RNA Biology 7:1, 33-42; January/February 2010; © 2010 Landes Bioscience

Computational prediction of sRNAs and their targets in bacteria

Rolf Backofen[1] and Wolfgang R. Hess[2]

[1]Faculty of Engineering; Department of Computer Sciences; Chair for Bioinformatics; and [2]Faculty of Biology; Genetics and Experimental Bioinformatics Group; University of Freiburg; Freiburg Initiative in Systems Biology; Freiburg, Germany

Key words: algorithms, bacteria, comparative genomics, compensatory mutations, non-coding RNA, regulatory RNA, small RNA

Abbreviations: sRNA, small RNA

There is probably no major adaptive response in bacteria which does not have at least one small RNA (sRNA) as part of its regulatory network controlling gene expression. Thus, prokaryotic genomes encode dozens to hundreds of these riboregulators. Whereas the identification of putative sRNA genes during initial genome annotation is not yet common practice, their prediction can be done subsequently by various methods and with variable efficacy, frequently relying on comparative genome analysis. A large number of these sRNAs interact with their mRNA targets by antisense mechanisms. Yet, the computational identification of these targets appears to be challenging because frequently the partial and incomplete sequence complementarity is difficult to evaluate. Here we review the computational approaches for detecting bacterial sRNA genes and their targets, and discuss the current and future challenges that this exciting field of research is facing.

Introduction: The Challenge of Computational Prediction of Regulatory Small RNAs and their Targets

Due to divergence in functions, sequences and structures, there are no common identifiers for bacterial sRNAs. Even widely accepted characteristics, such as small size (<200 nt) and lacking coding capacity, do not always apply, as is demonstrated by the example of *Staphylococcus aureus* RNAIII, which clearly is a regulatory RNA, but is 514 nt in size and does code for a short peptide as well.[1,2] Notwithstanding these facts, we will use the term "sRNA" in this review throughout for the class of bacterial regulatory RNA.

Parameters used in conventional genome annotation and gene modeling are meaningless for the prediction of sRNA genes and consequently there is no universal method for the detection of all classes of bacterial sRNAs. Nevertheless, after less than ten years of research in this field, advantages and shortfalls of the different approaches are becoming increasingly clear, which has led to various algorithms that provide reliable sRNA predictions in a more focused context, as well as program packages which can be applied to large scale biocomputational analysis. We will describe these approaches and associated studies (see **Table 1** for overview) in more detail in the first part of this review and summarize the evidence for a high number of sRNA genes in bacterial genomes which await functional characterization.

In the second part of this review, we summarize the state of the art of predicting the possible targets of bacterial sRNAs (see **Table 2** for overview). Given the high number of different sRNAs in the average bacterial genome, probably several hundred, the identification of their targets now becomes the really critical bottleneck for further progress in this field. Biocomputational predictions of sRNA targets are the key to efficiently elucidating sRNA functions and correctly assigning them within the cellular regulatory infrastructure. The main challenges are that targets are frequently encoded far away, at different genomic loci and that some sRNAs have single targets whereas other are true master regulators with a multitude of mRNAs under their control. A big problem is that the interacting sequence elements are mostly only short sequence stretches of imperfect similarity, which can reside in any part of the sRNA and can even be formed through the joining of sequence elements from two separate domains.

Experimental approaches for the discovery and characterization of bacterial regulatory sRNAs and their targets have been reviewed recently[3-7] and are outside the scope of this review.

Prediction of sRNAs

Prediction based on comparative genomics. The standard procedure for the prediction of bacterial sRNAs by comparative genomics consists of four steps. First, conserved sequences are identified in intergenic regions. Then, these are clustered and compared in pairwise or multiple alignments. Finally, these alignments are scored based on predicted RNA structural features, using RNAz,[8] eQRNA[9] or evofold,[10] for example.

The pioneering studies on sRNA prediction in *E. coli* (see **Table 1** for overview) were based on comparative genome analysis of closely related enterobacteria[11] and included, in one case, a search for promoters and Rho-independent terminators in intergenic regions.[12] The idea to score conservation of RNA secondary

*Correspondence to: Rolf Backofen and Wolfgang R. Hess; Email: backofen@informatik.uni-freiburg.de and wolfgang.hess@biologie.uni-freiburg.de
Submitted: 10/16/09; Revised: 11/12/09; Accepted: 11/13/09
Previously published online:
www.landesbioscience.com/journals/rnabiology/article/10655

Table 1. Genome-wide biocomputational sRNA screens in diverse bacteria

Species	Method for prediction and reference	Number of predicted/experimentally verified RNAs (when tested)
Enterobacteria		
E. coli (Salmonella typhimurium, S. typhi, S. paratyphi, Klebsiella pneumonia, Yersinia pestis)	Comparative genomics plus identification of σ70 promoters and Rho-independent terminators[12]	24/14
E. coli (Salmonella enteritidis, S. thyphi, S. typhimurium, S. paratyphi, Klebsiella pneumonia, Yersinia pestis)	Comparative genomics[11]	58/17
Escherichia coli	Promoter-terminator prediction, Gapped Markov Model Index[33]	87 sRNA, 46 antisense RNA candidates/8 and 4
E. coli (Salmonella typhimurium, Klebsiella pneumonia)	Comparative genomics and scoring the conservation of RNA secondary structures[13]	275
Escherichia coli	Identification of σ70 promoters and Rho-independent terminators plus further filtering steps using RNAMotif[22]	144 new sRNA candidates, 32 previously described sRNAs/7
Escherichia coli	Neural network classifier using only sequence and structure-based features derived from the genome[70]	601/3 (of 6 tested)
Escherichia coli	Boosted genetic programming[37]	135 novel sRNA candidates plus 152 that overlap predictions in the literature/12
Several prokaryotic and archaeal genomes	Machine learning approach; neural networks and support vector machines were used to extract the shared features of known sRNAs for the prediction of new candidates[36]	370 in *E. coli*
Vibrio cholerae	sRNAPredict (sequence conservation and prediction of Rho-independent terminators)[28]	32/6
Vibrio cholerae, V. parahaemolyticus, V. vulnificus	Search focused on σ54 promoter/Rho-independent terminators in intergenic regions and conservation in three genomes[25]	Four or five in each of the three genomes predicted and verified
Cyanobacteria		
3 Prochlorococcus & Synechococcus WH8102	Comparative genomics and scoring the thermodynamic stability values derived from consensus folding[15]	18 high-scoring sRNA candidates/7
Synechocystis sp. PCC6803, *Microcystis aeruginosa, Thermosynechococcus elongatus, Synechococcus elongates*	Comparative genomics, scoring by ALIFOLDZ and RNAz[17]	109 clusters of sRNA candidates in the four species/5 (in Synechocystis)
Synechocystis sp. PCC6803	Rho-independent terminators[34]	713 candidate terminators/11 antisense RNAs and 27 sRNAs
Proteobacteria		
Pseudomonas aeruginosa	Pattern searches for Fur-consensus binding sites in intergenic regions, combined with predictions for Rho-independent terminators[26]	Two (PrrF1 and PrrF2)/2
Pseudomonas aeruginosa (comparison of genomes from 10 different Pseudomonas)	Comparative genomics, scoring by RNAz[21]	115 (221) sRNA candidates of which 101 (85) were previously known and 14 (136) were novel based on NcDNAlign (MultiZ)
Pseudomonas aeruginosa	sRNAPredict2[29]	17
Pseudomonas aeruginosa (+5 additional Pseudomonas)	QRNA[71]	130/8
Sinorhizobium meliloti (+4 additional genomes)	Comparative genomics, Intergenic Sequence Inspector, QRNA, sRNAPredict2[72]	60/14
Sinorhizobium meliloti (+8 related α-proteobacteria)	Comparative genomics, scoring by eQRNA and RNAz[73]	32/8
Bacilli		
Bacillus subtilis (+7 different additional genomes)	Comparative genomics, QRNA for scoring[74]	8 (from 12 tested)

