Bioinformatics of Prokaryotic RNAs

Rolf Backofen[c,h], Fabian Amman[d,b], Fabrizio Costa[c], Sven Findeiß[a,b], Andreas S. Richter[f,c], Peter F. Stadler[d,e,g,h,b,i]

[a]*Bioinformatics and Computational Biology Research Group, University of Vienna, Währingerstraße 29, A-1090 Wien, Austria*
[b]*Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria*
[c]*Bioinformatics Group, Department of Computer Science, University of Freiburg, Georges-Köhler-Allee 106, D-79110 Freiburg, Germany*
[d]*Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany*
[e]*Max Planck Institute for Mathematics in the Sciences, Inselstraße 22, D-04103 Leipzig, Germany*
[f]*Max Planck Institute of Immunobiology and Epigenetics, Stübeweg 51, D-79108 Freiburg, Germany*
[g]*Fraunhofer Institute for Cell Therapy and Immunology – IZI, Perlickstraße 1, D-04103 Leipzig, Germany*
[h]*Center for non-coding RNA in Technology and Health, University of Copenhagen, Grønnegårdsvej 3, DK-1870 Frederiksberg C, Denmark*
[i]*Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA*

** All authors contributed equally.*
The authors declare that no competing interests exist.

Abstract

The genome of most prokaryotes gives rise to surprisingly complex transcriptomes comprising not only protein-coding mRNAs, often organized as operons, but also harbors dozens or even hundreds of highly structured small regulatory RNAs and unexpectedly large levels of anti-sense transcripts. Comprehensive surveys of prokaryotic transcriptomes and the need to characterize also its non-coding components is heavily dependent on computational methods and workflows, many of which have been developed or at least adapted specifically for the use with eubacterial and archaeal data. This review provides on overview on the state of the art of RNA bioinformatics focussing on applications to prokaryotes.

Keywords: RNA bioinformatics, TSS annotation, target prediction, gene finding, RNA–RNA interaction, secondary structure prediction

1. Introduction

During the last decade thousands of small RNAs (sRNAs) have been discovered in a widely diverse set of prokaryotes. Beyond the evolutionary ancient "housekeeping" RNA genes encoding tRNAs, rRNAs, RNAse P RNA and SRP RNA (as well as tmRNA and 6S RNA in Eubacteria), typical genomes harbour dozens or even hundreds of sRNAs with predominantly regulatory roles. Archaea in addition have small nucleolar RNAs (snoRNAs) directing chemical modifications of rRNAs and other RNA targets. Compared to protein-coding genes, most of the prokaryotic RNAs are still rather poorly characterized in terms of their structure, function, and phylogenetic distribution. In particular, with the advent of high throughput transcriptomics, large numbers of sRNA candidates have been detected, but so far have not received attention beyond a note of their genomic coordinates.

Computational approaches have been very successful in facilitating, extending, and complementing experimental investigations. In this contribution we review the state of the art and the limitations of RNA bioinformatics as applied to prokaryotes. Our presentation emphasizes in particular methods and tools that were developed or substantially improved within the Priority Program SPP 1258: Sensory and regulatory RNAs in Prokaryotes funded by the *Deutsche Forschungsgemeinschaft* from 2007-2013.

Email addresses: backofen@informatik.uni-freiburg.de (Rolf Backofen), afabian@bioinf.uni-leipzig.de (Fabian Amman), costa@informatik.uni-freiburg.de (Fabrizio Costa), sven@tbi.univie.ac.at (Sven Findeiß), arichter@ie-freiburg.mpg.de (Andreas S. Richter), studla@bioinf.uni-leipzig.de (Peter F. Stadler)

2. Structure Prediction

The complex three-dimensional structures of single-stranded nucleic acids are dominated by base pairing both in terms of the energy of folding and in the sense that much of the shape can be understood in terms of the co-planar arrangement of the bases. At the same time, the status of a nucleotide as either paired or unpaired can be interrogated experimentally by means of chemical or enzymatic probing. This makes *secondary structures* an important level of description. At the same time, arrangements of base pairs can be predicted with fair accuracy from the sequence based on a few simple model assumptions: (i) Every nucleotide pairs with at most one other pairing partner, (ii) stacking of co-planar base pairs stabilizes the structures while unpaired "loop" regions primarily account for destabilizing effects, (iii) stacking and loops contribute additively to the energy of folding. Sequence-dependent energy parameters have been derived from a plethora of thermodynamic measurements[1]. Additional constraints, in particular the exclusion of crossing base pairs, i.e., the suppression of pseudoknots, leads to exact dynamic programming algorithms that run in cubic time on quadratic memory[2,3]. Not only the groundstate "minimum free energy" (mfe) structures can be computed in this manner. McCaskill's algorithm[4], for example, computes the partition function of the Boltzmann ensemble and provides access to all equilibrium base pairing probabilities; so-called stochastic backtracing procedures can generate large Boltzmann-weighted samples[5,6]. The most prominent implementations of RNA folding algorithms are `mfold`[7] and the `ViennaRNA Package`[8,9].

The non-crossing condition is not always satisfied in particular in highly structured RNAs such as RNAse P RNA. The paradigm of secondary structure folding can be extended to pseudoknotted structures, albeit at the expense of much higher computational costs. Different classes of pseudoknot structures have been defined and can be computed by a large number of tools[10], see also[11,12,13,14,15,16].

The accuracy of secondary prediction from single sequences is far from perfect for a wide variety of reasons. Some derive from limitations of the secondary structure model, such as deviations from the additive model, insufficient knowledge of energy parameters, simplified parametrization of multi-loops, and the exclusion of non-standard base pairs. Although some of these shortcomings can be overcome by a more complex model e.g. based on the Leontis-Westhof representation[17] without sacrificing computational efficiency[18], the need to parametrize such an extended model becomes an obstacle in itself. A second set of limitions is biological in nature: salt condition and physiological temperature which the studied species favors may differ substantially from the standard conditions at which thermodynamic parameters have been measured. Even more importantly, RNA is rarely ever "naked" but bound to proteins that may affect the energetics of folding. In addition, the precise transcript might be known only partially, or structure motifs are embedded into a larger RNA. In these cases, one has to apply *local* structure prediction, which is an even harder problem[19].

There are two remedies for these problems: (i) instead of just a single sequence, evolutionary information on patterns of sequence conservation may be taken into account, or (ii) experimental evidence such as chemical probing or FRET data may be incorporated into structure prediction.

When accurate sequence alignments can be obtained, these may serve as basis for computing consensus structures. The simplest approach, implemented e.g. in `RNAalifold`[20,21] is to extend the RNA folding algorithms to compute a secondary structure that minimizes the average folding energy of the aligned sequences. A more sophisticated phylogenetic model replacing simple averaging is implemented in `PETfold`[22]. At lower levels of sequence conservation, folding and alignment must be computed simultaneously at much higher computational cost. Several practical approaches exist, from full-flegded implementations of the Sankoff algorithm[23], e.g. in `Foldalign`[24] and `Dynalign`[25], to computationally much more efficient approximations that restrict themselves to base pairs that are thermodynamically plausible for the individual sequences. Tools of the latter type are `LocaRNA` and its variants[26,27,28,29] and `SPARSE`[30]. A conceptually different approach taken by the `RNAshapes` package[31] makes use of coarse-grained structures. In all cases, the output consists of a sequence alignment annotated by a consensus structure — exactly the input required later on for homology search.

Experimental data can be integrated into structure prediction either as hard constraints (enforcing or prohibiting certain base pairs) or as soft constraints that distort the ensemble of structure by adding bonus energies or energy penalities to encouraged or discouraged structural elements, resp. Measurement of SHAPE[32], PARS[33], or other chemical or enzymatic probing methods can be converted into pseudo-energies added to paired or unpaired bases, leading to a distortion of the Boltzmann ensemble towards the experimental signal[34,35]. Most recently, more sophisticated approaches have appeared towards reconciliating experimental data with the thermodynamic folding approach.

RNAassist[36] formulates the problem in terms of simultaneously minimizing position-dependent energy penalities and the deviation of observed and predicted probabilities for unpaired nucleotides. SeqFold uses the experimental data to select locally stable secondary structure from the Boltzmann ensemble[37]. In ShapeKnots[38] an interative procedures is used to include pseudoknots and SHAPE information. It has been applied to e.g. investigate the structure of a SAM-I riboswitch.

3. Gene Finding and Transcriptomics

3.1. Homology Search

The initial gene annotation of a newly sequenced genome is created by comparison with known sequences of related organisms together with the application of *de novo* prediction methods, in particular the search of open reading frames of sufficient length. Since non-coding RNAs (ncRNAs) do not offer a similar generic sequence pattern, they are much harder to predict from scratch[39]. As a consequence, only a few well-known RNA genes such as tRNAs, RNAse P RNA, SPR RNA, and the ribosomal RNA subunits are annotated for most prokaryotic genomes.

