l=1,2} \Pr^{\mathrm{th}}[\mathcal{E}_l(\sigma_l^{\mathrm{p}}(s_l,\mathcal{A}_l)) \mid s_l] & \text{else.} \end{cases} \tag{4.2}$$

The structural ensembles $\mathcal{E}_1(\sigma^{\mathrm{p}})$ and $\mathcal{E}_2(\sigma^{\mathrm{p}})$ denote the set of structures that extend the partial structure σ^{p} using only base pairs of the first and second sequence, respectively, i.e. $\mathcal{E}_l(\sigma^{\mathrm{p}}) = \{\sigma' \supseteq \sigma^{\mathrm{p}} \mid \forall (i,j) \in \sigma' \backslash \sigma^{\mathrm{p}} : 1 \leq i < j \leq |s_l|\}$. The probabilities of single-stranded positions are defined by

$$\Pr_{\mathrm{ss}}^{2,\mathrm{th}}(i,s_1\&s_2) = \begin{cases} 0 & \text{if } \exists j \text{ with } (i,j) \in \sigma_1^{\mathrm{p}} \text{ or } (j,i) \in \sigma_1^{\mathrm{p}}, \\ 0 & \text{if } \exists j \text{ with } (i,j) \in \sigma_2^{\mathrm{p}} \text{ or } (j,i) \in \sigma_2^{\mathrm{p}}, \\ \Pr_{\mathrm{ss,raw}}^{2,\mathrm{th}}(i) \times \prod_{l=1,2} \Pr^{\mathrm{th}}[\mathcal{E}_l(\sigma_l^{\mathrm{p}}(s_l,\mathcal{A}_l)) \mid s_l] & \text{else.} \end{cases} \tag{4.3}$$

In the evolutionary model, reliabilities of base pairs and single-stranded positions are, analogously to Equations (4.2) and (4.3), defined by

$$\mathcal{R}_{\mathrm{bp}}^{2,\mathrm{evo}}(i,j,\mathcal{A}_1\&\mathcal{A}_2) = \begin{cases} \Pr^{\mathrm{evo}}[\mathcal{E}_1(\sigma_1^{\mathrm{p}}) \mid \mathcal{A}_1] & \text{if } (i,j) \in \sigma_1^{\mathrm{p}}, \\ \Pr^{\mathrm{evo}}[\mathcal{E}_2(\sigma_2^{\mathrm{p}}) \mid \mathcal{A}_2] & \text{if } (i,j) \in \sigma_2^{\mathrm{p}}, \\ \mathcal{R}_{\mathrm{bp,raw}}^{2,\mathrm{evo}}(i,j) \times \prod_{l=1,2} \Pr^{\mathrm{evo}}[\mathcal{E}_l(\sigma_l^{\mathrm{p}}) \mid \mathcal{A}_l] & \text{else,} \end{cases} \tag{4.4}$$

and

$$\mathcal{R}_{\mathrm{ss}}^{2,\mathrm{evo}}(i,\mathcal{A}_1\&\mathcal{A}_2) = \begin{cases} 0 & \text{if } \exists j \text{ with } (i,j) \in \sigma_1^{\mathrm{p}} \text{ or } (j,i) \in \sigma_1^{\mathrm{p}}, \\ 0 & \text{if } \exists j \text{ with } (i,j) \in \sigma_2^{\mathrm{p}} \text{ or } (j,i) \in \sigma_2^{\mathrm{p}}, \\ \mathcal{R}_{\mathrm{ss,raw}}^{2,\mathrm{evo}}(i) \times \prod_{l=1,2} \Pr^{\mathrm{evo}}[\mathcal{E}_l(\sigma_l^{\mathrm{p}}) \mid \mathcal{A}_l] & \text{else,} \end{cases} \tag{4.5}$$

respectively.

The constrained expected accuracy of a joint structure σ is then calculated from the reliabilities of all base pairs and single-stranded positions by

$$\begin{aligned}\mathrm{EA}(\sigma) &= \sum_{(i,j)\in\sigma}\left(\mathcal{R}^{2,\mathrm{evo}}_{\mathrm{bp}}(i,j,\mathcal{A}_1\&\mathcal{A}_2) + \frac{\beta}{n}\sum_{\substack{s_1\&s_2\in\\ \mathcal{A}_1\&\mathcal{A}_2}}\mathrm{Pr}^{2,\mathrm{th}}_{\mathrm{bp}}(i,j,s_1\&s_2)\right)\\ &\quad+\alpha\sum_{i\in\mathrm{ss}(\sigma)}\left(\mathcal{R}^{2,\mathrm{evo}}_{\mathrm{ss}}(i,\mathcal{A}_1\&\mathcal{A}_2) + \frac{\beta}{n}\sum_{\substack{s_1\&s_2\in\\ \mathcal{A}_1\&\mathcal{A}_2}}\mathrm{Pr}^{2,\mathrm{th}}_{\mathrm{ss}}(i,s_1\&s_2)\right) \qquad (4.6)\\ &= \sum_{(i,j)\in\sigma}\mathcal{R}^{2}_{\mathrm{bp}}(i,j) + \alpha\sum_{i\in\mathrm{ss}(\sigma)}\mathcal{R}^{2}_{\mathrm{ss}}(i),\end{aligned}$$

where α is a weighting factor for the single-stranded reliabilities, and $\mathrm{Pr}^{2,\mathrm{th}}_{\mathrm{bp}}(i,j,s_1\&s_2)$, $\mathrm{Pr}^{2,\mathrm{th}}_{\mathrm{ss}}(i,s_1\&s_2)$, $\mathcal{R}^{2,\mathrm{evo}}_{\mathrm{bp}}(i,j,\mathcal{A}_1\&\mathcal{A}_2)$ and $\mathcal{R}^{2,\mathrm{evo}}_{\mathrm{ss}}(i,\mathcal{A}_1\&\mathcal{A}_2)$ are defined as in Equations (4.2) to (4.5). The weighting factors α and β are, in the following, set to be equal to the default values of `PETfold`, i.e. $\alpha = 0.2$ and $\beta = 1$.

The constrained expected accuracy structure σ_{int} that maximises the score given in Equation (4.6) is calculated by a Nussinov-style algorithm [131], which evaluates base pairs and single-stranded bases with their respective reliabilities. The resulting structure σ_{int} might contain both intramolecular as well as intermolecular base pairs. The final consensus structure including the interaction is then given by

$$\sigma = \sigma_1^{\mathrm{p}} \cup \sigma_2^{\mathrm{p}} \cup \sigma_{\mathrm{int}}.$$

The algorithm presented above has a time complexity of $O(n\cdot k\cdot l^3)$, where n is the number of sequences in the alignments, k is the number of iterations in the adjustment of δ to ensure probable partial structures and l is the sum of the sequence lengths of both alignments.

4.2.2 Estimation of optimal parameters on sRNA–mRNA interactions

The prediction performance of `PETcofold` under various parameter settings was evaluated on a dataset of bacterial interactions. This dataset contained 13 different sRNAs and 32 sRNA–mRNA interactions with experimental support from the organisms *E. coli*, *Salmonella* and *S. aureus* (Table A.5).

For each sRNA, the RNA family sequence alignment was downloaded from the Rfam database 9.1 [55]. Alignments of the target mRNAs and their orthologous genes were created as follows. Genome sequences for all 69 species with available complete genome according to the Rfam annotation were downloaded from the EMBL (European Molecular Biology Laboratory) Nucleotide Sequence Database [33]. `OrthoMCL` [105] was used with default parameters to identify groups of orthologous genes separately for the *S. aureus* species (according to the Rfam family of RNAIII) and for all remaining species (according to the Rfam families of the *E. coli* and *Salmonella* sRNAs). For each target gene, a 250 nt subsequence was extracted (150 nt upstream and 100 nt downstream of the annotated translation start site). The sequence length of 250 nt was chosen because all interactions occurred from positions −132 to +56 relative to the start codon; flanking regions were included for the prediction of mRNA structures and for compensation of misannotated translation start sites. Sets of orthologous genes were compiled according to the `OrthoMCL` prediction excluding sequences of species that are not contained in the Rfam family of the sRNA interacting with the target. The sets of orthologous target sequences were locally aligned with method E-INS-i from the `MAFFT` package [89], which uses a generalised affine gap cost model.

The resulting dataset was processed by (i) homology reduction and (ii) removal of sequences that were very distant to the reference organism, i.e. the organism in which the interaction was detected. The latter aims to remove false positive predicted orthologs and was achieved by excluding all target sequences with less than 60 percent pairwise sequence identity ($PI_{\mathrm{ref}}^{\mathrm{int}}$) at the interaction site and within 10 nt of the flanking sequences compared with the reference sequence. This threshold was chosen as, in real applications, sequence-based alignments are often used as input, which perform satisfactorily for comparative RNA structure prediction when pairwise sequence identities are above 60 percent [54, 199]. The homology reduction was performed with the objective of avoiding a bias by overweighting redundant sequence information. To this end, target sequences were clustered with the `BLASTClust` tool [2] using a word size of 8 and an identity threshold of 100 percent over an area covering 90 percent of each sequence. The sequence with the lowest $PI_{\mathrm{ref}}^{\mathrm{int}}$ was taken from each cluster. For the final sRNA sequence alignments, all sequences of species without available complete genome or without occurrence in the target sequence set were removed from the Rfam sequence alignments, followed by removal of gap-only alignment columns.

The covariance of a dataset was evaluated by the number of consistent and compensatory base pair exchanges (CBP) within the interactions. For a specific interaction, the CBP is computed as the average number of consistent and compensatory base pairs in all interacting alignment columns, with a consistent or compensating base pair being distinct from the base pairing in the reference sequence. Thus, the maximal value of the CBP for two interacting alignment columns is 5. The CBP of a dataset was then computed as the mean CBP of all interactions within the dataset. For the sRNA–mRNA dataset introduced above, the CBP is only 0.114, i.e. little compensatory interaction base pair exchanges can be observed. Another covariance measure is the probability $\Pr[\sigma_{i,j} \mid \mathcal{A}, T, M]$ of a base pair (i, j) calculated by `Pfold`. It is more accurate because it takes into account, by the tree T, the evolutionary distance of the sequences in the alignment $\mathcal{A}$. The model M describes the substitution rates of base pairs and unpaired bases and the probabilities of secondary structure production rules [93]; however, the bias introduced by M can be ignored. To measure the covariance of a full interaction, we use the normalised `Pfold` reliability $\mathcal{R}(\text{int})$, which is the mean of all base pair probabilities in the interaction.

All `PETcofold` predictions were evaluated by calculating their correlations to the structures from literature ignoring non-canonical base pairs. We used the Matthews correlation coefficient [120] defined as

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP}+\text{FP})(\text{TP}+\text{FN})(\text{TN}+\text{FP})(\text{TN}+\text{FN})}},$$

where TP is the number of mutual base pairs in the two assignments (true positives), TN is the number of mutual pairs of bases that are not base pairing (true negatives), FP is the number of predicted base pairs that are not in the annotated assignment (false positives) and FN is the number of base pairs in the annotated assignment that are not predicted to pair (false negatives). The MCC should be maximised to achieve the best possible trade-off between sensitivity and PPV. The MCC of the prediction to the annotation is computed for each single sequence in the alignment. The mean of these single MCCs gives the mean MCC of a prediction. Mean and median MCC of a whole dataset are calculated from the mean MCCs of all interactions contained in the dataset.

The influence of `PETcofold`'s parameter setting on its prediction quality was analysed on the dataset of 32 bacterial sRNA–mRNA interactions as introduced above. `PETcofold` is mainly controlled by two parameters: the parameter δ sets the maximal intramolecular base pair reliability of bases to be free for intermolecular folding, and the parameter γ sets the minimal partial structure probability. For each of the two parameters, 11 values ranging from 0.0 to 1.0 in steps of 0.1 were tested yielding 121 parameter combinations. Furthermore, we tested the influence of the option `-noLP`, which disallows pairs that can only occur isolated in the thermodynamic part, i.e. it is used as an option for `RNA(co)fold`. Columns with more than 50 percent gaps were removed. Figure 4.3 shows a 3D plot of the mean interaction MCCs for all parameter combinations tested (without option `-noLP`). `PETcofold` yielded the best performance for $\delta = 0.9$ and γ ranging from 0.0 to 0.5 with a mean interaction MCC of 0.494 (median MCC: 0.546). When using the option `-noLP`, the best mean interaction MCC was 0.491 (median MCC: 0.526) for $\delta = \gamma = 0.9$. Consequently, the influence of this option on the mean MCC is only marginal.

The numerical experiments indicate that the use of intramolecular constraints improves the prediction of interaction sites. Nevertheless, the probability threshold δ for base pairs in the partial structure σ^{p} has to be fairly high to achieve the best MCCs. In other words, the structural mass of intramolecular partial structures has to be high to support the constrained expected accuracy scoring. When setting δ to 1, then no base will be constrained by base pairing for the intermolecular folding step. Consequently, no loop–loop interactions between the two single structures are allowed. We found that a δ of 0.9

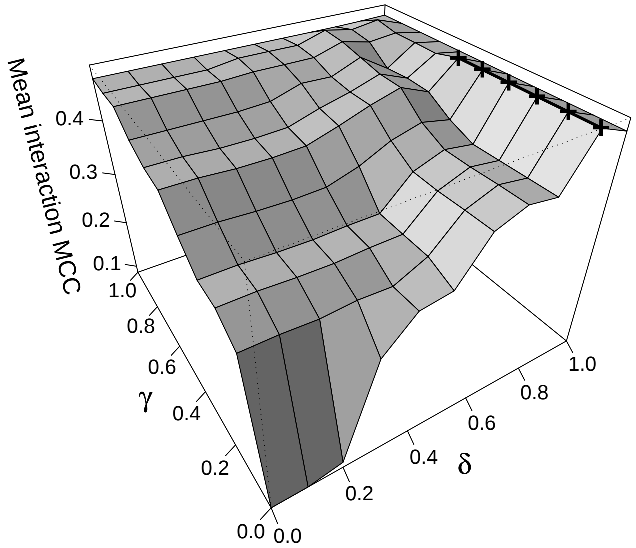


Figure 4.3. Performance of `PETcofold` while varying the parameters δ (maximal intramolecular base pair reliability) and γ (minimal partial structure probability). The 3D plot shows the mean MCC of 32 interactions. Predictions were carried out without the option `-noLP`. The maximal MCC is marked with '+'.

yields the best MCC. This setting forbids interactions in highly structured regions and, thus, accounts for the importance of interaction site accessibility.

