Bioinformatics of prokaryotic RNAs

Rolf Backofen1,2,4, *, Fabian Amman3,2, Fabrizio Costal1, Sven Findell1,5,6, Andreas S Richter3,4,6, and Peter F Stadler1,2,4,6,7

Bioinformatics Group; Department of Computer Science; University of Freiburg; Georges-Köhler-Allee 106; D-79110 Freiburg, Germany; *Center for non-coding RNA in Technology and Health; University of Copenhagen; Grøndalsgade 1; DK-1870 Frederiksberg C, Denmark; Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17; A-1090 Wien, Austria; Bioinformatics Group; Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, Halleinstr 16-18; D-04107 Leipzig, Germany; Bioinformatics and Computational Biology Research Group, University of Vienna, Währingerstraße 25; A-1090 Wien, Austria; Max Planck Institute for Immunobiology and Epigenetics, Stiftung Stiftung 51; D-79108 Freiburg, Germany; 2Max Planck Institute for Mathematics in the Sciences, Inselstraße 23; D-04103 Leipzig, Germany; 3Fraunhofer Institute for Cell Therapy and Immunology – IZI, Perlickstraße 1; D-04103 Leipzig, Germany; 4Santa Fe Institute, Santa Fe, NM USA; 5These authors contributed equally

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

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 their 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 bacterial and archaeal data. This review provides an overview on the state-of-the-art of RNA bioinformatics focusing on applications to prokaryotes.

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, RNA P RNA, and SRP RNA (as well as tmRNA and 6S RNA in bacteria), typical genomes harbor dozens or even hundreds of sRNAs with predominantly regulatory roles. Archaea, in addition, have homologs of the small nucleolar RNAs of Eukarya (snoRNAs), directing chemical modifications of RNAs and other RNA targets. Compared with protein-coding genes, most of the prokaryotic RNAs are well or even hundreds of sRNAs, but so far have not received much 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. Although we cover a broad variety of approaches, our presentation emphasizes 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. It is a successful example of a coordinated project in which many new or adapted bioinformatics tools have been developed specifically according to the needs of several experimental groups.

Structure Prediction

The complex three-dimensional structures formed by many functional 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. The problem of secondary structure prediction is well investigated and described elsewhere.1–3 The most prominent implementations of RNA folding algorithms are mfold4 and the ViennaRNA Package.5,6 Standard approaches consider only non-crossing structures, a condition that is not always satisfied. Different classes of pseudoknot structures have been defined7 and corresponding prediction algorithms have been implemented, albeit at the expense of higher computational complexity.8–10 The accuracy of secondary structure 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 parameterization of multi-loops, and the exclusion of non-standard base pairs. In addition, the precise transcript might be known only partially, or structure motifs are embedded into a larger RNA, which leads to the even harder problem of local structure prediction.11 There are two remedies for these problems: (1) instead of just a single sequence, evolutionary information on patterns of sequence conservation...
may be taken into account, or (2) 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 a basis for computing consensus structures. The simplest approach, implemented e.g., in RNAalifold,\textsuperscript{15,16} 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 averages, e.g., PETfold\textsuperscript{26} At lower levels of sequence conservation, folding and alignment must be computed simultaneously at a much higher computational cost. Several practical approaches exist, from full-fledged implementations of the Sankoff algorithm,\textsuperscript{3,4} e.g., in Foldalign\textsuperscript{7} and Dynalign,\textsuperscript{28} 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,\textsuperscript{29,30} and SPARSE.\textsuperscript{31} A conceptually different approach taken by the RNAshapes package\textsuperscript{32} 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 penalties to encouraged or discouraged structural elements, resp. Measurement of SHAPE,\textsuperscript{33} PARS,\textsuperscript{34} 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 toward the experimental signal.\textsuperscript{35,36} Most recently, more sophisticated approaches have appeared toward reconciling experimental data with the thermodynamic folding approach. RNAassist\textsuperscript{37} formulates the problems in terms of simultaneously minimizing position-dependent energy penalties 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.\textsuperscript{38} In ShapeKnots,\textsuperscript{39} an interactive procedure is used to include pseudoknots and SHAPE information. It has been applied to e.g., investigate the structure of a SAM-I riboswitch.

