Expanding the repertoire of bacterial (non-)coding RNAs

Der Fakultät für Mathematik und Informatik
der Universität Leipzig
eingereichte

Dissertation

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM

im Fachgebiet
Informatik

vorgelegt
von Dipl.-Inf. Sven Findeiß

geboren am 16. September 1982 in Zwickau

Leipzig, den 6. Februar 2011

Contents

Abstract

The detection of non-protein-coding RNA (ncRNA) genes in bacteria and their diverse regulatory mode of action moved the experimental and bio-computational analysis of ncRNAs into the focus of attention. Regulatory ncRNA transcripts are not translated to proteins but function directly on the RNA level. These typically small RNAs have been found to be involved in diverse processes such as (post-)transcriptional regulation and modification, translation, protein translocation, protein degradation and sequestration.

Bacterial ncRNAs either arise from independent primary transcripts or their mature sequence is generated via processing from a precursor. Besides these autonomous transcripts, RNA regulators (e.g. riboswitches and RNA thermometers) also form chimera with protein-coding sequences. These structured regulatory elements are encoded within the messenger RNA and directly regulate the expression of their "host" gene.

The quality and completeness of genome annotation is essential for all subsequent analyses. In contrast to protein-coding genes ncRNAs lack clear statistical signals on the sequence level. Thus, sophisticated tools have been developed to automatically identify ncRNA genes. Unfortunately, these tools are not part of generic genome annotation pipelines and therefore computational searches for known ncRNA genes are the starting point of each study. Moreover, prokaryotic genome annotation lacks essential features of protein-coding genes. Many known ncRNAs regulate translation via base-pairing to the 5' UTR (untranslated region) of mRNA transcripts. Eukaryotic 5' UTRs have been routinely annotated by sequencing of ESTs (expressed sequence tags) for more than a decade. Only recently, experimental setups have been developed to systematically identify these elements on a genome-wide scale in prokaryotes.

The first part of this thesis, describes three experimental surveys of exploratory field studies to analyze transcript organization in pathogenic bacteria. To identify ncRNAs in *Pseudomonas aeruginosa* we used a combination of an experimental RNomics approach and ncRNA prediction. Besides already known ncRNAs we identified and validated the expression of six novel RNA genes.

Global detection of transcripts by next generation RNA sequencing techniques unraveled an unexpectedly complex transcript organization in many bacteria. These ultra high-throughput methods give us the appealing opportunity to analyze the complete RNA output of any species at once. The development of the differential RNA sequencing (dRNA-seq) approach enabled us to analyze the primary transcriptome of *Helicobacter pylori* and *Xanthomonas campestris*. For the first time we generated a comprehensive and precise transcription start site (TSS) map for both species and provide a general framework for the analysis of dRNA-seq data. Focusing on computer-aided analysis we developed new tools to annotate TSS, detect small protein-coding genes and to infer homology of newly detected transcripts. We discovered hundreds of TSS in intergenic regions, upstream of protein-coding genes, within operons and antisense to annotated genes. Analysis of 5' UTRs (spanning from the TSS to the start codon of the adjacent protein-coding gene) revealed an unexpected size diversity ranging from zero to several hundred nucleotides. We identified and validated the expression of about 60 and about 20 ncRNA candidates in *Helicobacter* and *Xanthomonas*, respectively. Among these ncRNA candidates we found several small protein-coding genes that have previously evaded annotation in both species. We showed that the combination of dRNA-seq and computational analysis is a powerful method to examine prokaryotic transcriptomes.

Experimental setups are time consuming and often combined with huge costs. Another limitation of experimental approaches is that genes which are expressed in specific developmental stages or stress conditions are likely to be missed. Bioinformatic tools build an alternative to overcome such restraints. General approaches usually depend on comparative genomic data and evolutionary signatures are used to analyze the (non-)coding potential of multiple sequence alignments. In the second part of my thesis we present our major update of the widely used ncRNA gene finder `RNAz` and introduce `RNAcode`, an efficient tool to asses local protein-coding potential of genomic regions.

`RNAz` has been successfully used to identify structured RNA elements in all domains of life. However, our own experience and the user feedback not only demonstrated the applicability of the `RNAz` approach, but also helped us to identify limitations of the current implementation. Using a much larger training set and a new classification model we significantly improved the prediction accuracy of `RNAz`.

During transcriptome analysis we repeatedly identified small protein-coding genes that have not been annotated so far. Only a few of those genes are known to date and standard protein-coding gene finding tools suffer from the lack of training data. To avoid an excess of false positive predictions, gene finding software is usually run with an arbitrary cutoff of 40-50 amino acids and therefore misses the small sized protein-coding genes. We have implemented

RNAcode which is optimized for emerging applications not covered by standard protein-coding gene annotation software. In addition to complementing classical protein gene annotation, a major field of application of RNAcode is the functional classification of transcribed regions. RNA sequencing analyses are likely to falsely report transcript fragments (e.g. mRNA degradation products) as non-coding. Hence, an evaluation of the protein-coding potential of these fragments is an essential task. RNAcode reports local regions of high coding potential instead of complete protein-coding genes. A training on known protein-coding sequences is not necessary and RNAcode can therefore be applied to any species. We showed this with our analysis of the *Escherichia coli* genome where the current annotation could be accurately reproduced. We furthermore identified novel small protein-coding genes with RNAcode in this extensively studied genome. Using transcriptome and proteome data we found compelling evidence that several of the identified candidates are *bona fide* proteins.

In summary, this thesis clearly demonstrates that bioinformatic methods are mandatory to analyze the huge amount of transcriptome data and to identify novel (non-)coding RNA genes. With the major update of RNAz and the implementation of RNAcode we contributed to complete the repertoire of gene finding software which will help to unearth hidden treasures of the RNA World.

Acknowledgment

An dieser Stelle, möchte ich mich bei all denen bedanken, die zum Gelingen dieser Arbeit beigetragen haben.

Ein besonderer Dank geht an Prof. Peter F. Stadler für die Möglichkeit, frei und selbstständig an einem sehr interessanten Themengebiet arbeiten zu dürfen. Danken möchte ich dir auch für deine wohltuende, unkonventionelle Art und die damit verbundenen interessanten Diskussion beim Bier.

Für die gute Zusammenarbeit vielen Dank an: Jana Hertel, Manja Marz, Wolfgang Otto, Christine Schulz, Cynthia M. Sharma, Kristin Reiche, Cornelius Schmidtke, Andreas R. Gruber, Stefan Washietl, Stefan Kalkhof, David Langenberger, Alexander Donath, Stephan Bernhart, Steve Hoffmann, Marcus Lechner, Lydia Steiner, Sonja J. Prohaska, Jens Steuck sowie allen Co-Autoren die zum Gelingen dieser Arbeit beigetragen haben. Für das Korrekturlesen dieser Arbeit und die vielen hilfreichen Kommentare bedanke ich mich sehr bei Jana Hertel, Manja Marz und Stephan Bernhart.

Danken möchte ich natürlich auch all den Leipziger Bioinformatikern für den Spaß bei der Arbeit und all die leckeren Kuchen, die wir gemeinsam verkosten durften.

Außerdem möchte ich an dieser Stelle der “guten Seele” Petra Pregel für die tatkräftige Unterstützung bei der Planung von Dienstreisen, der gemeinsamen Organisation der Herbstseminare und für die schöne und abwechslungsreiche Zeit danken.

Ein großes Dankeschön an meine Familie, Mitbewohner und Freunde, für die vielen schönen gemeinsamen Stunden, die Ablenkung vom Altagstrott und dafür dass ihr mir immer wieder Mut und Rückhalt gebt.

!DANKE!

This thesis is based on the following publications:

- Sonnleitner E, Sorger-Domenigg T, Madej MJ, Findeiß S, Hackermüller J, Hüttenhofer A, Stadler PF, Bläsi U and Moll I. *Detection of small RNAs in Pseudomonas aeruginosa by RNomics and structure-based bioinformatic tools.* Microbiology, 2008 Oct; 154(10): 3175–3187
- Donath A, Findeiß S, Hertel J, Marz M, Otto W, Schulz C, Stadler PF and Wirth S. *Noncoding RNA In Evolutionary Genomics and Systems Biology.* Editor Gustavo Caetano-Anollés; Wiley-Blackwell, Hoboken; 2010; 251-283
- Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiß S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, Stadler PF and Vogel J. *The primary transcriptome of the major human pathogen Helicobacter pylori.* Nature, 2010 Mar; 464(7286): 250-5.
- Schmidtke C, Findeiß S, Sharma CM, Kuhfuß J, Hoffmann S, Vogel J, Stadler PF and Bonas U. *Genome-wide automated differential transcriptome analysis of the plant pathogen Xanthomonas identifies sRNAs with putative virulence functions.* in preparation
- Findeiß S, Schmidtke C, Stadler PF and Bonas U. *A novel family of plasmid-transferred anti-sense ncRNAs.* RNA Biol, 2010 Mar; 7(2): 120-4.
- Gruber AR, Findeiß S, Washietl S, Hofacker IL and Stadler PF. *RNAZ 2.0: Improved noncoding RNA detection.* Pac Symp Biocomput, 2010 Jan; 15:69-79.
- Washietl S, Findeiß S, Müller S, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF and Goldman N. *RNAcode: robust discrimination of coding and noncoding RNAs in comparative sequence data.* RNA, accepted
- Müller SA, Kohajda T, Findeiß S, Stadler PF, Washietl S, Kellis M, von Bergen M and Kalkhof S. *Optimization of Parameters for Coverage of Low Molecular Weight Proteins.* Anal Bioanal Chem, 2010 Dec; 398(7-8): 2867–2881.

Additional publications related to the presented studies in a more indirect way:

- Schilling D, Findeiß S, Richter AS, Taylor JA and Gerischer U. *The small RNA Aar in Acinetobacter baylyi: a putative regulator of amino acid metabolism.* Arch Microbiol, 2010 Sep; 192(9): 691-702
- Findeiß S, Langenberger D, Stadler PF and Hoffmann S. *Traces of Post-Transcriptional RNA Modifications in Deep Sequencing Data.* Biological Chemistry, accepted
- Lechner M, Findeiß S, Steiner L, Marz M, Stadler PF and Prohaska SJ. *Proteinortho: Detection of (Co-)Orthologs in Large-Scale Analysis* BMC Bioinformatics, submitted

1

The fascinating world of RNA

The term "RNA World" dates back to 1986 when Walter Gilbert proposed an RNA centered perspective of the origin of life [Gilbert 1986]. He described how the universe of today's known molecules might be originated from RNA. Particularly the idea that DNA, the universal information storage, and even proteins, the functional translation of DNA, have arisen from RNA raised a lot of discussions. The most prominent counterargument of the *RNA World* hypothesis is that a spontaneous emergence of complicated molecules such as RNA is highly unlikely. Hence, scientists struggle to show how ribonucleotides and from that RNA molecules could be synthesized. Only recently John D. Sutherland showed how ribonucleotides could have been assembled under plausible prebiotic conditions [Sutherland 2010]. Thus, a milestone to proof the *RNA World* hypothesis has been achieved by biochemical investigations.