The parameter γ adapts the probability of partial structures to cover a high mass of the entire ensemble of structures. This strategy avoids that the intramolecular folding step introduces constraints that are incompatible with reliable alternative structures supporting the interaction site. We achieved the best performance for $0 \leq \gamma \leq 0.5$.

4.2.3 Performance on simulated interactions with increased covariance

The dataset described in the previous subsection is based on sRNA–mRNA interactions that were experimentally validated in a reference organism. Although it is common practice to use homologous sequences from Rfam families and to computationally predict orthologous genes, this approach could be limited in two respects. From a biological point of view, it is assumed that the homologous sRNAs regulate the same targets by the same mechanism in all organisms that are included in the sequence sets. Although this assumption is true for some interactions, it does not apply to all of them (compare Section 3.2). Furthermore, the target orthologs might contain false positive predictions with different physiological functions and regulatory mechanisms. From a technical point of view, weak covariance at the interaction sites limits the full potential of `PETcofold`. For example, the most conserved region of the sRNA SgrS is involved in base pairing its target [78].

We therefore created a simulated dataset in which the degree of compensating base changes was controlled. We used `SISSI` [60] to simulate sequence data with site-specific in-

Table 4.1. Prediction performance of `PETcofold` on simulated sequence data with increased covariance.

Scaling factor	Mean branch length	$\mathcal{R}$(int)	CBP	Mean MCC
1	0.03	0.486	0.832	0.505
5	0.15	0.665	1.818	0.677
10	0.30	0.699	2.182	0.717
25	0.75	0.732	2.550	0.743
50	1.50	0.775	2.755	0.773
75	2.25	0.791	2.858	0.783
100	3.00	0.801	2.908	0.790
150	4.50	0.822	2.988	0.802
200	6.00	0.839	3.040	0.821

The scaling factor multiplies each branch length of the phylogenetic tree. Mean branch length denotes the mean of the mean branch lengths for all 32 phylogenetic trees. $\mathcal{R}$(int) denotes the mean base pair probability at the interaction sites (calculated by `Pfold`). CBP denotes the average number of consistent and compensatory interaction base pair exchanges in the simulated sequence data. The MCC evaluates only the interaction. `PETcofold` was called with parameters $\delta = 0.9$, $\gamma = 0.1$ and option `-noLP` to forbid lonely base pairs.

teractions annotated by the sRNA–mRNA interactions of our dataset along phylogenetic trees. To be biologically relevant, we estimated each phylogenetic tree from the corresponding alignment using a maximum likelihood method with an independent model. For tree reconstructions, `IQPNNI` [192] was used with the maximum likelihood approach of Felsenstein [49]; otherwise, the default settings were used. To specify the rate matrices, we simply counted the frequencies of the nucleotides {`A`, `C`, `G`, `U`} for sites evolving independently and doublet frequencies for the distant RNA interaction pairs. For each of the 32 alignments, we performed 20 simulations with the same length and in the context of the annotations of the sRNA–mRNA interactions using a Markov model of nucleotide sequence evolution [60] with rate matrix types of the model of Felsenstein [49]. We initially started simulations along the estimated phylogenetic trees. We then multiplied each branch length by a scaling factor to increase the covariance. The simulation runs were repeated under the same parameters with eight different scaling factors as shown in Table 4.1.

The average number of consistent and compensatory interaction base pair exchanges (CBP) computed from the simulated data with scaling factor 1 was 0.8 (Table 4.1), which was higher than the CBP of 0.1 in the real data. This was due to the fact that the simulations covered only a subset of the evolutionary constraints on the sequences. For example, only the interactions without the thermodynamic contribution of stacked base pairs were used as structural constraints. Other evolutionary constraints on the sequences were neglected. However, taking into account all aspects in the simulations with a corresponding maximum likelihood framework is beyond the scope of this study. At this point, we focused on evaluating `PETcofold`'s performance on datasets with increased covariance.

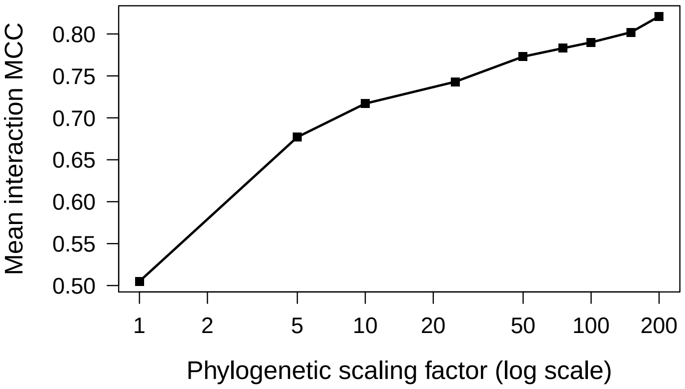


Figure 4.4. Prediction performance of `PETcofold` on phylogenetically simulated sequence data for 32 interactions. The covariance was increased by multiplying each branch length with a phylogenetic scaling factor. The prediction accuracy of `PETcofold` in terms of MCC correlates with the phylogenetic scaling factor, and, thus, with the covariance at the interaction sites. `PETcofold` was called with parameters $\delta = 0.9$, $\gamma = 0.1$ and option `-noLP`.

We applied `PETcofold` to these datasets and computed the mean interaction MCC of all 20 simulation runs for 9 different phylogenetic scaling factors. Figure 4.4 shows the mean interaction MCC plotted over the phylogenetic scaling factor for `PETcofold` predictions with $\delta = 0.9$, $\gamma = 0.1$ and the option `-noLP`. The mean MCC of the predicted interactions was 0.505 for scaling factor 1. The prediction accuracy of `PETcofold` in terms of MCC increased with increasing scaling factors. A mean MCC of 0.821 was achieved for scaling factor 200. Table 4.1 shows for all scaling factors the mean interaction MCC and the covariance in the data as evaluated by $\mathcal{R}$(int) and CBP. It can be clearly seen that the performance of `PETcofold` correlated with the covariance at the interaction sites.

4.3 Predicting conserved joint structures of sRNA–mRNA complexes

In the following, we explore the performance of `PETcofold` in the prediction of joint secondary structures of two interacting RNAs, i.e. prediction of both secondary structure and RNA–RNA interaction. A comparison to existing joint secondary structure prediction methods evaluates to what extend our novel comparative approach improves the prediction quality over single sequence-based methods. We analysed four sRNA–target mRNA complexes, each with a previously described interaction model based on structural mapping: MicA–*ompA* [187], OxyS–*fhlA* [6], RyhB–*fur* [191] and RyhB–*sodB* [57]. The interactions of RyhB–*sodB* and OxyS–*fhlA* involve one and two loop–loop interactions, respectively. Consequently, their joint structures contain pseudoknots between the single RNA structures and the RNA–RNA interaction.

The structure prediction is based on the dataset introduced in Section 4.2. To make the predictions comparable with the annotated structures, an mRNA subsequence as given in the proposed interaction complex model was used instead of the 250 nt subsequence. Regarding the sRNAs, the Rfam entry of RyhB (Rfam accession number RF00057) missed the first 29 nt of the *E. coli* RyhB sequence. However, this subsequence is involved in forming the secondary structure. Thus, homologs of the RyhB sRNA were searched with the semi-global alignment tool `GotohScan` [72] in all organisms that are contained in the RyhB alignment (using an E-value cut-off of 1×10^{-3}). Homologs were found in all of these organisms except *Vibrio cholerae* O395. A multiple sequence-structure alignment of the identified homologous RyhB sequences was computed with `LocARNA` [210]. The mRNA and sRNA alignments of all four examples were hand curated by removal of redundant sequences and of sequences that are very distinct from the reference organism *E. coli*, in which the interaction models were experimentally determined.

To predict the joint structures, `PETcofold` was called with parameters $\delta = 0.9$, $\gamma = 0.1$, without allowing lonely base pairs in the thermodynamic model (`-noLP`) and optionally with extension of constrained stems by inner and outer bp (`-extstem`). `PETcofold` was compared with the following single sequence-based methods: the sparsified version of `inteRNA` [157], `PairFold` [5], `RactIP` [88], all with default parameters, and `RNAcofold` [15] using parameters `-d2 -noLP`. All methods apart from `RactIP` predict minimum free energy joint secondary structures. `inteRNA` is based on the model by Chitsaz et al. [30], whereas `PairFold` and `RNAcofold` are based on folding of the concatenated input sequences using the model of Zuker and Stiegler [217]. `RactIP` uses integer linear programming to maximise an objective function that is based on internal and external base pair probabilities. All computations were performed on a machine with AMD Opteron 2356 processor (2.3 GHz) and 16 GB RAM.

`PETcofold` is a comparative approach that detects conserved joint secondary structures. Hence, to compare with the other, single sequence-based, approaches, we also determined conserved consensus structures from the results of `inteRNA`, `PairFold`, `RactIP` and `RNAcofold`. The consensus structure is defined by all base pairs that are conserved in a given percentage of single structures (here: 80 and 100 percent) of each of the sequences in the multiple alignment. All consensus structures were evaluated by calculating their correlations to the joint secondary structures from literature. A commonly used measure is the Matthews correlation coefficient (MCC) [120], see also Section 4.2. For RNA secondary structures, the geometric mean of sensitivity (SENS) and positive predictive value (PPV),

$$\sqrt{\mathrm{SENS} \times \mathrm{PPV}} = \sqrt{\mathrm{TP}/(\mathrm{TP}+\mathrm{FN}) \times \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})},$$

is a good approximation of the MCC [64] and is used here. Table 4.2 lists the approximated MCCs of all four sRNA–mRNA examples for our method `PETcofold` and the compared methods.

Table 4.2. Prediction performance and runtime of `PETcofold` and other joint structure prediction methods on four sRNA–mRNA examples.

sRNA–mRNA	MCC									
	`PETcofold`		`inteRNA`		`PairFold`		`RactIP`		`RNAcofold`	
		`-extstem`	80%	100%	80%	100%	80%	100%	80%	100%
MicA–*ompA*	0.87	0.83	0.49	0.51	0.86	0.74	0.57	0.57	0.80	0.67
OxyS–*fhlA*	0.80	0.82	0.64	0.64	0.61	0.61	0.48	0.48	0.61	0.61
RyhB–*fur*	0.13	0.13	0.12	0.00	0.21	0.21	0.19	0.00	0.21	0.21
RyhB–*sodB*	0.67	0.71	0.70	0.68	0.65	0.51	0.65	0.59	0.65	0.63
Average	0.62	0.62	0.49	0.46	0.58	0.52	0.47	0.41	0.57	0.53

sRNA–mRNA	Run time [s]					
	`PETcofold`		`inteRNA`	`PairFold`	`RactIP`	`RNAcofold`
		`-extstem`				
MicA–*ompA*	28.7	28.4	69493.1	3.2	3.0	0.2
OxyS–*fhlA*	20.6	19.3	129636.7	1.9	2.0	0.2
RyhB–*fur*	26.4	25.3	65599.2	2.6	2.7	0.2
RyhB–*sodB*	15.4	15.2	23579.3	1.7	2.0	0.1
Average	22.8	22.1	72077.1	2.4	2.5	0.2

The MCC evaluates the joint structure, i.e. both the interaction between the two RNAs and the secondary structure of each single RNA. `PETcofold` was called with parameters $\delta = 0.9$, $\gamma = 0.1$, option `-noLP` to forbid lonely base pairs and optionally with option `-extstem` for extension of constrained stems. `inteRNA`, `PairFold` and `RactIP` were called with default parameters. `RNAcofold` was called with options `-d2 -noLP`. The columns 80% and 100% give the result for the consensus structure with base pairs that occur in 80% and 100%, respectively, of the single structures. The runtime of all single sequence-based approaches is the sum for all input sequences.

For the interactions of OxyS–*fhlA* and RyhB–*sodB*, the MCC of the `PETcofold` predictions was slightly higher when using the option `-extstem`. The opposite applies to MicA–*ompA*. On the non-curated alignments, the MCC was up to 0.2 lower (data not shown), which emphasises the importance of high-quality input alignments for our method. However, the option `-extstem` seemed to improve the prediction when using low-quality input alignments by extension of imperfectly (structurally) conserved stems. When comparing `PETcofold` with the other methods, our method overall showed a better performance in predicting the joint structures. The prediction quality of the RyhB–*fur* joint structure is very low for all compared approaches (maximal MCC of 0.21), which implies that the published interaction model is not predictable with the evaluated computational approaches. Thus, the prediction for this example is not reliable. When excluding RyhB–*fur*, our approach gives consistently more reliable predictions than the single sequence-based methods (see Table 4.2). For comparison to the complex joint secondary structure prediction methods with high resource consumption (both high time and memory complexity), we were only able to compare to `inteRNA`, as this is currently the only method with a sufficiently low resource consumption. However, the evaluation shows that this (minimum free energy structure) prediction approach is not very reliable without homology information. Thus, one has to resort to the more complex partition function approaches, which, however, have drastically larger time and memory requirements. For example, it was not possible to obtain predictions from `RNArip` [80] with reasonable resources. In comparison, `PETcofold` gave reliable predictions within seconds, which makes it also fast enough for genome-scale applications.

The `PETcofold` parameter γ sets the threshold for the minimal partial structure probability (see Section 4.2). For the prediction of joint secondary structures, we used a rather restrictive γ of 0.1 to allow many constraints for intramolecular base pairs. We also tested other values for γ in the range of 0.3 to 0.9, but achieved no performance improvement by these values (see Table 4.3).