Gene Finding and Transcriptomics

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.\textsuperscript{40} As a consequence, only a few well-known RNA genes such as tRNAs, RNase P RNA, SPR RNA, and the tRNA subunits, are annotated for most prokaryotic genomes. Both homology search and many of the comparative genomics approaches discussed below are applicable not only to independent sRNAs but also to structured RNA elements, which includes, in particular, riboswitches.\textsuperscript{41} RNA thermometers,\textsuperscript{42} and several other cis-acting elements. For brevity, we will simply speak of ncRNAs in the following.

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).\textsuperscript{43} This number includes, however, a large number of CRISPR RNA repeats, many ribosomal RNA 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,\textsuperscript{44} only a very moderate number of prokaryotic RNA families have been described in detail and deposited to databases, see e.g. references 39–42. As a consequence, the majority of sRNA families remain in practice unavailable for genome annotation pipelines. For the same reason, it is impossible to give an accurate estimate on the total number of bacterial 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. Thus, it is not implicitly the method of choice.\textsuperscript{45} Specialized ncRNA sequence homology search derivates of blast are available, e.g., blastR.\textsuperscript{46} Semi-global dynamic programming algorithms such as Gotohscan\textsuperscript{47} 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,\textsuperscript{48} 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.\textsuperscript{49} 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 homology 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.\textsuperscript{50} 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\textsuperscript{51} or employing structural alignments.\textsuperscript{52} It remains unclear, however,
whether such techniques can substantially improve the sensitiv-
ity of homology search for distantly related species.

Feature-based gene prediction

dRNApredict\(^8\) uses typical features of prokaryotic RNAs:
elevated sequence conservation, putative promoter sequences,
and Rho-independent terminator elements. TranstermHP, for
instance, is used to predict Rho-independent terminators.\(^5\) Its
scoring function favors G/C-rich stem loops followed by a poly-T
track. It is obviously extremely difficult to detect correct termi-
nator elements in species with a high G/C-content and in those
that use structural elements deviating from the canonical termi-
nator structure. In order to increase sensitivity and specificity,
dRNApredict focuses on intergenic regions and analyzes the co-
ocurrence of several of the above-mentioned features. While this
strategy works quite well for well-characterized bacterial clades,
it is bound to fail in others. \textit{Xanthomonas} and \textit{Helicobacter}, for
example, lack typical promoter sequences and distinct terminator
hairpins.\(^44,53\)

Transcriptomics

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

(1) Library preparation: Transcriptome analyses consist of
\textit{"wet-lab"} experiments and \textit{"dry-lab"} data evaluation. Both com-
ponents 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 plat-
form, 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 per-
formed and, depending on the sequencing platform and sequenc-
ing 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 (http://www.bioinformatics.babра-
ham.ac.uk/projects/fastqc) is commonly used to initially check
and visualize the quality of the raw sequencing data. Software
suites such as the FASTX-Toolkit (http://hannonlab.cshl.edu/ fastx_toolkit) provide several tools to preprocess the raw sequenc-
ing reads by e.g., removal of the adaptor 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 versatil-
ity. 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.\(^6\) It is therefore
important to match the choice of mapping tool to the research
question.\(^6\) We used segemehl\(^66,67\) very successfully in a variety
of studies, ranging from dRNA-seq analysis to split read map-
ing in prokaryotes. In our hands, segemehl has proven to be
a flexible and highly accurate framework. This has also been
repeatedly shown in benchmarks using real live and simulated
data.\(^62\)

Once the mapping step is completed, mapping summary sta-
tistics help to verify whether all prior steps have been success-
ful. 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 overwhelming majority of the reads. There are,
however, biological relevant exceptions that usually end up in the
"sequencing trash bin." These include transcripts containing
self-splicing introns in bacteria, as well as enzymatically spliced
and circularized RNAs in archaea. A recent study showed that
such "atypical" transcript structures may be much more abund-
ant than expected.\(^61\) It remains, however, unclear to what extent
rare transcripts of this type are biologically relevant, how many
of them are technical artifacts, 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.\(^64\)

(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 typi-
cal prokaryotic genome, the non-coding portion is mainly com-
prised of reads that originate from the highly abundant RNAs
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 polycis-
tronic transcripts, i.e., transcripts that contain more than one
gene, remains uncertain. Therefore, the number of reads map-
ing to intergenic regions is overestimated due to this knowledge
 gap. The determination of polycystronic 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-charac-
terized cases such as the \textit{eag} pathogenicity island of \textit{H. pylori}.\(^7\)

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 polycis-
tronic transcripts but also internally and anti-sense 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 \textit{Xanthomonas} and \textit{Helicobacter}, for
instance, only traces of the -10 box are detectable, but no distinct
-35 box has been reported.\(^56\) It seems to be a matter of fact that
the current experimental setups enable the detection of TSS with
species-specific housekeeping promoters, but alternative 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. So-called leaderless transcripts that lack 5′ UTRs completely, i.e., translation start and TSS are mapped to (almost) the same position, are abundant in archaea,"""" but have been thought to be quite rare in bacteria. Surprisingly, 5′-triphosphate-sequencing-adaptor ligation. The 3′-phosphomonoester.