Beside this fundamental research the term *RNA World* is also used to emphasize the great variety of known RNA molecules. Thus, the detection and functional characterization of RNA molecules gives another perspective onto the fascinating world of RNA. It has to be distinguished between protein-coding and non-protein-coding RNAs (ncRNAs). Protein-coding RNAs, better known as messenger RNAs (mRNAs), are transcribed from DNA and are subsequently translated into proteins. For a long time RNA was only seen as a template for the functional instance protein. However, ncRNAs that do not encode a protein template gain more and more interest. Well known ncRNAs are ribosomal RNA (rRNA) and transfer RNA (tRNA). Both essential for the process of protein synthesis in all domains of life. Other omnipresent ncRNA examples are RNase P RNA and the Signal Recognition Particle

(SRP) RNA. While the first plays a major role in the generation of mature 5'-ends of tRNAs, which are processed from a precursor sequences, the latter is involved in the co-translational trans-location of proteins through membranes. Interestingly, all mentioned ncRNAs act in combination with proteins. Both, the protein and the RNA component(s) are essential for the function of the complex.

As this thesis will mainly focus on bacteria some ubiquitous prokaryotic regulators should be mentioned here as well. One is the transfer-messenger RNA (tmRNA). As the name suggests, this RNA molecule exhibits two domains. The first domain has a tRNA like shape and by that the tmRNA can enter the ribosome. The nascent polypeptide is transferred to the tmRNA and the second domain serves as mRNA template. By this two step mechanism the tmRNA can free stalled ribosomes from damaged mRNA transcripts. Another prokaryotic housekeeping RNA is 6S. It has a conserved secondary structure that mimics an open promoter complex and RNA polymerase (RNAP) binds 6S RNA instead of DNA. Prokaryotes utilize this RNAP–6S RNA storage to overcome periods of starvation. Those few ncRNA examples already show how essential and complex regulation by non-protein-coding molecules is.

In fact most of the recently detected bacterial ncRNAs seem to be linage-specific without detectable homologs in other species. The majority of functionally analyzed ncRNAs were initially discovered in the well studied model organism *Escherichia coli*. The multi-target regulating GcvB is such an example [Urbanowski *et al.* 2000, Sharma *et al.* 2007]. Homologs of this ncRNA have been identified in enterobacteria such as *Salmonella typhimurium* and *Vibrio cholerae* only. Another example is the cyanobacterial functional RNA Yfr1. A motif-based search revealed Yfr1 orthologs only within cyanobacteria lineages [Voss *et al.* 2007]. Comparing ncRNA expression between *Mycobacteria* species identified pathogen specific RNA molecules. DiChiara *et al.* [2010] found several ncRNAs expressed in *Mycobaterium tuberculosis* which could be functioning in mediation of virulence. Thus, the presence of ncRNAs makes the difference between (non-)pathogenic strains.

The gain and loss of ncRNAs is an ongoing process. As recent studies suggest we are far from a comprehensive understanding of the *RNA World* at the moment. If we understand, however, the functional roles RNA molecules can have we might elucidate their role in the origin of life.

Organization of this thesis

Chapter 2 reviews the biological background of ncRNA mediated regulation in prokaryotes. This chapter, furthermore, summarizes state of the art experimental and computational approaches for ncRNA gene detection.

In Chapter 3 the application of experimental RNomics and next generation sequencing techniques for the identification of RNA transcripts in the three model organisms *Pseudomonas aeruginosa*, *Helicobacter pylori* and *Xanthomonas campestris* is described.

The question: "How to assess non-coding and protein-coding potential of comparative genomic data" is addressed in Chapter 4. There, I describe the major update of the ncRNA gene finder `RNAz` and outline the general `RNAcode`-approach.

Chapter 5 summarizes and discusses the outcome of the presented studies and gives a brief outlook.

2

Background

The majority of annotated prokaryotic genes are protein-coding*. On DNA level the complete template of the transcribed gene is encoded. It encompasses the 5' untranslated region (UTR), the open reading frame (ORF) and the 3' UTR. A gene spans from the so called transcription start site (TSS) to the 3' end of the terminator sequence, see Figure 2.1. Only the ORF serves as the nucleotide template of the protein and its length is usually dividable by three. An ORF, typically, begins with an ATG triplet which encodes the amino acid methionine and ends with a canonical stop codon (TAG, TGA or TAA). Both UTRs have regulatory functions. The 5' UTR comprises regulatory elements, e.g. the Shine-Dalgarno (SD) sequence which is essential for ribosome binding and the subsequent translation. The 3' UTR on the other hand is important for transcript stability. The promoter region is located upstream of the protein-coding gene. It involves regulatory sequences, e.g. -10 (Pribnow box) and -35 region, which guide the RNAP complex to the proximity of the TSS. The minimal core-enzyme of RNAP in *E. coli* is a multi-subunit complex of two α, β, β' subunits and an additive ω factor. For the DNA specific binding the so-called holoenzyme is formed by the addition of a specific σ factor. In *E. coli* at least seven σ factors are known . Out of this set σ^{70} regulates almost all of the 'housekeeping' genes within the exponential growth [Maeda *et al.* 2000].

For ncRNA genes the mechanisms of transcription initiation and termination are similar. Position of promoter elements and terminator sequences define the size of the transcribed gene, see Figure 2.1. However, these RNAs are often referred to as non-protein-coding RNAs as they do not contain an ORF. In contrast to protein-coding genes these molecules act on RNA level (see below). Several ncRNAs, e.g. 6S and tmRNA, are processed by RNase

*96% of *E. coli* genes listed in NCBI database are protein-coding

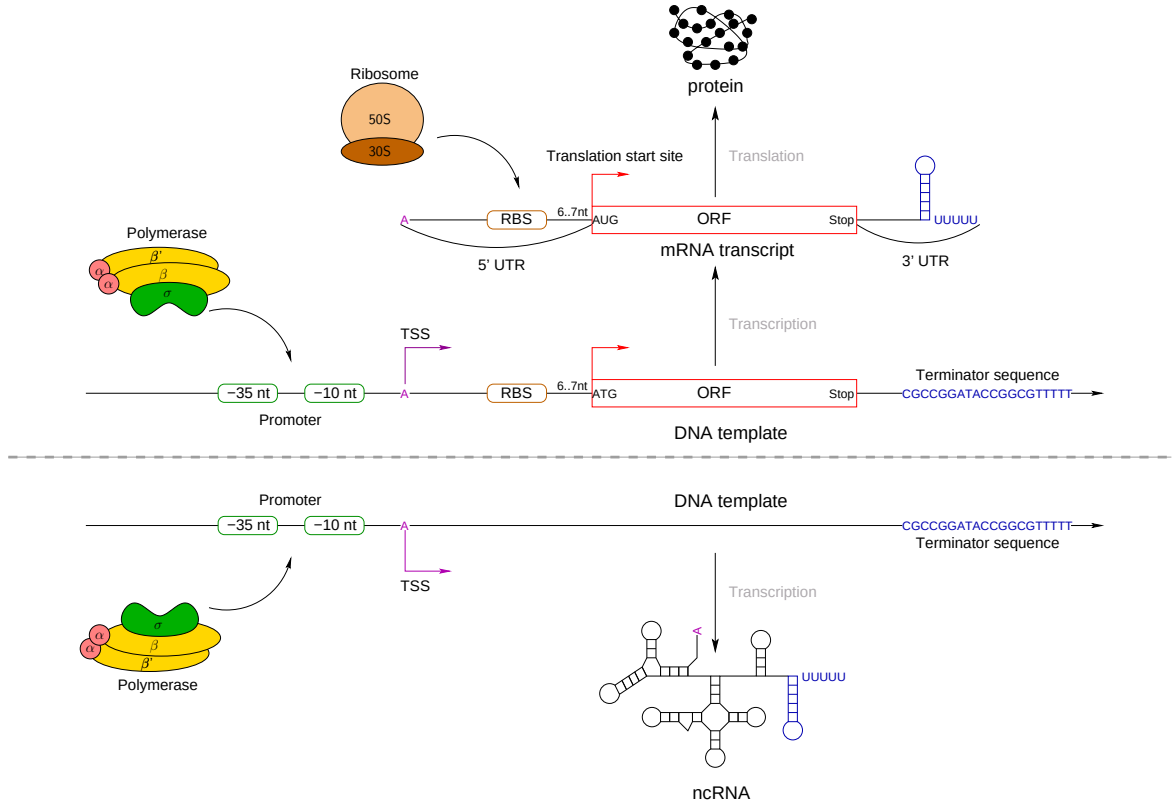


Figure 2.1. Schematic drawing of protein (top) and ncRNA synthesis (bottom). Specific sequence motifs, e.g. -35 and -10 region, in the promoter region guide the RNA Polymerase holoenzyme onto the DNA. The DNA sequence spanning from the TSS (transcription start site) to the end of terminator sequence is transcribed. In case of a protein-coding gene additional regulatory regions are encoded on DNA level and are copied to the mRNA transcript. The RBS (ribosome entry site) within the 5' UTR recruits the ribosome close to the translation start site which is the A of the canonical AUG start codon. An ORF (open reading frame) is the blueprint of the encoded protein. Starting with the AUG all coding base triplets are translated into amino acids until the stop codon. The translated protein forms the functional instance. An non-protein-coding gene is transcribed from its DNA template. Here, the functional instance is a (structured) ncRNA.

cleavage from a precursor sequence.

A special gene organization is the so called operon. In contrast to the above described monocystronic transcript organization encodes an operon a set of ORFs and/or ncRNAs which are simultaneously transcribed (polycistronic transcript). Hence, already one promoter and terminator sequence is sufficient to express a set of functional related genes at the same time.

2.1 Functional roles of RNA regulators

With respect to the location of their target, RNA regulators are often classified as *cis*- or *trans*-acting molecules. A *cis*-acting regulator is typically encoded within the same operon, gene or directly on the opposite strand (antisense) of its target gene. Whereas *trans*-acting means that the RNA regulator modulates the expression of a distant gene. I introduce these terms as they are often used in literature. However, such a classification becomes obsolete when typical *cis*-acting elements also regulate targets in *trans*. Only recently this has been shown for a riboswitch, a classical *cis*-acting regulator. Loh *et al.* [2009] evidenced the existence of a *trans*-acting riboswitch in *Listeria monocytogenes*. Furthermore, Han *et al.* [2010] discussed possible cross-recognition between different copies of the *Ibs-sib* systems, a classical *cis*-antisense regulated toxin-antitoxin module.