Figure 4.5 shows the annotation and `PETcofold` prediction for all four examples together with the sequence alignments used as input. In the case of MicA–*ompA*, `PETcofold` correctly predicted all interaction base pairs from the annotation. The interaction site is highly conserved in both RNAs and contains only one compensatory mutation for the pairing between alignment positions 16 and 200. The intramolecular structures contain compensatory mutations, for instance, between alignment positions 60 and 65. The OxyS–*fhlA* interaction involves two binding sites, which each reside in stem–loops such that OxyS and *fhlA* form a double kissing hairpin interaction. The *fhlA* interaction sites are located at the ribosome binding site and within the coding region, respectively. `PETcofold` was able to predict the intramolecular stem loops in both RNAs, but only the interaction that involves the 5' stem–loop in OxyS and the CDS site in *fhlA*. For RyhB–*sodB*, all interaction base pairs were predicted, even though the predicted interaction was longer

Table 4.3. Performance of `PETcofold` on prediction of four sRNA–mRNA joint secondary structures.

sRNA–mRNA	MCC of joint secondary structure									
	`PETcofold`					`PETcofold -extstem`				
γ	0.1	0.3	0.5	0.7	0.9	0.1	0.3	0.5	0.7	0.9
MicA–*ompA*	0.87	0.87	0.87	0.87	0.87	0.83	0.83	0.83	0.83	0.83
OxyS–*fhlA*	0.80	0.80	0.80	0.71	0.71	0.82	0.82	0.71	0.71	0.71
RyhB–*fur*	0.13	0.13	0.13	0.13	0.12	0.13	0.13	0.13	0.13	0.12
RyhB–*sodB*	0.67	0.67	0.67	0.67	0.67	0.71	0.71	0.71	0.70	0.70
Average	0.62	0.62	0.62	0.60	0.59	0.62	0.62	0.60	0.59	0.59

`PETcofold` was run with a value of 0.9 for the intramolecular base pair reliability threshold (δ), varying values for the minimal partial structure probability (γ), without allowing lonely base pairs (`-noLP`) and optionally with extension of constrained stems (`-extstem`).

than experimentally observed. The intramolecular stem–loops that reside the interaction sites were not predicted. The low prediction quality for RyhB–*fur* might be explained by two observations. First, the interaction site given in the literature overlaps with the highly probable terminator stem of RyhB, which was constrained for intermolecular folding. Second, a region of the interaction site that was subject to experimental validation (alignment positions 144 to 149) contains base pairs that are not supported by up to 3 out of 6 sequences of the alignment.

4.4 Discussion

We presented `PETcofold`, the first comparative method for the prediction of a joint secondary structure of two interacting RNAs. The method identifies evolutionary conserved structures and can exploit the information from compensating base changes in the intramolecular structures of the two RNAs and the interactions between them. Furthermore, `PETcofold` allows for the prediction of pseudoknots between intra- and intermolecular base pairs.

We have shown in controlled runs on simulated data that the covariance information improves the prediction ability for RNA–RNA interactions. We have also shown for four bacterial sRNAs that the addition of evolutionary information from multiple sequence alignments improves the performance in comparison to methods based on single sequences. This implies that single sequence-based methods could perform better if the comparative information is also taken into account. As for other RNA structure prediction methods making use of sequence-based alignments, it is well-documented that these work best with an average pairwise sequence identity above 60 percent [54, 199]. In the process of cleaning up the (already sparse) bacterial data, we took this into account by removing

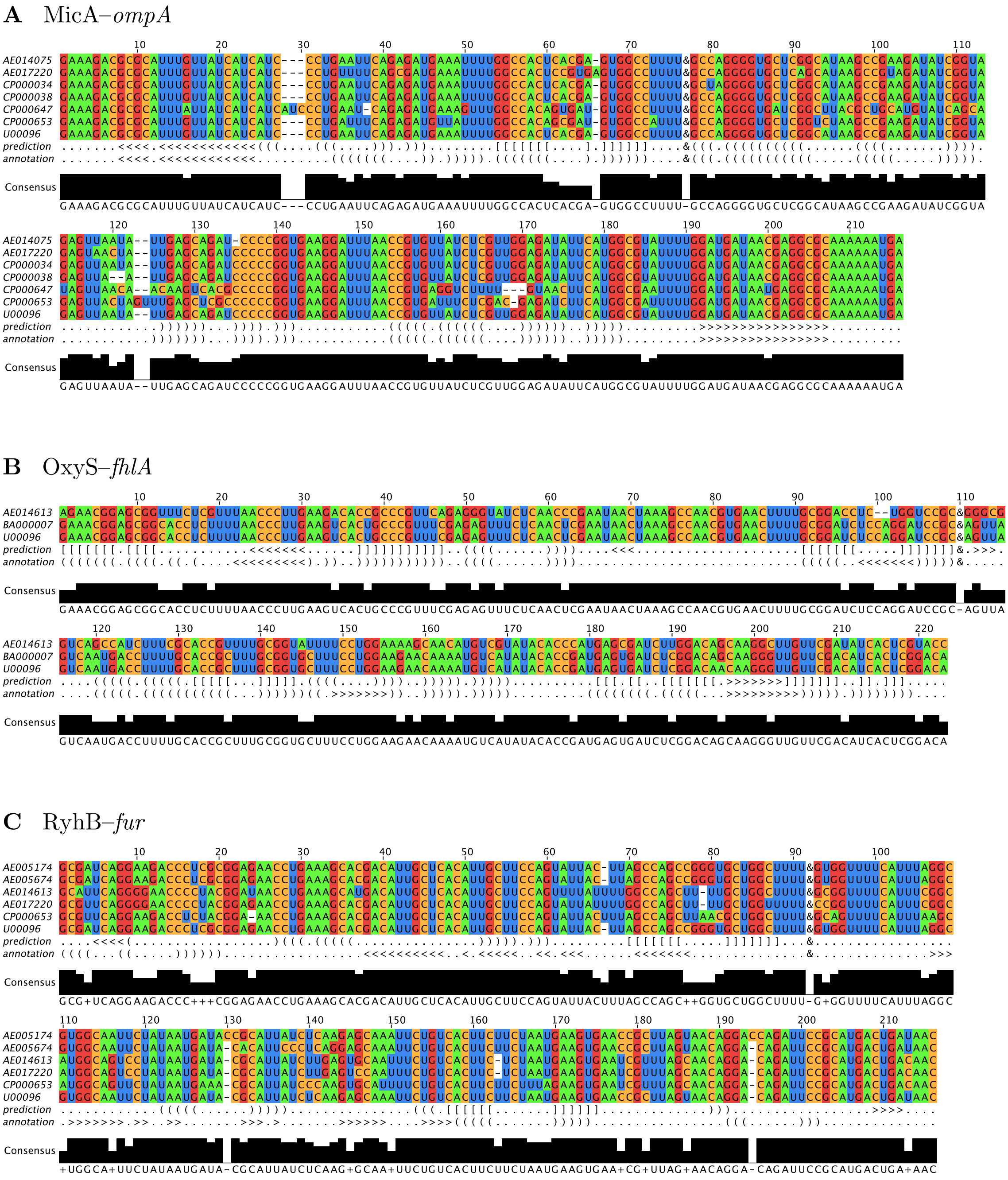


Figure 4.5. Joint secondary structures of the sRNA–mRNA interaction complexes of **(A)** MicA–*ompA*, **(B)** OxyS–*fhlA*, **(C)** RyhB–*fur* and **(D)** RyhB–*sodB*. (Figure is continued on the next page.)

(Continued)

D RyhB–*sodB*

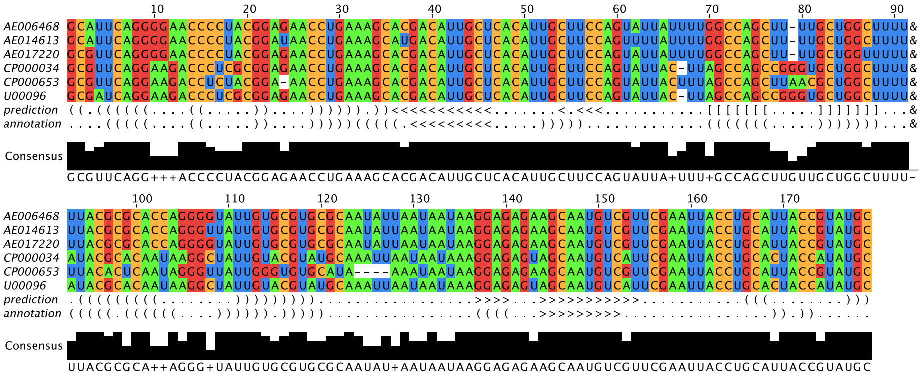

AE006468
AE014613
AE017220
CP000034
CP000653
U00096
prediction
annotation
Consensus
GCGUUCAGG+++ACCCCUACGGAGAACCUGAAAGCACGACAUUGCUCACAUUGCUUCCAGUAUUA+UUU+GCCAGCUUGUUGCUGGCUUUU-
UUACGCGCA++AGGG+UAUUGUGCGUGCGCAAUAU+AAUAAUAAGGAGAGAAGCAAUGUCGUUCGAAUUACCUGCAUUACCGUAUGC

Figure 4.5. Joint secondary structures of the sRNA–mRNA interaction complexes of **(A)** MicA–*ompA*, **(B)** OxyS–*fhlA*, **(C)** RyhB–*fur* and **(D)** RyhB–*sodB*. Each sequence alignment shows the two input alignments concatenated by the linker symbol '&', the structure predicted by `PETcofold` (with parameters $\delta = 0.9$, $\gamma = 0.1$ and options `-noLP` and, except for MicA–*ompA*, `-extstem`) and the interaction model from literature [6, 57, 187, 191]. Sequences are labelled with the genome accession numbers of the corresponding organisms. Angle brackets indicate intermolecular base pairs. Round and square brackets indicate intramolecular base pairs. Square brackets indicate positions that were constrained in step 1 of the `PETcofold` pipeline. For OxyS–*fhlA*, only columns with $< 50\%$ gaps are shown. The alignments were visualised with `Jalview` [204].

more distant sequences to keep the balance of accurate assignment and covariance patterns of base pairs. Future improvements can include explicit handling of redundant sequence information. Currently, redundant sequences contribute equally to the scoring to directly reflect the input data instead of overweighting outliers in a dense evolutionary tree. Thus, datasets with highly redundant sequences should be cleaned prior to usage of the program.

Many genomic screens for ncRNAs predict secondary structures with compensatory base pair changes on RNA alignments. These intramolecular structures can be used directly as input of step 1 of the `PETcofold` pipeline to identify highly reliable substructures, which are then constrained for step 2. For example, `PETcofold` could be applied to predict RNA–RNA interactions on the `CMfinder`-generated structure-based alignments of *de novo* predicted candidate structured RNAs in bacteria or in the ENCODE regions of vertebrates [185, 207].

The sparse amount of known examples of sequences with RNA–RNA interactions and the paucity of covariations were the reasons why we introduced simulated data. Some of the known examples of RNA–RNA interactions, e.g. from bacterial sRNA–mRNA and eukaryotic miRNA–mRNA interactions, tend to be rather conserved. Even in cases with little or no compensating base pairs (in interaction sites or in the intramolecular structure), any given variation will collectively contribute to the calculation of the reliabilities and thereby to the overall structure of the interaction complex. Hence, in the complete lack of covariation, PETcofold reduces to an improved energy folding approach, which also has impact on matching up the interacting base pairs. The sRNA–mRNA interactions studied here exhibit only a small number of compensatory base pair changes, which might be because sRNA sequences often show poor conservation across distant bacterial species. Thus, the regulators might be recently acquired and rapidly evolving [205]. Nevertheless, many of the homologous ncRNAs and mRNAs have been found based on sequence similarity, which leads to highly identical sequences and thereby also to highly conserved interactions. PETcofold considers sequence conservation, but its full power is only revealed when the input data contains structural covariance. In the future, we expect that deep sequencing approaches will give rise to many more characterised transcriptomes, which will increase the amount of data available for analysis including RNAs containing compensating base pair changes.

Certain RNA–RNA interactions, e.g. the functionally important interactions between the 16S and 23S rRNAs in the ribosome, involve non-canonical base pairs. Our method primarily incorporates canonical Watson–Crick base pairs and G-U wobble base pairs. However, the Pfold model includes equilibrium distributions for the frequencies of all possible 16 base pairs and substitution rates for all possible base pair substitutions (16×16 matrix), which have been estimated from given trusted alignments of tRNAs and rRNAs including non-canonical base pairs [93]. The probabilities of non-canonical base pairs are low compared to Watson–Crick and wobble base pairs and, thus, in practice, non-canonical base pairs are only found together with canonical base pairs. Nevertheless, the equilibrium distributions in the evolutionary model of PETcofold could be adapted to increase the impact of non-canonical base pairs by using a different training set.

In the second folding step of the PETcofold pipeline, a joint structure for the concatenated input alignments is predicted under the constraint that all positions participating in the partial structures as determined in the first folding step are single-stranded. As a result, the energy contributions from this cofolding step might be slightly biased. For example, an extension of a helix in a partial structure will be evaluated as an internal loop or hairpin. We partly solve this problem by extending reliable stems in the partial structures in the intramolecular folding step. Furthermore, the RNAcofold algorithm could be adapted to use new symbols for base pair constraints, which are currently handled as single-stranded in the recursions.

In some cases, the hierarchical folding approach predicts RNA–RNA interactions with reduced accuracy because the intramolecular constraints overlap with the interaction sites. In these cases, the interaction site accessibility model of PETcofold is too strict and overestimates the stability of intramolecular structures. As an example, one of the two OxyS–*fhlA* interaction sites was not predicted because the stem enclosing the second interaction site had a high reliability and, thus, got constrained (Figure 4.5B). To prevent too restrictive constraints that might conflict with reliable alternative structures, we introduced a threshold γ for the probability of the ensemble of structures that are compatible with the constraints. The interaction prediction for the example OxyS–*fhlA*, however, could not be improved by adjusting the value of this parameter. A future version of `PETcofold` could also take into account the cost of opening stems as done in, e.g. `RNAup` and `IntaRNA`.

Chapter 5

Conclusion

In this thesis, we developed two novel approaches that address the problem of computational RNA–RNA interaction prediction. Both methods were applied to the prediction of interactions between bacterial sRNAs and their target mRNAs. In addition, we identified sequence and structure features that are common to the majority of experimentally verified sRNA–mRNA interactions and successfully used these features to improve the genome-wide prediction of sRNA targets.