In contrast to translation start sites that can be identified by well-established gene annotation strategies,"""" surprisingly little is known about transcription start sites (TSS) in most bacte-
ria. Even though a thorough TSS annotation can serve as valuable source of information to (1) understand the architecture of polycistronic transcripts, (2) use it as a paramount hallmark for ncRNA gene annotation, and (3) determine the extent of the 5′ UTR, which often harbors regulatory elements such as ribo-
switches, RNA thermometer, and sRNA binding sites. The first successfully applied methods to annotate TSS were primer extension66 and RACE. Both techniques aim to find the 5′ end of primary TSS. Mono-nucleotides for transcription are provided"""", and the latter to be zeros with probability 1. The parameters spec-
fied from the underlying individual library with a Skellam distribution. However, to deduce the parameters from the underlying individual librar-
ies, a zero-inflated Poisson distribution is used instead of a mere Poisson distribution. This allows one to consider the region in focus as a mixture of transcribed and not transcribed segments, where the former are assumed to follow a Poisson distribution and the latter to be zeros with probability 1. The parameters spec-
ifying the Skellam distribution are solely derived from the read density in the transcribed region. The main advantage of TSSAR

The recent development in automated TSS annotation from 5′-phosphate-dependent exonuclease (TEX) to deplete the total RNA of frag-
mnts that are not protected from exonuclease degradation by a 5′-triphosphate. As a control, total RNA from the same extrac-
tion is processed the same way, but without the TEX treatment. Therefore, in the final analysis step, the differences between the treated (i.e. TEX) and untreated library have to be screened position-wise for sites with a compel-
ing enrichment of RNA-seq read starts in the plus vs. the minus library. That is why this method was named differential RNA-
seq (dRNA-seq).

The first applications of dRNA-seq were manually analyzed 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 subjective assessment of the analyzer, the results suffer from a certain lack of reproducibility and cons-
sistency. Therefore, soon after, the first statistical approaches to evaluate dRNA-seq data were proposed. Schmidtk, et al."""" 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 distri-
bution, 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."""" 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 strength of this method, as implemented in the program TTSpredator, lies in its 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 5′-phosphate data, TSSAR,"""" picks up the idea from Schmidtk, 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 librar-
ies, a zero-inflated Poisson distribution is used instead of a mere Poisson distribution. This allows one to consider the region in focus as a mixture of transcribed and not transcribed segments, where the former are assumed to follow a Poisson distribution and the latter to be zeros with probability 1. The parameters spec-
ifying the Skellam distribution are solely derived 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. A comparison of TTSpredator and TSSAR is given in Figure 1.

Similar to the eukaryotic research community, the understanding of prokaryotic genomes can benefit from shifting from only 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 one to characterize bacterial transcriptomes semi-automated in a high-throughput manner, a comprehensive transcript annotation becomes feasible.

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 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 an A/T-rich genome, and thereby, lacks recognizable terminator hairpins.

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 six of 16 possible 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, so that sequence patterns reflecting the structural constraints rapidly accumulate and become readily detectable already at 10% of sequence divergence. Qrna investigates pairwise 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 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. In RNAz, Figure 2, we have therefore taken a different approach. Two lines of evidence inform about conservation of RNA structures: (1) structural similarity above the level expected from placing the differences at random positions, (2) 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. It detects all types of conserved secondary structure elements, including bona fide sRNAs, riboswitches, RNA thermometers, structured cis-acting elements, as well as terminator hairpins. Since its initial publication, several improvements have been introduced. In particular, RNAz 2.0 makes use of improved consensus structure prediction for assessing structural conservation, 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-P23 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. First, it determines the phylogenetic distribution of conserved sequence elements as well as annotated protein-coding genes. Coherent phylogenetic distribution and co-occurrence in clusters of conserved non-coding elements and coding sequences then indicate that conserved, un-annotated sequences may harbor sRNAs or conserved UTR elements, including riboswitches. 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 toward sRNA prediction can be found in e.g., in reference 90.