The Rfam database, as the most extensive repository of structured RNAs, lists in its current version 11.0 a total of 605 RNA families with prokaryotic members (527 bacterial and 107 archaeal)[40]. This number includes, however, a large number of CRISPR RNA repeats, many riboswitches, mRNA elements, as well as ubiquitous RNA families such as tRNAs or RNAse P. There is, at present, no comprehensive repository of prokaryotic small RNAs. The overwhelming majority of sRNAs discovered after the publication of a reference genome are documented only in the main text of publications or in supplemental material. Despite community efforts and incentives such as free open access publication of RNA family descriptions in this journal[41], only a very moderate number prokaryotic RNA families have been described in detail and deposited to databases, see e.g.[42,43,44,45]. As a consequence, the majority of sRNA families remains in practise unavailable for genome annotation pipelines. For the same reason it is impossible to give an accurate estimate on the total number of eubacterial or archaeal sRNA families or to globally assess their phylogenetic distributions with any degree of certainty.

The most widely used tool for homology search is blast. For highly diverged sequences blast typically reports several small fragments instead of the full length match to the query sequence. Semi-global dynamic programming algorithms such as Gotohscan[46] are a viable alternative given the small genome size of prokaryotes. This program reports full length hits, makes subsequent processing of the predicted homologs much easier and is particularly well-suited for ncRNAs[47], which — in contrast to protein-coding genes — are typically short and evolve rapidly at the sequence level. These properties generally limit the sensitivity of purely sequence-based methods. The information content of the query can be increased by making use of secondary structure conservation as well. Covariance models (CMs), a generalization of HMMs to tree-like structures provide a convenient technical basis[48]. They have to be trained from multiple sequence alignments annotated by a consensus structure. In contrast to blast, which is content with a single query sequence, CMs require a collection of evolutionarily related and alignable homologs as a starting point. With infernal 1.1 a highly efficient implementation of a search tool for CMs has become available that is suitable for large-scale applications[49]. Most covariance models, in particular the models of the Rfam families, are dominated by sequence information. At least in this regime, infernal is the most effective tool available. Phylogenetic distance, and hence decreasing sequence conservation, eventually limits applicability of homology search. It is possible in principle to include thermodynamic stability, either using the idea of thermodynamic matchers[50] or employing structural alignments[29]. It remains unclear, however, whether such techniques can substantially improve the sensitivity of homology search for distantly related species.

3.2. Feature-Based Gene Prediction

sRNApredict[51] uses typical features of prokaryotic sRNAs: elevated sequence conservation, putative promoter sequences, and Rho-independent terminator elements. TranstermHP, for instance, is used to predict Rho-independent terminators[52]. Its scoring function favours G/C-rich stem loops followed by a poly-T track. It is obviously extremely difficult to detect correct terminator elements in species with a high G/C-content and in those that use structural elements deviating from the canonical terminator structure. In order to increase sensitivity and specificity, sRNApredict focuses on intergenic regions and analyzes the co-occurrence of several of the above-mentioned features. While this strategy works quite well for well-characterized eubacterial clades, it is bound to fail in others. *Xanthomonas* and *Helicobacter*, for example, lack typical promoter sequences and distinct terminator hairpins[53,47].

3.3. Transcriptomics

Bacterial (and archaeal) transcriptomics can almost always be performed with a reference genome in place. This simplifies the work flow, which is basically composed of the following steps.

(1) Library preparation: Transcriptome analyses consist of "wet-lab" experiments and "try-lab" data evaluation. Both components greatly influence the final outcome and it is therefore recommended to design the experimental setup in a cooperative way, such that practical and theoretical issues are discussed at the very beginning. Selection of an appropriate sequencing platform, e.g., 454 or Illumina, and the enrichment or depletion of certain RNA classes are only two of many design decisions that depend on the research question. The actual experiments are performed and, depending on the sequencing platform and sequencing depth, several gigabytes of RNA transcript data are reported.

(2) Quality check: Sequencing machines typically output `FASTQ`-formatted files. This extended version of `FASTA` files is augmented by quality information for each called nucleotide along the sequence. `FastQC`[1] is commonly used to initially check and visualize the quality of the raw sequencing data. Software suites such as the `FASTX-Toolkit`[2] provide several tools to preprocess the raw sequencing reads by e.g. removal of the adapter and bar code sequences that have been attached during library preparation, or by filtering of low complexity reads. These steps can have a drastic influence on the mapping quality.

(3) Read mapping: A large number of software tools for read mapping has become available that differ widely in their algorithmic basis, memory consumption, speed, and versatility. Mapping strategies furthermore differ in their treatment of reads that map equally good to multiple genomic locations and in their handling of insertions and deletions[54,55,56,57,58]. It is therefore important to match the choice of mapping tool to the research question[59]. Once the mapping step is completed, mapping summary statistics help to verify whether all prior steps have been successful. Transcriptome studies that investigate prokaryotes usually assume that reads map without interruption ("split-free") and with near perfect sequence identity to the genome. This is, indeed, the case for the overhelming majority of the reads. There are, however, biological relevant exceptions that usually end up in the "sequencing trash bin". Examples include transcripts containing self-splicing introns in Eubacteria, as well as enzymatically spliced and circularized RNAs in Archaea. A recent study showed that such "atypical" transcript structures may be much more abundant than expected[60]. It remains, however, unclear to what extent rare transcripts of this type are biologically relevant, how many of them are technical artefacts and to what extent one detects true cellular RNAs that are nevertheless functionally irrelevant. Post-transcriptional modifications may furthermore lead to large local error rates[61].

(4) Transcript annotation and classification: The transcripts are then evaluated with respect to the genomic loci they have been mapped to. This covers in general a classification into protein-coding, non-coding and intergenic regions. For a typical prokaryotic genome, the non-coding portion is mainly comprised of reads that originate from the highly abundant tRNAs and rRNAs and from a few well-characterized house keeping genes such as tmRNA and 6S RNA. In most prokaryotes, only the open reading frames of protein-coding genes are annotated, while regulatory regions of mRNA transcripts, i.e., their UTRs (untranslated regions) are missing and the structure of polycistronic transcripts, i.e., transcripts that contain more than one gene, remains uncertain. Thereby the number of reads mapping to intergenic regions is overestimated due to this knowledge gap. The detection of polycistronic transcripts can be achieved by using a high sequencing depth close to saturation. The exact determination of transcriptional units is, however, challenging as gap-free expression cannot be found even for well-characterized cases such as the *cag* pathogenicity island of *H. pylori*[53]. Another difficult task is the precise mapping of the genomic positions where transcription is initiated. This challenge has been addressed by specific sequencing library preparation steps; the evaluation of the resulting read patterns is described in more detail in the next subsection on transcription start site (TSS) annotation. The determined TSS maps revealed an unexpected complexity of the transcription unit organization. Transcription is initiated as expected ahead of annotated genes and polycistronic transcripts but also internally and antisense to them and therefore almost everywhere along the genome. Upstream of the determined TSS, promoter sequence motifs are expected. Textbook knowledge describing two conserved elements, i.e., the -10 and -35 box, has been revised as these motifs are extremely variable between species. In *Xanthomonas* and *Helicobacter*, for instance, only traces of the -10 box are detectable but no distinct -35 box has been reported[53,47]. It seems to be a matter of fact that the current experimental

[1] `http://www.bioinformatics.babraham.ac.uk/projects/fastqc`

[2] `http://hannonlab.cshl.edu/fastx_toolkit`

setups enable the detection of TSS with species-specific housekeeping promoters, but alternative σ binding motifs are still hidden. The sequence between an annotated TSS and the start of a nearby downstream protein-coding gene gives rise to its 5' UTR. Again, surprising results such as a large number of leaderless transcripts, i.e., translation start and TSS are mapped to (almost) the same position, and 5' UTRs lacking Shine-Dalgarno sequence patterns have been reported[53,47]. Beside the possibility to gain new insights into protein-coding genes, most prokaryotic transcriptome studies are set up to detect novel non-coding RNA genes. These are typically identified by the analysis of read accumulations in intergenic regions or anti-sense to annotated genes. The existence of transcription units that might correspond to non-coding genes is verified by independent experiments such as northern blotting and their exact size is determined by RACE. A single study reveals dozens of novel RNA genes that need to be further characterized. Common tasks are the detection of homologous sequences, structural conservation analysis, evaluation of their coding potential and target prediction. For a detailed description of these evaluations, we refer to the Sections 3.1, 3.4 and 4, respectively.

TSS Annotation

In contrast to translation start sites that can be identified by well-established gene annotation strategies[62,63], surprisingly little is known about transcription start sites (TSS) in most bacteria. Even though a thorough TSS annotation can serve as valuable source of information to (i) understand the architecture of polycistronic transcripts, (ii) use it as a paramount hallmark for ncRNA gene annotation, and (iii) determine the extend of the 5'UTR, which often harbors regulatory elements such as riboswitches, RNA thermometer, and sRNA binding sites.

The first successfully applied methods to annotate TSS were primer extension[64] and RACE[65]. Both techniques aim to find the 5' end of partly characterized genes, but suffer from two major drawbacks. Firstly, with these techniques it is not possible to distinguish between 5' ends of an RNA formed by a transcription initiation event or by an RNA cleavage event, which often occurs in the course of RNA processing. Secondly, both techniques are difficult to scale up to a genome-wide high-throughput application. Therefore, two RNA-seq based methods for reliable annotation of TSS in bacterial genomes were developed recently[66,53]. Both methods exploit the phosphorylation pattern unique to primary TSS. Mono-nucleotides for transcription are provided to the RNA polymerase in the form of nucleotide triphosphates, which are broken down in the process of transcription elongation and the released energy is used to form a phosphodiester bond between the newly conjoined nucleosides. As a consequence, the first nucleotide still has a triphosphate attached to its 5' carbon atom. In contrast, if the phosphodiester bond of two consecutive nucleosides is broken by endonucleolytic cleavage, the remaining fragment is a 5'-phosphomonoester.