In the first part of this thesis, we presented a fast and general approach for the prediction of RNA–RNA interactions. Our approach `IntaRNA` was designed to predict mRNA target sites for base-pairing ncRNAs like eukaryotic miRNAs or bacterial sRNAs, but the method can also be applied to predict other types of RNA–RNA interactions. The prediction of target sites with `IntaRNA` is based on two assumptions: (i) the interaction contains a seed region that is thought to initiate interaction formation and (ii) the accessibility of the interaction sites is important for target recognition. Our evaluation on a dataset of experimentally proven sRNA–mRNA interactions demonstrated that the incorporation of these two requirements substantially improves the prediction quality of `IntaRNA`. The target sites of sRNAs are typically located in 5' UTRs and the beginning of the CDS. Our tool is fast enough to search these regions on a genome-wide scale within minutes to hours, depending on the length of the sRNA sequence and the number of annotated genes in the respective species. The runtime of `IntaRNA` is dominated by the calculation of the interaction site accessibilities. Multiple runs of genome-wide target predictions in the same species can therefore be sped up significantly when the accessibilities of all annotated genes have already been precomputed. Due to its high accuracy and decent runtime, the tool `IntaRNA` is among the best of its class and is frequently used to identify target candidates of functionally uncharacterised sRNAs. Another reason for the standing of our tool is its accessibility via an easy to use web interface.

Motivated by the high demand for accurate large-scale sRNA target identification approaches, we systematically analysed functional sRNA–mRNA interactions in the second part of this thesis to explore new features that potentially improve the specificity of genome-wide target predictions. We first compiled a set of 74 highly reliable sRNA–mRNA interactions in *E. coli* and *Salmonella*. All interactions were experimentally verified by *in vitro* probing or mutational studies. We then collected a set of all mRNAs including full-length 5' UTRs, which is required for an accurate calculation of structural RNA properties. Based on these data, we generated a negative dataset that closely resembled the functional interactions. When comparing positive and negative dataset, we found to our surprise that only interaction sites in sRNAs, but not in targets, displayed significant sequence conservation. As a result of the missing target site conservation, we observed no general conservation of complementarity between sRNAs and targets. The accessibility of the interaction sites in general and of seed regions in particular was significantly higher in both sRNAs and targets. We also observed that the sequence composition of the interaction sites and their flanking regions agreed with the binding preference of the RNA chaperone Hfq. Both a computational benchmark and a case study in the cyanobacterium *Prochlorococcus* MED4 confirmed the importance of interaction site accessibility and seed regions in general and the importance of seed conservation in sRNAs. Most notably, we demonstrated that the incorporation of appropriate seed constraints considerably reduces the number of target candidates and that an accessibility scoring improves the ranking of true positives. The finding on characteristical sequence composition at and around the interaction sites suggests the use of a machine learning approach to classify putative sRNA targets and discriminate between functional and non-functional ones [214].

In the third part of this thesis, we presented a comparative approach for the prediction of joint secondary structures of two interacting RNAs. Given two multiple alignments each representing a conserved RNA, our method `PETcofold` can take covariance information in intra- and intermolecular base pairs into account to predict secondary structures and interactions of the two RNAs. We showed for four sRNA–mRNA examples that our alignment-based method is able to predict joint secondary structures with a higher accuracy than methods based on single sequences. Another evaluation used a dataset with phylogenetically simulated sequences enriched for covariance patterns at the interaction sites, for which we observed a better performance with increased amounts of covariance. As for other alignment-based RNA structure prediction methods, the prediction performance of `PETcofold` crucially depends on the quality of the input alignments. It is, however, challenging that the conservation of target complementarity can range from marginal to full conservation even for different targets of the same sRNA as shown in our analysis of sRNA–mRNA interaction features. This is especially problematic as it is not known a priori whether the interaction between a specific sRNA and mRNA is well conserved or not. A very promising alternative to alignment-based approaches is provided by the following strategy: instead of using fixed RNA sequence alignments, a set of homologous sRNA sequences and sets of homologous mRNA sequences are used as input. Interactions are predicted separately in each of the phylogenetically related species and significances in terms of p-values are calculated for each putative interaction. For each set of homologous mRNAs, the p-values of all single interactions are then combined into a single p-value as an overall significance measure for this target. Preliminary results indicate that this strategy can significantly improve the specificity of genome-wide sRNA target predictions. Furthermore, this approach seems to be relatively robust to missing interaction conservation in a subset of the species (Georg, J., Wright, P. R., Richter, A. S., Hess, W. R., and Backofen, R. In preparation).

Long RNA–RNA interactions form helices similar to the DNA double strand [206]. Due to the turn of the helix, (i) the length of the interaction between the two RNAs is constrained and (ii) a certain number of unpaired bases on either side of the interaction are necessary to enable the first enclosing intramolecular base pair [138]. In practice, this means that computational methods can predict longer helices due to base pair complementarity, but in the three-dimensional topology these base pairings would be spatially unfeasible. However, taking topological constraints into account would be excessively time-costly. Our RNA–RNA interaction prediction methods `IntaRNA` and `PETcofold`, and all other methods capable of genome-scale analysis, work only at the level of secondary structures. However, in the future it would be useful to additionally accommodate the constraints posed by the three-dimensional RNA structure.

To conclude, we developed a new approach for the fast and accurate prediction of RNA–RNA interactions that incorporates interaction site accessibility and seed regions, and we developed the first comparative method for the prediction of secondary structures and interactions of two multiple alignments of RNA sequences. Furthermore, we performed one of the first systematic studies on structural accessibility and conservation of interacting sRNAs and mRNAs. Finally, the identification of two novel targets of the cyanobacterial sRNA Yfr1 by a combination of seed-based target prediction and experimental target verification proved the practicality of a combination of computational and experimental methods, especially in organisms where genetic manipulation constitutes a great challenge.

Appendix A

Datasets

A.1 Evaluation of bacterial RNA–RNA interaction features

Table A.1. Dataset of sRNA–mRNA interactions in *E. coli*. Target interaction site positions are given as distance to the annotated translation start site. Interactions are given in bracket notation, where the '&' symbol concatenates the sRNA with its target, matching brackets represent base pairs between the two sequences and dots represent unpaired positions.

sRNA	Target	sRNA site	Target site	Interaction	Validation	Ref.
ArcZ	*rpoS*	66 – 91	-120 – -99	`(((((((..(((.((((.(((.((((&)))))))))))))).)))))))`	compensatory mutations	[113]
ChiX	*chbC*	38 – 58	-69 – -49	`(((((((..((((((((((((&))))))))))))..)))))))`	sRNA mutations	[132]
ChiX	*chiP*	45 – 56	-19 – -8	`((((((((((((&))))))))))))`	compensatory mutations	[149]
ChiX	*dpiB*	46 – 57	-37 – -26	`((((((((((((&))))))))))))`	compensatory mutations	[112]
CyaR	*luxS*	35 – 49	-12 – 3	`(....(.((((((((&)))))))).)....)`	compensatory mutations	[38]
CyaR	*nadE*	35 – 49	-11 – 3	`(...(((((.(((((&))))))))))...)`	compensatory mutations	[38]
CyaR	*ompX*	38 – 48	-9 – 2	`(.(((((((((&))))))))).)`	compensatory mutations	[38]
CyaR	*yqaE*	31 – 50	-4 – 16	`(.(.(((((((((......(&)......))))))))).).)`	compensatory mutations	[38]
DsrA	*hns*	31 – 43	7 – 19	`(((((((((((((&)))))))))))))`	compensatory mutations	[102]
DsrA	*rpoS*	8 – 32	-119 – -95	`(.((((((((..(((((((((((((&)))))))))))))..)))))))).)`	compensatory mutations	[109]
FnrS	*folE*	1 – 12	-27 – -15	`(((((((.((((&))))..)))))))`	sRNA mutations	[44]
FnrS	*folX*	1 – 6	-7 – -2	`((((((&))))))`	sRNA mutations	[44]
FnrS	*gpmA*	38 – 57	-13 – 4	`((((((...(((((.(((((&)))))))))).))))))`	compensatory mutations	[44]

FnrS	*maeA*	31 – 65	-21 – 10	`(((.(.(((((.(((.((((((..(((((((((((&`
				`)))))))))))))))))))).))))).))))`
					compensatory mutations	[44]

FnrS	*sodB*	1 – 8	13 – 20	`((((((((&))))))))`	compensatory mutations	[44]

GcvB	*sstT*	64 – 99	-34 – 2	`(((.(...(((.(.((((((((((......((.(((&`
				`))).))......)))))))))).).)))...).)))`
					compensatory mutations	[146]

GlmZ	*glmS*	150 – 169	-40 – -22	`((((((((.....(((((((&)))))))....))))))))`	compensatory mutations	[189]
MicA	*ompA*	8 – 24	-21 – -6	`((((.((((((((((((&))))))))))))))))`	compensatory mutations	[187]
MicA	*phoP*	6 – 31	-15 – 8	`((((((((((.((((..((.((((((&))))))))).))))))))))))))`	compensatory mutations	[35]

MicC	*ompC*	1 – 30	-41 – -15	`((((((((((((((((........((((((&`
				`)))))).....))))))))))))))))`
					compensatory mutations	[28]

MicF	*ompF*	1 – 33	-16 – 10	`(((.((((((((......((((((((((.((((&`	*in vitro* probing	[159]

(Continued)

sRNA	Target	sRNA site	Target site	Interaction	Validation	Ref.
				)))).)))))))))))))))))))))		
OmrA	*cirA*	2 – 24	-35 – -10	((((((((((((...((((((((&))))))))....))).)))))).)))	compensatory mutations	[70]
OmrA	*csgD*	2 – 20	-79 – -61	((((((((((((((.((((&)))).))))))))))))))	compensatory mutations	[77]
OmrA	*ompR*	1 – 19	-29 – -11	(((((((((.(..(.((((&)))).)..).)))))))))	compensatory mutations	[70]
OmrA	*ompT*	1 – 33	-12 – 20	(((((((((.(((.(((((((((...(((..((&))..))).)))))).))).))).)))))))))	compensatory mutations	[70]
OmrB	*cirA*	2 – 24	-35 – -10	((((((((((((..(((((((((&)))))))))...))).)))))).)))	compensatory mutations	[70]
OmrB	*csgD*	2 – 20	-79 – -61	(((((((((((((((((((&)))))))))))))))))))	compensatory mutations	[77]
OmrB	*ompR*	1 – 19	-29 – -11	(((((((((.(..(..(((&)))..)..).)))))))))	compensatory mutations	[70]
OmrB	*ompT*	1 – 32	-12 – 20	(((((((((.(((.(.(((((((....(((((&)))))...)))))).).).))).)))))))))	compensatory mutations	[70]
OxyS	*fhlA*	22 – 30	34 – 42	(((((((((&)))))))))	compensatory mutations	[6]
		98 – 104	-15 – -9	(((((((&)))))))		
RprA	*rpoS*	33 – 62	-117 – -94	(((((..((((((.........((((((((&))))))))...))))))..)))))	compensatory mutations	[110]
RyhB	*cysE*	34 – 46	-4 – 9	(((((((((((((&)))))))))))))	compensatory mutations	[158]
RyhB	*fur*	38 – 76	-96 – -47	((((((((((..(.(((((..((.(((.....(((((((&))).))))))).))..))......))).).......))))))....))))	compensatory mutations	[191]
RyhB	*iscS*	40 – 68	-26 – 3	((((((((..............(((((((&))))))).............))))))))	*in vitro* probing	[40]
RyhB	*shiA*	44 – 55	-59 – -48	((((((((((((&))))))))))))	compensatory mutations	[142]
RyhB	*sodB*	38 – 46	-4 – 5	(((((((((&)))))))))	*in vitro* probing	[57]
SgrS	*manX*	159 – 172	24 – 37	((..((((((((((&))))))))))..))	compensatory mutations	[153]
SgrS	*ptsG*	168 – 187	-28 – -9	((((..((((((((.(((((&))))).))))))))..))))	compensatory mutations, *in vitro* probing	[90, 153]
Spot42	*galK*	20 – 61	-19 – 21	((((.(.((((((((((.((((.....((((..(((((((((&))))))))).))))..))))...)))))))))).).))))	*in vitro* probing	[126]
Spot42	*gltA*	4 – 13	-131 – -122	((((((((((&))))))))))	sRNA mutations	[13]

(Continued)

sRNA	Target	sRNA site	Target site	Interaction	Validation	Ref.
Spot42	*nanC*	1 – 17	-33 – -16	`(((((((((((((((((&)))).)))))))))))))`	compensatory mutations	[13]
Spot42	*srlA*	20 – 34	-15 – -1	`((((((.((((((((&)))))))).))))))`	compensatory mutations	[13]
Spot42	*sthA*	48 – 55	15 – 22	`((((((((&))))))))`	compensatory mutations	[13]

Spot42	*xylF*	1 – 33	2 – 40	`((.((((.((((((((.((((((((...(((((&`
				`)))))....)))))))).)).....)))))).)))).))`
					sRNA mutations	[13]

Table A.2. Dataset of sRNA–mRNA interactions in *Salmonella*. Target interaction site positions are given as distance to the annotated translation start site. Interactions are given in bracket notation, where the '&' symbol concatenates the sRNA with its target, matching brackets represent base pairs between the two sequences and dots represent unpaired positions.

sRNA	Target	sRNA site	Target site	Interaction	Validation	Ref.
ArcZ	*sdaC*	62 – 71	-13 – -3	((((((((((&))))).)))))	compensatory mutations	[134]
ArcZ	STM3216	63 – 87	-25 – -5	(((((((((.((((.....((((((&)))))).)))).)))))))))	compensatory mutations	[134]
ArcZ	*tpx*	66 – 83	10 – 26	((((((.......(((((&)))))......))))))	compensatory mutations	[134]
ChiX	*chbC*	35 – 55	-66 – -46	(((((((..((((((((((((&))))))))))))..)))))))	mRNA mutations	[50]
ChiX	*chiP*	42 – 53	-19 – -8	((((((((((((&))))))))))))	compensatory mutations	[50]
CyaR	*ompX*	35 – 66	-30 – 3	((..(((((((((..((((....((((..(((&))).))))).....)).))...))))))))).))	compensatory mutations	[133]
GcvB	*argT*	75 – 91	-57 – -42	(((((((.(((((((((&))))))))))))))))	*in vitro* probing	[163]
GcvB	*cycA*	72 – 85	-34 – -19	((((((((((((.(&).)).)))))))).))	*in vitro* probing	[165]
		138 – 161	-24 – -8	((((((((.......(((((((((&)))))))))))))))))		
GcvB	*dppA*	65 – 82	-30 – -14	((((((((.(((((((((&)))))))))))))))))	*in vitro* probing	[163]
GcvB	*gltI*	65 – 76	-38 – -27	((((((((((((&))))))))))))	*in vitro* probing	[163]
GcvB	*livJ*	63 – 87	-51 – -28	((((((.((((.(((((((.(((((&))))).))))))))))).))))))	*in vitro* probing	[163]
GcvB	*livK*	65 – 77	-29 – -17	(((((((((((((&)))))))))))))	*in vitro* probing	[163]
GcvB	*oppA*	65 – 89	-8 – 16	((((((((((...((((((((((((&))))))))))))..))))))))))	*in vitro* probing	[163]
GcvB	STM4351	69 – 79	-54 – -43	(((((((((((&)).)))))))))	*in vitro* probing	[163]
InvR	*ompD*	33 – 42	56 – 65	((((((((((&))))))))))	*in vitro* probing	[139]
MicA	*lamB*	8 – 36	-9 – 18	(((((...((.((((((((..((((((((&)))))))).)))))))).))..)))))	compensatory mutations	[19]
MicC	*ompD*	1 – 12	67 – 78	((((((((((((&))))))))))))	compensatory mutations	[140]
RybB	*chiP*	1 – 7	12 – 18	(((((((&)))))))	compensatory mutations	[11]
RybB	*fadL*	1 – 8	49 – 56	((((((((&))))))))	compensatory mutations	[135]
RybB	*ompA*	1 – 13	21 – 32	((((((((...((&))..))))))))	compensatory mutations	[135]
RybB	*ompC*	1 – 10	-50 – -41	((((((((((&))))))))))	*in vitro* probing	[11, 135]
RybB	*ompD*	1 – 9	18 – 26	(((((((((&)))))))))	compensatory mutations	[11, 135]
		1 – 10	10 – 20	((((((((((&))).)))))))		