Table 1. Genome-wide biocomputational sRNA screens in diverse bacteria (continued)

Staphylococcus aureus N315	Intergenic Sequence Inspector[39]	191/12
Streptococcus pneumoniae	Search for CiaR binding sites in intergenic regions, combined with predictions for Rho-independent terminators[27]	5
Actinomycetes		
Streptomyces coelicolor (+7 bacterial genomes for comparative genomics)	Hidden Markov model to combine primary sequence data (dinucleotide frequency information and Rho-independent terminators) with comparative genomics[30]	114 sRNA candidates, 20 tested/6

Table 2. Summary of target prediction tools and some related software mentioned in the text

Model and reference	Program download	URL for web-accessible tools
TargetRNA[45,75]	-	http://snowwhite.wellesley.edu/targetRNA/
Sequence-based scoring combined with stacking[46]	Available as available as a supplementary material file (Document S3) from the publisher's website: http://nar.oxfordjournals.org/cgi/content/full/gkl1096/DC1	-
IntaRNA[61]	http://www.bioinf.uni-freiburg.de/Software/	http://rna.informatik.uni-freiburg.de
RNAplex[52]	http://www.tbi.univie.ac.at/~htafer/	-
sRNATarget[58,59]	http://www.biosun.org.cn/srnatarget/	http://ccb.bmi.ac.cn/srnatarget/
RNAup[60]	http://www.tbi.univie.ac.at/~ulim/RNAup/index.html	http://rna.tbi.univie.ac.at/cgi-bin/RNAup.cgi
RNAhybrid[50]	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/	http://bibiserv.techfak.uni-bielefeld.de/rnahybrid/

structure rather than of primary sequence was introduced by Rivas et al. (2001).[13] Structure prediction was combined with comparative analysis of closely related enterobacterial genomes; mutational patterns were evaluated in pairwise sequence alignments and classified into three classes: compensatory base mutations, which are typical of conserved RNA secondary structure; synonymous mutations, mainly third codon position substitutions as observed among conserved protein-coding regions; and mutations at random, characteristic for other conserved sequence elements, such as transcription factor binding sites. This approach yielded 275 candidate loci for structural RNAs and these ideas were combined in the program QRNA.[13]

Incorporating thermodynamic stability values. Biocomputational approaches were also successfully applied to scan the genomes of bacteria not closely related to *E. coli*. At the time of these studies, more efficient scoring methods were sought to apply the powerful methods of comparative analysis to eukaryotic genomes. In particular, it was found that thermodynamic stability values derived from the consensus folding of aligned sequences allow effective prediction of functional RNAs.[8,14] Based on such a strategy, Axmann et al. (2005)[15] scored alignments of intergenic regions extracted from the genomes of four closely related marine cyanobacteria of the Prochlorococcus-Synechococcus lineage using ALIFOLDZ.[8,14] Expression analysis of the highest-scoring candidate regions under various growth and stress conditions confirmed seven new sRNAs in Prochlorococcus sp. MED4, several of which had homologues in the other three strains and which, therefore, were called Yfr1-7 for cYanobacterial Functional RNAs (**Table 1**). One of these sRNAs, Yfr7, later turned out to be the cyanobacterial ortholog of the 6S RNA.[16]

More recently, this approach was extended to the biocomputational prediction of sRNA genes and other sequence/structure-conserved elements in intergenic regions of the unicellular model cyanobacteria Synechocystis PCC6803, *Synechococcus elongatus* PCC6301, and *Thermosynechococcus elongatus* BP1, plus the toxic *Microcystis aeruginosa* NIES843.[17] After the removal of transposon-associated repeats, 341 sequences were left, belonging to 109 clusters of related RNA elements in the four species. Experimental analysis of selected sRNA candidates in Synechocystis PCC6803 validated five of the new sRNAs,[17] including new members of the Yfr2 family that accumulate to very high RNA concentrations and occur in up to nine copies per genome in some marine Synechococcus species.[18]

Applying RNAz for scoring and using ten different Pseudomonas genome sequences for comparative genomics, 115 sRNA candidate loci were predicted in *Pseudomonas aeruginosa* when using NcDNAlign[19] to construct multiple sequence alignments and 221 candidate loci when the less restrictive MultiZ[20] was used (**Table 1**).[21]

Non-comparative screens. A systematic scan for promoters and Rho-independent terminators in intergenic regions as a tool for sRNA discovery was first employed by Argaman et al. (2001)[12] as part of a more complex search strategy. Chen et al. (2002)[22] demonstrated that many additional sRNAs could be found when screening intergenic regions of *E. coli* for possible σ70-type promoter and Rho-independent terminator pairs that lie within a distance of 45–350 bp. The authors used RNAMotif, an algorithm that searches nucleic acid databases for RNA structure motifs[23] to find the terminators and to identify possible open reading frames. Applying several filters reduced the rate of false-positives and removed known sRNA genes, ending up with 144 candidates. From these, seven out of eight tested candidates were confirmed in northern blots to be new sRNA species. Interestingly, only ten of the forty sRNAs known at the time were recognized by the search

algorithm. RNAMotif was also successfully used to evaluate the distribution of the sRNA Yfr1 throughout the cyanobacterial radiation.[24]

When screening for transcriptional signals, it is a logical step to include as an additional search criterion a certain transcription factor binding site and consequently restrict the computational prediction of sRNA candidates to a distinct regulon. Following this idea, intergenic regions in *Vibrio cholerae* were scanned for pairs of relatively well-defined σ54 binding sites and Rho-independent terminators.[25] The rationale for this approach was the observation that the mRNAs for the quorum-sensing master regulators LuxR and HapR were destabilized in an Hfq-dependent fashion, and that this process would be controlled by the sigma factor σ54. To restrict the analysis further, the candidate sRNAs were required to be conserved in three related Vibrio species. Indeed, four (*Vibrio cholera*) and five sRNAs (*V. parahaemolyticus*, *V. vulnificus*) were finally found and shown to constitute an ultrasensitive regulatory switch that controls the transition into the high cell density, quorum-sensing mode of the cell.[25]

That it is feasible to include in the search the binding sites of real transcription factors (as opposed to a sigma factor recognition sequence) was demonstrated by the identification of sRNAs controlled through Fur, the Ferric uptake regulator, and CiaR, the response regulator which is part of the two-component regulatory system CiaRH controlling beta-lactam resistance, maintenance of cell integrity, competence and virulence in Streptococcus. A pattern search for Fur-consensus binding sites in intergenic regions of the *Pseudomonas aeruginosa* genome, combined with the prediction of Rho-independent terminators, yielded the two sRNAs PrrF1 and PrrF2, which are more than 95% identical to each other.[26] Similarly, in *Streptococcus pneumoniae* five sRNAs were fortuitously identified during analysis for CiaR binding sites.[27] These small non-coding RNAs, designated csRNAs for cia-dependent small RNAs, are 87 to 151 nt in size, and show a high degree of sequence similarity.