Discrimination between coding and non-coding regions poses technical as well as biological challenges not addressed by standard gene finders. 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, Figure 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. pylori. 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-hemolysin, and the Bacillus SR1 RNA involved in the regulation of arginine catabolism. The detection of such cases in genome-wide surveys remains difficult, although software for similar tasks has become available.

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 with the observed sequence conservation of the nucleic acid sequences. Well-conserved structured RNAs, such as Xanthomonas sX13, which is involved in virulence-specific gene expression and Hfq mRNA regulation, can easily be identified 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. Dual transcripts such as E. coli s117 are detectable by both RNAz and RNAcode.
In particular, RNAdecoder searches for conserved RNA structure within DNA regions known to be protein-coding; it suffers from very high FDRs, however. 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-function RNAs has been conducted in prokaryotes so far.

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. Rfam’s notion of a clan agglomerates families that clearly share a common ancestor but are too divergent to be reasonably aligned or groups of families that could be aligned, but have clearly distinct functions. At an even higher level, an RNA class further groups together ncRNA families or clans 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 and EvoFold, 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 (1) the nucleotide sequence, (2) the connectivity graph of base pairing interactions, or (3) 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. 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%. 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 simultaneously with the construction of alignments as described in the section on structure prediction. This approach was successfully used to classify all known CRISPR repeats. However, these alignment-based methods do not scale to efficiently cluster hundreds of thousands of candidate ncRNAs predicted by e.g. RNAz screens.

With GraphClust, 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 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 which structural features defined as small-localized subgraphs are extracted as outlined in Figure 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, which could be used for clustering.

For large data sets (i.e., > 10^6 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. 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 data set for similar sequences that were missed by graph-based pre-clustering. 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.
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 than the state-of-the-art clustering using LocARNA. Thus, GraphClust can successfully determine RNA classes for bacterial ncRNAs, even when the precise transcript boundaries are unknown.

### RNA-RNA Interactions

Models for predicting sRNA–mRNA interactions

The rise of high-throughput methods, first tiling arrays and now RNA-seq, to characterize transcriptomes, led to an explosion in the number of identified sRNAs in prokaryotes; more than 100 sRNAs have been reported in most species (e.g., refs. 105–108). Most sRNAs studied to date form base pair interactions with mRNAs to post-transcriptionally regulate their targets’ translation and stability. The functional characterization 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 uncharacterized 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. 110 and 111). Table 1 summarizes web-based applications designed for genome-wide sRNA target predictions.

#### Table 1. Web server for genome-scale prediction of sRNA target genes

| Name          | Features for target prediction | Classifier | Functional enrichment | URL of web server          | References |
|---------------|-------------------------------|------------|-----------------------|----------------------------|------------|
| CopraRNA      | X X X                          | - X        | http://rna.informatik.uni-freiburg.de/CopraRNA | 112           |
| IntaRNA       | - X X                          | - X        | http://rna.informatik.uni-freiburg.de/IntaRNA | 113, 114      |
| RNApredator   | - X X                          | - X        | http://rna.iibi.univie.ac.at/RNApredator | 115, 116      |
| sRNATarget    | - - X X                        | - X        | http://sribms.ac.cn/srntarget | 117           |
| sTarPicker    | - - X X                        | - X        | http://sribms.ac.cn/starpicker | 118           |
| TargetRNA2    | X X X                          | - -        | http://snowwhite.wellesley.edu/targetRNA |               |

All web servers are based on computational methods that score the sRNA–target interaction by their hybridization energy and by additional features as indicated in the table. Some servers directly allow for functional enrichment analysis of the highest-ranking target predictions.
Methods of the second class determine a joint secondary structure of two RNAs, i.e., a common structure 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, which is extended to handle the loop containing the concatenation point energetically as an external loop. Tools implementing this idea are, for example, PairFold and RNAfold. The sRNATarget web server computes the mfe structure of the concatenated sequence to derive interaction features, such as length-normalized free energy, seed match length, and A/U-content in single-stranded regions. A naïve 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. 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 intra- and intermolecular base pairs by the structural accessibility of the interaction features, such as length-normalized free energy, seed match length, and A/U-content in single-stranded regions. A naïve 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. 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 intra- and intermolecular base pairs by the structural accessibility of the interaction features, such as length-normalized free energy, seed match length, and A/U-content in single-stranded regions. A naïve 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. 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.
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. 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.150

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.151 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.152

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 not 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: (1) the scores are not normalized and depend, e.g., on the G/C-content of the organism, and (2) 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 similar \( P \) values from different organisms, one could use the score distribution of all genome-wide predicted interactions for an organism as background to calculate the \( P \) values. For the second point, one first has to determine how \( P \) values from different organisms can be combined. Even though 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 Hartung151 was used since it additionally allows to weight the \( P \) values, thus correcting for the evolutionary distance of the species.