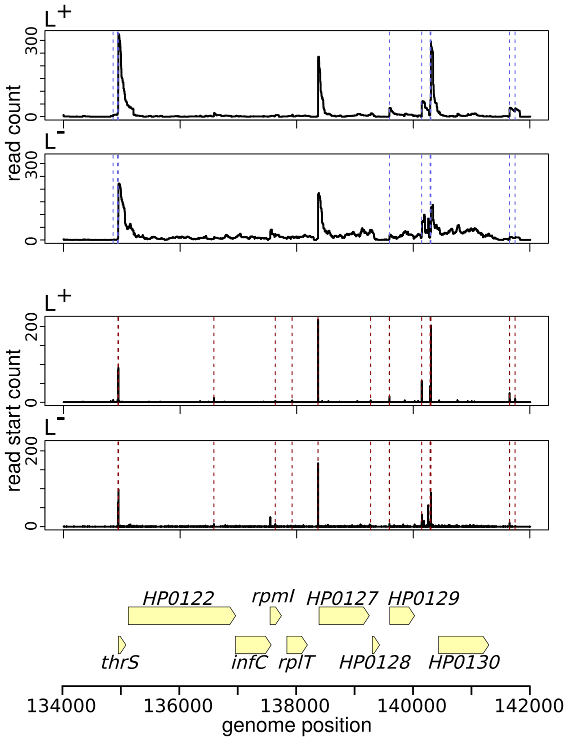

Figure 1: Comparison of automated TSS annotation from dRNA-seq data with `TTSpredator` and `TSSAR`. The upper plot pair shows the mapped read coverage in the treated (L^{+}) and untreated (L^{-}) library for an exemplary region from *H. pylori* dRNA-seq data[53]. Blue dashed lines indicate TSS annotated by `TTSpredator` (using default parameter). The middle plot pair shows essentially the same data, but only the read start coverage is plotted. This is how `TSSAR` looks at the data. Dashed red lines indicate TSS annotated by `TSSAR` (p-value cutoff of 10^{-4}). The bottom part shows the positions of the annotated genes in the considered region. The read coverage plots indicate that the data produced by dRNA-seq is more complex than it might appear from the method description. A refined data analysis is needed as simple global cut-off approches, whether for the difference or the ratio between the two libraries, cannot cope with the dynamics along different genomic regions.

In the method developed by Wurtzel et al.[66], the total RNA is treated with *Tobacco Acid Pyrophosphatase* (TAP), which removes the 5'-triphosphate and hence makes the RNA susceptible for the subsequent 5'-

sequencing-adapter ligation. The 3'-adapter is attached by a random primer. In contrast to a library which is not TAP-treated, reads associated with primary TSS are enriched in the TAP-treated library.

An alternative method[53] uses the *Terminator-5'-phosphate-dependent exonuclease* (TEX) to deplete the total RNA of fragments that are not protected from exonuclease degradation by a 5'-triphosphate. As a control, total RNA from the same extraction is processed the same way, but without the TEX treatment. Therefore, in the final anlysis step the differences between the treated (a.k.a. plus) library and the untreated (a.k.a. minus) library have to be screened position-wise for sites with a compelling enrichment of RNA-seq read starts in the plus versus the minus library. That is why this method was named differential RNA-seq (dRNA-seq).

The first applications of dRNA-seq were manually analysed by visualizing the reads and assessing the enrichment. Since such a screening is very time-consuming and tedious on genome-scale, and since it involves the subjectiv assessment of the analyzer, the results suffer from a certain lack of reproducibility and consistency. Therefore, soon after, the first statistical approaches to evaluate dRNA-seq data were proposed. Schmidtke et al.[47] modeled the density of read starts within the genome locally by applying a sliding window approach. Within each window, the distribution of read start counts per position are assumed to follow a Poisson distribution. As a consequence, the differences between the two libraries can be modeled by the Skellam distribution, which allows to calculate the probability to encounter the observed enrichment by chance.

Alternatively, global thresholds are applied to discriminate between significant read enrichment and background noise[70,71]. To gain specificity, the TSS calling is split into two steps. First, the relative read coverage increase in the treated library from position $i - 1$ to position i is evaluated. If this increase surpasses a defined threshold, the position is further evaluated whether the ratio of observed transcription initiation between treated and untreated library exceeds a defined threshold. If both tests are passed, the position is annotated as a TSS. The strenght of this method, as implemented in the program `TTSpredator`, lies in the ability to regard dRNA-seq data from different strains and/or growth conditions and dynamically adjust the thresholds if strong signals are observed in one sample. This circumvents a strict a priori threshold definition, which might be difficult to find for a new data set with different sequencing depth, genome size and TEX treatment efficiency.

The most recent development in automated TSS annotation from dRNA-seq data, `TSSAR`[72], picks up the idea from Schmidtke et al. to model the differences between the treated and untreated library with a Skellam distribution. However, to deduce the parameters from the underlying individual libraries, a zero-inflated Poisson distribution is used instead of a mere Poisson distribution. This allows to consider the region in focus as a mixture of transcribed and not transcribed segments, where the later are assumed to follow a Poisson distribution and the former to follow a uniform zero distribution. The parameters specifying the Skellam distribution are solely deduced from the read density in the transcribed region. The main advantage of `TSSAR` is the statistical sound analysis resulting in a robust enrichment p-value for each genomic position, which in turn leads to little dependency to a priori defined parameters that can greatly depend on the details of the experimental design and execution. Furthermore, `TSSAR` is provided as an easy-to-use web service, making its application rather convenient.

Similar to the eukaryotic research community, the understanding of prokaryotic genomes can benefit from shifting from the established protein-coding gene centered genome annotation to the incorporation of more information on transcripts, with all their diversity in function and architecture. With the recent developments both in wet-lab experiments and computational analysis that allow to characterize bacterial transcriptomes semi-automated in a high-throughput manner, a comprehensive transcript annotation becomes feasible. A comparison of `TSSpredator` and `TSSAR` is shown in Fig. 1.

3.4. Comparative Genomics

Non-coding RNAs are in many cases detectable by comparative genomics alone, i.e., without the benefit of either known homologs or expression data. `SIPHT`[73] makes use of invariant features of many bacterial genes. It identifies candidate loci based on sequence conservation in intergenic regions combined with predicted Rho-independent terminators (downstream) and predicted transcription factor binding sites (upstream). The software also evaluates homology with known sRNAs and cis-regulatory RNA elements. The tool is not directly applicable to some genera such as *Helicobacter*, which has a `A/T`-rich genome and thereby lacks recognizable terminator hairpins[53].

Stabilizing selection acting to preserve secondary structure elements imposes constraints on variations that become fixed in a population and hence are observable as differences between orthologous sequences from evolutionarily related organism. In particular, evolutionarily conserved base pairs admit only 6 of 16 pos-

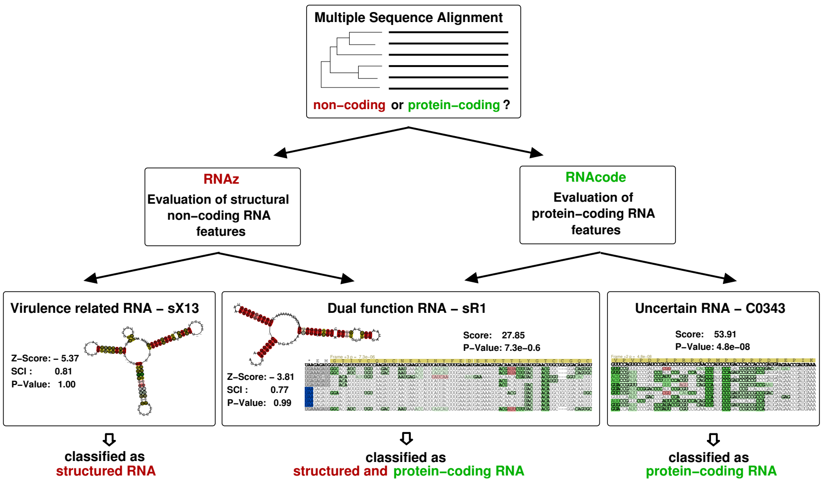

Figure 2: Evolutionary signals are used to classify multiple sequence alignments into non- or protein-coding. `RNAz` combines structural and thermodynamic descriptors and measures of sequence conservation to detect excess conservation of secondary structure, while `RNAcode` identifies increased conservation of putative ORFs compared to the observed sequence conservation of the nucleic acid sequences. Well-conserved structured RNAs such *Xanthomonas* sX13, which is involved in virulence specific gene expression and *hfq* mRNA regulation, can easily be identified[67] with `RNAz`. The *E. coli* transcript C0343, originally annotated as a small RNA, does not exhibit typical features of a structured RNA. Instead, `RNAcode` reveals a well-conserved short coding sequence[68]. Dual transcripts such as *B. subtilis* sR1[69] are detectable by both `RNAz` and `RNAcode`.

sible nucleotide pairs: `GC`, `CG`, `AU`, `UA`, `GU`, and `UG`. Computer simulations have indicated that RNA sequences still evolve in a drift-like manner even under very strong selection on their secondary structure[74,75] so that sequence patterns reflecting the structural constraints rapidly accumulate and become readily detectable already at 10% of sequence divergence.