These studies clearly demonstrated that systematic screens for transcription factor binding sites can be very useful when focusing on a certain class of sRNAs. They not only added the additional criterion of searching for a specific transcription factor binding site to the portfolio of parameters, but also provided impressive functional information, because they started right from the beginning with a hypothesis for the respective sRNA's function.

Joining the search for termination signals with comparative genomics. In several earlier studies, the set of possible transcriptional units in intergenic regions was filtered against the sequences of closely related bacterial genomes.[12,25] Consequently, Livny et al. (2005)[28] reasoned that sRNA genes can be predicted solely by relying on sequence conservation of intergenic regions and predictions of Rho-independent terminators, without any further information. Implementing this idea in a program called sRNAPredict2, they identified 32 novel sRNAs in *Vibrio cholera*, from which nine were tested and six confirmed,[28] and 17 sRNAs in *Pseudomonas aeruginosa* (**Table 1**).[29] This is of more general relevance because the successful use of bioinformatics to identify sRNAs is frequently rendered impossible in non-enterobacteria due to the lack of information on promoter consensus sequences and transcription factor binding sites. Also building off of this idea, Swiercz et al. (2008)[30] developed sRNAFinder. This program uses a hidden Markov model to integrate primary sequence data with comparative genomics information when predicting sRNA genes. The primary sequence data include dinucleotide frequency information and Rho-independent terminators. Comparative genomics information includes evidence of compensatory basepair mutations that conserve RNA secondary structure. The authors focused on the multicellular, differentiating actinomycete *Streptomyces coelicolor*, and compared the sequences against seven different genomes. A total of 114 sRNA candidates were predicted, 20 tested and 6 experimentally confirmed (**Table 1**).

Applying sRNAPredict2 to the growing database of bacterial genome sequences, Livny et al. (2007)[31] predicted more than 2,700 previously unannotated candidate sRNA loci. Further following this approach, SIPHT was developed to conduct searches for putative sRNA-encoding genes in all 932 bacterial replicons present in the NCBI database at that time. These searches yielded nearly 60% of previously confirmed sRNAs, hundreds of previously annotated cis-encoded regulatory RNA elements such as riboswitches, and over 45,000 novel candidate intergenic loci which await experimental verification.[32]

Finding cis-encoded antisense RNAs. Computational screens have been used successfully for the prediction of non-coding RNA in various eubacteria, but much less often to find antisense RNAs. Yachie et al. (2006)[33] combined predictions for intergenic non-coding RNAs and antisense RNAs in *E. coli*, resulting in 87 sRNA and 46 antisense RNA candidates, of which, respectively, eight and four candidates could be experimentally verified. The employed prediction strategy was a combination of promoter and Rho-independent terminator prediction using the newly introduced Gapped Markov Model Index (GMMI), followed by experimental analysis. The GMMI takes into account sequence patterns, nucleotide biases and higher order base relations, as they occur, for example, through base pairing in structured RNA molecules. This is a reasonable approach for the prediction of intergenic sRNAs, yet it is less suitable for a prediction focusing on antisense RNAs, as these function mainly by mere sequence complementarity rather than specific sequence or structural features. When rigorously tested by using tiling microarrays, an approach to finding antisense RNAs based on Rho-independent terminators turned out to be less productive. In the cyanobacterium Synechocystis 6803, Georg et al. (2009)[34] found that 11 out of 73 asRNAs, but 27 out of 60 intergenic sRNAs with high microarray expression levels had been predicted using this approach. Thus, this strategy worked relatively well for sRNAs transcribed from intergenic regions (45% correctly predicted), whereas the rate of true positives for the antisense RNAs was only 15%.[34] As the most likely reason for this difference, the authors assumed either sequence constraints from the protein-coding region or more frequent or more complex RNA processing in the case of antisense RNAs.[34] It is also known that some trans-acting sRNAs, such as 4.5S, 6S and DicF are processed from poly- or di-cistronic transcripts, and therefore lack an Rho-independent terminator.[35] It is possible that such 3'-processed RNA species are more frequent among antisense RNAs, making their correct computational prediction technically challenging.

Large scale approaches. Serious effort has been invested in automating sRNA identification and targeting a large number of genomes at once. In a machine learning approach, neural networks and support vector machines were used to extract the shared features of known sRNAs for the prediction of new candidates in several prokaryotic and archaeal genomes.[36] This approach depends less on prior knowledge of the specific RNA gene features of a given organism. The underlying algorithm employed both compositional parameters (nucleotide and dinucleotide composition) and structural motif parameters to discriminate functional RNAs from random non-coding sequences. The screen by Carter et al. (2001)[36] predicted 370 novel sRNA candidates in the *E. coli* genome.

Automatic discovery of sequence patterns by boosted genetic programming was used to create sRNA classifiers to distinguish non-coding functional RNA sequences from other intergenic sequences.[37] This approach resulted in the prediction of 135 novel sRNA candidates and of 152 loci that overlapped previous predictions in the literature. In this study, twelve of sixteen candidates were experimentally shown to be actual sRNAs and six of the twelve verified candidates had not been predicted in any of the previous studies. The relatively high confirmation rate was taken as evidence that many of the predicted sRNAs actually exist, and that in a well-studied model such as *E. coli* many more sRNA genes are still to be characterized.[37]

An automated sRNA screening procedure for the extraction, selection and visualization of candidate intergenic regions has been implemented in the software package 'Intergenic Sequence Inspector', or ISI.[38] This program filters intergenic regions according to variable input parameters, including length or GC content, and can select those with significant sequence conservation among phylogenetically related bacteria. In the gram-positive bacterium *Staphylococcus aureus*, ISI identified 191 sRNA candidates, which were rigorously tested by microarrays and northern blots, leaving a minimum of 12 expressed sRNA genes in *S. aureus*.[39]

Computational Prediction of sRNA Targets

In the first part of this review, we summarized the various approaches to the prediction of sRNAs and the evidence for a high number of sRNA genes in bacterial genomes awaiting functional characterization. The critical bottleneck is the identification of the targets of these sRNAs. Experimental approaches for the detection of sRNA targets include standard genetic screens, gene knockouts and overexpression of the sRNA of interest, followed by proteomics and microarray analysis, the co-immunoprecipitation of direct interaction partners and the characterization of relevant ribonucleoprotein particles. However, all of these methods require a large amount of effort and are very time-consuming. Therefore, methods for highly sensitive biocomputational target prediction, followed by focused experimental analysis, are highly desirable. Previous reviews, which considered such aspects to some extent, were provided by Pichon and Felden (2008)[40] and Vogel and Wagner (2007).[5]

Target prediction based on sequence. Concerning computational target prediction, the initial step is always (see **Table 2** for overview on the different methods) the search for regions both in the mRNA and the sRNA that are complementary to each other. However, the "strength" of complementarity of these regions is measured in many different ways. In contrast to microRNA, there are currently few additional steps (such as conservation of specific regions or the enforcement of a seed region) that are used on a regular basis. It appears that sRNA-mRNA interactions are even more flexible than those between miRNAs and mRNAs. This fact makes it hard to determine single significant features for sRNA-mRNA interactions, which would be comparable to the 6–8 nt long seed region found to be important for the prediction of miRNA targets.[41,42]

For the evaluation of complementarity, pure sequence-based methods like BLAST[43] can be used to search for long stretches of complementarity. However, it is important to consider also the non-Watson-Crick G-U-pairs, which can be done using GUUGle.[44] Therefore, the most simple mode of TargetRNA,[45] namely the individual base pair model, can be considered an entirely sequence-based approach. Here A-U and G-C base pairs are given the same score. For this reason, Mandin et al. (2007)[46] introduced a similar model but differing in the scoring of individual base pairs, which is inspired by the strength of the respective base pair. In addition, the scoring takes stacking into account. Thus, the scoring of a perfect duplex corresponds to the associated energy. This difference is especially important for genomes with low GC-content, such as Listeria.[47] The main advantage of these approaches is their simplicity, since the computational costs usually grow at most geometrically with the input length. Another advantage is that one can easily calculate the significance of the found matches, which will be discussed later.