In the context of my thesis the term RNA regulator covers both independent ncRNAs and structured elements (e.g. riboswitches, RNA thermometer) that are encoded within mRNA transcripts. Non-coding RNA regulators are involved in various cellular processes: mRNA stability and translation, protein translocation, protein degradation and sequestration, transcriptional regulation as well as RNA processing and modification. Most known bacterial RNA regulators arise from independent genes and have a typical length of 50-600 nucleotides (nts) [Pichon & Felden 2008, Repoila & Darfeuille 2009]. Small RNA (sRNA) is therefore a commonly used synonym for these bacterial molecules. The principle modes of action are summarized in the following sections.

2.1.1 Translation regulation

The canonical model of sRNA mediated translation regulation is to control the access of the Shine-Dalgarno sequence of the respective target mRNA (Figure 2.2). Translation repression is typically achieved by sequestration of the Shine-Dalgarno sequence via imperfect base pairing between sRNA and mRNA. Well studied examples are MicA, MicC and MicF

sRNAs that regulate the expression outer membrane proteins, OmpA, OmpC and OmpF, respectively. Furthermore RybB, RseX and IpeX sRNAs are likely to use the same mechanism to regulate the expression of ompC (reviewed in [Guillier *et al.* 2006]). The mRNA of ompA is an additional target of RseX [Douchin *et al.* 2006]. Interestingly, in *V. cholerae* ompA is found to be regulated by the VrrA RNA which has been shown to use the canonical translation repression mechanism [Song *et al.* 2008]. Other sRNA-mRNA target pairs are OxyS-fhlA, Spot42-galK, DsrA-hns, SgrS-ptsG, SR1-ahrC, RNAIII-spa and many more as reviewed by Repoila & Darfeuille [2009]. The mechanism of translation inhibition by base pairing is also known for antisense RNA systems (reviewed by Brantl [2002]). Since the sRNA is encoded directly on the opposite strand of the target gene an intrinsic feature of these systems are stretches of complementary nucleotides between sRNA and mRNA. Toxin-antitoxin pairs are specific antisense systems (reviewed by Gerdes & Wagner [2007], Fozo *et al.* [2008a]). The toxin-encoding stable mRNA encodes a protein that rapidly leads to cell death unless its translation is suppressed by a short-lived small RNA. These typically plasmid-encoded modules prevent the growth of plasmid-free offspring thus ensuring the persistence of the plasmid in the population. *E. coli's* hok/Sok system is only one but possibly the best studied example. Bacterial mRNA stability and therefore the half-life of the molecule is strongly affected by the association with ribosomes [Deana & Belasco 2005]. When sRNA-mRNA complex formation inhibits ribosome binding RNase dependent degradation of the repressed target is quickly initiated (see Section 2.1.5).

Translation activation is typically supported by the opening of local Shine-Dalgarno blocking structures, see Figure 2.2. One well studied example is the activation of *E. coli's* major stress and stationary phase sigma factor, RpoS. The mRNA comprises an extreme long 5' UTR (600 nt) which folds into an translational inactive structure. Two sRNAs, DsrA and RprA, are known to initiate a refolding of rpoS and thereby activate translation of the mRNA. Although both sRNAs bind to the same region within the mRNA they act under different conditions. DsrA regulates at low growth temperatures whereas RprA is induced upon cell surface stress (reviewed by Repoila *et al.* [2003]). Other examples of sRNA-mRNA pairs where translation is activated upon base pair interaction are RyhB-shiA, GlmY/GmZ-gmlS and RNAIII-hla (recently reviewed by Repoila & Darfeuille [2009]). Antisense conformations are also known to activate translation. GadY RNA, for instance, is transcribed from the opposite strand of the *gadX* 3' UTR. Upon base pairing accumulation of the GadX protein has been observed [Opdyke *et al.* 2004]. Further studies showed that the polycistronic mRNA, gadXW, is rapidly cleaved after GadY binding [Tramonti *et al.* 2008]. The cleavage has a stabilizing effect on gadX mRNA and seems to be responsible for the observed GadX protein accumulation.

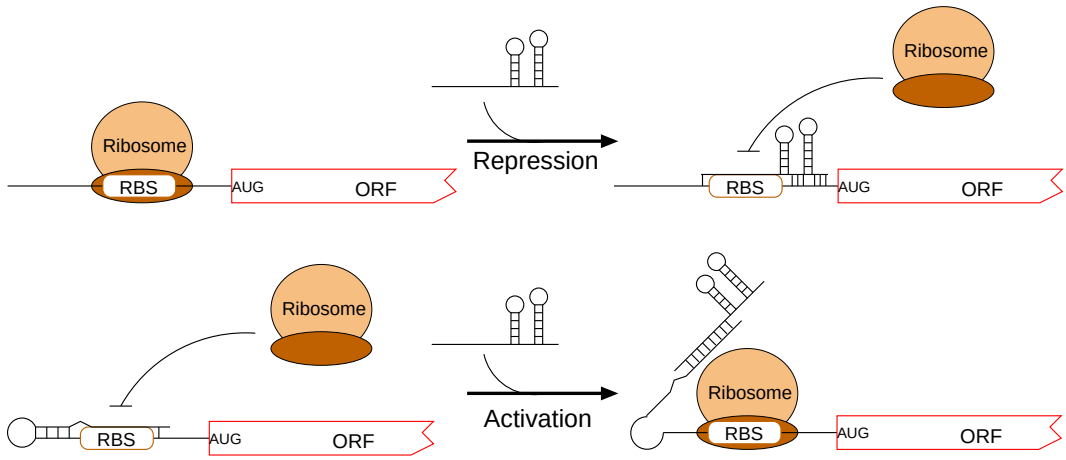


Figure 2.2. Small RNAs can either repress (top) or enhance (bottom) translation of their target. In both cases an (imperfect) sRNA-mRNA interaction regulates the accessibility of the ribosome binding site (RBS). Translation repression: In absence of the sRNA regulator the ribosome binds and translates the mRNA transcript. If the sRNA becomes available it binds close to the RBS and therefore inhibits ribosome binding. Translation activation: Under non active conditions the mRNA folds into an inhibitory structure (blocked RBS). By base pairing with a complementary region of the mRNA, the sRNA disrupts the inhibitory secondary structure. The ribosome gets access to the RBS and translation is initiated.

The fact that several sRNAs (e.g. RybB and DsrA) are listed as translational repressor and activator already indicates the complexity of sRNA mediated regulation. The expression of one mRNA can be modulated by several sRNAs as exemplified by ompC and its sRNA regulators MicC, RseX, RybB and Ipex. Furthermore is the opposite, one sRNA regulates different mRNA targets, true. GcvB, for instance, is known to regulate the expression of at least seven mRNA targets [Sharma *et al.* 2007]. Functional redundancy of four sRNAs, Qrr1-4, has been described in *Vibrio* species. The known targets luxO, hapR and vca0939 are regulated by each of the Qrr sRNAs [Svenningsen *et al.* 2009]. As exemplified by GadY, not only monocystronic mRNAs but also polycistronic transcripts are sRNA targets. Thus the expression of several genes in polycistronic transcripts can be regulated by a single sRNA molecule. It remains to be seen whether these examples are exceptions to the rule. However, they support the assumption that sRNA mediated translation regulation might be far more complex than previously expected.

2.1.2 Protein sequestration

Small RNAs regulate also non-mRNAs. The most prominent examples of sRNAs that directly interact with cellular proteins are summarized below. All three examples modulate protein activity by mimicking the structure of the actual RNA or DNA target.

6S RNA, one of the first known sRNAs, was detected and sequenced in the late 1960s [Hindley 1967, Brownlee 1971]. The ubiquitous RNA molecule interacts with the RNAP holoenzymes and therefore regulates the transcription of many "housekeeping" genes in bacteria. Generally one copy of the 6S RNA was found within almost all eubacteria and two copies are present in Gram-positive bacteria. Further analysis of the two copies in *Bacillus subtilis* revealed a distinct expression during different growth phases [Barrick *et al.* 2005]. Axmann *et al.* [2007] reported two alternative structural conformations of the 6S RNA molecule in *Cyanobacteria* species.

6S RNA has a well characterized secondary structure that folds into a three domain ('closing stem', 'central bubble' and 'terminal loop') long hairpin (Figure 2.3). In all known 6S RNAs the 'central bubble' is flanked by G-C rich stems [Barrick *et al.* 2005]. In *E. coli* bio-genesis of 6S RNA relies on co-ordinated transcription and cleavage of a dicystronic transcript [Hsu *et al.* 1985, sun Kim & Lee 2004].

The 6S RNA is linked with two functions: i) inhibition of σ^{70}-dependent transcription and ii) release of σ^{70}-RNAP complex during the outgrowth. The amount of 6S RNA molecules within the cells increases 10-fold from exponential to stationary growth and up to 75% of all

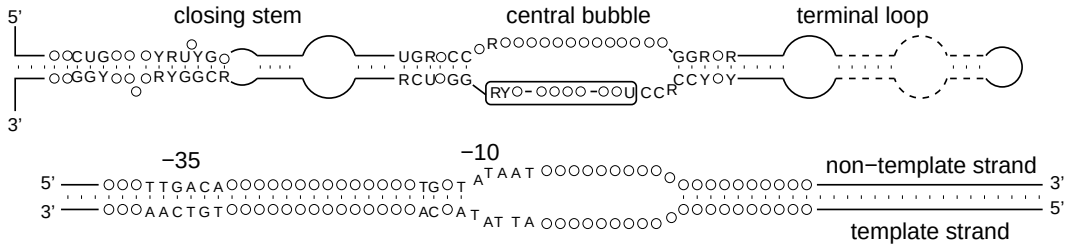


Figure 2.3. Comparison of 6S RNA consensus secondary structure (top) and an open promoter complex (bottom). 6S RNA consists of three structurally conserved domains, namely "closing stem", "central bubble" and "terminal loop". Nucleotides correspond to highly conserved regions and open circles are conserved in more than 60% of all species. Adapted from Barrick *et al.* [2005].

σ^{70}-RNAP-holoenzymes are bound by 6S RNA molecules [Wassarman & Storz 2000, Wassarman & Saecker 2006]. Since this mechanism is common during the (late) stationary phase, when nucleotides become scarce 6S RNA seems to store the RNAP-holoenzyme. When nucleotides become available RNAP uses the 6S RNA molecule as template for the 14-20 nt long pRNA [Wassarman & Saecker 2006]. This short RNA molecule binds to the template region of 6S RNA and mediates the release of the RNAP-holoenzyme. The 6S-pRNA complex is only observed during the outgrowth and seems to be subjected to rapid degradation [Wassarman 2007].

The transfer-messenger RNA (tmRNA), also known as 10Sa RNA or SsrA, is part of a complex that acts as an unique translation quality-control and ribosome rescue system in all eubacteria and some eukaryotic organelles. Initially transcribed as a precursor, the tmRNA molecule needs to be processed before it becomes functional. In *E. coli*, this precursor has a length of 457 nts and is finally cleaved by cellular ribonucleases to a mature tmRNA molecule with a length of 363 nts. The 5' and the 3' ends of the tmRNA form a tRNA-like domain, including an acceptor stem, a D-loop without stem, and a T-arm. Instead of the anticodon stem-loop structure, typical for normal tRNAs, a long stem connects the tRNA-like domain to the rest of the tmRNA molecule. A peptide reading frame, which ends with a stop codon forms the second essential domain of the tmRNA molecule.