`qrna`[76] investigates pair-wise alignments. The algorithm is based on stochastic context free grammars and estimates the posterior probabilities for an input alignment to be structured RNA, protein-coding, or neither. Its first application to *E. coli*[77] resulted in the prediction of several dozens of novel ncRNAs, many of which have been validated. Multiple sequence alignments convey much more information on substitution patterns than pairwise alignments but are also much harder to simulate as a detailed stochastic model as in `evofold`[78]. In `RNAz`[79], Fig. 2, we have therefore taken a different approach. Two lines of evidence inform about conservation of RNA structures: (i) structural similarity above the level expected from placing the differences at random positions[80], (ii) a lower free energy of folding than expected for the same sequence composition. Instead of an explicit stochastic model, `RNAz` uses machine learning to distinguish between true ncRNAs and decoys with the same dinucleotide content and the same gap pattern as the input alignments. The software is primarily designed for the large genomes of higher eukaryotes but has been employed successfully also for many prokaryotes[81,82,83,84]. It detects all types of conserved secondary structure elements, including *bona fide* sRNAs, riboswitches and RNA thermometers, as well as terminator hairpins. Since its initial publication several improvements have been introduced. In particular, `RNAz` 2.0[85] makes use of improved consensus structure prediction for assessing structural conservation[21], it explicitly accounts for dinucleotide distribution, and it has been retrained on a much larger training set including many prokaryotic RNAs. Nevertheless, `RNAz` still suffers from relatively large false discovery rates (FDR) and a limited accuracy in particular of the boundaries of its predicted structures. Reevaluating the `RNAz` predictions with structure-based alignment reliability scores computed by `LocARNA-P`[28] not only improves the boundary prediction by more than a factor of three but also halves the FDR.

A completely different comparative approach is taken by `NAPP`[86]. First it determines the phylogenetic distribution of conserved sequence elements as well as annotated protein-coding genes. Coherent phylogenetic distribution and co-occurrences with certain groups of proteins then indicate that conserved, un-annotated sequences may harbour sRNAs. An advantage of this approach is that the association with known proteins at least hints at potential functions of the candidate sRNA. A comparison of different computation approaches towards sRNA prediction can be found e.g. in ref.[86].

Discrimination between coding and non-coding regions poses technical as well as biological challenges not addressed by standard gene finders[87]. Ironically, authors working on non-coding RNAs repeatedly had to implement *ad hoc* solutions to detect coding regions. While longer protein-coding sequences are easily recognized by the absence of stop codons and characteristic, often species-specific patterns of codon usage, it is impossible to reliably detect short peptides of 20 amino acids or less in a single sequence. In complete analogy to RNA secondary structures, however, conservation of peptide sequences constrains the variation of the underlying nucleic acid sequence in characteristic ways. Most obviously, third codon positions are expected to be much more variable. `RNAcode`[68], Fig. 2, is based on this idea and evaluates for all six possible reading frames whether the amino acids obtained by translating a putative codon is more conserved than expected by the conservation at nucleic acid level. Translated into log-odds scores these estimates form the basis of a dynamic programming algorithm that identifies statistically significant conserved peptides in the alignment of nucleic acid sequences. The method was applied e.g. to identify very small peptides as well as annotation errors in *H. phylori*[53,88].

A particular difficulty is posed by transcripts that function both as sRNA by virtue of a conserved secondary structure and at the same time code for a conserved peptide. Well-known examples from the realm of prokaryotes is the *Staphylococcus aureus* RNAIII, which regulates target genes as sRNA and encodes the 26 amino acid sequence of delta-haemolysin[89], and the *Bacillus* SR1 RNA involved in the regulation of arginine catabolism[69]. The detection of such cases in genome-wide surveys remains difficult although software for similar tasks has become available. In particular `RNAdecoder`[90] searches for conserved RNA structure within DNA regions known to be protein-coding; it suffers from very high FDRs, however[91]. The intersection of `RNAz` and `RNAcode` predictions can provide at least plausible candidates but is certainly not ideal either. To the best of our knowledge no systematic survey for dual RNAs has be conducted in prokaryotes so far.

3.5. Estimation of RNA Families and Classes

The Rfam database divides ncRNAs according to inherent functional, structural, or compositional similarities in more than 2200 different RNA families[40]. At a higher level, an *RNA class*[92] further groups together ncRNAs whose members have no clear homology at the sequence level, and presumably do not derive from a common ancestor, but still share common structural properties as a consequence of functional analogy. Prominent examples are microRNAs (miRNAs) and the two distinct classes of snoRNAs (box H/ACA and box C/D).

Current methods for the *de novo* annotation of ncRNAs rely on unsupervised techniques, such as clustering, to group similar RNAs and subsequent computation of the consensus structure. Using methods implemented in tools like `RNAz`[79] and `EvoFold`[78], further characteristics that are indicative of functional ncRNA genes are evaluated.

In this framework, the initial clustering phase is a crucial step and in order to be successful it requires the specification of an appropriate distance or similarity notion that can characterize the functional properties of RNA sequences. The distance measures of course depend on the level of information available and ultimately on the representation used to encode the RNA molecules. These representations can be based on (i) the nucleotide sequence, (ii) the connectivity graph of base pairing interactions, or (ii) the full three-dimensional conformation. The third option is not yet viable as there is a lack of both experimental techniques to determine 3D conformations of functional RNAs in a large scale setting (i.e., for machine learning approaches), and of efficient, and sufficiently accurate, modeling techniques to compute these conformations.

Frequently only sequence information is used since it is directly available from sequencing experiments, of relatively low noise, and it can be manipulated efficiently and with ease by computers[93,94]. By construction, any pure sequence-based approach is restricted to RNA families and must fail to detect functional similarity in case of low sequence identity. Indeed, family assignments of structured RNAs obtained from sequence alignments are often wrong when pairwise sequence identities drops below 60%[92]. Much lower similarity levels are quite common within a single RNA class. There is therefore a pressing need for similarity and distance notions that efficiently take into account both sequence and structure.

One possible solution is to do structure prediction simultaneous with the construction of alignments[27,24] as described in Section 2. This approach was successfully used to classify all known CRISPR repeats[95]. However, these alignment-based methods do not scale to efficiently cluster hundred of thousands of candidate ncRNAs predicted e.g. by `RNAz` screens.

With `GraphClust`[96] a very different approach has become available. It avoids the alignment phase and the explicit computation of a distance matrix altogether. At the same time it is not restricted to a single structural hypothesis. In order to deal with structural alternatives, abstract shape analysis[97] is used to summarize the ensemble of predicted structures. It provides an *a priori* classification of structures and allows the efficient retrieval of a single representative secondary structure per class, so that each sequence is represented by a small set of sufficiently different secondary structures. Each structure is then interpreted as a labeled graph from

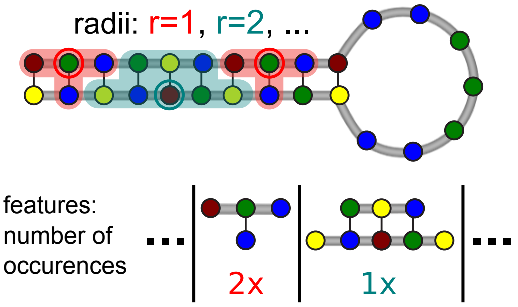

Figure 3: Features describing a secondary structure graph. Each graph is described by the set of all neighborhood subgraphs up to a maximal radius *r* around a certain nucleotide.

which structural features defined as small localized subgraphs are extracted as outlined in Fig. 3. The resulting sparse feature vectors for each structure amount to a direct generalization of the well-known *k*-mer similarity from strings to labeled graphs[98], which could be used for clustering.

For large datasets (i.e., $>10^4$ sequences) one cannot afford the quadratic complexity of clustering algorithms that rely on a pairwise distance or similarity information. Instead, `GraphClust` formulates the clustering problem in terms of approximate nearest neighbor queries which can be answered with a sub-linear complexity using locality sensitive hashing[99]. The similarity of the *k*-nearest neighbors can then be used to estimate how compact or dense each neighborhood is within the set of feature vectors so that the most compact non-overlapping neighborhoods can be selected as candidate clusters.

Each of these candidate clusters is then refined using alignment techniques designed to discard incompatible RNA sequences. A corresponding covariance model is employed to scan the original dataset for similar sequences that were missed by graph-based preclustering. The entire procedure is then iterated on the remaining instances producing in each round a user-defined number of clusters that can later be merged to decrease the final cluster fragmentation.

`GraphClust` was successfully applied to cluster bacterial ncRNAs. Using a benchmark set of 363 ncRNAs, `GraphClust` detected 43 high-quality clusters representing 38 families[96]. In this benchmark, additional genomic context was added to simulate the application scenario of unknown precise transcript boundaries. The quality of clustering (measured with the F-measure or with the Rand index) was higher then the state-of-the-art clustering using `LocARNA`. Thus, `GraphClust` can successfully determine RNA classes for bacterial ncRNAs, even when the precise transcript is unknown.