Thermodynamic scoring of mRNA-sRNA mixed duplexes. The next step in complexity involves approaches that do not score the base pairs of the interaction independently but uses a scoring system that is known from the prediction of RNA secondary structures, namely the scoring of stacked base pairs and internal loops. This leads to the thermodynamic scoring of mixed duplexes consisting of mRNA and sRNA sequences, and can be considered a restricted and specialized version of full RNA-secondary structure folding (like Mfold[48] and RNAfold[49]). The first approach in this direction was RNAhybrid[50] (also implemented as RNAduplex in the Vienna RNA Package), whose main application area is the prediction of microRNA targets. Here the scoring of a base pair (i, k), where i is a position in the mRNA, and k is a position in the sRNA, depends on the immediately following base pair (i', k'), where $i' > i$ and $k < k'$ (assuming that both mRNA and sRNA are notated 5' → 3', and the interaction is anti-parallel as usual). If $i' = i + 1$ and $k' = k - 1$, then the two base pairs form a stack, which is usually energetically favorable. Otherwise, the two base pairs close an internal loop or bulge. The energy parameters for this scoring are the same as in RNA secondary structure folding, and represent free energies (in kcal/mole) that were derived from experimental data using the nearest neighbor model by Mathews et al. (1999).[51] Later, a similar approach was used in TargetRNA with applications to the prediction of sRNA targets. Both approaches use a restriction on the length of internal loops in the mixed duplexes, since long internal loops are energetically

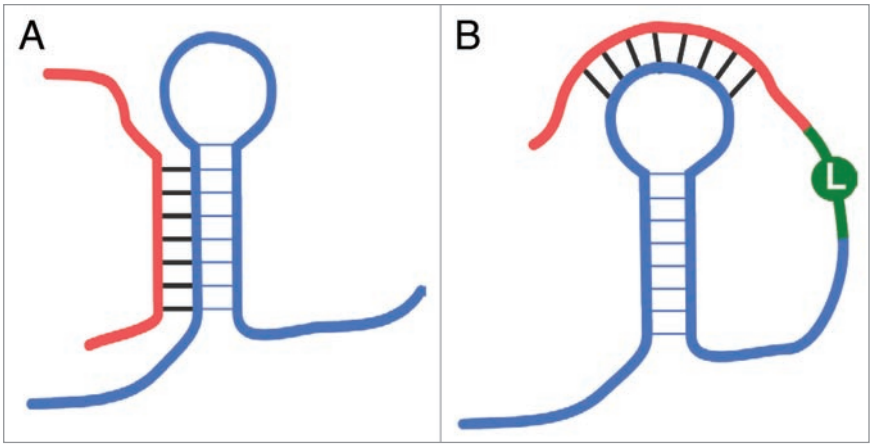

Figure 1. (A) Physiologically impossible structure that might be predicted by simple duplex scoring. (B) A non-nested structure that cannot be predicted by concatenation approaches.

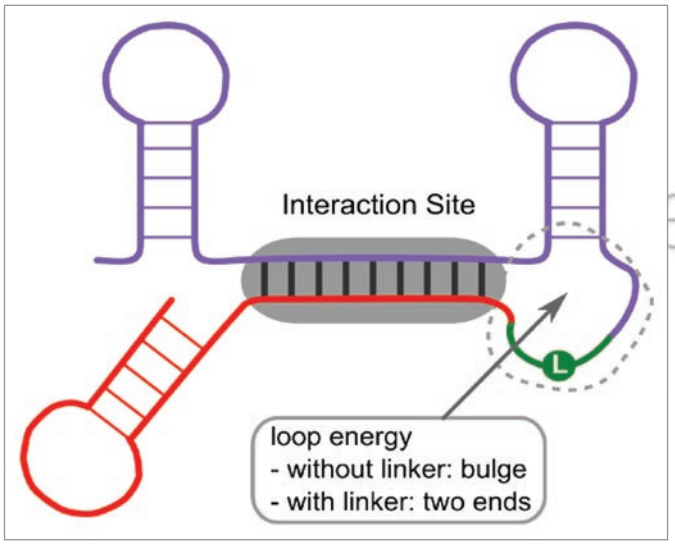

Figure 2. Joint structure of two RNAs using concatenation with a linker element (in green). Without special treatment of the linker element, the associated loop would be scored as a bulge, giving rise to a high positive energy contribution. With the special treatment of the linker, this sequence has only external bases with associated dangling end contributions.

unfavorable and increase computational complexity, where the maximal loop length L contributes quadratically to the run time. RNAplex[52] has a similar energy model as RNAduplex and RNAhybrid, except for internal loops. Whereas explicit energy tables are used for small internal loops, big internal loops are usually evaluated using a logarithmic length term and an asymmetry penalty. In RNAplex, the length dependent term is replaced by an affine gap penalty, which removes the quadratic factor introduced by the maximal loop length L.

Compared to entirely sequence-based approaches, this simple energy model offers several advantages. First, it provides a much more realistic model of RNA-RNA interaction, as compared to approaches based on sequence complementarity, and allows the user to take many variables into account, such as temperature, which is an important parameter when considering the stability of duplexes. Second, these approaches are very fast because their computational complexity is comparable to simple local sequence alignment. Third, one can easily calculate the significance (i.e., p-values) of the found hits, again due to their similarity to local sequence alignment. It is well known that the score of an optimal local alignment follows an extreme value distribution where local scores can be regarded as the maximum of a set of independent variables.[53] Therefore, assuming an extreme value distribution with a location parameter u and a scale parameter s for the length-normalized hybridization scores, one can estimate these two parameters by fitting the extreme value distribution to an empirical distribution generated from normalized hybridization scores for a large set of randomly generated sequences, where the sequences are generated using the actual dinucleotide frequency of the mRNA space of interest. It is important to use dinucleotide instead of mono-nucleotide shuffling because the energy of duplexes depends on the dinucleotide frequencies, due to base pair stacking. The scores have to be normalized according to length because longer putative target and sRNA sequences will tend to have more negative energies.

The main disadvantage of these approaches is that they neglect intra-molecular base pairs. This can have two effects. First, neglecting these pairings could predict biologically impossible interactions where one of the interacting regions is buried in a stable intra-molecular structure (**Fig. 1**). Second, methods that ignore these pairings tend to predict interactions that are too long because it is usually more favorable to extend interactions if the effect of breaking intra-molecular base pairs is ignored.