Sometimes, messenger RNAs lack appropriate termination signals. These "nonstop" mRNAs are unable to promote the release of the nascent protein and the recycling of stalled ribosomes (see Dulebohn *et al.* [2007] for a recent review). Together with SmpB (small protein B), EF-Tu (Elongation factor Tu) and GTP the tmRNA forms the "stalled ribosome

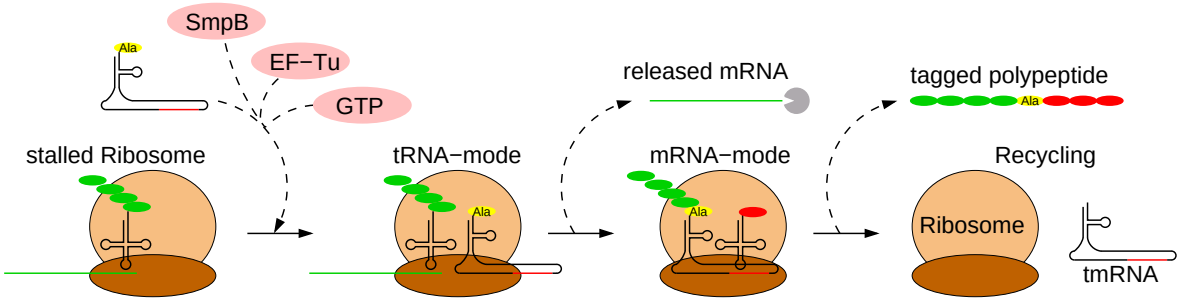


Figure 2.4. The "stalled ribosome recognition complex" (EF-Tu, SmpB, GTP and tmRNA) recognizes nonstop mRNA-Ribosome complexes. tmRNA enters the ribosome and the nascent polypeptide is transferred (tRNA-mode). The ribosome continuous translation using the mRNA-domain as template (tRNA-mode). The aberrant mRNA is released and directly degraded by RNase R (Pakman symbol). Finally translation terminates and both tmRNA and ribosome are free. The polypeptide is released and recognized, probably by the attached tag (red), for degradation.

recognition complex". This complex recognizes ribosomes stalled at the 3' end of nonstop mRNAs (Figure 2.4). Acting in a tRNA mode tmRNA enters the ribosome [Valle *et al.* 2003, Sundermeier *et al.* 2008] and the nascent polypeptide is transferred to the tmRNA. At this point, the tmRNA switches from a tRNA- to an mRNA-like mode. The ribosome continues the translation process with the mRNA-domain as a surrogate template, whereas the damaged mRNA is selectively recognized and degraded by RNase R (reviewed by Richards *et al.* [2008]). The translation terminates at a tmRNA-encoded stop codon, releasing the nascent polypeptide with the 11-amino acid degradation tag at its C-terminus. This permits recycling of the ribosomal subunits into the cellular pool and an efficient decay of the tagged peptide which otherwise might have deleterious effects for the cell. In conclusion, tmRNAs perform three key functions: i) promoting the degradation of aberrant mRNAs ii) rescuing stalled ribosomes and iii) tagging incomplete polypeptide chains.

Proteins of the Csr (Carbon storage regulator) system and its homolog, Rsm (Repressor of secondary metabolites), are involved in the regulation of various biological processes, e.g. quorum sensing, pathogenesis, motility and biofilms, by mRNA binding (reviewed by [Timmermans & Melderen 2010]). Antagonists are sRNAs, which sequester and therefore inhibit the regulatory function of the respective protein (reviewed by Babitzke & Romeo [2007]). CsrA is the central RNA binding protein of the Csr system and binds to GGA motifs of its mRNA targets. Homologs involved in virulence are RsmA in *Pseudomonas* and RsmE

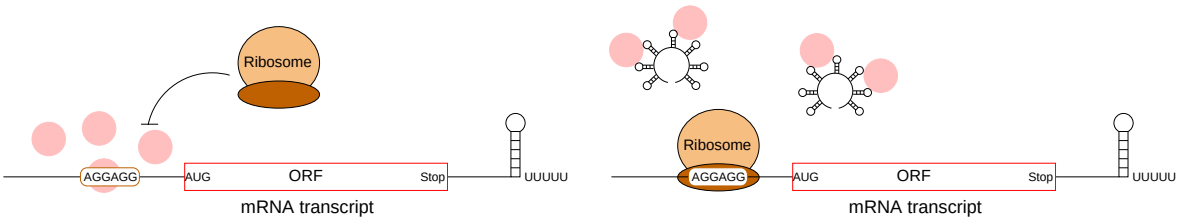


Figure 2.5. Schematic drawing of direct protein regulation as known from the Csr and Rsm systems. In absence of the sRNA molecules the proteins (pink circles) bind to the mRNA and therefore, repress translation. Upon sRNA expression, multiple stem-loops of the sRNA molecule interact with the protein and thus inhibit protein activity. This allows ribosome binding and enables translation.

in *Erwinia* species. In *E. coli* two sRNAs, namely CsrB and CsrC, are the antagonists of the CsrA protein. These sRNAs form complex structures with repeated loop-exposed GGA motifs and therefore mimic mRNA targets of CsrA, see Figure 2.5. Similar to *E. coli* a single sRNA homolog (RsmB) has been detected in *E. carotovora*. However, two sRNAs (RsmY and RsmZ) are present in *P. aeruginosa* and up to three RNA molecules are encoded in the genome of *V. cholerae* (CsrB, CsrC and CsrD) as well as *P. fluroscens* (RsmX, RsmY and RsmZ) [Lapouge *et al.* 2008]. Although length, sequence and number of exposed GGA motifs varies between these sRNA homologs they show similarity at the structural level [Valverde *et al.* 2004].

2.1.3 RNAs with dual-function

Although non-coding and protein-coding RNA genes are commonly differentiated by their coding capacity, currently a few sRNAs with a dual-function are known. They encode a typically small sized ORF and act in addition as antisense sRNA.

One example is the *Staphylococus* lineage specific RNAIII. The 500 nt long transcript folds into a complex 14-stem-loop structure and harbors the template of the δ-hemolysin protein. Acting as an sRNA, RNAIII has at least four targets whose translation is either up- (*hla*) or down-regulated (*spa, rot, coa*). RNAIII binds each target by one or two loop-loop interactions in the vicinity of the Shine-Dalgarno sequence. The RNAIII regulatory network implements a key process in virulence gene expression of *Staphylococcus* [Boisset *et al.* 2007, Toledo-Arana *et al.* 2007].

The 227 nt long SgrS RNA which is also translated into the 43 amino acid (aa) long SgrT peptide was first identified in *E. coli*. So far, *sgrS* genes have been detected in several

enterobacteria. SgrS RNA is expressed under glucose-phosphate stress. In an Hfq and RNase E dependent manner SgrS represses the translation of the glucose-phosphate transporter PtsG. Base-pairing between SgrS and its target mRNA occurs downstream of the *sgrT* ORF. Interestingly, expression of the SgrT protein also regulates glucose transport. Regulation of both the SgrT peptide and the SgrS RNA are equally efficient. Although the phenotypes of SgrS- and SgrT-only mutants are the same, different mechanisms seem to be used to promote recovery from stress and negatively affect glucose transport [Wadler & Vanderpool 2007]. Homologs of SgrS are broadly distributed in enteric bacteria but have diverged in size and sequence [Horler & Vanderpool 2009]. Moreover, in a few homologs the sgrT is missing or rendered non-functional by mutation of the start codon.

The third example is SR1 originally found in *B. subtilis* [Licht *et al.* 2005]. Experiments clearly showed a function of SR1 in the arginine catabolism pathway by RNA-RNA interaction with the ahrC mRNA, thus confirming its nature as functional sRNA [Heidrich *et al.* 2006; 2007]. Only recently, Gimpel *et al.* [2010] evidenced that the *gapA* operon is regulated by a short peptide encoded by SR1. Although intensive studies it took several years to recognize and proof the dual-function of SR1.

2.1.4 Riboswitches and RNA Thermometers

These typically *cis*-acting elements, encoded in 5' and 3' UTRs, represent a direct link between the genetic information (mRNA) and the environmental conditions. Binding of a small molecule (ligand; metabolite) causes an allosteric rearrangement of the riboswitch structure and therefore the mRNA conformation changes. Known ligands are amino acids (e.g. lysine), nuclear bases (e.g. guanine and adenine) and sugar (e.g. glucose-6-phosphate). Additionally, Mg^{2+} stabilizes these interactions or is directly sensed by the riboswitch [Coppins *et al.* 2007, Cochrane *et al.* 2007]. Depending on the genomic context a riboswitch (e.g. TPP) could act in one case as activator and in another as repressor [Cheah *et al.* 2007].

Almost all known riboswitches comprise two essential elements. The first one is a conserved binding site (aptamer) which senses the ligand. The second functional element of a riboswitch is the so called expression platform. This element varies widely in sequence and structure between homologs [Barrick & Breaker 2007]. Upon ligand binding structural changes of the expression platform directly regulate the transcription and/or translation of the downstream ORF (Figure 2.6). In Gram-positive bacteria riboswitches regulate the formation of (anti)terminator stems. Depending on whether the aptamer is loaded with the ligand or not, a terminator structure is formed that prematurely terminates the mRNA transcription. Two scenarios of riboswitch mediated translation attenuation are possible. Either the aptamer

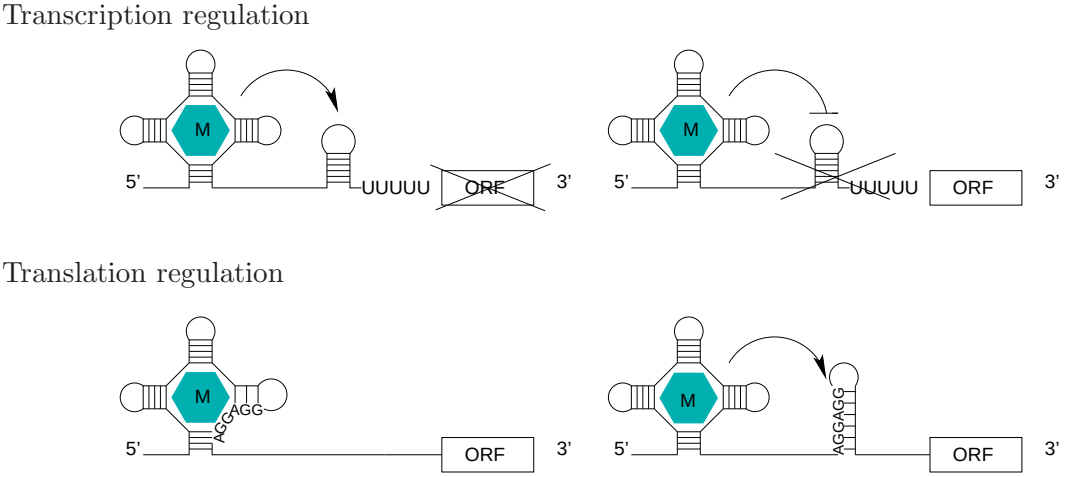


Figure 2.6. Riboswitch mediated gene regulation in prokaryotes. Most switches are located in the 5' UTR. They either regulate transcription (top) or translation (bottom). Transcription is typically regulated by formation of a transcription terminator (OFF switch) or blockade of an anti-terminator hairpin (ON switch). Translation attenuation is accomplished by structural sequestration of the Shine-Dalgarno (AGGAGG) sequence.

region or an additional stable hairpin structure sequesters the ribosome binding site. Structure formation inhibits ribosome binding and therefore translation of the downstream ORF is repressed [Rodionov *et al.* 2002].