(Continued)

sRNA	Target	sRNA site	Target site	Interaction	Validation	Ref.
RybB	*ompF*	1 – 9	-46 – -38	`(((((((((&)))))))))`	compensatory mutations	[135]
RybB	*ompN*	1 – 16	5 – 20	`((((.(((((((((((&))))))))))).))))`	compensatory mutations	[20]
RybB	*ompS*	1 – 14	7 – 20	`((((..((((((((&))))))))..))))`	sRNA deletion mutant	[135]
RybB	*ompW*	1 – 13	3 – 20	`(((((((.(((((&)))))......)))))))`	compensatory mutations	[135]
RybB	*tsx*	1 – 16	-26 – -7	`(((((((...((((((&))))))).......)))))))`	compensatory mutations	[135]

Table A.3. Enterobacterial organisms used for conservation analysis of sRNA–mRNA interactions and their respective NCBI RefSeq database genome accession numbers [145].

Organism	RefSeq genome accession number
Citrobacter koseri	NC_009792
Citrobacter rodentium	NC_013716
Cronobacter sakazakii	NC_009778
Enterobacter sp. 638	NC_009436
Escherichia coli K-12	NC_000913
Escherichia fergusonii	NC_011740
Klebsiella pneumoniae	NC_009648
Pectobacterium carotovorum	NC_012917
Photorhabdus luminescens	NC_005126
Proteus mirabilis	NC_010554
Salmonella Typhimurium	NC_003197
Salmonella Typhi	NC_003198
Serratia proteamaculans	NC_009832
Shigella boydii	NC_007613
Shigella dysenteriae	NC_007606
Shigella flexneri	NC_004337
Shigella sonnei	NC_007384
Sodalis glossinidius	NC_007712
Yersinia enterocolitica	NC_008800
Yersinia pestis	NC_003143
Yersinia pseudotuberculosis	NC_006155

Table A.4. Dataset of non-functional sRNA–mRNA interactions. Interaction site positions in the mRNA are given as distance to the annotated translation start site. Interactions are given in bracket notation, where the '&' symbol concatenates the sRNA and the mRNA, matching brackets represent base pairs between the two sequences and dots represent unpaired positions. The last column gives for each non-functional interaction the true target of the corresponding verified interaction.

Organism	sRNA	mRNA	sRNA site	mRNA site	Predicted non-functional interaction	True target
E. coli	ArcZ	*ypdB*	26 – 51	-133 – -112	`((((((....(((.((((((((((((&)))).)))))))))))))))))`	*rpoS*
E. coli	ChiX	*ltaE*	4 – 24	-371 – -351	`((((((((((((((..(((((&))))).)))))).))))))))`	*chbC*
E. coli	ChiX	*ybaK*	12 – 23	-144 – -133	`((((((((((((&))))))))))))`	*chiP*
E. coli	ChiX	*ybaK*	2 – 13	-134 – -123	`((((((((((((&))))))))))))`	*dpiB*
E. coli	CyaR	*hofO*	69 – 83	-182 – -168	`(....(.((((((((&)))))))).)....)`	*luxS*
E. coli	CyaR	*livF*	66 – 80	-127 – -114	`(((...(((.(((((&))))).)))..)))`	*nadE*
E. coli	CyaR	*pabC*	60 – 79	-7 – 13	`((((......((.((.((((&)))).)).))......))))`	*yqaE*
E. coli	CyaR	*rlpA*	49 – 59	-256 – -246	`(.(((((((((&))))))))).)`	*ompX*
E. coli	DsrA	*barA*	73 – 85	34 – 46	`(((((((((((((&)))))))))))))`	*hns*

E. coli	DsrA	*ykgE*	56 – 83	-58 – -27	`((((...(((((((((((((.((..(((&`
					`)))))..)))))))).....)))))...))))`
						rpoS

E. coli	FnrS	*lspA*	110 – 115	-132 – -127	`((((((&))))))`	*folX*
E. coli	FnrS	*rmuC*	86 – 105	-12 – 5	`((.(((((((((...(((((&))))).)))))))))))`	*gpmA*
E. coli	FnrS	*rsmH*	15 – 22	132 – 139	`((((((((&))))))))`	*sodB*
E. coli	FnrS	*ubiE*	27 – 38	-60 – -48	`(((((((.((((&))))..)))))))`	*folE*

E. coli	FnrS	*yaeI*	77 – 116	-115 – -73	`(.....(.(((((..((.((((((((((.((.((((((((&`
					`))))))))...........)).)))))))))).)).)))))))`
						maeA

E. coli	GcvB	*yfeY*	6 – 41	-44 – -9	`((.(..(((((((....((((....(((.((((.((&`
					`)).)))).)))))))..)))).)))..)......))`
						sstT

E. coli	GlmZ	*gabT*	65 – 84	-174 – -156	`((((...(((((((..((((&)))))))))))....))))`	*glmS*
E. coli	MicA	*aceF*	45 – 61	-380 – -365	`(((((((((((.(((((&))))))))))))))))`	*ompA*

E. coli	MicA	*tauC*	36 – 68	-146 – -113	`(((((.(((......((((.(((.((..(((((&`
					`)))))..))...))).)))).....))).)))))`
						phoP

E. coli	MicC	*yiiD*	41 – 70	-50 – -24	`(((((..((((((((.((((...((..(((&)))..)).)))).))).))))))))))`	*ompC*
E. coli	MicF	*ileS*	56 – 90	-353 – -316	`((((..(((.(((((..((((((((..((((...(&`	*ompF*

(Continued)

Organism	sRNA	mRNA	sRNA site	mRNA site	Predicted non-functional interaction	True target
					`)...)))).))))......))))....))))))))))))`	
E. coli	OmrA	*gspM*	43 – 75	-41 – -10	`(.((((((((..(((((((..((((.(((((.(&`	*ompT*
					`))))))).)))))))))).))))))...)).)`	
E. coli	OmrA	*insI*	27 – 45	-18 – 1	`(((((...((.((((((((&)).))))))))...)))))`	*ompR*
E. coli	OmrA	*marA*	25 – 47	-70 – -45	`((((((((..(((((((((.(((&).)))))))).)))...)).))))))`	*cirA*
E. coli	OmrA	*ygcO*	64 – 82	-41 – -23	`((((.((((((((((((((&))))))))).))))))))))`	*csgD*
E. coli	OmrB	*hofC*	36 – 54	-59 – -41	`(((((.(.((..((((.((&))..))))..)).))))))`	*ompR*
E. coli	OmrB	*mcbA*	31 – 63	-56 – -23	`(((((.(.((...(((.........((((((((&`	*csgD*
					`))))))))............))).)).).)))))`	
E. coli	OmrB	*pyrB*	35 – 57	-146 – -121	`((((((((((((((((((((..(&).)).))))..))))).)))))))))`	*cirA*
E. coli	OmrB	*yfdQ*	35 – 66	-85 – -54	`(((((...((((((((.(((.(.((((((.((&`	*ompT*
					`)).))))))).)))...)))).)))).)))))`	
E. coli	OxyS	*apaH*	7 – 13	-30 – -24	`(((((((&)))))))`	*fhlA*
E. coli	OxyS	*caiT*	73 – 81	76 – 84	`(((((((((&)))))))))`	*fhlA*
E. coli	RprA	*paaC*	67 – 96	-37 – -14	`(((((..((..((.(((((......(((((&)))))...)))))))..)))))))`	*rpoS*
E. coli	RyhB	*sdhC*	48 – 60	-161 – -149	`(((((((((((((&)))))))))))))`	*cysE*
E. coli	RyhB	*yagV*	77 – 85	-397 – -389	`(((((((((&)))))))))`	*sodB*
E. coli	RyhB	*yeaW*	3 – 37	-131 – -75	`(.(((((..((((.(((((((((((.(((.(((.(&`	*fur*
					`))))..))).)).........))))).)).))....))))).....)))))......)`	
E. coli	RyhB	*yicS*	1 – 29	-180 – -152	`(((....(........((((((.(.((((&`	*iscS*
					`))))).)))))).........)....)))`	
E. coli	RyhB	*yigE*	77 – 88	-188 – -177	`((((((((((((&))))))))))))`	*shiA*
E. coli	SgrS	*hlyE*	66 – 79	43 – 56	`((..((((((((((&))))))))))..))`	*manX*
E. coli	SgrS	*putP*	64 – 83	-121 – -102	`((.((((((..(((((((((&)))))))))..)))))).))`	*ptsG*
E. coli	Spot42	*aat*	43 – 59	-28 – -11	`(((((((((((((((((&))))))).))))))))))`	*nanC*
E. coli	Spot42	*arnT*	80 – 89	-37 – -28	`((((((((((&))))))))))`	*gltA*
E. coli	Spot42	*cynX*	82 – 89	18 – 25	`((((((((&))))))))`	*sthA*
E. coli	Spot42	*rhlE*	65 – 106	-40 – -1	`((.(((((((((((((((..((.((....((((((..(((((&`	*galK*

(Continued)

Organism	sRNA	mRNA	sRNA site	mRNA site	Predicted non-functional interaction	True target
					))))))..).))))).))))..))))))))).)))))).))	
E. coli	Spot42	*xylH*	61 – 75	-123 – -109	((((((.((((((((&)))))))).))))))	*srlA*
E. coli	Spot42	*yggU*	64 – 96	15 – 53	(((((((...(((((((((.((((.((.(((((&))))).)).))))))...))....)))))..)))))))	*xylF*
Salmonella	ArcZ	STM3651	92 – 116	-125 – -105	((..((((((....(((((((((((&)))))))))))))))))..))	STM3216
Salmonella	ArcZ	*pduJ*	73 – 82	-178 – -168	((((((((((&))))).)))))	*sdaC*
Salmonella	ArcZ	*purA*	5 – 22	108 – 124	((((((.......(((((&)))))......))))))	*tpx*
Salmonella	ChiX	*atpB*	4 – 24	-177 – -157	((((((((((((..(((((((&))))))).)))))).))))))	*chbC*
Salmonella	ChiX	*ybaK*	2 – 13	-135 – -124	((((((((((((&))))))))))))	*chiP*
Salmonella	CyaR	STM1787	1 – 32	-110 – -78	((((....(((..((((...((((((.(((((&))))).)))))).).))).....)))...))))	*ompX*
Salmonella	GcvB	STM1049	57 – 80	-165 – -149	(((((((((((.......((((((&)))))))))))))))))	*cycA*
Salmonella	GcvB	STM2598	49 – 60	-169 – -158	((((((((((((&))))))))))))	*gltI*
Salmonella	GcvB	STM2768	20 – 37	-142 – -126	((((((.(((((((((((&)))))))))))))))))	*dppA*
Salmonella	GcvB	STM4002	49 – 65	-159 – -144	(((((((.(((((((((&))))))))))))))))	*argT*
Salmonella	GcvB	STM4032.2N	36 – 60	-20 – 4	(((((((((.(((.(((.(((((((&))))))).))).))))))))))))	*livJ*
Salmonella	GcvB	STM4032.2N	37 – 61	-21 – 3	((((((((.(((.(((.((((((((&)))))))).))).)))))))))))	*oppA*
Salmonella	GcvB	*flgJ*	32 – 44	-139 – -127	(((((((((((((&)))))))))))))	*livK*
Salmonella	GcvB	*rfaJ*	49 – 62	-422 – -407	((((((((((((.(&).)).)))))))).))	*cycA*
Salmonella	GcvB	*sspH2*	40 – 50	-112 – -101	(((((((((((&)).)))))))))	STM4351
Salmonella	InvR	*dps*	50 – 59	63 – 72	((((((((((&))))))))))	*ompD*
Salmonella	MicA	STM0952	41 – 69	-52 – -26	((((.(((((((((((((((.....((((&))))))).))))))))))))...))))	*lamB*
Salmonella	MicC	*yaeH*	92 – 103	19 – 30	((((((((((((&))))))))))))	*ompD*
Salmonella	RybB	STM1632	32 – 40	-40 – -32	(((((((((&)))))))))	*ompF*
Salmonella	RybB	STM1636	30 – 45	-192 – -173	(((((((...((((((&)))))).......)))))))	*tsx*
Salmonella	RybB	STM3356	62 – 77	3 – 18	((((.(((((((((((&))))))))))).))))	*ompN*
Salmonella	RybB	*pmrF*	66 – 79	60 – 73	((((..((((((((&))))))))..))))	*ompS*
Salmonella	RybB	*pps*	19 – 28	69 – 79	((((((((((&))).)))))))	*ompD*

(Continued)

Organism	sRNA	mRNA	sRNA site	mRNA site	Predicted non-functional interaction	True target
Salmonella	RybB	*pyrG*	44 – 56	24 – 35	`((((((((...((&))..))))))))`	*ompA*
Salmonella	RybB	*rstB*	65 – 74	-78 – -69	`((((((((((&))))))))))`	*ompC*
Salmonella	RybB	*secF*	20 – 32	131 – 148	`(((((((.(((((&)))))......)))))))`	*ompW*
Salmonella	RybB	*stfC*	36 – 42	45 – 51	`(((((((&)))))))`	*chiP*
Salmonella	RybB	*wcaG*	32 – 40	44 – 52	`(((((((((&)))))))))`	*ompD*
Salmonella	RybB	*yabI*	45 – 52	92 – 99	`((((((((&))))))))`	*fadL*