Approaches based on RNA secondary structures. The observations outlined in the previous chapter lead to the introduction of several other approaches that incorporate the effects of the internal structures of both mRNA and sRNA. There are two classes of approaches. The first class of approaches, with pairfold,[54] RNAcofold,[55] and the method presented by Dirks et al. (2007)[56] as part of the NUpack package being representatives, consider joint structures of mRNA and sRNA that are generated by concatenating the two sequences using a special linker character. Then, a modified version of the usual RNA-folding algorithm (as in MFOLD[48] and RNAfold[49]) is applied. Basically, the recursive structure is the same but loops that contain the linker symbol are treated specially. This is because an internal loop containing the linker element is in fact not an internal loop, but consists only of external bases (**Fig. 2**). As a result, these approaches predict joint structures which are nested in the sequence, stemming from the concatenation of the two input sequences, since this is a restriction crucial for the recursive calculation of the joint structure. In the following, we will refer to this class of approaches as *concatenation approaches*. **Figure 1** on the right side shows an example of a possible joint structure that cannot be predicted using these approaches.

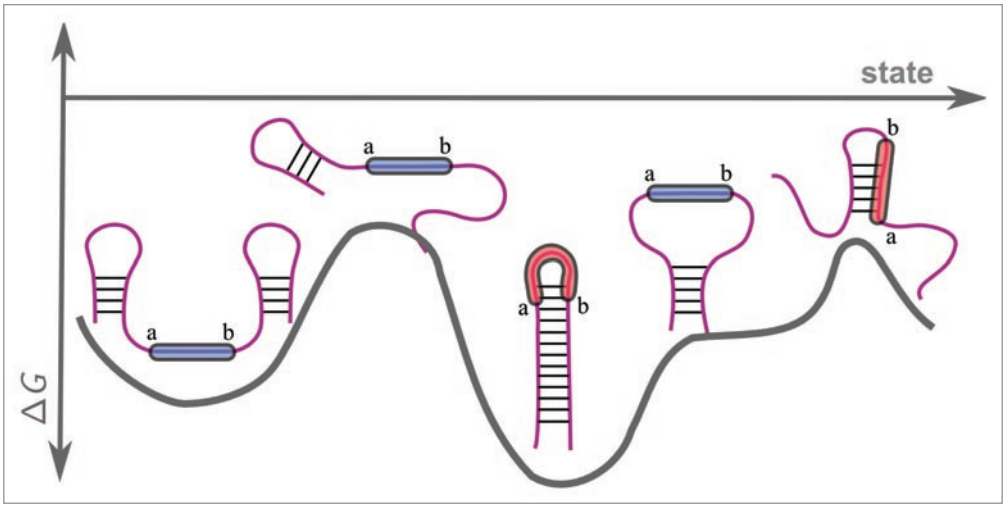

Figure 3. Energy landscape and accessibility. Given a putative interaction site between positions *a* and *b*, there are several structures where this site is single-stranded (denoted by a blue oval), whereas others cover the interaction site. The latter ones cannot be structures that are adopted in a joint structure. The partition function $Z^{sg(a,b)}$ for the ensemble of structures where the subsequence between *a* and *b* is single stranded would be the sum of all Boltzmann-weighted energies for the structures with horizontal ovals.

Concatenation approaches. An advantage of the *concatenation approach* is that all techniques regularly used in plain RNA-secondary structure prediction can be transferred to the cofolding approach. Hence, it is also possible to calculate the partition function of all joint structures, as well as base pair probabilities (intramolecular as well as base pairs between the two sequences) using a variant of McCaskill's approach.[57] The partition function *Z(S)* for a sequence *S* (which might be composed of two sequences joined with a linker symbol) is the sum of all Boltzmann-weighted energies of all structures *R* the sequence *S* can take, i.e.,

$$Z(S) = \Sigma_{R \text{ structure of } S} e^{-E(S)/RT}$$

Once the partition function can be calculated, as in the case of RNA folding, then the Boltzmann probability of a specific structure can be calculated by

$$e^{-E(S)/RT}/Z$$

Even more importantly, the probability of a base pair (*i; j*) (where *i, j* are positions in either of the sequences) can be calculated using a modification of the partition function computation to sum up the Boltzmann-weighted energies for all structures that contain the base pair, i.e., to calculate

$$Z^{(i,j)} = \Sigma_{R \text{ contains } (i,j)} e^{-E(S)/RT};$$

The probability of the base pair (*i,j*) is then given by $Z^{(i,j)}/Z$. Furthermore, in the case of two interacting molecules A and B where one can calculate the partition functions for the single molecules, for the homo-dimers (A bound to A and B bound to B) and for the hetero-dimer as this is the case for the *concatenation approaches*, it is possible to calculate concentration-dependent melting temperatures, which can then be compared to experimental data.

An interesting application of the concatenation approach is presented in Cao et al. (2009)[58] and Zhao et al. (2008).[59] It addresses the problem that most practical approaches for interaction prediction restrict the search to a region around the start codon. However, many different settings concerning the length and position of this region are also possible. Here a classification method was used to solve this problem. First, multiple overlapping regions were considered, and the joint minimum free energy structure for all the overlapping regions was calculated using a simple *concatenation approach* (without special treatment of the linker sequence). The minimum free energy structure of a region is then used to extract different features such as percent composition of bases in interior loops, bulge loops, etc. In addition, sequence features like percentage of A + U bases (since Hfq is supposed to bind to AU-rich regions) were used. In total, 10 features for 1,000 overlapping sequences in the region +30 to -30 around the start codon were computed, giving rise to a secondary structure profile. These features were then used to train two classifiers (based on Naive-Bayes and SVM) on a data set of 46 positive samples and 86 negative samples.

Handling accessibility in mRNA-sRNA interactions. Because the concatenation approaches are predicting joint structures that are nested, they cannot handle important structural elements like double kissing hairpins. For that reason, another class of approaches have been introduced that can handle this class of interactions. The basic idea here is not to predict a single joint structure (or an ensemble of joint structures) but to investigate first the ensemble properties of the single sequences that are important for a putative interaction. Basically, an interaction site must be accessible (i.e., not covered by intra-molecular base pairs) because the positions in the interaction site will be bound by the interaction partner. Thus, for any two positions

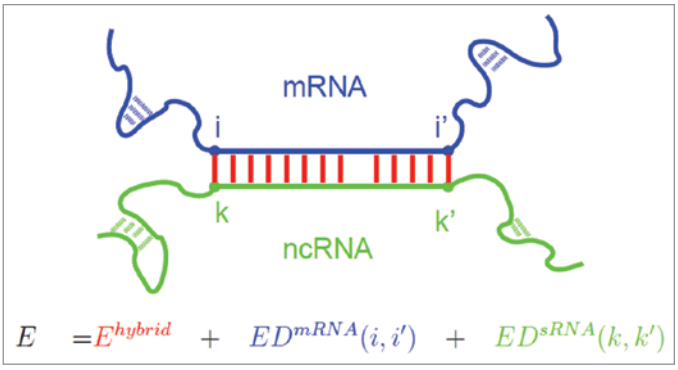

Figure 4. Evaluation of an interaction in RNAup and IntaRNA. The ED-values are precalculated for all possible regions in both sequences.