4. RNA-RNA Interactions

4.1. Models for Predicting sRNA–mRNA Interactions

The rise of high-throughput methods, first tiling arrays and now RNA-seq, to characterize transcriptomes had led to an explosion in the number of identified sRNAs in prokaryotes; more than hundred sRNAs have been reported in most species (e.g.[100,101,102,103]). Most sRNAs studied to date form base pair interactions with mRNAs to post-transcriptionally regulate their targets' translation and stability[104]. The functional characterisation of novel sRNAs thus involves identification of their interaction partners together with the precise interaction sites. A promising strategy to cope with the steadily increasing number of discovered but uncharacterised sRNAs is computational prediction of candidate sRNA targets, followed by experimental verification using transcriptomics and proteomics approaches.

Computational methods for predicting RNA–RNA interactions fall into four main classes. The following section gives an overview of the available methods and tools with an emphasis on sRNA–mRNA interaction prediction (previously also reviewed in refs.[105,106]). Table 1 summarises web-based applications designed for genome-wide sRNA target predictions.

The first class of methods evaluates the stability of the duplex formed between two RNA molecules aiming to find the loci in both partners that yield the energetically most favourable hybridisation. Only base pairs between the two RNAs are evaluated, while their intramolecular structure is ignored. The most popular tools of this type are `RNAhybrid`[114], `RNAduplex` and `RNAplex`[115], and `DINAMelt`[116,117]. Methods of this class are primarily tailored for predicting potential binding sites of short RNAs (like eukaryotic miRNAs) in large target RNAs as they tend to maximise the hybridisation length. The prediction is based on a modified version of the secondary structure prediction algorithm of ref.[3] that omits multiloops. A simplified loop energy model was introduced by `RNAplex`. This tool also allows to favour shorter interactions by per-nucleotide penalties. The web server `TargetRNA`[118,119] was specifically designed for the prediction of bacterial sRNA targets; it provides two scoring schemes: (i) scoring of individual base pairs by a local alignment-like algorithm[120] or (ii) duplex mfe similar to `RNAhybrid`. Recently, its successor `TargetRNA2` was released (unpublished).

Methods of the second class determine a joint secondary structure of two RNAs, i.e., a common structure

Name	Features for target prediction			Classifier	Functional enrichment	URL of web server	References
	Conservation	Accessibility	Seed region				
`CopraRNA`	X	X	X	-	X	`http://rna.informatik.uni-freiburg.de/CopraRNA`	107
`IntaRNA`	-	X	X	-	X	`http://rna.informatik.uni-freiburg.de/IntaRNA`	108,109
`RNApredator`	-	X	-	-	X	`http://rna.tbi.univie.ac.at/RNApredator`	110,111
`sRNATarget`	-	-	X	X	-	`http://ccb.bmi.ac.cn/srnatarget`	112
`sTarPicker`	-	X	X	X	-	`http://ccb.bmi.ac.cn/starpicker`	113
`TargetRNA2`	X	X	X	-	-	`http://snowwhite.wellesley.edu/targetRNA`	

Table 1: Web server for genome-scale prediction of sRNA target genes. All web server are based on computational methods that score the sRNA–target interaction by their hybridisation energy and by additional features as indicated in the table. Some server directly allow for functional enrichment analysis of the highest-ranking target predictions.

including both intra- and intermolecular base pairs. The two input RNA sequences are concatenated and then folded by an RNA folding algorithm such as Zuker's algorithm[3], which is extended to handle the loop containing the concatenation point energetically as an external loop. Tools implementing this idea are, for example, `PairFold`[121] and `RNAcofold`[122]. The `sRNATarget` web server[112,123] computes the mfe structure of the concatenated sequence to derive interaction features such as length-normalised free energy, seed match length and `A/U`-content in single-stranded regions. A naive Bayes classifier based on these features is then applied to discriminate sRNA–mRNA interactions from non-interacting sRNAs and mRNAs. The main disadvantage of all concatenation-based approaches is their restriction on the allowed interaction types. The underlying RNA folding algorithm can only predict pseudoknot-free secondary structures, although many interaction sites are actually located in loop regions[124]. Interactions between two stem loops (loop–loop interactions) represent a pseudoknot in the context of the concatenated sequences and, therefore, cannot be predicted by these approaches.

The third class comprises interaction prediction methods that model the competition between formation of duplex and intramolecular base pairs by the structural accessibility of the interaction sites. This strategy is supported by two systematic studies which showed that functional interaction sites are typically well-accessible in both sRNAs and their target mRNAs[125,126]. The tools `IntaRNA`[109] and `RNAup`[127,128] calculate the thermodynamics of RNA–RNA interactions as sum of two energy contributions: (i) the energy required to make the sRNA and target interaction sites accessible, which is calculated from the ensemble of all secondary structures, and (ii) the hybridisation energy of the two interacting subsequences. `IntaRNA` additionally incorporates seed regions, i.e., regions of (nearly) perfect sequence complementarity, that are thought to initiate interaction formation. The `IntaRNA` web server[108] allows for genome-scale sRNA target predictions followed by functional enrichment analysis of top target predictions and visualization of putative interaction regions. `RNAplex` optionally approximates interaction site accessibility by position-specific per-nucleotide penalties[111]. An sRNA target prediction web server on top of `RNAplex` is implemented by the software `RNApredator`[110]. The web server `sTarPicker` combines ideas from accessibility-based and concatenation-based approaches[113]. Putative seed interactions are extended by computing a joint secondary structure of sRNA and mRNA. The predictions are then classified into true and false interaction predictions based on the interaction features `A/U`-content, hybridisation energy, accessibility and seed length. All methods represented by this class can predict complex interactions like loop–loop interactions, but interactions are restricted to one locus. For RNA–RNA interactions involving two or more interaction sites as, e.g., OxyS–*fhlA*[129] and RNAIII–*rot*[89], only one of the interaction sites can be predicted. Whether formation of interactions at multiple loci is a common principle and frequently required for regulation by sRNAs *in vivo* is still an open question. The sRNA RNAIII, for example, binds its target *coa* in *Staphylococcus aureus* both via an imperfect duplex and a loop–loop interaction, but the former interaction alone is sufficient for *in vivo* repression[130].

Several tools of the third class have been successfully applied to identify sRNA targets in various prokaryotic species. `IntaRNA`, for example, aided in finding that the cyanobacterial sRNA Yfr1 inhibits translation of two outer membrane proteins[131] and that the sRNA PhrS stimulates translation of the quorum-sensing regulator *pqsR* in *Pseudomonas*[132]. But sRNA–mRNA interactions are not restricted to the bacterial domain of life. Jäger et al.[133], for example, showed by a combination of computational and experimental approaches that the archaeal sRNA$_{162}$ targets both a *cis*- and a *trans*-encoded mRNA via two distinct domains.

Methods of the final class can predict more complex joint secondary structures and also allow for multiple interaction sites. The IRIS tool[134] introduced a model that maximises the number of base pairs. Alkan et al.[135] then presented a more realistic energy model. The type of joint structures considered in this study were the basis for several subsequent approaches to predict mfe structures[136,137,138], to compute the partition function of joint secondary structures[139,140] and to sample joint secondary structures[141]. All these algorithms have a high time and space complexity, in practise precluding genome-wide application. Except for IRIS, all methods of this class are also not able to handle pseudoknotted structures or crossing interactions. Consequently, they still cannot predict instances like the two loop–loop interactions between RNAIII and *rot* in *Staphylococcus aureus* as these constitute a crossing interaction[89].

4.2. Comparative sRNA Target Prediction

Genome-scale prediction of sRNA target genes is a computationally challenging task and all methods presented above suffer from a high false positive rate. Starting from the observation that the target binding site in the sRNA is marked by high sequence conservation across related species[125,126], comparative target prediction for conserved sRNAs appears to be a promising strategy to reduce the number of false positive predictions.

PETcofold was the first comparative method for the prediction of RNA–RNA interactions and joint secondary structures[142,143,144]. Using two multiple alignments of RNA sequences as input, PETcofold predicts conserved RNA–RNA interactions and RNA structures taking into account covariance information arising from compensatory base pair exchanges. Such an alignment-based strategy will predominantly report duplexes in which the interaction base pairing is conserved across species. Its applicability is, therefore, limited to a subclass of interactions that exhibit broad evolutionary conservation. The same constraint applies to other comparative joint secondary structures prediction approaches such as ripalign[145].

Interactions with conserved base pairing pattern cover only a subset of all observed interactions; conservation of target complementarity can range from marginal to full conservation even for different targets of the same sRNA[126]. This observation is particularly challenging for alignment-based approaches as it is not known *a priori* whether the interaction between a specific sRNA and mRNA is well conserved or not. CopraRNA introduced a very promising alternative strategy overcoming fixed input sequence alignments[107].

As for other comparative approaches, CopraRNA's main idea is to combine the target prediction in several species. But in contrast to the above-mentioned approaches, CopraRNA does neither enforce conservation of the interaction site nor of the interaction pattern. Rather, it performs target prediction in each organism independently and then combines the evidence for all these predictions (see Fig. 4). The basic assumption is that only the target regulation by the sRNA is required to be conserved, but the specific base pairing pattern can be variable and the interaction site might have even been shifted, especially in the mRNA. For a functional interaction, it is often sufficient to have a binding in proximity to the ribosomal binding site without the necessity of a fixed position.