Open Questions

Many questions and computational problems remain open. Although experimental and computational methods are now in place to identify transcription start sites, the corresponding termination sites still cannot be determined reliably, in particular, when they are not associated with Rho-independent terminator structures. Even less is known 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.153 The still unknown rules governing the binding of Hfq to specific sRNAs in what appears to be a highly dynamic molecular mechanism154 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 interactions, this must also include the determination of protein–RNA interactions. Furthermore, not only the sRNA targets, but also the transcriptional regulation of the sRNA itself has to be understood. This would allow one to apply the systems biology toolbox to explore the dynamics of the full gene regulatory network, which is most likely to be altered by the introduction of sRNAs into the network.

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

Disclosure of Potential Conflicts of Interest
No potential conflicts of interest were disclosed.

Acknowledgments
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.
Bon M, Orland H. TT2NE: a novel algorithm to parallelize the prediction of RNA secondary structures. J Mol Biol 2012; 423:1059-66; PMID:22373923; http://dx.doi.org/10.1016/j.jmb.2012.06.007

Liu B, Mathews DH, Turner DH. RNA pseudoknots: secondary structure prediction for aligned RNA sequences. J Mol Biol 2012; 423:1059-66; PMID:22373923; http://dx.doi.org/10.1016/j.jmb.2012.06.007

Seemann SE, Gorodkin J, Backofen R. Fast and accurate RNA secondary structure prediction. RNA 2013; 19:165-72; PMID:23591180; http://dx.doi.org/10.1261/rna.039041.111
Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate analysis. J Comput Biol 2012; 19:796-813; 2011; 27:2790-6; PMID:21856737; http://dx.doi.org/10.1093/nar/gkr1382.

Hackermüller J, Reinhardt R, et al. The primary transcriptome of the major human pathogen Helicobacter pylori. Nature 2004; 434:265-9; PMID:15086605; http://dx.doi.org/10.1038/nature03457.

Hatem A, Bozdag M, Hering O, Lange C, Soppa J. Benchmarking short sequence mapping tools. BMC Bioinformatics 2013; 14:184; PMID:23758764; http://dx.doi.org/10.1186/1471-2105-14-184.

Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, Doose G, Alexis M, Kirsch R, Findeiß S, Langenberger D, Stadler PF, Hoffmann S, Hanauer D, Machné R, Mörl M, Hoffmann S, Stadler PF. RNAz 2.0: improved noncoding RNA prediction of full-length cDNAs from rare transcripts: Differentiating protein-coding and noncoding RNA. BMC Bioinformatics 2011; 12:207; PMID:21524280; http://dx.doi.org/10.1186/1471-2105-12-207.

Gerischer U. The small RNA Aar in Acinetobacter baumannii. RNA Biology 2010; 7:109-15; PMID:20034236; http://dx.doi.org/10.4161/rna.7.1.11106.

Sonnleitner E, Sorger-Domenigg T, Madej MJ, Nawrocki EP. Structural RNA homology search for a single-base resolution map of an orphaned small RNA. RNA 2011; 17:578-94; PMID:21357752; http://dx.doi.org/10.1261/rna.154015.110.

Schneckenhuber Y, Brenneis M, Hering O, Lange C, Soppa J, Hackermüller J, Schweizer A, Lottspeich F, Krauss D, Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Findeiss S, Langenberger D, Stadler PF, Hoffmann S, Hanauer D, Machné R, Mörl M, Hoffmann S, Stadler PF, Fontana W, Stadler PF, Huynen MA, Stadler PF, Hoffmann S, Florin P. Thermodynamic matchers: strengthening the signal of noncoding RNA in comparative genomics data. RNA 2011; 17:578-94; PMID:21357752; http://dx.doi.org/10.1261/rna.154015.110.

Degen G, Hörnig A, Fischedick K, Heinrichs N, Reimold B, Nusskern D, Kikot, K. Identification of a small RNA with two remarkably conserved features. Nucleic Acids Res 2011; 39:5319-29; PMID:21639586; http://dx.doi.org/10.1093/nar/gkr895.