A.2 Performance evaluation of PETcofold

Table A.5. Dataset of bacterial sRNAs and their target mRNAs used in the performance evaluation of PETcofold. Rfam acc. denotes the Rfam accession number of the sRNA [55].

sRNA	Rfam acc.	Target	Organism	Ref.
CyaR	RF00112	*luxS*	*E. coli*	[38]
CyaR	RF00112	*nadE*	*E. coli*	[38]
CyaR	RF00112	*ompX*	*E. coli*	[38]
CyaR	RF00112	*yqaE*	*E. coli*	[38]
CyaR	RF00112	*ompX*	*Salmonella*	[133]
DsrA	RF00014	*hns*	*E. coli*	[102]
DsrA	RF00014	*rpoS*	*E. coli*	[109]
GcvB	RF00022	*sstT*	*E. coli*	[146]
GcvB	RF00022	*argT*	*Salmonella*	[163]
GcvB	RF00022	*dppA*	*Salmonella*	[163]
GcvB	RF00022	*gltI*	*Salmonella*	[163]
GcvB	RF00022	*livJ*	*Salmonella*	[163]
GcvB	RF00022	*livK*	*Salmonella*	[163]
GcvB	RF00022	*oppA*	*Salmonella*	[163]
GcvB	RF00022	STM4351	*Salmonella*	[163]
GlmZ	RF00083	*glmS*	*E. coli*	[189]
MicA	RF00078	*ompA*	*E. coli*	[187]
MicA	RF00078	*lamB*	*Salmonella*	[19]
MicC	RF00121	*ompC*	*E. coli*	[28]
MicC	RF00121	*ompD*	*Salmonella*	[140]
MicF	RF00033	*ompF*	*E. coli*	[159]
OmrA	RF00079	*cirA*	*E. coli*	[70]
OmrA	RF00079	*ompR*	*E. coli*	[70]
OmrA	RF00079	*ompT*	*E. coli*	[70]
OxyS	RF00035	*fhlA*	*E. coli*	[6]
RNAIII	RF00503	SA1000	*S. aureus*	[18]
RNAIII	RF00503	SA2353	*S. aureus*	[18]
RNAIII	RF00503	*spa*	*S. aureus*	[81]
RprA	RF00034	*rpoS*	*E. coli*	[110]
RyhB	RF00057	*fur*	*E. coli*	[191]
RyhB	RF00057	*sodB*	*E. coli*	[57]
SgrS	RF00534	*ptsG*	*E. coli*	[90]

Bibliography

[1] Alkan, C., Karakoç, E., Nadeau, J. H., Sahinalp, S. C., and Zhang, K. RNA-RNA interaction prediction and antisense RNA target search. *J Comput Biol*, 13(2):267–82, 2006.

[2] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, 25(17):3389–402, 1997.

[3] Amaral, P. P., Dinger, M. E., Mercer, T. R., and Mattick, J. S. The eukaryotic genome as an RNA machine. *Science*, 319(5871):1787–9, 2008.

[4] Ameres, S. L., Martinez, J., and Schroeder, R. Molecular basis for target RNA recognition and cleavage by human RISC. *Cell*, 130(1):101–12, 2007.

[5] Andronescu, M., Zhang, Z. C., and Condon, A. Secondary structure prediction of interacting RNA molecules. *J Mol Biol*, 345(5):987–1001, 2005.

[6] Argaman, L. and Altuvia, S. *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.

[7] Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E. G., Margalit, H., and Altuvia, S. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr Biol*, 11(12):941–50, 2001.

[8] Axmann, I. M., Kensche, P., Vogel, J., Kohl, S., Herzel, H., and Hess, W. R. Identification of cyanobacterial non-coding RNAs by comparative genome analysis. *Genome Biol*, 6(9):R73, 2005.

[9] Backofen, R. and Hess, W. R. Computational prediction of sRNAs and their targets in bacteria. *RNA Biol*, 7(1):33–42, 2010.

[10] Baker, J. L., Sudarsan, N., Weinberg, Z., Roth, A., Stockbridge, R. B., and Breaker, R. R. Widespread genetic switches and toxicity resistance proteins for fluoride. *Science*, 335(6065):233–5, 2012.

[11] Balbontín, R., Fiorini, F., Figueroa-Bossi, N., Casadesús, J., and Bossi, L. Recognition of heptameric seed sequence underlies multi-target regulation by RybB small RNA in *Salmonella enterica*. *Mol Microbiol*, 78(2):380–94, 2010.

[12] Bartel, D. P. MicroRNAs: target recognition and regulatory functions. *Cell*, 136(2): 215–33, 2009.

[13] Beisel, C. L. and Storz, G. The base-pairing RNA Spot 42 participates in a multi-output feedforward loop to help enact catabolite repression in *Escherichia coli*. *Mol Cell*, 41(3):286–97, 2011.

[14] Bernhart, S. H., Hofacker, I. L., and Stadler, P. F. Local RNA base pairing probabilities in large sequences. *Bioinformatics*, 22(5):614–5, 2006.

[15] Bernhart, S. H., Tafer, H., Mückstein, U., Flamm, C., Stadler, P. F., and Hofacker, I. L. Partition function and base pairing probabilities of RNA heterodimers. *Algorithms Mol Biol*, 1(1):3, 2006.

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

[17] Blattner, F. R., Plunkett III, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., and Shao, Y. The complete genome sequence of *Escherichia coli* K-12. *Science*, 277(5331):1453–62, 1997.

[18] Boisset, S., Geissmann, T., Huntzinger, E., Fechter, P., Bendridi, N., Possedko, M., Chevalier, C., Helfer, A. C., Benito, Y., Jacquier, A., Gaspin, C., Vandenesch, F., and Romby, P. *Staphylococcus aureus* RNAIII coordinately represses the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism. *Genes Dev*, 21(11):1353–66, 2007.

[19] Bossi, L. and Figueroa-Bossi, N. A small RNA downregulates LamB maltoporin in *Salmonella*. *Mol Microbiol*, 65(3):799–810, 2007.

[20] Bouvier, M., Sharma, C. M., Mika, F., Nierhaus, K. H., and Vogel, J. Small RNA binding to 5' mRNA coding region inhibits translational initiation. *Mol Cell*, 32(6): 827–37, 2008.

[21] Breaker, R. R. Prospects for riboswitch discovery and analysis. *Mol Cell*, 43(6): 867–79, 2011.

[22] Breaker, R. R. Riboswitches and the RNA world. *Cold Spring Harb Perspect Biol*, 4(2):a003566, 2012.

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

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

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

[26] Cheah, M. T., Wachter, A., Sudarsan, N., and Breaker, R. R. Control of alternative RNA splicing and gene expression by eukaryotic riboswitches. *Nature*, 447(7143): 497–500, 2007.

[27] Chen, S., Lesnik, E. A., Hall, T. A., Sampath, R., Griffey, R. H., Ecker, D. J., and Blyn, L. B. A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. *Biosystems*, 65(2-3):157–77, 2002.

[28] Chen, S., Zhang, A., Blyn, L. B., and Storz, G. MicC, a second small-RNA regulator of Omp protein expression in *Escherichia coli. J Bacteriol*, 186(20):6689–97, 2004.

[29] Chevalier, C., Boisset, S., Romilly, C., Masquida, B., Fechter, P., Geissmann, T., Vandenesch, F., and Romby, P. *Staphylococcus aureus* RNAIII binds to two distant regions of *coa* mRNA to arrest translation and promote mRNA degradation. *PLoS Pathog*, 6(3):e1000809, 2010.

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

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

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

[33] Cochrane, G., Akhtar, R., Bonfield, J., Bower, L., Demiralp, F., Faruque, N., Gibson, R., Hoad, G., Hubbard, T., Hunter, C., Jang, M., Juhos, S., Leinonen, R., Leonard, S., Lin, Q., Lopez, R., Lorenc, D., McWilliam, H., Mukherjee, G., Plaister, S., Radhakrishnan, R., Robinson, S., Sobhany, S., Hoopen, P. T., Vaughan, R., Zalunin, V., and Birney, E. Petabyte-scale innovations at the European Nucleotide Archive. *Nucleic Acids Res*, 37(Database issue):D19–25, 2009.

[34] Consortium, A. F. B., Backofen, R., Bernhart, S. H., Flamm, C., Fried, C., Fritzsch, G., Hackermüller, J., Hertel, J., Hofacker, I. L., Missal, K., Mosig, A., Prohaska, S. J., Rose, D., Stadler, P. F., Tanzer, A., Washietl, S., and Will, S. RNAs everywhere: genome-wide annotation of structured RNAs. *J Exp Zoolog B Mol Dev Evol*, 308B(1):1–25, 2007.

[35] Coornaert, A., Lu, A., Mandin, P., Springer, M., Gottesman, S., and Guillier, M. MicA sRNA links the PhoP regulon to cell envelope stress. *Mol Microbiol*, 76(2): 467–79, 2010.

[36] Crick, F. Central dogma of molecular biology. *Nature*, 227(5258):561–3, 1970.

[37] Darfeuille, F., Unoson, C., Vogel, J., and Wagner, E. G. H. An antisense RNA inhibits translation by competing with standby ribosomes. *Mol Cell*, 26(3):381–92, 2007.

[38] De Lay, N. and Gottesman, S. The Crp-activated small noncoding regulatory RNA CyaR (RyeE) links nutritional status to group behavior. *J Bacteriol*, 191(2):461–76, 2009.

[39] Delihas, N. Annotation and evolutionary relationships of a small regulatory RNA gene *micF* and its target *ompF* in *Yersinia* species. *BMC Microbiol*, 3:13, 2003.

[40] Desnoyers, G., Morissette, A., Prévost, K., and Massé, E. Small RNA-induced differential degradation of the polycistronic mRNA *iscRSUA*. *EMBO J*, 28(11): 1551–61, 2009.

[41] Dimitrov, R. A. and Zuker, M. Prediction of hybridization and melting for double-stranded nucleic acids. *Biophys J*, 87(1):215–26, 2004.

[42] Do, C. B., Woods, D. A., and Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–8, 2006.

[43] Dühring, U., Axmann, I. M., Hess, W. R., and Wilde, A. An internal antisense RNA regulates expression of the photosynthesis gene *isiA*. *Proc Natl Acad Sci USA*, 103 (18):7054–8, 2006.

[44] Durand, S. and Storz, G. Reprogramming of anaerobic metabolism by the FnrS small RNA. *Mol Microbiol*, 75(5):1215–31, 2010.

[45] Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet*, 2 (12):919–29, 2001.

[46] Eggenhofer, F., Tafer, H., Stadler, P. F., and Hofacker, I. L. RNApredator: fast accessibility-based prediction of sRNA targets. *Nucleic Acids Res*, 39(Web Server issue):W149–54, 2011.

[47] Ellwanger, D. C., Büttner, F. A., Mewes, H.-W., and Stümpflen, V. The sufficient minimal set of miRNA seed types. *Bioinformatics*, 27(10):1346–50, 2011.

[48] Ezz El-Din El-Shaer, M. A new heuristic algorithm for IntaRNA for improved RNA–RNA interaction prediction. Bachelor thesis, German University in Cairo, March 2011.

[49] Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*, 17(6):368–76, 1981.

[50] Figueroa-Bossi, N., Valentini, M., Malleret, L., Fiorini, F., and Bossi, L. Caught at its own game: regulatory small RNA inactivated by an inducible transcript mimicking its target. *Genes Dev*, 23(17):2004–15, 2009.

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

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

[53] Fröhlich, K. S. and Vogel, J. Activation of gene expression by small RNA. *Curr Opin Microbiol*, 12(6):674–82, 2009.

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

[55] Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., Finn, R. D., Griffiths-Jones, S., Eddy, S. R., and Bateman, A. Rfam: updates to the RNA families database. *Nucleic Acids Res*, 37(Database issue):D136–40, 2009.

[56] Gaspin, C. and Westhof, E. An interactive framework for RNA secondary structure prediction with a dynamical treatment of constraints. *J Mol Biol*, 254(2):163–74, 1995.

[57] Geissmann, T. A. and Touati, D. Hfq, a new chaperoning role: binding to messenger RNA determines access for small RNA regulator. *EMBO J*, 23(2):396–405, 2004.

[58] Georg, J. and Hess, W. R. *cis*-antisense RNA, another level of gene regulation in bacteria. *Microbiol Mol Biol Rev*, 75(2):286–300, 2011.

[59] Gerdes, K., Rasmussen, P. B., and Molin, S. Unique type of plasmid maintenance function: postsegregational killing of plasmid-free cells. *Proc Natl Acad Sci USA*, 83(10):3116–20, 1986.

[60] Gesell, T. and von Haeseler, A. *In silico* sequence evolution with site-specific interactions along phylogenetic trees. *Bioinformatics*, 22(6):716–22, 2006.

[61] Gesteland, R. F., Cech, T. R., and Atkins, J. F., editors. *The RNA world*. Cold Spring Harbor Laboratory Press, 3rd edition, 2006.

[62] Goericke, R. and Welschmeyer, N. A. The marine prochlorophyte *Prochlorococcus* contributes significantly to phytoplankton biomass and primary production in the Sargasso Sea. *Deep Sea Res Part I Oceanogr Res Pap*, 40(11-12):2283–2294, 1993.

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

[64] Gorodkin, J., Stricklin, S. L., and Stormo, G. D. Discovering common stem-loop motifs in unaligned RNA sequences. *Nucleic Acids Res*, 29(10):2135–44, 2001.

[65] Gottesman, S. Microbiology: Dicing defence in bacteria. *Nature*, 471(7340):588–9, 2011.

[66] Gottesman, S. and Storz, G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harb Perspect Biol*, 3(12):a003798, 2011.

[67] Griffin, B. E. Separation of 32P-labelled ribonucleic acid components. The use of polyethylenimine-cellulose (TLC) as a second dimension in separating oligoribonucleotides of '4.5 S' and 5 S from *E. coli*. *FEBS Lett*, 15(3):165–168, 1971.

[68] Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R., and Hofacker, I. L. The Vienna RNA websuite. *Nucleic Acids Res*, 36(Web Server issue):W70–4, 2008.