The gmlS riboswitch is a so far unique RNA element which is present in certain Gram-positive bacteria. Instead of a conformational change it stimulates a self-cleaving ribozyme activity that acts on the GmlS mRNA [Klein & Ferré-D'Amaré 2006, Cochrane *et al.* 2007]. Located within the 5' UTR of the glucoseamine-6-phosphate (GclN6P) synthetase gene the ribozyme activity is triggered by the binding of GclN6P and Mg^{2+}. Interestingly, in Gram-negative bacteria glmS expression is regulated by two sRNAs GlmY and GlmZ, respectively. Both sRNAs are highly similar in sequence and structure but they form a regulatory hierarchy instead of simple redundancy (reviewed by Görke & Vogel [2008]).

While riboswitches sense specific molecules, so-called RNA thermometers sense temperature as a physical stimulus [Narberhaus 2010]. This provides a simple and direct regulatory mechanism without the aid of an additional interaction partner. RNA thermometers are usually located in the 5' UTR of heat shock (e.g. *E. coli ipbA*) and virulence (e.g. *Listeria monocytogenes pfrA*) genes [Johansson *et al.* 2002, Waldminghaus *et al.* 2009]. These thermosensors typically inhibit ribosome binding in the low temperature regime. An increase of temperature

destabilizes the structure, leads to refolding of the 5' UTR and permits translation initiation. This refolding is enabled by non-canonical and temperature-labile base pairs around the Shine-Dalgarno sequence. Two well-known examples are fourU and ROSE-like thermometer (reviewed by Narberhaus [2010]).

2.1.5 An RNA chaperone and its helper

Bacterial sRNAs often act in concert with protein components as ribonucleoprotein complexes (RNP). Many of such RNP complexes and their mediated functions are known (reviewed by Pichon & Felden [2007]). One key player is the sm-like protein Hfq which is present in about a half of all Gram-positive as well as Gram-negative bacteria and at least in one archaeon *Methanococcus jannaschii* [Nielsen *et al.* 2007]. The abundant thermostable 70-110 amino acid long protein forms homohexamers (in *E. coli* ∼10.000 copies per cell) of which 80-90% are found in association with ribosomes [Brennan & Link 2007]. Hfq has two distinct binding sites: the proximal site binds to sRNA as well as mRNA whereas the distal site interacts with poly(A) tails. This construction enables Hfq to act in many ways as a pleiotropic post-transcriptional regulator. In *E. coli* approximately a fourth of the known sRNAs bind to Hfq and at least 50 proteins seem to be regulated. Interestingly, existence of a Hfq coding gene does not imply a functional role within the species.

A chaperone-like function is indicated by the fact that Hfq supports the formation of sRNA-mRNA heteroduplexes (Figure 2.7). Although short and imperfect base pairing between sRNAs and their mRNA targets are enhanced by the RNA chaperone the regulatory mechanism of Hfq is still unclear. It is, furthermore, ambiguous if one Hfq hexamer binds both RNA molecules simultaneously or if two Hfq molecules are necessary to bring sRNA and mRNA together [Storz *et al.* 2004]. As exemplified by RNAIII Hfq binding does not necessarily imply an effect on the sRNA-mRNA complex formation [Boisset *et al.* 2007]. Even in *hfq*-deletion mutants both RNA molecules interact rapidly. A chaperone activity is, however, evidenced as Hfq mediated structural changes of sRNAs (e.g. OxyS) and mRNAs (e.g. *sodB*) have been reported [Moll *et al.* 2003, Geissmann & Touati 2004]. Furthermore, Hfq is required for the inter-cellular stability of several sRNAs [Valentin-Hansen *et al.* 2004]. Interestingly, Lee & Feig [2008] reported an interaction of Hfq and one of the most abundant RNA species in a cell, tRNAs. The authors showed previously unrecognized phenotypes associated with mis-translation and significantly reduced translational fidelity of *hfq* deletion mutants. Besides this sRNA dependent regulation, Hfq also acts alone as translational repressor of mRNAs [Urban & Vogel 2008]. Given all these information is it not surprising that the expression of the *hfq* gene is auto-regulated by its own product [Vecerek *et al.* 2005].

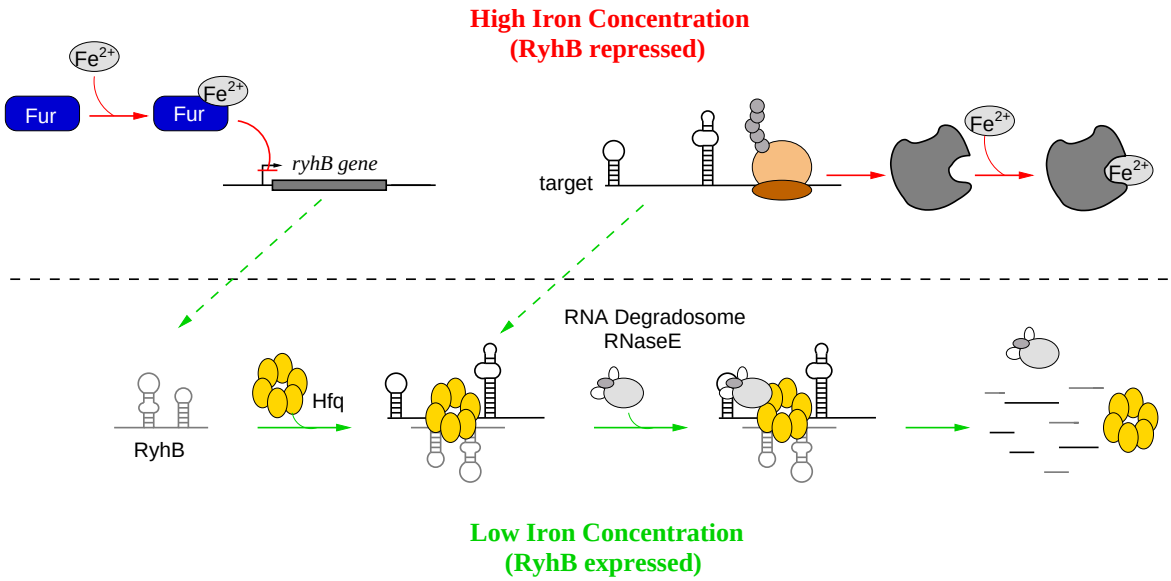


Figure 2.7. Mechanism of Hfq mediated RyhB-mRNA complex formation. When sufficient iron is available (top) Fur represses RyhB expression and non-essential iron-using proteins are expressed. When iron becomes scarce (bottom) Fur repression is repealed and RyhB is rapidly expressed. Hfq stabilizes RyhB which subsequently pairs with an mRNA target. The RNA degradosome recognizes the sRNA-mRNA complex and degrades both RNAs simultaneously. Adapted from Massé *et al.* [2007].

Additionally, it is known that Hfq is interacting with proteins like poly(A) polymerase I (PAP I), polynucleotide phosphorylase (PNP) and ribonucleases (RNases) which form complexes that regulate mediated decay of several mRNAs [Brennan & Link 2007]. RNases are key regulators that influence processing and turnover of many RNA molecules. In fact, essential steps of tRNA maturation are processing of the 5' and 3' ends. The 5' end of a tRNA precursor is processed by RNase P enzyme which is generally composed of RNA and protein subunits. However, catalytic function can be conducted by the RNA subunit alone. This indicates that function of tRNA 5' end processing resides within the RNA subunit [Guerrier-Takada *et al.* 1983]. The 3' end of tRNA precursor sequences is either cleaved by RNase Z or RNase E (reviewed in Redko *et al.* [2007]). Besides its function in tRNA maturation RNase E is one of the main enzyme components of the degradosome. This multiprotein complex is involved in the degradation of many RNAs. Most of the mRNA decay processes in *E. coli* are assumed to begin at typically single-stranded RNase E cleavage sites. Moreover, RNase E in complex with Hfq and certain sRNA (e.g. SgrS, RyhB) leads to translational repression and degradation of the corresponding mRNA target (e.g. ptsG and sodB), see Figure 2.7 [Viegas & Arraiano 2008]. Another very important and highly conserved ribonuclease is RNase III.

This RNase plays multiple roles in rRNA and mRNA processing. RNase III preferentially binds to double-stranded RNA sequences. Thus, sRNA-mRNA duplexes form a perfect substrate. One example is RNAIII based gene regulation in *S. aureus* which directly depends on the action of RNase III [Boisset *et al.* 2007].

2.2 Identification of (non-)coding transcripts

In recent years computational and experimental approaches have been used to identify RNA regulators in various model organisms in all domains of life. Surveys that aim to identify novel transcripts try to complete the so called transcriptome of an organism. Wang *et al.* [2009] define the transcriptome as "... the complete set of transcripts in a cell, and their quantity, for a specific developmental stage or physiological condition". In the following two sections methods for RNA identification by computational and experimental means are summarized. I will mainly focus on those approaches that have been used in projects I was involved in and refer to detailed reviews for other methods.

2.2.1 Computational approaches

The functionality of ncRNAs is often strongly dependent on their structure. The primary structure of an RNA molecule is simply its sequence. Most ncRNAs, however, are single-stranded molecules that form internal interactions between their nucleotides guanine (G), cytosine (C), adenine (A) and uracil (U). So called base pairs are typically formed between G *and* C, A *and* U as well as G *and* U. These non-crossing base pairs stabilize and determine the (pseudo-knot free) secondary structure of the ncRNA. The three dimensional structure of an RNA molecule strongly depends on the secondary structure and especially the unbound nucleotides therein. Thus, the secondary structure is an fairly good approximation of the actual functional structure formed by an RNA molecule within the cell.

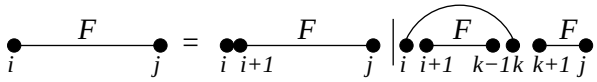


Figure 2.8. Secondary structure prediction using base pair maximization. For each nucleotide i of a given sequence $F_{i,j}$ two alternatives are possible. If i is unpaired the problem can be reduced to sequence $F_{i+1,j}$. If i is paired with another base k the folding problem is divided in two sub-problems: sequence $F_{i+1,k-1}$ which is enclosed by the base pair and the sequence $F_{k+1,j}$ behind the base pair.