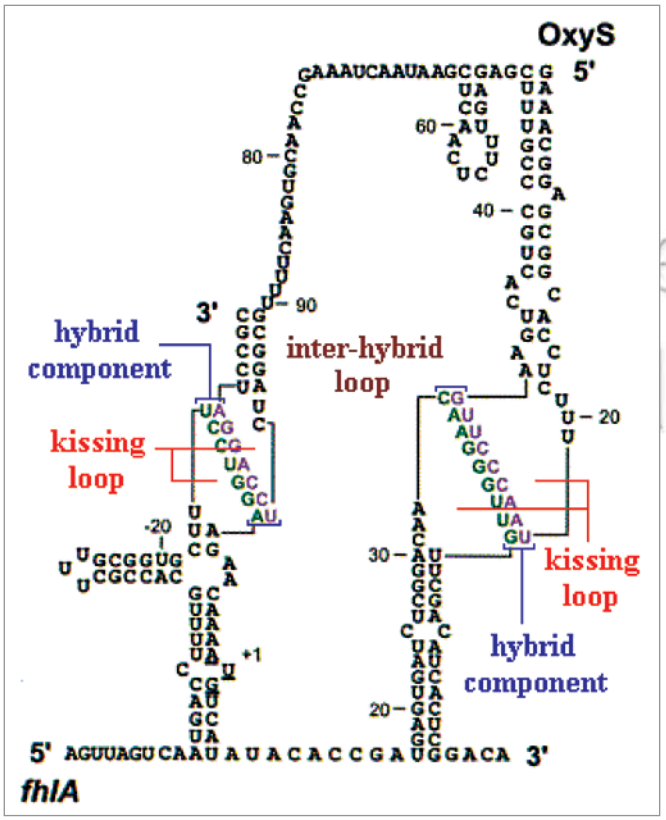

Figure 5. Interaction components of OxyS and fhlA as presented in Argaman et al. (2000).[76]

$a < b$ in a sequence, one computes the energy that is required to make the sequence stretch between a and b free of intra-molecular base pairs. Then, one calculates the partition function $Z^{sg(a,b)}$ for the ensemble of structures that leave the putative interaction site single-stranded (**Fig. 3**). Next, one calculates the ensemble energy by the formula $E^{sg(a,b)} = -RT\ ln(Z^{sg(a,b)})$. Defining the energy of the ensemble of all structures by $E_{all} = -RT\ ln(Z)$, where Z is the total partition function, we get the energy $ED(sg(a,b))$ that is required to make the interaction site accessible as

$$ED(a,b) = E^{sg(a,b)} - E_{all}$$

Note that the above term is positive and thus can be considered a penalty. Now, approaches like RNAup[60] and IntaRNA[61] use pre-calculated ED-values for all possible interaction regions to calculate a combined energy of the ED-values and the energy given by the duplex. Thus, an interaction of two regions *i..i'* of the first sequence with a region *k..k'* of the second region is evaluated as shown in **Figure 4**. The ED-values for all regions in one sequence can be pre-calculated with basically the same complexity as the calculation of base pair probabilities in normal RNA folding using the RNAplfold approach.[62,63] For the combined energy E, the recursion is basically similar to the entirely sequence-based approach in RNAhybrid, with the exception that one has to do it for all possible right end points separately, since one needs to know the complete regions for determining the ED-values. This approach, which is used in RNAup, leads to a quadratic overhead depending on the maximal length of interaction site considered. In IntaRNA, this overhead is avoided by applying a heuristic approach. Nevertheless, the prediction quality of RNAup and IntaRNA are basically equal because IntaRNA uses an additional seed condition.

When comparing the concatenation approaches (RNAcofold, PairRNA and NUpack) and the approaches working with accessibility (RNAup and IntaRNA), both restrict the set of joint structures that are taken into account. In the case of the concatenation approaches, one can predict only joint structures where the interaction arcs (i.e., the base pairs between two RNAs) are not covered by intra-molecular base pairs. An equivalent condition is that the interaction arcs may only occur at external positions. A position k is external in a structure if there is no base pair (i, j) in the structure that covers k (i.e., where $i < k < j$). Given a joint structure, the external positions are those that are external in the two substructures generated by restricting the joint structure to the single sequences. The approaches using accessibility, on the other hand, assume a single interaction site, which may not contain any intra-molecular base pair or end of an intra-molecular base pair. The reason for this restriction is simply that the unrestricted problem (i.e., finding the best joint structure of two interacting RNAs without any restriction on the type of structures) is computationally a very hard problem. Just recently, it was shown in Alkan et al. (2006)[64] that the general problem is NP-complete, which means in practice that an exact algorithm would require exponential time [The precise definition of NP-complete is more complex. NP is a class of problems that are currently believed to be different from the class P of problems that can be solved in polynomial time. Unless NP = P (which is believed to be very unlikely), there cannot be an algorithm that exactly solves the general interaction problem in polynomial time for all instances. However, there might be algorithms that solve the problem in reasonable time for most practical instances]. However, there are known classes of interactions like the interaction of OxyS and fhlA, which have two or more kissing hairpins (**Fig. 5**). This kind of interaction cannot be predicted by the concatenation approaches, nor can it be predicted by approaches

using the accessibility of one single interaction site. For that reason, new methods have been introduced that extend the class of allowed joint structures. The IRIS tool[65] introduced a new recursive scheme which allowed for the first time considering more than one kissing hairpin. It uses an energy model that maximizes the number of base pairs. Then, the extension to a more realistic energy model was considered,[64] inspired by the standard nearest neighbor energy model of single RNA sequence folding.

Furthermore, a precise definition of the class of structures treated by each approach was given.[64] Both approaches can handle the OxyS-*fhlA* interaction, and both approaches predict a single structure with the minimum free energy (the structure with the maximal number of base pairs, in the case of IRIS). However, as already observed in the folding of a single RNA, the MFE structure is often wrong. The standard way to overcome this problem is to use a partition function variant, as already described above, for the concatenation approaches. Since one has to calculate the sum over all possible joint structures, it is necessary to reformulate the recursion equations such that every joint structure is decomposed in a unique way. This problem was solved independently by Chitsaz et al. (2009)[66] and Huang et al. (2009).[67] Thus, both approaches allow calculation of important quantities like melting temperatures and base pair probabilities. As demonstrated,[66] the melting temperatures calculated by the algorithm are in good agreement with the experimentally measured ones, as exemplified in the case of OxyS:fhlA interaction for the wild-type and three mutated constructs. The above described tools for the prediction of a joint structure still have a very high computational complexity [the computation time is on the order of $O(n^6)$, where n is the length of the input sequence(s)]. Thus, there were attempts to reduce this complexity by considering an approximation to the original problem. A very intuitive way is to use accessibility (a.k.a. RNAup/IntaRNA), but allowing the use of more than one interaction site. The ED-value for measuring the energy required to make a site accessible can be calculated from the probability that this site is single-stranded. Now it is immediately clear that these probabilities are not independent for different interaction sites. Thus, conditional probabilities have to be used instead. Although this problem is deemed to be too complex to be calculated, a Bayesian approximation of these conditional probabilities was introduced by Chitsaz et al. (2009)[68] and Salari et al. (2009).[69] Their approximations allowed a fast calculation of these conditional probabilities, and resulted in a fast heuristic method to predict the specific (multiple) binding sites of two interacting RNAs.