[69] Güell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., Rode, M., Suyama, M., Schmidt, S., Gavin, A.-C., Bork, P., and Serrano, L. Transcriptome complexity in a genome-reduced bacterium. *Science*, 326(5957):1268–71, 2009.

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

[71] Hansel, A., Pattus, F., Jürgens, U. J., and Tadros, M. H. Cloning and characterization of the genes coding for two porins in the unicellular cyanobacterium *Synechococcus* PCC 6301. *Biochim Biophys Acta*, 1399(1):31–9, 1998.

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

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

[74] Hindley, J. Fractionation of 32P-labelled ribonucleic acids on polyacrylamide gels and their characterization by fingerprinting. *J Mol Biol*, 30(1):125–36, 1967.

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

[76] Hofacker, I. L., Fekete, M., and Stadler, P. F. Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, 319(5):1059–66, 2002.

[77] Holmqvist, E., Reimegård, J., Sterk, M., Grantcharova, N., Römling, U., and Wagner, E. G. H. Two antisense RNAs target the transcriptional regulator CsgD to inhibit curli synthesis. *EMBO J*, 29(11):1840–50, 2010.

[78] Horler, R. S. P. and Vanderpool, C. K. Homologs of the small RNA SgrS are broadly distributed in enteric bacteria but have diverged in size and sequence. *Nucleic Acids Res*, 37(16):5465–76, 2009.

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

[80] Huang, F. W. D., Qin, J., Reidys, C. M., and Stadler, P. F. Target prediction and a statistical sampling algorithm for RNA-RNA interaction. *Bioinformatics*, 26(2): 175–81, 2010.

[81] Huntzinger, E., Boisset, S., Saveanu, C., Benito, Y., Geissmann, T., Namane, A., Lina, G., Etienne, J., Ehresmann, B., Ehresmann, C., Jacquier, A., Vandenesch, F., and Romby, P. *Staphylococcus aureus* RNAIII and the endoribonuclease III coordinately regulate *spa* gene expression. *EMBO J*, 24(4):824–35, 2005.

[82] Hussein, R. and Lim, H. N. Direct comparison of small RNA and transcription factor signaling. *Nucleic Acids Res*, 2012. Advance Access published May 22, 2012, doi:10.1093/nar/gks439.

[83] Hüttenhofer, A. and Noller, H. F. Footprinting mRNA-ribosome complexes with chemical probes. *EMBO J*, 13(16):3892–901, 1994.

[84] Jabbari, H., Condon, A., and Zhao, S. Novel and efficient RNA secondary structure prediction using hierarchical folding. *J Comput Biol*, 15(2):139–63, 2008.

[85] Jäger, D., Sharma, C. M., Thomsen, J., Ehlers, C., Vogel, J., and Schmitz, R. A. Deep sequencing analysis of the *Methanosarcina mazei* Gö1 transcriptome in response to nitrogen availability. *Proc Natl Acad Sci USA*, 106(51):21878–82, 2009.

[86] Johansson, J., Mandin, P., Renzoni, A., Chiaruttini, C., Springer, M., and Cossart, P. An RNA thermosensor controls expression of virulence genes in *Listeria monocytogenes*. *Cell*, 110(5):551–61, 2002.

[87] Jousselin, A., Metzinger, L., and Felden, B. On the facultative requirement of the bacterial RNA chaperone, Hfq. *Trends Microbiol*, 17(9):399–405, 2009.

[88] Kato, Y., Sato, K., Hamada, M., Watanabe, Y., Asai, K., and Akutsu, T. RactIP: fast and accurate prediction of RNA-RNA interaction using integer programming. *Bioinformatics*, 26(18):i460–i466, 2010.

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

[90] Kawamoto, H., Koide, Y., Morita, T., and Aiba, H. Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq. *Mol Microbiol*, 61(4):1013–22, 2006.

[91] Kertesz, M., Iovino, N., Unnerstall, U., Gaul, U., and Segal, E. The role of site accessibility in microRNA target recognition. *Nat Genet*, 39(10):1278–84, 2007.

[92] Kettler, G. C., Martiny, A. C., Huang, K., Zucker, J., Coleman, M. L., Rodrigue, S., Chen, F., Lapidus, A., Ferriera, S., Johnson, J., Steglich, C., Church, G. M., Richardson, P., and Chisholm, S. W. Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. *PLoS Genet*, 3(12):e231, 2007.

[93] Knudsen, B. and Hein, J. RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics*, 15(6):446–54, 1999.

[94] Knudsen, B. and Hein, J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, 31(13):3423–8, 2003.

[95] Kolbe, D. L. and Eddy, S. R. Local RNA structure alignment with incomplete sequence. *Bioinformatics*, 25(10):1236–43, 2009.

[96] Kozak, M. Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, 361:13–37, 2005.

[97] Kretschmer-Kazemi Far, R. and Sczakiel, G. The activity of siRNA in mammalian cells is related to structural target accessibility: a comparison with antisense oligonucleotides. *Nucleic Acids Res*, 31(15):4417–24, 2003.

[98] Kröger, C., Dillon, S. C., Cameron, A. D. S., Papenfort, K., Sivasankaran, S. K., Hokamp, K., Chao, Y., Sittka, A., Hébrard, M., Händler, K., Colgan, A., Leekitcharoenphon, P., Langridge, G. C., Lohan, A. J., Loftus, B., Lucchini, S., Ussery, D. W., Dorman, C. J., Thomson, N. R., Vogel, J., and Hinton, J. C. D. The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. *Proc Natl Acad Sci USA*, 109(20):E1277–86, 2012.

[99] Lall, S., Grün, D., Krek, A., Chen, K., Wang, Y.-L., Dewey, C. N., Sood, P., Colombo, T., Bray, N., MacMenamin, P., Kao, H.-L., Gunsalus, K. C., Pachter, L., Piano, F., and Rajewsky, N. A genome-wide map of conserved microRNA targets in *C. elegans*. *Curr Biol*, 16(5):460–71, 2006.

[100] Lange, S. J., Maticzka, D., Möhl, M., Gagnon, J. N., Brown, C. M., and Backofen, R. Global or local? Predicting secondary structure and accessibility in mRNAs. *Nucleic Acids Res*, 2012. Advance Access published February 28, 2012, doi:10.1093/nar/gks181.

[101] Laursen, B. S., Sørensen, H. P., Mortensen, K. K., and Sperling-Petersen, H. U. Initiation of protein synthesis in bacteria. *Microbiol Mol Biol Rev*, 69(1):101–23, 2005.

[102] Lease, R. A., Cusick, M. E., and Belfort, M. Riboregulation in *Escherichia coli*: DsrA RNA acts by RNA:RNA interactions at multiple loci. *Proc Natl Acad Sci USA*, 95(21):12456–61, 1998.

[103] Lerner, M. R., Boyle, J. A., Mount, S. M., Wolin, S. L., and Steitz, J. A. Are snRNPs involved in splicing? *Nature*, 283(5743):220–4, 1980.

[104] Li, A. X., Marz, M., Qin, J., and Reidys, C. M. RNA-RNA interaction prediction based on multiple sequence alignments. *Bioinformatics*, 27(4):456–63, 2011.

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

[106] Long, D., Lee, R., Williams, P., Chan, C. Y., Ambros, V., and Ding, Y. Potent effect of target structure on microRNA function. *Nat Struct Mol Biol*, 14(4):287–94, 2007.

[107] Luo, K. Q. and Chang, D. C. The gene-silencing efficiency of siRNA is strongly dependent on the local structure of mRNA at the targeted region. *Biochem Biophys Res Commun*, 318(1):303–10, 2004.

[108] Maglott, D., Ostell, J., Pruitt, K. D., and Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res*, 39(Database issue):D52–7, 2011.

[109] Majdalani, N., Cunning, C., Sledjeski, D., Elliott, T., and Gottesman, S. DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription. *Proc Natl Acad Sci USA*, 95(21): 12462–7, 1998.

[110] Majdalani, N., Hernandez, D., and Gottesman, S. Regulation and mode of action of the second small RNA activator of RpoS translation, RprA. *Mol Microbiol*, 46(3): 813–26, 2002.

[111] Malys, N. and McCarthy, J. E. G. Translation initiation: variations in the mechanism can be anticipated. *Cell Mol Life Sci*, 68(6):991–1003, 2011.

[112] Mandin, P. and Gottesman, S. A genetic approach for finding small RNAs regulators of genes of interest identifies RybC as regulating the DpiA/DpiB two-component system. *Mol Microbiol*, 72(3):551–65, 2009.

[113] Mandin, P. and Gottesman, S. Integrating anaerobic/aerobic sensing and the general stress response through the ArcZ small RNA. *EMBO J*, 29(18):3094–107, 2010.

[114] Mann, M., Smith, C., Rabbath, M., Edwards, M., Will, S., and Backofen, R. CPSP-web-tools: a server for 3D lattice protein studies. *Bioinformatics*, 25(5):676–7, 2009.

[115] Marín, R. M. and Vaníček, J. Efficient use of accessibility in microRNA target prediction. *Nucleic Acids Res*, 39(1):19–29, 2011.

[116] Markham, N. R. and Zuker, M. DINAMelt web server for nucleic acid melting prediction. *Nucleic Acids Res*, 33(Web Server issue):W577–81, 2005.

[117] Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Anderson, I., Lykidis, A., Mavromatis, K., Ivanova, N. N., and Kyrpides, N. C. The integrated microbial genomes system: an expanding comparative analysis resource. *Nucleic Acids Res*, 38(Database issue):D382–90, 2010.

[118] Massé, E. and Gottesman, S. A small RNA regulates the expression of genes involved in iron metabolism in *Escherichia coli*. *Proc Natl Acad Sci USA*, 99(7):4620–5, 2002.

[119] Mathews, D., Sabina, J., Zuker, M., and Turner, D. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, 288(5):911–40, 1999.

[120] Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim Biophys Acta*, 405(2):442–51, 1975.

[121] McCaskill, J. S. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, 29(6-7):1105–19, 1990.

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

[123] Mironov, A. S., Gusarov, I., Rafikov, R., Lopez, L. E., Shatalin, K., Kreneva, R. A., Perumov, D. A., and Nudler, E. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell*, 111(5):747–56, 2002.

[124] Mitschke, J., Georg, J., Scholz, I., Sharma, C. M., Dienst, D., Bantscheff, J., Voß, B., Steglich, C., Wilde, A., Vogel, J., and Hess, W. R. An experimentally anchored map of transcriptional start sites in the model cyanobacterium *Synechocystis* sp. PCC6803. *Proc Natl Acad Sci USA*, 108(5):2124–9, 2011.

[125] Mizuno, T., Chou, M. Y., and Inouye, M. A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA). *Proc Natl Acad Sci USA*, 81(7):1966–70, 1984.

[126] Møller, T., Franch, T., Udesen, C., Gerdes, K., and Valentin-Hansen, P. Spot 42 RNA mediates discoordinate expression of the *E. coli* galactose operon. *Genes Dev*, 16(13):1696–706, 2002.

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

[128] Mückstein, U., Tafer, H., Bernhart, S. H., Hernandez-Rosales, M., Vogel, J., Stadler, P. F., and Hofacker, I. L. Translational control by RNA-RNA interaction: Improved computation of RNA-RNA binding thermodynamics. In Elloumi, M., Küng, J., Linial, M., Murphy, R., Schneider, K., and Toma, C., editors, *Bioinformatics Research and Development*, volume 13 of *Communications in Computer and Information Science*, pages 114–127. Springer-Verlag Berlin Heidelberg, 2008.

[129] Nakamura, T., Naito, K., Yokota, N., Sugita, C., and Sugita, M. A cyanobacterial non-coding RNA, Yfr1, is required for growth under multiple stress conditions. *Plant Cell Physiol*, 48(9):1309–18, 2007.

[130] Nguyen, T. X., Alegre, E. R., and Kelley, S. T. Phylogenetic analysis of general bacterial porins: a phylogenomic case study. *J Mol Microbiol Biotechnol*, 11(6): 291–301, 2006.

[131] Nussinov, R., Pieczenik, G., Griggs, J. R., and Kleitman, D. J. Algorithms for loop matchings. *SIAM J Appl Math*, 35(1):68–82, July 1978.

[132] Overgaard, M., Johansen, J., Møller-Jensen, J., and Valentin-Hansen, P. Switching off small RNA regulation with trap-mRNA. *Mol Microbiol*, 73(5):790–800, 2009.

[133] Papenfort, K., Pfeiffer, V., Lucchini, S., Sonawane, A., Hinton, J. C. D., and Vogel, J. Systematic deletion of *Salmonella* small RNA genes identifies CyaR, a conserved CRP-dependent riboregulator of OmpX synthesis. *Mol Microbiol*, 68(4):890–906, 2008.

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

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

[136] Pasquinelli, A. E. MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. *Nat Rev Genet*, 13(4):271–82, 2012.

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

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

[139] Pfeiffer, V., Sittka, A., Tomer, R., Tedin, K., Brinkmann, V., and Vogel, J. A small non-coding RNA of the invasion gene island (SPI-1) represses outer membrane protein synthesis from the *Salmonella* core genome. *Mol Microbiol*, 66(5):1174–91, 2007.

[140] Pfeiffer, V., Papenfort, K., Lucchini, S., Hinton, J. C. D., and Vogel, J. Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. *Nat Struct Mol Biol*, 16(8):840–6, 2009.

[141] Poritz, M. A., Bernstein, H. D., Strub, K., Zopf, D., Wilhelm, H., and Walter, P. An *E. coli* ribonucleoprotein containing 4.5S RNA resembles mammalian signal recognition particle. *Science*, 250(4984):1111–7, 1990.

[142] Prévost, K., Salvail, H., Desnoyers, G., Jacques, J.-F., Phaneuf, É., and Massé, E. The small RNA RyhB activates the translation of *shiA* mRNA encoding a permease of shikimate, a compound involved in siderophore synthesis. *Mol Microbiol*, 64(5): 1260–73, 2007.

[143] Prévost, K., Desnoyers, G., Jacques, J.-F., Lavoie, F., and Massé, E. Small RNA-induced mRNA degradation achieved through both translation block and activated cleavage. *Genes Dev*, 25(4):385–96, 2011.

[144] Pruitt, K. D., Tatusova, T., and Maglott, D. R. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, 35(Database issue):D61–5, 2007.