Schneckenhuber Y, Brenneis M, Hering O, Lange C, Soppa J, Hackermüller J, Schweizer A, Lottspeich F, Krauss D, Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Findeiss S, Langenberger D, Stadler PF, Hoffmann S, Hanauer D, Machné R, Mörl M, Hoffmann S, Stadler PF, Fontana W, Stadler PF, Huynen MA, Stadler PF, Hoffmann S, Florin P. Thermodynamic matchers: strengthening the signal of noncoding RNA in comparative genomics data. RNA 2011; 17:578-94; PMID:21357752; http://dx.doi.org/10.1261/rna.154015.110.

Degen G, Hörnig A, Fischedick K, Heinrichs N, Reimold B, Nusskern D, Kikot, K. Identification of a small RNA with two remarkably conserved features. Nucleic Acids Res 2011; 39:5319-29; PMID:21639586; http://dx.doi.org/10.1093/nar/gkr895.

Schneckenhuber Y, Brenneis M, Hering O, Lange C, Soppa J, Hackermüller J, Schweizer A, Lottspeich F, Krauss D, Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Findeiss S, Langenberger D, Stadler PF, Hoffmann S, Hanauer D, Machné R, Mörl M, Hoffmann S, Stadler PF, Fontana W, Stadler PF, Huynen MA, Stadler PF, Hoffmann S, Florin P. Thermodynamic matchers: strengthening the signal of noncoding RNA in comparative genomics data. RNA 2011; 17:578-94; PMID:21357752; http://dx.doi.org/10.1261/rna.154015.110.

Degen G, Hörnig A, Fischedick K, Heinrichs N, Reimold B, Nusskern D, Kikot, K. Identification of a small RNA with two remarkably conserved features. Nucleic Acids Res 2011; 39:5319-29; PMID:21639586; http://dx.doi.org/10.1093/nar/gkr895.

Schneckenhuber Y, Brenneis M, Hering O, Lange C, Soppa J, Hackermüller J, Schweizer A, Lottspeich F, Krauss D, Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, Findeiss S, Langenberger D, Stadler PF, Hoffmann S, Hanauer D, Machné R, Mörl M, Hoffmann S, Stadler PF, Fontana W, Stadler PF, Huynen MA, Stadler PF, Hoffmann S, Florin P. Thermodynamic matchers: strengthening the signal of noncoding RNA in comparative genomics data. RNA 2011; 17:578-94; PMID:21357752; http://dx.doi.org/10.1261/rna.154015.110.

Degen G, Hörnig A, Fischedick K, Heinrichs N, Reimold B, Nusskern D, Kikot, K. Identification of a small RNA with two remarkably conserved features. Nucleic Acids Res 2011; 39:5319-29; PMID:21639586; http://dx.doi.org/10.1093/nar/gkr895.
Gardner PP, Wilm A, Washietl S. A benchmark of genome-scale de novo assembly methods. *Nature Methods* 2013; 10:1285-90; PMID:24308310.

Heyne S, Costa F, Rose D, Backofen R. GraphClust: a computational method for alignment-free structural clustering of local RNA secondary structures. *Nucleic Acids Res* 2013; 41:8034-44; PMID:23987502.

Boisset S, Geissmann T, Huntzinger E, Fechter P, Backofen R. IntaRNA: efficient prediction of bacterial sRNA targets. *Bioinformatics* 2008; 24:2657-63; PMID:18434344.

Rosales M, Vogel J, Stadler PF, Hofacker IL. Thermodynamics of RNA-RNA binding. *Bioinformatics* 2006; 22:1177-84; PMID:16165007.

Eggenhofer F, Tafer H, Bernstein PF, Hofacker IL. RNAplex: a fast tool for thermodynamic modeling of RNA duplex formation. *Bioinformatics* 2004; 20:1912-19; PMID:15168939.

Hein J. A comparative method for finding and folding RNA secondary structures. *Proc Natl Acad Sci U S A* 2006; 103:19869-74; PMID:16988872.

Mückstein U, Tafer H, Bernhart SH, Hernandez-Jover T, Hofacker IL. RNAconcuts: an interactive web interface for RNA secondary structure prediction. *Bioinformatics* 2013; 29:2058-2060; PMID:23465770.

Mückstein U, Tafer H, Bernhart SH, Hernandez-Jover T, Hofacker IL. RNAconcuts: an interactive web interface for RNA secondary structure prediction. *Bioinformatics* 2013; 29:2058-2060; PMID:23465770.