In order to combine the single evidences of an interaction from each organism, one could naively use the average of all calculated scores. This approach has, however, two caveats: (i) the scores are not normalized and depend, e.g., on the G/C-content of the organism, and (ii) closely related species are likely to have similar scores due to their similarity in sequence composition. Concerning the first point, a way to normalize the score is to use *p*-values instead of raw scores. Since each sRNA has typically only few functional interactions (for example a total of 21 direct targets has previously been reported for the well-characterized sRNA GcvB[146]), one can use the score distribution of all genome-wide predicted interactions for a given sRNA in one organism as background to calculate the *p*-values. For the second point, one first has to determine how *p*-values from different organism can be combined. Albeit intuitively a good solution, the product of *p*-values does not constitute a *p*-value anymore as it is not uniform across the background. For that purpose, one has to use a transformation. In CopraRNA the inverse normal method of Hartung[147] was used since it additionally allows to weight the *p*-values, thus correcting for the evolutionary distance of the species.

5. Open Questions

Many questions and computational problems remain open. Although experimental and computational methods are now in place to identify transcription start sites,

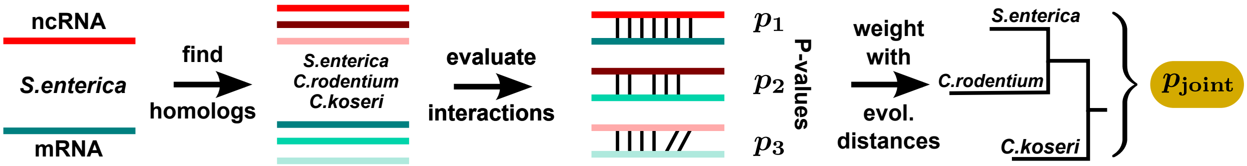

Figure 4: Comparative prediction of sRNA targets as implemented in the `CopraRNA` pipeline. For a given pair of sRNA and mRNA sequences, the associated homologs are selected. In the next step, the best interaction in each species is determined and scored by its p-value. Finally, all species-specific p-values are combined into a single joint p-value while taking the evolutionary distances into account.

the corresponding termination sites still cannot be determined reliably, in particular when they are not associated with Rho-independent terminator structures. Even less is know about other forms of RNA processing such as cleavage and editing: Where does it occur? How do processing patterns look like in RNA-seq data?

Although it has become clear that sRNAs are abundant in most prokaryotes, we still lack a clear picture of their phylogenetic distribution. In particular distant homologies have remained largely unexplored. The abundance of pseudoknots and complex interaction structures is still unknown, at least in part due to the high computational cost but also the limited reliability of prediction algorithms in particular when applied to single sequences. The RNA chaperone Hfq facilitates pairing of sRNA and target mRNA in diverse bacterial lineages[148]. The still unknown rules governing the binding of Hfq to specific sRNAs in what appears to be a highly dynamic molecular mechanism[149] are likely to provide a dramatic improvement for predicting functional sRNA–mRNA interactions and thus for the functional annotation of sRNAs. Eventually, the goal would be to complete the whole bacterial gene regulatory network. Due to their influence on RNA-RNA interaction, this must also include the determination of proteins-RNA interactions. Furthermore, not only the sRNA targets, but also the transcriptional regulation of the sRNA itself has to be understood. This would allow to apply the systems biological tool box and explore the dynamics of the full gene regulatory network, which are most likely to be altered by the introduction of sRNAs into the network.

Recent time has seen the development of plethora of high-throughput approaches like CLIP-seq to investigate this network. It can also be seen that these new experimental techniques require also a constant development of appropriate bioinformatics tools. The constant mutual development of experimental techniques and associated bioinformatics method was well established in the Priority Program SPP 1258, which thus can serve as a blueprint for similar collaborative projects.

Acknowledgements

This work was supported by the DFG under the auspices of Priority Program SPP 1258: Sensory and regulatory RNAs in Prokaryotes, grant nos. BA 2168/2 and STA 850/7.

References

[1] Turner DH and Mathews DH. NNDB: the nearest neighbor parameter database for predicting stability of nucleic acid secondary structure. Nucleic Acids Res., 2010, 38:280–282.

[2] Waterman MS. Secondary structure of single - stranded nucleic acids. Adv. Math. Suppl. Studies, 1978, 1:167–212. Studies on foundations and combinatorics.

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

[4] McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers, 1990, 29:1105–1119.

[5] Tacker M, Stadler PF, Bornberg-Bauer EG, Hofacker IL, and Schuster P. Algorithm independent properties of RNA structure prediction. Eur. Biophy. J., 1996, 25:115–130.

[6] Ding Y and Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res., 2003, 31:7280–7301.

[7] Zuker M. `Mfold` web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res., 2003, 31:3406–3415.

[8] Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, and Schuster P. Fast folding and comparison of RNA secondary structures. Monatsh. Chem., 1994, 125:167–188.

[9] Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, and Hofacker IL. ViennaRNA Package 2.0. Alg. Mol. Biol., 2011, 6:26.

[10] Liu B, Mathews D, and Turner DH. RNA pseudoknots: folding and finding. F1000 Biol Rep., 2010, 2:8.

[11] Möhl M, Will S, and Backofen R. Fixed parameter tractable alignment of RNA structures including arbitrary pseudoknots. In Proceedings of the 19th Annual Symposium on Combinatorial Pattern Matching (CPM 2008), LNCS, pages 69–81. Springer-Verlag, 2008.

[12] Bindewald E, Kluth T, and Shapiro BA. CyloFold: secondary structure prediction including pseudoknots. Nucleic Acids Res., 2010, 38:W368–W372.

[13] Möhl M, Will S, and Backofen R. Lifting prediction to alignment of RNA pseudoknots. J Comput Biol, 2010, 17(3):429–42.

[14] Bon M and Orland H. TT2NE: a novel algorithm to predict RNA secondary structures with pseudoknots. Nucleic Acids Res, 2011, 39:e93.

[15] Reidys CM, Huang FWD, Andersen JE, Penner RC, Stadler PF, and Nebel ME. Topology and prediction of RNA pseudoknots. Bioinformatics, 2011, 27:1076–1085. Addendum in: Bioinformatics 28:300 (2012).

[16] Möhl M, Salari R, Will S, Backofen R, and Sahinalp SC. Sparsification of RNA structure prediction including pseudoknots. Algorithms Mol Biol, 2010, 5(1):39.

[17] Leontis NB and Westhof E. Geometric nomenclature and classification of rna base pairs. RNA, 2001, 7:499–512.

[18] Höner zu Siederdissen C, Berhart SH, Stadler PF, and Hofacker IL. A folding algorithm for extended RNA secondary structures. Bioinformatics, 2011, 27:i129–i137. ISMB.

[19] Lange SJ, Maticzka D, Möhl M, Gagnon JN, Brown CM, and Backofen R. Global or local? Predicting secondary structure and accessibility in mRNAs. Nucleic Acids Res, 2012, 40(12):5215–26.

[20] Hofacker IL, Fekete M, and Stadler PF. Secondary structure prediction for aligned RNA sequences. J Mol Biol, 2002, 319(5):1059–66.

[21] Bernhart SH, Hofacker IL, Will S, Gruber AR, and Stadler PF. `RNAalifold`: improved consensus structure prediction for RNA alignments. BMC Bioinformatics, 2008, 9:474.

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

[23] Sankoff D. Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. SIAM J. Appl. Math., 1985, 45:810–825.

[24] Torarinsson E, Havgaard JH, and Gorodkin J. Multiple structural alignment and clustering of RNA sequences. Bioinformatics, 2007, 23(8):926–32.

[25] Harmanci AO, Sharma G, and Mathews DH. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. BMC Bioinformatics, 2007, 8:130.

[26] Backofen R and Will S. Local sequence-structure motifs in RNA. Journal of Bioinformatics and Computational Biology (JBCB), 2004, 2(4):681–698.

[27] Will S, Reiche K, Hofacker IL, Stadler PF, and Backofen R. Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol, 2007, 3(4):e65.

[28] Will S, Joshi T, Hofacker IL, Stadler PF, and Backofen R. LocARNA-P: Accurate boundary prediction and improved detection of structured RNAs for genome-wide screens. RNA, 2012, 18:900–914.

[29] Will S, Siebauer MF, Heyne S, Engelhardt J, Stadler PF, Reiche K, and Backofen R. `LocARNAscan`: Incorporating thermodynamic stability in sequence and structure-based RNA homology search. Alg. Mol. Biol., 2013, 8:14.

[30] Will S, Schmiedl C, Miladi M, Möhl M, and Backofen R. SPARSE: Quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. In Deng M, Jiang R, Sun F, and Zhang X, eds., Proceedings of the 17th International Conference on Research in Computational Molecular Biology (RECOMB 2013), volume 7821 of Lect. Notes Comp. Sci., pages 289–290, Berlin, Heidelberg, 2013. Springer.

[31] Reeder J and Giegerich R. RNA secondary structure analysis using the RNAshapes package. Curr Protoc Bioinformatics, 2009, 26:12.8.1–12.8.17.

[32] Low JT and Weeks KM. SHAPE-directed RNA secondary structure prediction. Methods, 2010, 52:150–158.

[33] Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, Chang HY, and Segal E. Genome-wide measurement of RNA secondary structure in yeast. Nature, 2010, 467:103–107.