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

[146] Pulvermacher, S. C., Stauffer, L. T., and Stauffer, G. V. The small RNA GcvB regulates *sstT* mRNA expression in *Escherichia coli*. *J Bacteriol*, 191(1):238–48, 2009.

[147] Raghavan, R., Groisman, E. A., and Ochman, H. Genome-wide detection of novel regulatory RNAs in *E. coli*. *Genome Res*, 21(9):1487–97, 2011.

[148] Rajewsky, N. microRNA target predictions in animals. *Nat Genet*, 38 Suppl:S8–13, 2006.

[149] Rasmussen, A. A., Johansen, J., Nielsen, J. S., Overgaard, M., Kallipolitis, B., and Valentin-Hansen, P. A conserved small RNA promotes silencing of the outer membrane protein YbfM. *Mol Microbiol*, 72(3):566–77, 2009.

[150] Rehmsmeier, M., Steffen, P., Höchsmann, M., and Giegerich, R. Fast and effective prediction of microRNA/target duplexes. *RNA*, 10(10):1507–17, 2004.

[151] Repoila, F., Majdalani, N., and Gottesman, S. Small non-coding RNAs, co-ordinators of adaptation processes in *Escherichia coli*: the RpoS paradigm. *Mol Microbiol*, 48(4):855–61, 2003.

[152] Ribes, V., Römisch, K., Giner, A., Dobberstein, B., and Tollervey, D. *E. coli* 4.5S RNA is part of a ribonucleoprotein particle that has properties related to signal recognition particle. *Cell*, 63(3):591–600, 1990.

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

[154] Rivas, E., Klein, R. J., Jones, T. A., and Eddy, S. R. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr Biol*, 11(17):1369–73, 2001.

[155] Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S., Chain, P., Ahlgren, N. A., Arellano, A., Coleman, M., Hauser, L., Hess, W. R., Johnson, Z. I., Land, M., Lindell, D., Post, A. F., Regala, W., Shah, M., Shaw, S. L., Steglich, C., Sullivan, M. B., Ting, C. S., Tolonen, A., Webb, E. A., Zinser, E. R., and Chisholm, S. W. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature*, 424(6952):1042–7, 2003.

[156] Salari, R., Backofen, R., and Sahinalp, S. C. Fast prediction of RNA-RNA interaction. *Algorithms Mol Biol*, 5:5, 2010.

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

[158] Salvail, H., Lanthier-Bourbonnais, P., Sobota, J. M., Caza, M., Benjamin, J.-A. M., Mendieta, M. E. S., Lépine, F., Dozois, C. M., Imlay, J., and Massé, E. A small RNA promotes siderophore production through transcriptional and metabolic remodeling. *Proc Natl Acad Sci USA*, 107(34):15223–8, 2010.

[159] Schmidt, M., Zheng, P., and Delihas, N. Secondary structures of *Escherichia coli* antisense *micF* RNA, the 5'-end of the target *ompF* mRNA, and the RNA/RNA duplex. *Biochemistry*, 34(11):3621–31, 1995.

[160] Seemann, S. E., Gorodkin, J., and Backofen, R. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. *Nucleic Acids Res*, 36 (20):6355–62, 2008.

[161] Shao, Y., Chan, C. Y., Maliyekkel, A., Lawrence, C. E., Roninson, I. B., and Ding, Y. Effect of target secondary structure on RNAi efficiency. *RNA*, 13(10):1631–40, 2007.

[162] Sharma, C. M. and Vogel, J. Experimental approaches for the discovery and characterization of regulatory small RNA. *Curr Opin Microbiol*, 12(5):536–46, 2009.

[163] Sharma, C. M., Darfeuille, F., Plantinga, T. H., and Vogel, J. A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. *Genes Dev*, 21(21):2804–17, 2007.

[164] Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiß, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F., and Vogel, J. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, 464(7286):250–5, 2010.

[165] Sharma, C. M., Papenfort, K., Pernitzsch, S. R., Mollenkopf, H.-J., Hinton, J. C. D., and Vogel, J. Pervasive post-transcriptional control of genes involved in amino acid metabolism by the Hfq-dependent GcvB small RNA. *Mol Microbiol*, 81(5):1144–65, 2011.

[166] Sharp, P. A. The centrality of RNA. *Cell*, 136(4):577–80, 2009.

[167] Shine, J. and Dalgarno, L. The 3'-terminal sequence of Escherichia coli 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci USA*, 71(4):1342–6, 1974.

[168] Simonetti, A., Marzi, S., Jenner, L., Myasnikov, A., Romby, P., Yusupova, G., Klaholz, B. P., and Yusupov, M. A structural view of translation initiation in bacteria. *Cell Mol Life Sci*, 66(3):423–36, 2009.

[169] Sittka, A., Lucchini, S., Papenfort, K., Sharma, C. M., Rolle, K., Binnewies, T. T., Hinton, J. C. D., and Vogel, J. Deep sequencing analysis of small noncoding RNA and mRNA targets of the global post-transcriptional regulator, Hfq. *PLoS Genet*, 4 (8):e1000163, 2008.

[170] Smith, T. F. and Waterman, M. S. Identification of common molecular subsequences. *J Mol Biol*, 147(1):195–7, 1981.

[171] Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G. R., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res*, 12(10):1611–8, 2002.

[172] Starczynowski, D. T., Morin, R., McPherson, A., Lam, J., Chari, R., Wegrzyn, J., Kuchenbauer, F., Hirst, M., Tohyama, K., Humphries, R. K., Lam, W. L., Marra, M., and Karsan, A. Genome-wide identification of human microRNAs located in leukemia-associated genomic alterations. *Blood*, 117(2):595–607, 2011.

[173] Starmer, J., Stomp, A., Vouk, M., and Bitzer, D. Predicting Shine-Dalgarno sequence locations exposes genome annotation errors. *PLoS Comput Biol*, 2(5):e57, 2006.

[174] Steglich, C., Futschik, M. E., Lindell, D., Voß, B., Chisholm, S. W., and Hess, W. R. The challenge of regulation in a minimal photoautotroph: non-coding RNAs in *Prochlorococcus*. *PLoS Genet*, 4(8):e1000173, 2008.

[175] Steglich, C., Lindell, D., Futschik, M., Rector, T., Steen, R., and Chisholm, S. W. Short RNA half-lives in the slow-growing marine cyanobacterium *Prochlorococcus*. *Genome Biol*, 11(5):R54, 2010.

[176] Storz, G., Vogel, J., and Wassarman, K. M. Regulation by small RNAs in bacteria: expanding frontiers. *Mol Cell*, 43(6):880–91, 2011.

[177] Stougaard, P., Molin, S., and Nordström, K. RNAs involved in copy-number control and incompatibility of plasmid R1. *Proc Natl Acad Sci USA*, 78(10):6008–12, 1981.

[178] Tafer, H. and Hofacker, I. L. RNAplex: a fast tool for RNA-RNA interaction search. *Bioinformatics*, 24(22):2657–63, 2008.

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

[180] Takyar, S., Hickerson, R. P., and Noller, H. F. mRNA helicase activity of the ribosome. *Cell*, 120(1):49–58, 2005.

[181] Tjaden, B. TargetRNA: a tool for predicting targets of small RNA action in bacteria. *Nucleic Acids Res*, 36(Web Server issue):W109–13, 2008.

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

[183] Toledo-Arana, A., Dussurget, O., Nikitas, G., Sesto, N., Guet-Revillet, H., Balestrino, D., Loh, E., Gripenland, J., Tiensuu, T., Vaitkevicius, K., Barthelemy, M., Vergassola, M., Nahori, M.-A., Soubigou, G., Régnault, B., Coppée, J.-Y., Lecuit, M., Johansson, J., and Cossart, P. The *Listeria* transcriptional landscape from saprophytism to virulence. *Nature*, 459(7249):950–6, 2009.

[184] Tomizawa, J., Itoh, T., Selzer, G., and Som, T. Inhibition of ColE1 RNA primer formation by a plasmid-specified small RNA. *Proc Natl Acad Sci USA*, 78(3):1421–5, 1981.

[185] Torarinsson, E., Yao, Z., Wiklund, E. D., Bramsen, J. B., Hansen, C., Kjems, J., Tommerup, N., Ruzzo, W. L., and Gorodkin, J. Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res*, 18(2):242–51, 2008.

[186] Turner, D. H. and Mathews, D. H. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. *Nucleic Acids Res*, 38 (Database issue):D280–2, 2010.

[187] Udekwu, K. I., Darfeuille, F., Vogel, J., Reimegård, J., Holmqvist, E., and Wagner, E. G. H. Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA. *Genes Dev*, 19(19):2355–66, 2005.

[188] Urban, J. H. and Vogel, J. Translational control and target recognition by *Escherichia coli* small RNAs *in vivo*. *Nucleic Acids Res*, 35(3):1018–37, 2007.

[189] Urban, J. H. and Vogel, J. Two seemingly homologous noncoding RNAs act hierarchically to activate *glmS* mRNA translation. *PLoS Biol*, 6(3):e64, 2008.

[190] Vaulot, D., Marie, D., Olson, R. J., and Chisholm, S. W. Growth of *Prochlorococcus*, a photosynthetic prokaryote, in the equatorial Pacific Ocean. *Science*, 268(5216): 1480–1482, 1995.

[191] Večerek, B., Moll, I., and Bläsi, U. Control of Fur synthesis by the non-coding RNA RyhB and iron-responsive decoding. *EMBO J*, 26(4):965–75, 2007.

[192] Vinh, L. S. and von Haeseler, A. IQPNNI: moving fast through tree space and stopping in time. *Mol Biol Evol*, 21(8):1565–71, 2004.

[193] Vogel, J. and Luisi, B. F. Hfq and its constellation of RNA. *Nat Rev Microbiol*, 9 (8):578–89, 2011.

[194] Vogel, J. and Sharma, C. M. How to find small non-coding RNAs in bacteria. *Biol Chem*, 386(12):1219–38, 2005.

[195] Vogel, J., Argaman, L., Wagner, E. G. H., and Altuvia, S. The small RNA IstR inhibits synthesis of an SOS-induced toxic peptide. *Curr Biol*, 14(24):2271–6, 2004.

[196] Voß, B., Gierga, G., Axmann, I. M., and Hess, W. R. A motif-based search in bacterial genomes identifies the ortholog of the small RNA Yfr1 in all lineages of cyanobacteria. *BMC Genomics*, 8:375, 2007.

[197] Voß, B., Georg, J., Schön, V., Ude, S., and Hess, W. R. Biocomputational prediction of non-coding RNAs in model cyanobacteria. *BMC Genomics*, 10:123, 2009.

[198] Wagner, E. G. H. Kill the messenger: bacterial antisense RNA promotes mRNA decay. *Nat Struct Mol Biol*, 16(8):804–6, 2009.

[199] Washietl, S. and Hofacker, I. L. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol*, 342(1):19–30, 2004.

[200] Washietl, S., Hofacker, I. L., Lukasser, M., Hüttenhofer, A., and Stadler, P. F. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat Biotechnol*, 23(11):1383–90, 2005.

[201] Wassarman, K. M. 6S RNA: a small RNA regulator of transcription. *Curr Opin Microbiol*, 10(2):164–8, 2007.

[202] Wassarman, K. M. and Storz, G. 6S RNA regulates *E. coli* RNA polymerase activity. *Cell*, 101(6):613–23, 2000.

[203] Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G., and Gottesman, S. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev*, 15(13):1637–51, 2001.

[204] Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. Jalview Version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, 25(9):1189–91, 2009.

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

[206] Watson, J. D. and Crick, F. H. C. Molecular structure of nucleic acids. A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[207] Weinberg, Z., Barrick, J. E., Yao, Z., Roth, A., Kim, J. N., Gore, J., Wang, J. X., Lee, E. R., Block, K. F., Sudarsan, N., Neph, S., Tompa, M., Ruzzo, W. L., and Breaker, R. R. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res*, 35(14):4809–19, 2007.

[208] Weinberg, Z., Wang, J. X., Bogue, J., Yang, J., Corbino, K., Moy, R. H., and Breaker, R. R. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. *Genome Biol*, 11(3):R31, 2010.

[209] Wiedenheft, B., Sternberg, S. H., and Doudna, J. A. RNA-guided genetic silencing systems in bacteria and archaea. *Nature*, 482(7385):331–8, 2012.

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

[211] Will, S., Joshi, T., Hofacker, I. L., Stadler, P. F., and Backofen, R. LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs. *RNA*, 18(5):900–14, 2012.

[212] Winkler, W., Nahvi, A., and Breaker, R. R. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, 419(6910):952–6, 2002.

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

[214] Zhao, Y., Li, H., Hou, Y., Cha, L., Cao, Y., Wang, L., Ying, X., and Li, W. Construction of two mathematical models for prediction of bacterial sRNA targets. *Biochem Biophys Res Commun*, 372(2):346–50, 2008.

[215] Zieve, G. W. Two groups of small stable RNAs. *Cell*, 25(2):296–7, 1981.

[216] Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res*, 31(13):3406–15, 2003.

[217] Zuker, M. and Stiegler, P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–48, 1981.

Abbreviations

asRNA	*cis*-encoded antisense RNA
bp	base pair(s)
CBP	consistent/compensatory base pair exchanges
CDS	coding sequence
CRISPR	clustered regularly interspaced short palindromic repeats
E. coli	*Escherichia coli*
ED	energy difference (free energy that is required to make a subsequence single-stranded)
EF	expected fraction of unpaired bases
fMet	formylmethionine
FN	false negatives
FP	false positives
GFP	green fluorescent protein
GTP	guanosine triphosphate
IF	(translation) initiation factor
MCC	Matthews correlation coefficient
mfe	minimum free energy
miRNA	microRNA
mRNA	messenger RNA
NCBI	National Center for Biotechnology Information
ncRNA	non-coding RNA
nt	nucleotide(s)
ORF	open reading frame
PPV	positive predictive value
PU	probability that a subsequence is unpaired
RBS	ribosome binding site
RNA-seq	(high-throughput) RNA sequencing
rRNA	ribosomal RNA
S. aureus	*Staphylococcus aureus*
SD	Shine–Dalgarno

SENS	sensitivity
siRNA	small interfering RNA
snRNA	small nuclear RNA
sRNA	small (bacterial) RNA
TN	true negatives
TP	true positives
TPP	thiamine pyrophosphate
tRNA	transfer RNA
TSS	transcription start site
UTR	untranslated region