Müller SA, Findeiß S, Pernitzsch SR, Stadler PF, Hofacker IL. Freiburg RNA Tools: a web server integrating INTARNA, EXPARNA and LOCARNA. *Nucleic Acids Res* 2013; 41:W37-43; PMID:23756039.

Smith C, Heyne S, Richter AS, Will S, Backofen R. hRINTARNA: a tool for predicting targets of small RNAs. *Bioinformatics* 2011; 27:1934-40; PMID:21593134.

Mehlhorn K, Sivasankaran SK, Hokamp K, Sittka A, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe N, Li W. sRNATarget: a web server for predicting targets of bacterial small RNAs. *Nucleic Acids Res* 2005; 33:2433-9; PMID:16016936.

Kozak Y, Goyal V, Jarzembowski M, Giegerich R, Voss B, Rehmsmeier M. Abstract models for prediction of bacterial sRNA targets. *Bioinformation* 2006; 2:19-30; PMID:17453240.

Raghava G, Grötker T. OGRMEME: an algorithm for detection of novel regulatory RNAs in E. coli. *Bioinformatics* 2001; 17:279-86; PMID:11179510.

Franz M, Böck S, von Haeseler A, Stadler PF. A comparative algorithm for finding and folding RNA secondary structures. *Proc Natl Acad Sci U S A* 2010; 107:2234-9; PMID:20415380.

Mehlhorn K, Sivasankaran SK, Hokamp K, Sittka A, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe N, Li W. sRNATarget: a web server for predicting targets of bacterial small RNAs. *Nucleic Acids Res* 2005; 33:2433-9; PMID:16016936.

Raghava G, Grötker T. OGRMEME: an algorithm for detection of novel regulatory RNAs in E. coli. *Bioinformatics* 2001; 17:279-86; PMID:11179510.

Franz M, Böck S, von Haeseler A, Stadler PF. A comparative algorithm for finding and folding RNA secondary structures. *Proc Natl Acad Sci U S A* 2010; 107:2234-9; PMID:20415380.

Mehlhorn K, Sivasankaran SK, Hokamp K, Sittka A, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe N, Li W. sRNATarget: a web server for predicting targets of bacterial small RNAs. *Nucleic Acids Res* 2005; 33:2433-9; PMID:16016936.

Raghava G, Grötker T. OGRMEME: an algorithm for detection of novel regulatory RNAs in E. coli. *Bioinformatics* 2001; 17:279-86; PMID:11179510.

Franz M, Böck S, von Haeseler A, Stadler PF. A comparative algorithm for finding and folding RNA secondary structures. *Proc Natl Acad Sci U S A* 2010; 107:2234-9; PMID:20415380.

Mehlhorn K, Sivasankaran SK, Hokamp K, Sittka A, Griffiths-Jones S, Finn RD, Nawrocki EP, Kolbe N, Li W. sRNATarget: a web server for predicting targets of bacterial small RNAs. *Nucleic Acids Res* 2005; 33:2433-9; PMID:16016936.

Raghava G, Grötker T. OGRMEME: an algorithm for detection of novel regulatory RNAs in E. coli. *Bioinformatics* 2001; 17:279-86; PMID:11179510.

Franz M, Böck S, von Haeseler A, Stadler PF. A comparative algorithm for finding and folding RNA secondary structures. *Proc Natl Acad Sci U S A* 2010; 107:2234-9; PMID:20415380.
137. Cherbuli C, Bézier S, Remelli C, Mangin C, Frechère P, Goinsanne T, Vaudensac F, Rouxly P. Staphylosporon aneurum RNAse H1 in two distal regions of the rna5 gene structure and promote mRNA degradation. PLoS Pathog 2010; 6:e1000899. PMID:20930067; http://dx.doi.org/10.1371/journal.ppat.1000899

138. Richter AS, Schlüchtern C, Buczekel R, Stüfler C. Sequence-based INTRNA prediction combined with GTP-dependent system identifies miRNA targets in the small RNA M. Bioinformatics 2010; 26:5-15. PMID:19907075; http://dx.doi.org/10.1093/bioinformatics/btp692

139. Seemann SE, Gonzalez N, Sorge-Dawonning T, Hush S, Richter AS, Buczekel R, Williams P, Hainschole A, Hau D, Blue U. The small RNA PheI stimulates synthesis of the pseudomonas aeruginosa quinolone signal. Mol Microbiol 2013; 89:898-909. PMID:23701551; http://dx.doi.org/10.1111/j.1365-2958.2013.07926.x