To understand how most ncRNA gene finders work one has to understand how the secondary structure of a given sequence can be calculated. The first attempt by Nussinov *et al.* [1978] was to maximize the number of base pairs of a given RNA sequence $F_{i,j}$. Starting with the first nucleotide i there are only two possibilities: either it is paired or unpaired. If the latter is the case the problem can be reduced to sequence $F_{i+1,j}$. Otherwise the nucleotide forms a

base pair with another one. In that case the folding problem is split in two sub-problems: i) the sequence which is enclosed by the formed base pair and ii) the sequence that is behind the base pair (Figure 2.8). With this decomposition the structure with the maximal number of base pairs can be computed by dynamic programming.

Secondary structure prediction

The structure with the maximum number of possible base pairs is not necessarily the one an RNA molecule adopts in the cell. Since molecular stability is driven by the energy the *thermodynamically* most stable structure might get closer to RNA behavior in nature. Actually, the so called MFE (Minimum Free Energy) structure is computed with respect to the nearest-neighbor model. This model assumes that a local structural motif is only dependent on the nucleotides forming the motif, the adjacent nucleotides and their interactions. The thermodynamic parameters of these structural motifs have been experimentally estimated since 1971 [Tinoco *et al.* 1971] and were further improved by Mathews *et al.* [1999; 2004].

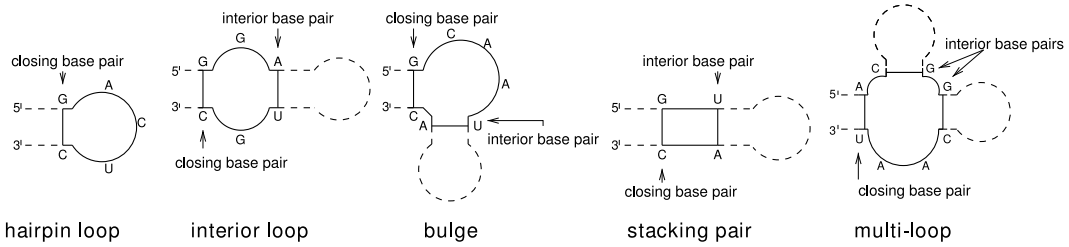
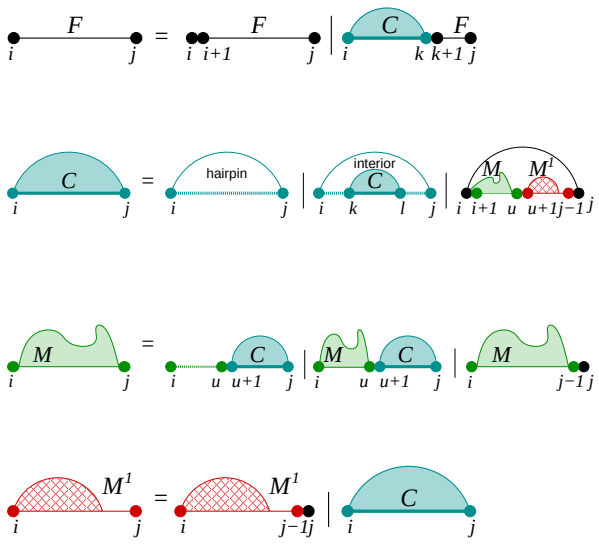


Figure 2.9. Any RNA secondary structure is a composition of three loop types which are differentiated by their degree (number of adjacent base pairs): i) hairpin loop with degree 1, ii) interior loop with special cases bulge and stacking pair and degree 2 iii) multiloop with degree ≥ 3. Adapted from Flamm *et al.* [2004]

The secondary structure of any RNA molecule can be decomposed into structural motifs for which thermodynamic parameters have been estimated, see Figure 2.9. Namely any secondary structure can be decomposed into the following loop types which are differentiated by the number of adjacent base pairs (i.e. degree): A *hairpin* is a loop with degree 1 and has a minimum loop length of 3 nucleotides. *Interior loops* have degree 2 and vary in loop size. *Stacked pairs* (degree 2, loop size 0) and *bulges* (degree 2, loop size > 0) are special cases of interior loops. In case of *bulges* only one side shows unpaired bases. *Multiloop* structures are enclosed by at least three base pairs (degree ≥ 3). Nussinov's dynamic programming approach can be extended by this loop decomposition. The corresponding recursions are given and illustrated in Figure 2.10.



$$F_{i,j} = \min \begin{cases} F_{i+1,j}, \\ \min\limits_{i<k\leq j} C_{i,k} + F_{k+1,j} \end{cases}$$

$$C_{i,j} = \min \begin{cases} \mathcal{H}(i,j), \\ \min\limits_{i<k<l<j} C_{k,l} + \mathcal{I}(i,j;k,l), \\ \min\limits_{i<u<j} M_{i+1,u} + M^1_{u+1,j-1} + a \end{cases}$$

$$M_{i,j} = \min \begin{cases} \min\limits_{i<u<j} (u-i+1)c + C_{u+1,j} + b, \\ \min\limits_{i<u<j} M_{i,u} + C_{u+1,j} + b, \\ M_{i,j-1} + c \end{cases}$$

$$M^1_{i,j} = \min \begin{cases} M^1_{i,j-1} + c, \\ C_{i,j} + b \end{cases}$$

$$F_{i,i} = 0, C_{i,i} = M_{i,i} = M^1_{i,i} = \infty$$

Figure 2.10. Loop decomposition of an RNA secondary structure. Similar to the base pair maximization approach the recursion starts with a given sequence $F_{i,j}$. The first base of this sequence is either unpaired, the problem is reduced by one base to $F_{i+1,j}$, or it is paired with another base k. This formed base pair splits the folding problem in two sub-problems: one is the sequence $F_{k+1,j}$ behind the base pair (i,k) and the other is a sub-structure $C_{i,k}$. The formed base pair can enclose a hairpin $\mathcal{H}(i,j)$, an interior loop $\mathcal{I}(i,j;k,l)$ which again covers a sub-structure $C_{k,l}$ or a multiloop structure $M_{i+1,u}$ $M^1_{u+1,j-1}$. $M_{i,j}$ is used to compute the free energy of the substructure given the constraint that the sequence is part of a multiloop and contains at least one component. The multiloop contribution $M_{i,j}$ can be decomposed into an unpaired stretch of nucleotides of length $u-i+1$ and a sub-structure $C_{u+1,j}$, a multiloop $M_{i,u}$ plus a sub-structure $C_{u+1,j}$ or into a multiloop $M_{i,j-1}$, which is truncated by one nucleotide, the unpaired base j. $M_{i,j}$ is used to compute the free energy of the substructure given the constraint that the sequence is part of a multiloop and contains exactly one component. The multiloop type $M^1_{i,j}$ can either be reduced by one base to $M^1_{i,j-1}$ or it is equivalent to $C_{i,j}$. a, b and c contain contribution of closing, branches and unpaired positions, respectively. Illustration (left) is adapted from Hofacker & Stadler [2007] and corresponding recursions (right) are taken from Hofacker *et al.* [1994].

Current RNA folding programs have a prediction accuracy (measured as the fraction of correctly predicted base pairs) of 50-70% on single RNA sequences [Eddy 2004]. Again, this is not entirely correct. The point is that there are several structures in the equilibrium and the predicted MFE structure is not necessarily the one an RNA adopts within the cell. Depending on environmental conditions, e.g. temperature and pH, the correct structure might be among the sub-optimal ones. Base pair probabilities can help to estimate a more accurate model of alternative foldings that are likely to occur in the living cell. The probability of a particular structure S is proportional to its Bolzmanfactor $exp\,(-E(S)/RT)$, where $E(S)$ is the energy of secondary structure S, R Bolzmann's constant in molar units and T is the temperature in Kelvin. The partition function defines the ensemble of structures:

$$Z = \sum_{S} e^{\left(\frac{-E(S)}{RT}\right)} \tag{2.1}$$

The partition function $Z_{i,j}$ over all structures on a sub-sequence $x[i \dots j]$ can be inferred from the MFE algorithm by replacing minimum operations with sums and additions with multiplications [McCaskill 1990]:

$$\begin{aligned}
Z_{i,j} &= Z_{i+1,j} + \sum_{i<k\leq j} Z^C_{i,k} Z_{k+1,j} \\
Z^C_{i,j} &= e^{-\beta\mathcal{H}(i,j)} + \sum_{i<k<l<j} Z^C_{k,l} e^{-\beta\mathcal{I}(i,j,k,l)} + \sum_{i<u<j} Z^M_{i+1,u} Z^{M^1}_{u+1,j-1} e^{-\beta a} \\
Z^M_{i,j} &= \sum_{i<u<j} e^{-\beta(u-i+1)c} Z^M_{u+1,j} + \sum_{i<u<j} Z^M_{i,u} Z^C_{u+1,j} e^{-\beta b} + Z^M_{i,j-1} e^{-\beta c} \\
Z^{M^1}_{i,j} &= Z^{M^1}_{i,j-1} e^{-\beta c} + Z^C_{i,j} e^{-\beta b} \\
Z_{i,i} &= 1, Z^C_{i,i} = Z^M_{i,i} = Z^{M^1}_{i,i} = 0;
\end{aligned}$$

Here $\beta = 1/RT$ denotes the inverse thermal energy. To estimate the probability that two nucleotides form a base pair one has to calculate the fraction of all structures that contain the base pair over all possible structures. More precisely this can be computed with:

$$p_{i,j} = \frac{\hat{Z}_{i,j} Z_{i+i,j-1} exp(\frac{-\beta_{i,j}}{RT})}{Z} \tag{2.2}$$

where $\hat{Z}_{i,j}$ is the partition function of all structures outside of the base pair (i,j). Graphical representations of secondary structures are depicted in Figure 2.11.