[34] Deigan KE, Li TW, Mathews DH, and Weeks KM. Accurate SHAPE-directed RNA structure determination. Proc Natl Acad Sci USA, 2009, 106:97–102.

[35] Zarringhalam K, Meyer MM, Dotu I, Chuang JH, and Clote P. Integrating chemical footprinting data into RNA secondary structure prediction. PLoS One, 2012, 7:e45160.

[36] Washietl S, Hofacker IL, Stadler PF, and Kellis M. RNA folding with soft constraints: Reconciliation of probing data and thermodynamic secondary structure prediction. Nucleic Acids Res., 2012, 40:4261–4272.

[37] Ouyang Z, Snyder MP, and Chang HY. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. Genome Res., 2013, 23:377–387.

[38] Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, and Weeks KM. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. Proc Natl Acad Sci USA, 2013, 110:5498–5503.

[39] Consortium AFB, Backofen R, Bernhart SH, Flamm C, Fried C, Fritzsch G, Hackermüller J, Hertel J, Hofacker IL, Missal K, et al. RNAs everywhere: genome-wide annotation of structured RNAs. J Exp Zoolog B Mol Dev Evol, 2007, 308B(1):1–25.

[40] Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, and Bateman A. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res, 2013, 41(Database issue):D226–32.

[41] Gardner PP and Gardner AG. A home for RNA families at RNA Biology. RNA Biology, 2009, 6:2–4.

[42] Gierga G, Voss B, and Hess WR. The Yfr2 ncrna family, a group of abundant RNA molecules widely conserved in cyanobacteria. RNA Biology, 2009, 6:222–227.

[43] Findeiß S, Schmidtke C, Stadler PF, and Bonas U. A novel family of plasmid-transferred anti-sense ncRNAs. RNA Biology, 2010, 7:120–124.

[44] del Val C, Romero-Zaliz R, Torres-Quesada O, Peregrina A, Toro N, and Jiménez-Zurdo JI. A survey of sRNA families in α-proteobacteria. RNA Biology, 2012, 9:119–129.

[45] Steif A and Meyer IM. The hok mRNA family. RNA Biology, 2012, 9:1399–1404.

[46] Hertel J, de Jong D, Marz M, Rose D, Tafer H, Tanzer A, Schierwarter B, and Stadler PF. Non-coding rna annotation of the genome of trichoplax adhaerens. Nucleic Acids Research, 2009, 37:1602–1615.

[47] Schmidtke C, Findeiß S, Sharma CM, Kuhfuß J, Hoffmann S, Vogel J, Stadler PF, and Bonas U. Genome-wide transcriptome analysis of the plant pathogen *Xanthomonas* identifies sRNAs with putative virulence functions. Nucleic Acids Res., 2012, 40:2020–2031.

[48] Eddy SR and Durbin R. RNA sequence analysis using covariance models. Nucleic Acids Res., 1994, 22:2079–2088.

[49] Nawrocki EP and Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, 2013, 29:2933–2935.

[50] Höchsmann T, Höchsmann M, and Giegerich R. Thermodynamic matchers: strengthening the significance of RNA folding energies. Comput Syst Bioinformatics Conf, 2006, pages 111–121.

[51] Livny J, Fogel MA, Davis BM, and Waldor MK. sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. Nucleic Acids Res., 2005, 33:4096–4105.

[52] Kingsford CL, Ayanbule K, and Salzberg SL. Rapid, accurate, computational discovery of Rho-independent transcription terminators illumniates their relationship to DNA uptake. Genome Biology, 2007, 8:R22.

[53] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiß S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt RR,

et al. The primary transcriptome of the major human pathogen *Helicobacter pylori*. Nature, 2010, 464:250–255.

[54] Trapnell C and Salzberg SL. How to map billions of short reads onto genomes. Nat Biotechnol, 2009, 27:455–457.

[55] Ruffalo M, LaFramboise T, and Koyutürk M. Comparative analysis of algorithms for next-generation sequencing read alignment. Bioinformatics, 2011, 27:2790–2796.

[56] Schbath S, Martin V, Zytnicki M, Fayolle J, Loux V, and Gibrat J. Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. J Comput Biol, 2012, 19:796–813.

[57] Hatem A, Bozdağ D, Toland AE, and Çatalyürek V. Benchmarking short sequence mapping tools. BMC Bioinformatics, 2013, 14:184.

[58] Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, RGASP Consortium, Rätsch G, Goldman N, Hubbard TJ, Harrow J, et al. Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods, 2013, 10:1185–1191.

[59] Caboche S, Audebert C, Lemoine Y, and Hot D. Comparison of mapping algorithms used in high-throughput sequencing: application to Ion Torrent data. BMC Genomics, 2014. in revision.

[60] Doose G, Alexis M, Kirsch R, Findeiß S, Langenberger D, Machné R, Mörl M, Hoffmann S, and Stadler PF. Mapping the RNA-Seq trash bin: Unusual transcripts in prokaryotic transcriptome sequencing data. RNA Biol, 2013, 10:1204–1210.

[61] Findeiß S, Langenberger D, Stadler PF, and Hoffmann S. Traces of post-transcriptional RNA modifications in deep sequencing data. Biological Chemistry, 2011, 392.

[62] Meyer F, Goesmann A, McHardy AC, Bartels D, Bekel T, Clausen J, Kalinowski J, Linke B, Rupp O, Giegerich R, et al. GenDB – an open source genome annotation system for prokaryote genomes. Nucleic Acids Research, 2003, 31(8):2187–2195.

[63] Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Cruveiller S, Lajus A, Pascal G, Scarpelli C, and Medigue C. MaGe: a microbial genome annotation system supported by synteny results. Nucleic acids research, 2006, 34(1):53–65.

[64] Duchêne M, Schweizer A, Lottspeich F, Krauss G, Marget M, Vogel K, Von Specht B, and Domdey H. Sequence and transcriptional start site of the *Pseudomonas aeruginosa* outer membrane porin protein F gene. Journal of bacteriology, 1988, 170(1):155–162.

[65] Frohman MA, Dush MK, and Martin GR. Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proceedings of the National Academy of Sciences, 1988, 85(23):8998–9002.

[66] Wurtzel O, Sapra R, Chen F, Zhu Y, Simmons BA, and Sorek R. A single-base resolution map of an archaeal transcriptome. Genome research, 2010, 20(1):133–141.

[67] Schmidtke C, Abendroth U, Brock J, Serrania J, Becker A, and Bonas U. Small RNA sX13: a multifaceted regulator of virulence in the plant pathogen Xanthomonas. Plos Pathogens, 2013, 9.

[68] Washietl S, Findeiß S, Müller S, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, and Goldman N. RNAcode: robust prediction of protein coding regions in comparative genomics data. RNA, 2011, 17:578–594.

[69] Gimpel M, Preis H, Barth E, Gramzow L, and Brantl S. SR1 – a small RNA with two remarkably conserved functions. Nucleic Acids Res., 2012, 40:11659–11672.

[70] Herbig A, Sharma C, and Nieselt K. Automated transcription start site prediction for comparative transcriptomics using the supergenome. EMBnet. journal, 2013, 19(A):pp–19.

[71] Dugar G, Herbig A, Förstner KU, Heidrich N, Reinhardt R, Nieselt K, and Sharma CM. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple *Campylobacter jejuni* isolates. PLoS genetics, 2013, 9(5):e1003495.

[72] Amman F, Wolfinger MT, Lorenz R, Hofacker IL, Stadler PF, and Findeiß S. TSSAR: TSS annotation regime for dRNA-seq data. BMC Bioinformatics, submitted.

[73] Livny J, Teonadi H, Livny M, and Waldor MK. High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. PLoS One, 2008, 3:e3197.

[74] Schuster P, Fontana W, Stadler PF, and Hofacker IL. From sequences to shapes and back: A case study in RNA secondary structures. Proc. Roy. Soc. Lond. B, 1994, 255:279–284.

[75] Huynen MA, Stadler PF, and Fontana W. Smoothness within ruggedness: the role of neutrality in adaptation. Proc. Natl. Acad. Sci. (USA), 1996, 93:397–401.

[76] Rivas E and Eddy SR. Noncoding RNA gene detection using comparative sequence analysis. BMC Bioinformatics, 2001, 2:8.

[77] Rivas E, Klein RJ, Jones TA, and Eddy SR. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. Curr. Biol., 2001, 11:1369–1373.

[78] Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, and Haussler D. Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. PLoS Comput Biol, 2006, 2(4):e33.

[79] Washietl S, Hofacker IL, and Stadler PF. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA, 2005, 102(7):2454–9.

[80] Gruber AR, Bernhart S, Hofacker IL, and Washietl S. Strategies for measuring evolutionary conservation of RNA secondary structures. BMC Bioinformatics, 2008, 9:122.

[81] Sonnleitner E, Sorger-Domenigg T, Madej MJ, Findeiß S, Hackermüller J, Hüttenhofer A, Stadler PF, Bläsi U, and Moll I. Detection of small non-coding RNAs in *Pseudomonas aeruginosa* by RNomics and structure-based bioinformatics tools. Microbiology, 2008, 154:3175–3187.

[82] Schilling D, Findeiss S, Richter AS, Taylor JA, and Gerischer U. The small RNA Aar in Acinetobacter baylyi: a putative regulator of amino acid metabolism. Arch Microbiol, 2010, 192:691–702.