140. Jäger D, Penkschau SR, Richter AS, Busukri R, Sharma CM, Schütze R-A. Anarchal RNA impairs air- and water-encoded miRNAs via two distinct domains. Nucleic Acids Res 2012; 40:10964-70. PMID:22932321; http://dx.doi.org/10.1093/nar/gks867

141. Pevzner DA, Berg BH, Betlach KM. RNase H: intermolecular RNA interaction search. Genome Inform 2006; 19:82-92. PMID:17946953

142. Akka C, Karska É, Studijn HM, Sabatini SC, Zhang K. RNA-RNA interaction prediction and autonomous RNA target search. J Comput Biol 2006; 13:267-82. PMID:16997295; http://dx.doi.org/10.1089/cmb.2006.13.267

143. Chiriac H, Backofen R, Sabatini SC. miRNA: Fast RNA-RNA binding site prediction. In Salitz S and Wieser 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.

144. Salaiti B, Backofen R, Sabatini SC. Fast prediction of RNA-RNA interaction. Algorithms Mol Biol 2010; 5:5. PMID:20608063; http://dx.doi.org/10.1186/1748-7188-5-5

145. Salak M, Abeli M, Welf S, Sabatini SC, Backofen R. Time and space efficient RNA-RNA interaction prediction via sparse folding. In Rieser B, ed., Proc. of RECOMB 2010, volume 6448 of Lecture Notes in Computer Science, pages 475–480. Springer-Verlag Berlin Heidelberg, 2010.

146. Chiriac H, Salak R, Sabatini SC, Backofen R. A partition function algorithm for interfering nuclear acid species. Bioinformatics 2009; 25:2067-76. PMID:19478001; http://dx.doi.org/10.1093/bioinformatics/btp212

147. Huang FWD, Qin J, Rivas CM, Sudiie F. Partition function and base pairing probabilities for RNA-RNA interaction prediction. Bioinformatics 2009; 25:2046-54. PMID:19787482; http://dx.doi.org/10.1093/bioinformatics/btp388

148. Huang FWD, Qin J, Rivas CM, Suijie F. Target prediction and a maximum sampling algorithm for RNA-RNA interaction prediction. Bioinformatics 2010; 26:175-83. PMID:19843495; http://dx.doi.org/10.1093/bioinformatics/btp389

149. Seemann SE, Richter AS, Gendell T, Backofen R, Gorschek E. PETFold: predicting conserved interactions and structures of two multiple alignments of RNA sequences. Bioinformatics 2011; 27:211-21. PMID:21088026; http://dx.doi.org/10.1093/bioinformatics/btr038

150. Seemann SE, Richter AS, Gorschek E. RNA-RNA interaction prediction based on the prediction of structures and RNA-RNA interactions. Algorithms Mol Biol 2010; 5:21. PMID:20942644; http://dx.doi.org/10.1186/1748-7188-5-22

151. Seemann SE, Menzel P, Backofen R, Gorschek E. The PETFold and PETFold web servers for intramolecular and intermolecular structures of multiple RNA sequences. Nucleic Acids Res 2011; 39:W577-81. PMID:21609946; http://dx.doi.org/10.1093/nar/gkq248

152. Li AX, Marx M, Qin J, Rivas CM. RNA-RNA interaction prediction based on multiple sequence alignments. Bioinformatics 2011; 27:406-9. PMID:21546094; http://dx.doi.org/10.1093/bioinformatics/btr159

153. Sharma CM, Panagiotou K, Panitschek MR, Mullenhof HG, Hanne K, Vögel J. Partition-based transcriptional control of genes encoded in animal and microbial pathogens by the Hfq-dependent GcvB small RNA. Mol Microbiol 2011; 81:1144-65. PMID:21696468; http://dx.doi.org/10.1111/j.1365-2958.2011.07751.x

154. Hartung J. A note on conducting dependent tests of significance. Biometrics 1999; 45:809-15. http://dx.doi.org/10.2307/2534082 (SICI)1521-4036(199911)45:4<809::AID-BIOM25>3.0.CO;2-T

155. Vögel J, Lies A, H. Hfq and itsconstellation of RNA. Nucleic Acids Res 2011; 39:570-89. PMID:21706622; http://dx.doi.org/10.1093/nar/gkq305

156. Fendeer A, H elf J, Hamburger K, Zimmermann B, Wagnor TF. RNA activity cycle on the TetR-family protein. Nucleic Acids Res 2010; 38:2420-1. http://dx.doi.org/10.1093/nar/gkq410