Concluding Remarks

In recent years, the comparative genomics-based prediction of sRNA genes has become a standard method used to search for such genes within bacterial genomes (see **Table 1** for overview). However, comparison to transcriptomic datasets suggests that despite the impressively high number of correctly predicted candidates, diverse sRNA genes have still been excluded from these predictions, whatever the scope or the approach taken. Although none of the existing sRNA gene finders is able to identify all the experimentally validated sRNA genes, they provide a fairly good starting point for an analysis of the sRNA complement in a prokaryotic genome.

The current focus is on developing tools to correctly predicting sRNA targets. Several such tools have been designed (see **Table 2** for overview) but there are still considerable developments to be made. This problem is even more complex than that of finding sRNAs due to several reasons. First, target interactions show surprising variability. Second, experimental testing of target prediction is still a very laborious task. Third, the problem of RNA interaction prediction is computationally complex, requiring restrictions in the respective models. This leads to the introduction of a variety of procedures. Up to now, four prediction models have been presented, of increasing complexity and prediction quality (**Table 2**), and substantial improvements are likely to occur in the near future. Once sRNAs and targets are known, the next challenge will be the successful integration of regulatory RNA in the existing models of regulatory networks. This would allow the unraveling of pathways that involve sRNA-induced regulation. It would also be a further important step for the functional characterization of the many found sRNAs.

Acknowledgements

We wish to thank for their continuous and generous support the German Federal Ministry of Education and Research (BMBF grant 0313921 to the "Freiburg Initiative in Systems Biology") and the German Research Foundation (DFG) Focus program SPP1258 "Sensory and regulatory RNAs in Prokaryotes" (grants HE 2544/4-1 and BA 2168/2-1).

References

1. Benito Y, Kolb FA, Romby P, Lina G, Etienne J, Vandenesch F. Probing the structure of RNAIII, the *Staphylococcus aureus agr* regulatory RNA, and identification of the RNA domain involved in repression of protein A expression. RNA 2000; 6:668-79.
2. Boisset S, Geissmann T, Huntzinger E, Fechter P, Bendridi N, Possedko M, et al. *Staphylococcus aureus* RNAIII coordinately represses the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism. Genes Dev 2007; 21:1353-66.
3. Sharma CM, Vogel J. Experimental approaches for the discovery and characterization of regulatory small RNA. Curr Opin Microbiol 2009; 12:536-46.
4. Vogel J, Sharma CM. How to find small non-coding RNAs in bacteria. Biol Chem 2005; 386:1219-38.
5. Vogel J, Wagner EG. Target identification of small noncoding RNAs in bacteria. Curr Opin Microbiol 2007; 10:262-70.
6. Altuvia S. Identification of bacterial small noncoding RNAs: experimental approaches. Curr Opin Microbiol 2007; 10:257-61.
7. Hüttenhofer A, Vogel J. Experimental approaches to identify non-coding RNAs. Nucleic Acids Res 2006; 34:635-46.
8. Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA 2005; 102:2454-9.
9. Rivas E. Evolutionary models for insertions and deletions in a probabilistic modeling framework. BMC Bioinformatics 2005; 6:63.
10. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, et al. Identification and classification of conserved RNA secondary structures in the human genome. PLoS Comput Biol 2006; 2:33.
11. Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S. Identification of novel small RNAs using comparative genomics and microarrays. Genes Devel 2001; 15:1637-51.
12. Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EG, Margalit H, et al. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. Curr Biol 2001; 11:941-50.
13. Rivas E, Klein RJ, Jones TA, Eddy SR. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. Curr Biol 2001; 11:1369-73.