It is assumed that homologous RNAs have common functions and it is therefore expected that they adopt similar structures. The underlying sequence, however, might be diverged. A

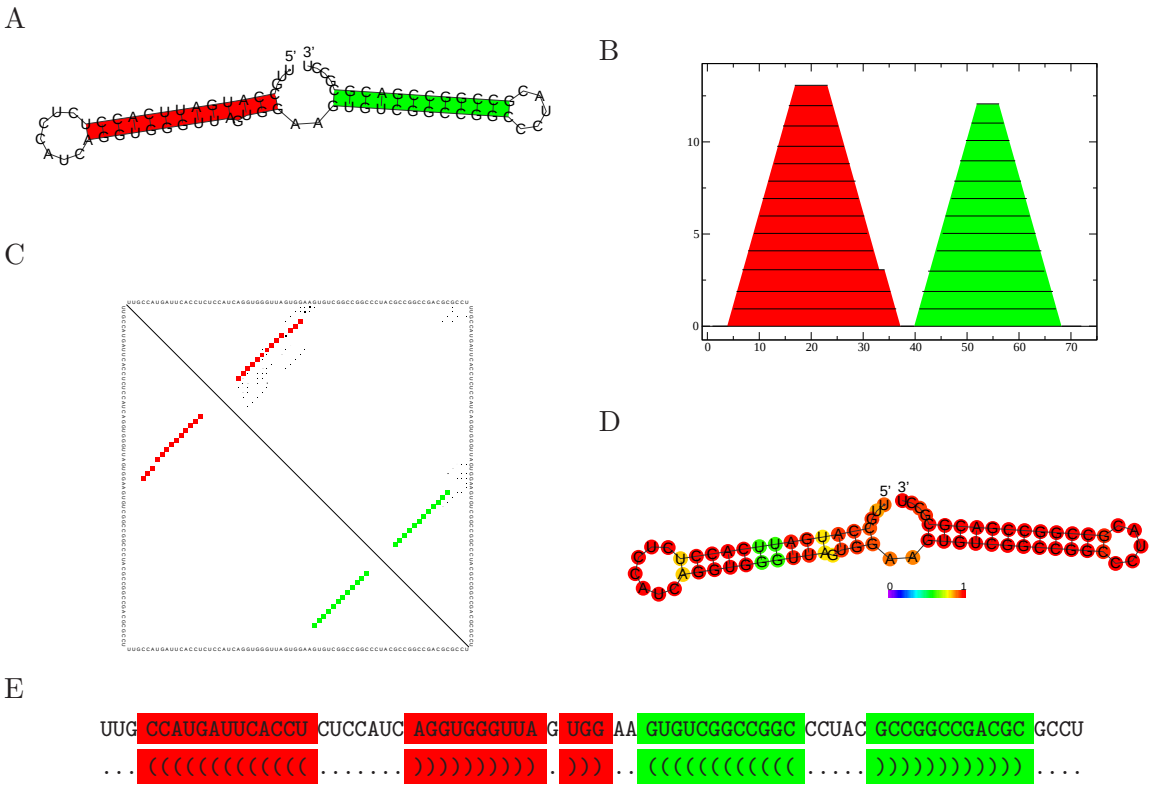


Figure 2.11. Graphical representation of an RNA secondary structures. A) MFE structure B) mountain plot C) dot plot D) probability representation of the MFE structure E) dot bracket string notation. The color code in D) indicates the probability that a base is (un)-paired. In the remaining plots the 5' and 3' stem loop of the RNA are highlighted in red and green, respectively. All plots represent the structure of PtaRNA1, see Section 3.3.

G-C base pair in one sequence might be mutated to an A-U pair in another sequence. This indicates a selective pressure that preserves the structure of the RNA molecule even though mutations changed the underlying sequence. Structure maintaining double mutations are termed *compensatory mutations*. Single nucleotide mutations that change for instance a G-C to a G-U base pair might hold biological information as well. These structure maintaining single base mutations are called *consistent mutations*. The the consensus secondary can be calculated by searching the best structure that the set of related RNA molecules can adopt. Predictions are necessary for co-variation analysis since the number of biologically validated structures is small.

If the sequences are similar enough classical sequence based methods, e.g. `ClustalW` [Thompson *et al.* 1994], can be applied in order to calculate the multiple sequence alignment of RNA homologs. Given a sequence based alignment approaches, such as `RNAalifold` [Hofacker *et al.*

2002, Hofacker 2007, Bernhart *et al.* 2008], can be used for subsequent co-variation analysis. Briefly, the folding algorithm implemented in `RNAalifold` combines thermodynamic information and co-variation of alignment columns into one scoring scheme. It counts compensatory as well as consistent mutations between sequence pairs. `RNAalifold` modifies the standard energy model by introducing a conservation score:

$$\gamma'(i,j) = \frac{1}{2} \sum_{\alpha,\beta \in \mathcal{A};\ \alpha \neq \beta} \begin{cases} h(\alpha_i, \beta_i) + h(\alpha_j, \beta_j) & if (\alpha_i, \alpha_j) \in \mathcal{B} \wedge (\beta_i, \beta_j) \in \mathcal{B} \\ 0 & \text{otherwise} \end{cases} \tag{2.3}$$

where α and β are sequences of the alignment $\mathcal{A}$, $\mathcal{B}$ is the set of canonical base pairs (AU,UA,GC,CG,GU,UG) and the hamming distance $h(a,b)$ equals to 0 if $a = b$ and 1 otherwise. The full co-variation score $\gamma(i,j)$ of `RNAalifold` includes additional penalties for sequences in which the base pair (i,j) cannot be formed:

$$\gamma(i,j) = \gamma'(i,j) + \delta \sum_{\alpha \in \mathcal{A}} \begin{cases} 0 & if (\alpha_i, \alpha_j) \in \mathcal{B} \\ 0.25 & if \alpha_i \wedge \alpha_j \text{ are gaps} \\ 1 & \text{otherwise} \end{cases} \tag{2.4}$$

The parameter δ weights the penalty value and therefore the importance of counter examples found within the alignment. Of course `RNAalifold` is only one example for a great variety of programs, e.g. `Pfold` [Knudsen & Hein 2003], `KnetFold` [Bindewald & Shapiro 2006] and `BayesFold` [Knight *et al.* 2004], available to date.

Anyway, homologous RNAs often show poor sequence similarity. Gardner *et al.* [2005] proofed that sequence alignments of structured RNAs fail if pair wise sequence identities drop below 60%. An alternative is to take the plain sequences and try to calculate the consensus secondary structure from scratch. Most popular approaches are variants of the Sankoff algorithm [Sankoff 1985]. In 1985 Sankoff proposed an algorithm that aligns sequence and structural features simultaneously. The Sankoff algorithm in its complete form requires $\mathcal{O}(n^6)$ CPU time and $\mathcal{O}(n^4)$ memory, where n is the length of the RNA sequences to be aligned. Because of that many different and sparsified variants of the original algorithm have been proposed [Mathews & Turner 2002, Hofacker *et al.* 2004, Havgaard *et al.* 2005, Will *et al.* 2007]. Essentially, these implementations use heuristics to reduce the search space by restricting possible consensus structures and possible alignments. `LocARNA`, for instance, implements a derivative of the Sankoff algorithm that reduces CPU time and memory consumption to $\mathcal{O}(n^4)$ and $\mathcal{O}(n^2)$, respectively [Will *et al.* 2007]. Using base pair probabilities the program efficiently calculates pairwise sequence-structure alignments. The implemented probability model, therefore, takes all possible structures of the sequences to be aligned into account.

Based on the `LocARNA` algorithm multiple sequence-structure alignments can be calculated with the high level programs `mLocarna`, `LocARNATE` and `RNAclust.pl` [Will *et al.* 2007, Otto *et al.* 2008]. Nonetheless, these sparsified implementations still have a high CPU and memory consumption.
Other programs, such as `RNA Sampler` [Xu *et al.* 2007] and `CMfinder` [Yao *et al.* 2006] are available as well. These methods are based on different assumptions and avoid the high time and memory complexity of the Sankoff algorithm.

For detailed reviews about the topic of RNA folding, consensus secondary structure prediction and a more comprehensive list and description of individual programs I refer to recent reviews and the references therein [Hofacker & Stadler 2007, Machado-Lima *et al.* 2008, Bernhart & Hofacker 2009].

Gene finding

When talking about gene finding one has to distinguish between approaches that search for homologs of a given sequence and methods for *de novo* gene prediction.

The task for tools of the first kind is the detection of a query sequence within a much larger database. Usually the database is a genome which is queried for a single sequence or a set of already known homologs. Typically, genome annotation surveys and evolutionary analysis of specific ncRNAs are the context of such homology searches. The first scenario aims to identify as many known ncRNA genes within a (newly sequenced) genome as possible. Reliable and comprehensive annotation of known ncRNA genes is essential to all subsequent studies. A common source of already known ncRNA sequences is the Rfam database [Gardner *et al.* 2010]. This database stores alignments, consensus secondary structures, covariance models (CMs) and a short description for hundreds of RNA families. Additionally, family specific databases, such as tRNAdb [Jühling *et al.* 2009], tmRDB and SRPDB [Andersen *et al.* 2006], build a profound data source.

On the other hand the genomes of several species are screened for a single ncRNA of interest in evolutionary surveys. Here the research focus is to identify as many homologs of the given sequence as possible. The set of homologs retrieved can be used to analyze their evolutionary origin, to predict the consensus secondary structure and to define core motifs, e.g. structural elements which are characteristic for this set of sequences.

The most prominent homology search method is `NCBI-Blast` [Altschul *et al.* 1990; 1997]. Different variants of this fast and efficient local alignment algorithm have been implemented.

Whereas protein homologs can be searched with specific implementations, i.e. `blastp`, `psi-blast`, `phi-blast`, the programs `blastn` and `megablast` are suitable to search nucleotide sequences within a sequence database. However, `Blast` implements a local alignment algorithm and resulting hits can be much shorter than the initial query sequence. Thus, a query 200 nucleotides long might result in several short fragmented `Blast` hits. These fragments are often to short to be recognized as significant by the user. Furthermore, Freyhult *et al.* [2007] noted a limited accuracy of `Blast` searches in their benchmark study. Especially, for ncRNAs the identification of full length homologs that might vary on sequence level but still fold into the same structure is an essential task. In 1982, Gotoh proposed a semi-global alignment algorithm which was not implemented until recently. This was mainly caused by the high time and memory consumption of $\mathcal{O}(n \times m)$, where n and m are the length of the database and the query sequence, of the algorithms full version. A direct implementation of the proposed algorithm would not be sufficient for long database sequences. The program `GotohScan` implements a sparse version of the algorithm. Briefly, for each position of the database sequence the score of the best semi-global alignment ending in k is calculated. Since only local optima of the resulting score distribution are of interest the time and memory consumption can be reduced to $\mathcal{O}(m^2)$. This makes `GotohScan` applicable even for the analysis of large eukaryotic genomes [Hertel *et al.* 2009]. Both, `Blast` and `GotohScan`, are sequence based gene finders. If structural information is available more sophisticated methods can be used.

Several family specific ncRNA gene finders have been implemented [Lowe & Eddy 1997, Laslett *et al.* 2002, Laslett & Canback 2004, Hertel & Stadler 2006, Yusuf *et al.* 2010]. `tRNAscan-SE` is a dedicated tool for tRNA gene identification. Features like the characteristic clover-leaf structure and the common tRNA length between 74 and 90 nts are used. With a sensitivity of 99-100% and very low rate of less than 0.00007 false positives per Mb `tRNAscan-SE` is one of the most accurate tools. `ARAGORN` is an implementation of a tmRNA gene finder. Again, characteristic features like the tRNA sub-structure formed by the 5' and 3' end of the 350-400 nucleotide long ncRNA and the presents of the tag peptide within a defined structural context are applied.