[83] del Val C, Rivas E, Torres-Quesada O, Toro N, and Jimnez-Zurdo JI. Identification of differentially expressed small non-coding RNAs in the legume endosymbiont Sinorhizobium meliloti by comparative genomics. Mol Microbiol, 2007, 66:1080–1091.

[84] Hot D, Slupek S, Wulbrecht B, D'Hondt A, Hubans C, Antoine R, Locht C, and Lemoine Y. Detection of small RNAs in Bordetella pertussis and identification of a novel repeated genetic element. BMC Genomics, 2011, 12:207.

[85] Gruber AR, Findeiß S, Washietl S, Hofacker IL, and Stadler PF. `RNAz 2.0`: improved noncoding RNA detection. Pac. Symp. Biocomput., 2010, 15:69–79.

[86] Ott A, Idali A, Marchais A, and Gautheret D. NAPP: the nucleic acid phylogenetic profile database. Nucleic Acids Res., 2012, 40:D205–D209.

[87] Dinger ME, Pang KC, Mercer TR, and Mattick JS. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. PLoS Comput Biol, Nov 2008, 4:e1000176.

[88] Müller SA, Findeiß S, Pernitzsch SR, Stadler PF, Hofacker IL, Sharma CM, von Bergen M, and Kalkhof S. Proteogenomic analysis of the *Helicobacter pylori* strain 26695 genome. J. Proteomics, 2013, 86:27–42.

[89] Boisset S, Geissmann T, Huntzinger E, Fechter P, Bendridi N,

Possedko M, Chevalier C, Helfer AC, Benito Y, Jacquier A, et al. *Staphylococcus aureus* RNAIII coordinately represses the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism. Genes Dev, 2007, 21(11):1353–66.

[90] Pedersen JS, Meyer IM, Forsberg R, Simmonds P, and Hein J. A comparative method for finding and folding RNA secondary structures within protein-coding regions. Nucleic Acids Res., 2004, 32:4925–4936.

[91] Findeiß S, Engelhardt J, Prohaska SP, and Stadler PF. Protein-coding structured RNAs: A computational survey of conserved RNA secondary structures overlapping coding regions in drosophilids. Biochimie, 2011, 93:2019–2023.

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

[93] Shi Y, Tyson GW, and DeLong EF. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. Nature, 2009, 459(7244):266–9.

[94] Kunin V, Sorek R, and Hugenholtz P. Evolutionary conservation of sequence and secondary structures in CRISPR repeats. Genome Biol, 2007, 8(4):R61.

[95] Lange SJ, Alkhnbashi OS, Rose D, Will S, and Backofen R. CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems. Nucleic Acids Res, 2013.

[96] Heyne S, Costa F, Rose D, and Backofen R. GraphClust: alignment-free structural clustering of local RNA secondary structures. Bioinformatics, 2012, 28(12):i224–i232.

[97] Giegerich R, Voss B, and Rehmsmeier M. Abstract shapes of RNA. Nucleic Acids Res, 2004, 32(16):4843–51.

[98] Costa F and Grave KD. Fast neighborhood subgraph pairwise distance kernel. In Proceedings of the 26 th International Conference on Machine Learning, pages 255–262. Omnipress, 2010.

[99] Indyk P and Motwani R. Approximate nearest neighbors: towards removing the curse of dimensionality. In Proceedings of the thirtieth annual ACM symposium on Theory of computing, pages 604–613. ACM, 1998.

[100] Kröger C, Dillon SC, Cameron ADS, Papenfort K, Sivasankaran SK, Hokamp K, Chao Y, Sittka A, Hébrard M, Händler K, et al. The transcriptional landscape and small RNAs of *Salmonella enterica* serovar Typhimurium. Proc Natl Acad Sci USA, 2012, 109(20):E1277–86.

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

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

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

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

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

[106] Tjaden B. Biocomputational identification of bacterial small rnas and their target binding sites. In Mallick B and Ghosh Z, eds., Regulatory RNAs, pages 273–293. Springer Berlin Heidelberg, 2012.

[107] Wright PR, Richter AS, Papenfort K, Mann M, Vogel J, Hess WR, Backofen R, and Georg J. Comparative genomics boosts target prediction for bacterial small RNAs. Proc Natl Acad Sci USA, 2013.

[108] Smith C, Heyne S, Richter AS, Will S, and Backofen R. Freiburg RNA Tools: a web server integrating IntaRNA, ExpaRNA and LocARNA. Nucleic Acids Res, 2010, 38 Suppl:W373–7.

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

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

[111] Tafer H, Amman F, Eggenhofer F, Stadler PF, and Hofacker IL. Fast accessibility-based prediction of RNA-RNA interactions. Bioinformatics, 2011, 27(14):1934–40.

[112] 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, 2009, 3(8):364–6.

[113] Ying X, Cao Y, Wu J, Liu Q, Cha L, and Li W. sTarPicker: A Method for Efficient Prediction of Bacterial sRNA Targets Based on a Two-Step Model for Hybridization. PLoS One, 2011, 6(7):e22705.

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

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

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

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

[118] Tjaden B, Goodwin SS, Opdyke JA, Guillier M, Fu DX, Gottesman S, and Storz G. Target prediction for small, noncoding RNAs in bacteria. Nucleic Acids Res, 2006, 34(9):2791–802.

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

[120] Smith TF and Waterman MS. Identification of common molecular subsequences. J Mol Biol, 1981, 147(1):195–7.

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

[122] Bernhart SH, Tafer H, Muckstein U, Flamm C, Stadler PF, and Hofacker IL. Partition function and base pairing probabilities of RNA heterodimers. Algorithms Mol Biol, 2006, 1(1):3.

[123] 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, 2008, 372(2):346–50.

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

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

[126] Richter AS and Backofen R. Accessibility and conservation: General features of bacterial small RNA-mRNA interactions?

RNA Biol, 2012, 9(7):954–65.

[127] Mückstein U, Tafer H, Hackermüller J, Bernhart SH, Stadler PF, and Hofacker IL. Thermodynamics of RNA-RNA binding. Bioinformatics, 2006, 22(10):1177–82.

[128] Mückstein U, Tafer H, Bernhart SH, Hernandez-Rosales M, Vogel J, Stadler PF, and Hofacker IL. 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, eds., Bioinformatics Research and Development, volume 13 of Communications in Computer and Information Science, pages 114–127. Springer-Verlag Berlin Heidelberg, 2008.

[129] 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, 2000, 300(5):1101–12.

[130] 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, 2010, 6(3):e1000809.

[131] Richter AS, 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, 2010, 26(1):1–5.

[132] Sonnleitner E, Gonzalez N, Sorger-Domenigg T, Heeb S, Richter AS, Backofen R, Williams P, Huttenhofer A, Haas D, and Blasi U. The small RNA PhrS stimulates synthesis of the Pseudomonas aeruginosa quinolone signal. Mol Microbiol, 2011, 80(4):868–85.

[133] Jäger D, Pernitzsch SR, Richter AS, Backofen R, Sharma CM, and Schmitz RA. An archaeal sRNA targeting *cis*- and *trans*-encoded mRNAs via two distinct domains. Nucleic Acids Res, 2012, 40(21):10964–79.

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

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

[136] Chitsaz H, Backofen R, and Sahinalp SC. biRNA: Fast RNA-RNA binding sites prediction. In Salzberg S and Warnow T, eds., Proc. of the 9th Workshop on Algorithms in Bioinformatics (WABI), volume 5724 of Lecture Notes in Computer Science, pages 25–36. Springer Berlin / Heidelberg, 2009.

[137] Salari R, Backofen R, and Sahinalp SC. Fast prediction of RNA-RNA interaction. Algorithms Mol Biol, 2010, 5:5.

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

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

[140] Huang FWD, Qin J, Reidys CM, and Stadler PF. Partition function and base pairing probabilities for RNA-RNA interaction prediction. Bioinformatics, 2009, 25(20):2646–54.

[141] Huang FWD, Qin J, Reidys CM, and Stadler PF. Target prediction and a statistical sampling algorithm for RNA-RNA interaction. Bioinformatics, 2010, 26(2):175–81.

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

[143] Seemann SE, Richter AS, Gorodkin J, and Backofen R. Hierarchical folding of multiple sequence alignments for the prediction of structures and RNA-RNA interactions. Algorithms Mol Biol, 2010, 5:22.

[144] Seemann SE, Menzel P, Backofen R, and Gorodkin J. The PETfold and PETcofold web servers for intra- and intermolecular structures of multiple RNA sequences. Nucleic Acids Res, 2011.

[145] Li AX, Marz M, Qin J, and Reidys CM. RNA-RNA interaction prediction based on multiple sequence alignments. Bioinformatics, 2011, 27(4):456–63.

[146] Sharma CM, Papenfort K, Pernitzsch SR, Mollenkopf HJ, Hinton JCD, and Vogel J. Pervasive post-transcriptional control of genes involved in amino acid metabolism by the Hfq-dependent GcvB small RNA. Mol Microbiol, 2011, 81(5):1144–65.

[147] Hartung J. A note on combining dependent tests of significance. Biom J, 1999, 41(7):849–55.

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

[149] Fender A, Elf J, Hampel K, Zimmermann B, and Wagner EGH. RNAs actively cycle on the Sm-like protein Hfq. Genes Dev, 2010, 24(23):2621–6.