14. Washietl S, Hofacker IL. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. J Mol Biol 2004; 342:19-30.
15. Axmann IM, Kensche P, Vogel J, Kohl S, Herzel H, Hess WR. Identification of cyanobacterial non-coding RNAs by comparative genome analysis. Genome Biol 2005; 6:73.
16. Axmann IM, Holtzendorff J, Voss B, Kensche P, Hess WR. Two distinct types of 6S RNA in Prochlorococcus. Gene 2007; 406:69-78.
17. Voss B, Georg J, Schön V, Ude S, Hess WR. Biocomputational prediction of non-coding RNAs in model cyanobacteria. BMC Genomics 2009; 10:123.
18. Gierga G, Voss B, Hess WR. The Yfr2 ncRNA family, a group of abundant RNA molecules widely conserved in cyanobacteria. RNA Biol 2009; 6:222-7.
19. Rose D, Hertel J, Reiche K, Stadler PF, Hackermuller J. NcDNAlign: plausible multiple alignments of non-protein-coding genomic sequences. Genomics 2008; 92:65-74.
20. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res 2004; 14:708-15.
21. Sonnleitner E, Sorger-Domenigg T, Madej MJ, Findeiss S, Hackermuller J, Huttenhofer A, et al. Detection of small RNAs in *Pseudomonas aeruginosa* by RNomics and structure-based bioinformatic tools. Microbiology 2008; 154:3175-87.
22. Chen S, Lesnik EA, Hall TA, Sampath R, Griffey RH, Ecker DJ, et al. A bioinformatics based approach to discover small RNA genes in the *Escherichia coli* genome. Biosystems 2002; 65:157-77.
23. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. RNAMotif, an RNA secondary structure definition and search algorithm. Nucleic Acids Res 2001; 29:4724-35.
24. Voss B, Gierga G, Axmann IM, Hess WR. A motif-based search in bacterial genomes identifies the ortholog of the small RNA Yfr1 in all lineages of cyanobacteria. BMC Genomics 2007; 8:375.
25. Lenz DH, Mok KC, Lilley BN, Kulkarni RV, Wingreen NS, Bassler BL. The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in *Vibrio harveyi* and *Vibrio cholerae*. Cell 2004; 118:69-82.
26. Wilderman PJ, Sowa NA, FitzGerald DJ, FitzGerald PC, Gottesman S, Ochsner UA, et al. Identification of tandem duplicate regulatory small RNAs in *Pseudomonas aeruginosa* involved in iron homeostasis. Proc Natl Acad Sci USA 2004; 101:9792-7.
27. Halfmann A, Kovacs M, Hakenbeck R, Bruckner R. Identification of the genes directly controlled by the response regulator CiaR in *Streptococcus pneumoniae*: five out of 15 promoters drive expression of small non-coding RNAs. Mol Microbiol 2007; 66:110-26.
28. Livny J, Fogel MA, Davis BM, Waldor MK. sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. Nucleic Acids Res 2005; 33:4096-105.
29. Livny J, Brencic A, Lory S, Waldor MK. Identification of 17 *Pseudomonas aeruginosa* sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2. Nucleic Acids Res 2006; 34:3484-93.
30. Swiercz JP, Hindra, Bobek J, Bobek J, Haiser HJ, Di Berardo C, et al. Small non-coding RNAs in *Streptomyces coelicolor*. Nucleic Acids Res 2008; 36:7240-51.
31. Livny J. Efficient annotation of bacterial genomes for small, noncoding RNAs using the integrative computational tool sRNAPredict2. Methods Mol Biol 2007; 395:475-88.
32. Livny J, Teonadi H, Livny M, Waldor MK. High-throughput, kingdom-wide prediction and annotation of bacterial non-coding RNAs. PLoS One 2008; 3:3197.
33. Yachie N, Numata K, Saito R, Kanai A, Tomita M. Prediction of non-coding and antisense RNA genes in *Escherichia coli* with Gapped Markov Model. Gene 2006; 372:171-81.
34. Georg J, Voss B, Scholz I, Mitschke J, Wilde A, Hess WR. Evidence for a major role of antisense RNAs in cyanobacterial gene regulation. Mol Sys Biol 2009; 5:305.
35. Wassarman KM, Zhang A, Storz G. Small RNAs in *Escherichia coli*. Trends Microbiol 1999; 7:37-45.
36. Carter RJ, Dubchak I, Holbrook SR. A computational approach to identify genes for functional RNAs in genomic sequences. Nucleic Acids Res 2001; 29:3928-38.
37. Saetrom P, Sneve R, Kristiansen KI, Snove O Jr, Grunfeld T, Rognes T, et al. Predicting non-coding RNA genes in *Escherichia coli* with boosted genetic programming. Nucleic Acids Res 2005; 33:3263-70.
38. Pichon C, Felden B. Intergenic sequence inspector: searching and identifying bacterial RNAs. Bioinformatics 2003; 19:1707-9.
39. Pichon C, Felden B. Small RNA genes expressed from *Staphylococcus aureus* genomic and pathogenicity islands with specific expression among pathogenic strains. Proc Natl Acad Sci USA 2005; 102:14249-54.
40. Pichon C, Felden B. Small RNA gene identification and mRNA target predictions in bacteria. Bioinformatics 2008; 24:2807-13.
41. Baek D, Villen J, Shin C, Camargo FD, Gygi SP, Bartel DP. The impact of microRNAs on protein output. Nature 2008; 455:64-71.
42. Selbach M, Schwanhausser B, Thierfelder N, Fang Z, Khanin R, Rajewsky N. Widespread changes in protein synthesis induced by microRNAs. Nature 2008; 455:58-63.
43. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990; 215:403-10.
44. Gerlach W, Giegerich R. GUUGle: a utility for fast exact matching under RNA complementary rules including G-U base pairing. Bioinformatics 2006; 22:762-4.
45. Tjaden B, Goodwin SS, Opdyke JA, Guillier M, Fu DX, Gottesman S, et al. Target prediction for small, noncoding RNAs in bacteria. Nucleic Acids Res 2006; 34:2791-802.
46. Mandin P, Repoila F, Vergassola M, Geissmann T, Cossart P. Identification of new noncoding RNAs in *Listeria monocytogenes* and prediction of mRNA targets. Nucleic Acids Res 2007; 35:962-74.
47. Glaser P, Frangeul L, Buchrieser C, Rusniok C, Amend A, Baquero F, et al. Comparative genomics of Listeria species. Science 2001; 294:849-52.
48. Zuker M. Prediction of RNA secondary structure by energy minimization. Meth Mol Biol 1994; 25:267-94.
49. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer S, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatshefte Chemie 1994; 125:167-88.
50. Rehmsmeier M, Steffen P, Hochsmann M, Giegerich R. Fast and effective prediction of microRNA/target duplexes. RNA 2004; 10:1507-17.
51. Mathews D, Sabina J, Zuker M, Turner D. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 1999; 288:911-40.
52. Tafer H, Hofacker IL. RNAplex: a fast tool for RNA-RNA interaction search. Bioinformatics 2008; 24:2657-63.
53. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 1990; 87:2264-8.
54. Andronescu M, Zhang ZC, Condon A. Secondary structure prediction of interacting RNA molecules. J Mol Biol 2005; 345:987-1001.
55. Bernhart SH, Tafer H, Muckstein U, Flamm C, Stadler PF, Hofacker IL. Partition function and base pairing probabilities of RNA heterodimers. Algorithms Mol Biol 2006; 1:3.
56. Dirks RM, Bois JS, Schaeffer JM, Winfree E, Pierce NA. Thermodynamic Analysis of Interacting Nucleic Acid Strands. SIAM Review 2007; 49:65-88.
57. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers 1990; 29:1105-19.
58. Cao Y, Zhao Y, Cha L, Ying X, Wang L, Shao N, et al. sRNATarget: a web server for prediction of bacterial sRNA targets. Bioinformation 2009; 3:364-6.
59. Zhao Y, Li H, Hou Y, Cha L, Cao Y, Wang L, et al. Construction of two mathematical models for prediction of bacterial sRNA targets. Biochem Biophys Res Commun 2008; 372:346-50.
60. Muckstein U, Tafer H, Hackermuller J, Bernhart SH, Stadler PF, Hofacker IL. Thermodynamics of RNA-RNA binding. Bioinformatics 2006; 22:1177-82.
61. Busch A, Richter AS, Backofen R. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. Bioinformatics 2008; 24:2849-56.
62. Bernhart SH, Hofacker IL, Stadler PF. Local RNA base pairing probabilities in large sequences. Bioinformatics 2006; 22:614-5.
63. Bompfunewerer AF, Backofen R, Bernhart SH, Hertel J, Hofacker IL, Stadler PF, et al. Variations on RNA folding and alignment: lessons from Benasque. J Math Biol 2008; 56:129-44.
64. Alkan C, Karakoc E, Nadeau JH, Sahinalp SC, Zhang K. RNA-RNA interaction prediction and antisense RNA target search. J Comput Biol 2006; 13:267-82.
65. Pervouchine DD. IRIS: intermolecular RNA interaction search. Genome Inform 2004; 15:92-101.
66. Chitsaz H, Salari R, Sahinalp SC, Backofen R. A partition function algorithm for interacting nucleic acid strands. Bioinformatics 2009; 25:365-73.
67. Huang FW, Qin J, Reidys CM, Stadler PF. Partition Function and Base Pairing Probabilities for RNA-RNA Interaction Prediction. Bioinformatics 2009; 25:2646-54.
68. Chitsaz H, Backofen R, Sahinalp SC. biRNA: Fast RNA-RNA binding sites prediction. In: Salzberg S, Warnow T, eds. Proc of the 9th Workshop on Algorithms in Bioinformatics (WABI) 2009; 25-36.
69. Salari R, Backofen R, Sahinalp SC. Fast prediction of RNA-RNA Interaction. In: Salzberg S, Warnow T, eds. Proc of the 9th Workshop on Algorithms in Bioinformatics (WABI) 2009; 261-72.
70. Tran TT, Zhou F, Marshburn S, Stead M, Kushner SR, Xu Y. De Novo computational prediction of non-coding RNA genes in prokaryotic genomes. Bioinformatics 2009; 25:2897-905.
71. González N, Heeb S, Valverde C, Kay E, Reimmann C, Junier T, et al. Genome-wide search reveals a novel GacA-regulated small RNA in Pseudomonas species. BMC Genomics 2008; 9:167.
72. Ulvé VM, Sevin EW, Chéron A, Barloy-Hubler F. Identification of chromosomal alpha-proteobacterial small RNAs by comparative genome analysis and detection in *Sinorhizobium meliloti* strain 1021. BMC Genomics 2007; 8:467.
73. del Val C, Rivas E, Torres-Quesada O, Toro N, Jimenez-Zurdo JI. Identification of differentially expressed small non-coding RNAs in the legume endosymbiont *Sinorhizobium meliloti* by comparative genomics. Mol Microbiol 2007; 66:1080-91.
74. Silvaggi JM, Perkins JB, Losick R. Genes for small, noncoding RNAs under sporulation control in *Bacillus subtilis*. J Bacteriol 2006; 188:532-41.
75. Tjaden B. TargetRNA: a tool for predicting targets of small RNA action in bacteria. Nucleic Acids Res 2008; 36:109-13.
76. Argaman L, 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; 300:1101-12.