The `fragrep` tool is a more general approach that implements a fragmented pattern search. Basically, the program takes an alignment and conserved blocks of interest as input. These user defined blocks are converted into position frequency matrices (PFMs), while poorly conserved regions are treated as simple distance constraints. A matching algorithm, originally used for transcription factor binding site detection, has been adapted to scan the resulting abstraction of the input alignment within a given database sequence. The newest version of `fragrep` extends the fragmented pattern matching tool by a structure-search approach [Mosig

et al. 2009].

Several analysis have shown that stochastic context free grammars (SCFGs) are well suited to combine sequence and secondary structure conservation into one scoring scheme [Sakakibara *et al.* 1994, Eddy & Durbin 1994, Brown 2000]. To scan for novel members of an RNA family, the `Infernal` [Nawrocki *et al.* 2009] package uses CMs, a specific type of SCFGs. CMs are typically build from Stockholm formatted multiple sequence alignments. This alignment format comes with a structure annotation marking which positions of the alignment are paired and unpaired. Within the CM position-specific log-odd scores are assigned to single-stranded regions, base paired regions and insertions and deletions. Given a CM the program `cmsearch` of the `Infernal` package can be applied to scan a database for putative homologs of the respective ncRNA. Due to the structural information used recent benchmarks demonstrated that the latest `Infernal` 1.0 version clearly outperforms `Blast` in terms of sensitivity [Nawrocki *et al.* 2009]. It has to be mentioned, however, that `Blast` is still orders of magnitude faster than `Infernal`.

In order to detect structured RNA genes within a given sequence the basic concept of all *de novo* prediction methods is to find descriptors that discriminate the structured element from the genomic background. Note that all methods described below search for stable RNA structures rather than for complete RNA genes. In particular a precise identification of gene boundaries and the detection of unstructured RNA genes is not feasible with these approaches. As an example, RNA regulators, such as riboswitches, which are a highly structured part of a much larger gene, are detectable.

An old but essentially still correct assumption was introduced by Le *et al.* [1988]. In principle the authors proposed that structured RNAs have a lower MFE than random sequences with the same nucleotide frequency. First the the mono-nucleotide composition of random sequences was kept constant. Rivas & Eddy [2000] showed, however, that the difference of the MFE of the known ncRNAs and their random mono-nucleotide background is not significant enough. Since MFE calculation is based on base stackings the next step was to preserve the di-nucleotide frequency while generating random sequences [Workman & Krogh 1999]. Indeed, the implementation of `RANDFOLD` [Bonnet *et al.* 2004] and subsequent analysis by Clote *et al.* [2005] showed that structured RNAs have a significantly lower MFE than the di-nucleotide preserved shuffled background sequences.

The MFE value is directly affected by the length of the given sequence. In fact additional or missing nucleotides of a real RNA structure that are (un-)paired can have direct effect on the estimation of the energy. Thus, approaches that use a fix-sized window to scan a

complete genome might miss ncRNA structures. The program `RNAplfold` [Bernhart *et al.* 2006] circumvents this problem by calculating averaged probability values for each base pair (i, j). Given a sequence interval of length L the implemented method estimates the probability p_{ij}^{L} using a modified version of the standard partition function calculation [McCaskill 1990]. Finally, the averaged probability π_{ij}^{L} is calculated considering all possible structures in all possible sequence intervals of length L that cover the bases i and j.

The G+C content of RNA structures was reported to be on average 50% [Rivas & Eddy 2000]. If all four nucleotides are equally distributed along and within a genome, 50% of G+C bases would not be surprising. In fact that is not the case. Especially the G+C content of prokaryotes varies between species. For instance *H. pylori* is a very AT-rich organism with a G+C content of ∼40% whereas *Xanthomonas* species have a high G+C content of ∼65%. Thus, analysis of GC-rich islands successfully identified ncRNAs [Klein *et al.* 2002, Meyer *et al.* 2009]. Approaches like `sRNAscanner` [Sridhar *et al.* 2010] and `sRNAPredict2` [Livny *et al.* 2006] use regulatory elements for the identification of ncRNAs or scan specific genomic regions for new candidates. As an example, the well studied bacteria *E. coli* was screened using the knowledge of consensus promoter motifs and the structural features of terminator sequences. Furthermore, the search space was restricted to intergenic regions*[Argaman *et al.* 2001]. These single sequence approaches are often combined with conservation analysis of candidate sequences [Wassarman *et al.* 2001, Livny *et al.* 2008].

The common task of comparative gene finding approaches is to classify a multiple sequence alignment as (structured) RNA element or something else. The reasoning for this type of approaches is similar to that of consensus structure prediction: related molecules having the same function are thought to show similar structures. Hence, it is not surprising that tools like `RNAalifold` and `Dynalign`, which are used for consensus structure prediction have been extended to gene finders [Washietl & Hofacker 2004, Uzilov *et al.* 2006]. `Alifoldz` is the extension of the `RNAalifold` program. Here, the `RNAalifold` energy E_{ali} of the given alignment is compared to that of randomized alignments. For the set of randomized alignments the mean μ and standard deviation σ of all energies are calculated. These values are subsequently combined into a z-score:

$$z = \frac{E_{ali} - \mu}{\sigma} \tag{2.5}$$

The more negative the z-score is the more significant is the difference of the given alignment to the randomized background. However, the z-score calculation crucially depends on the method used to generate randomized alignments. In contrast to single sequences, alignments cannot easily be shuffled, while simultaneously preserving di-nucleotide content, gap struc-

*intergenic regions are genomic intervals (between genes) that do not encode known genes.

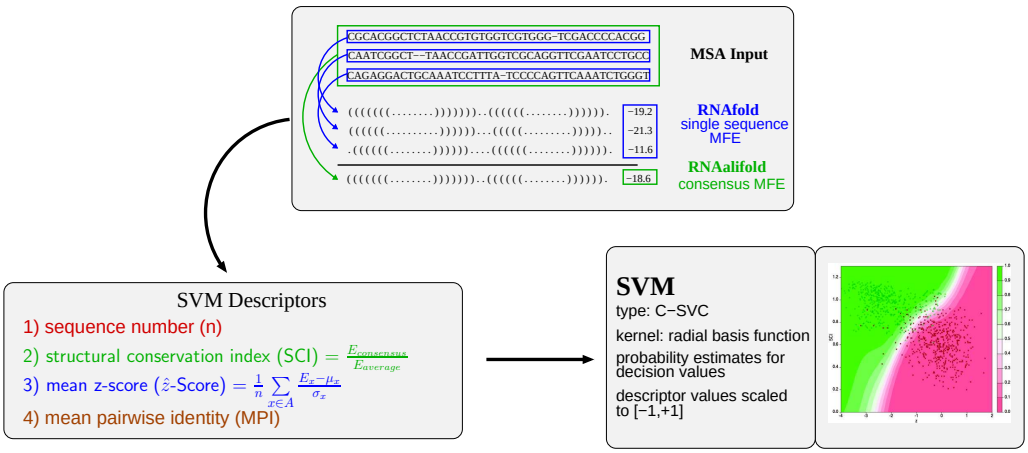


Figure 2.12. `RNAz` performs the classification of the multiple sequence alignment (MSA) input by means of a support vector machine (SVM). The MSA is analyzed and four SVM descriptors are calculated. The structural conservation index (SCI) and the mean pairwise identity (MPI) are measurements for structure and sequence conservation, respectively. Thermodynamic stability is expressed in terms of a averaged z-score. The SVM outputs, based on the degree of thermodynamic stability and structural conservation, the probability of structured RNA elements within the alignment. Contour plot is taken from [Washietl *et al.* 2005]. It shows the classification of native (green dots) and randomly shuffled (red dots) tRNAs and 5S RNAs based on z-score and SCI by using a SVM.

ture and the local degree of conservation. Washietl & Hofacker [2004] introduced a shuffling procedure that sorts alignment columns with the same gap structure and local conservation pattern into individual groups. The conservation pattern of each column in the initial alignment column is memorized. Randomized alignments are generated by sampling each column from the respective group. An alternative to explicit shuffling procedures is the simulation of sequence evolution along a phylogenetic tree [Gesell & Washietl 2008, Varadarajan *et al.* 2008].

`RNAz` is another program that takes advantage of `RNAalifold` [Washietl *et al.* 2005]. This approach uses a support vector machine (SVM) and features deduced from known and conserved ncRNA families to classify a given alignment as structured RNA or something else. Figure 2.12 depicts the general work-flow of `RNAz`. A detailed introduction and recent improvements of the `RNAz` approach are described in Section 4.1 of this thesis.

The combination of three different classification models is implemented in QRNA and its successor eQRNA [Rivas & Eddy 2001, Rivas 2005]. A pair SCFG is used to model the probability that the given alignment is the result of secondary structure evolution. A pair

hidden Markov model (HMM) checks if the input data has protein-coding potential and a different pair HMM represents the null model of an unconstrained sequence evolution. Finally, (e)QRNA decides which of the three models describes the given alignment best.

There are many more approaches available that address the task of gene finding by computational means. Again I refer to recent reviews and the references therein [Machado-Lima *et al.* 2008, Bernhart & Hofacker 2009].

2.2.2 Experimental approaches

The term "experimental RNomics" has been coined for identification of transcripts on a genome-wide scale [Hüttenhofer *et al.* 2002]. Three different methods of the first generation of sequencing approaches are described here: i) direct RNA sequencing ii) parallel cloning and iii) array techniques. Figure 2.13 depicts the general work-flow of all three approaches. Information summarized for all three methods is mainly taken from reviews [Hüttenhofer *et al.* 2002, Vogel & Sharma 2005, Stoughton 2005, Hüttenhofer & Vogel 2006, Sorek & Cossart 2010] and the references therein.

Direct RNA sequencing uses the possibility to separate RNA molecules on a gel. In brief, RNA molecules are sorted based on their size and charge. Total RNA is isolated from a cell and placed into a chamber of the gel. An electric field makes the molecules move through the gel material. Larger RNA molecules move more slowly than smaller ones and because of that bands which correspond to different sized RNA molecule fractions are separated on the gel. Ideally each band corresponds to one type of ncRNA molecules. Of course in diverse RNA families such as tRNAs this is not the case. A 2D gel electrophoresis, where different RNA molecules with similar size are separated, can address this particular issue. After visualization and excision of the band, the extracted RNA molecules are radioactively labeled at the 5' or 3' end with γ-[^{32}P]ATP and [^{32}P]pCp, respectively. The next step is either enzymatic or chemical sequencing. While enzymatic sequencing generates diverse fragments of an RNA molecule by base specific RNase induced cleavage, chemical sequencing specifically modifies each kind of RNA base and a strand scission generates labeled fragments. No matter how fragments of the RNA were generated fractionation again is achieved on a gel. The RNA fragments carry the radioactive label on one end and a specific nucleotide on the other. Thus, four lanes on one gel are used to separate fragments by their terminal nucleotide and size. Using autoradiography the labeled fragments are visualized and the RNA sequence can be read directly from the autoradiograph. As direct RNA sequencing is the most traditional method details of the experimental procedure have been modified and adopted over time.

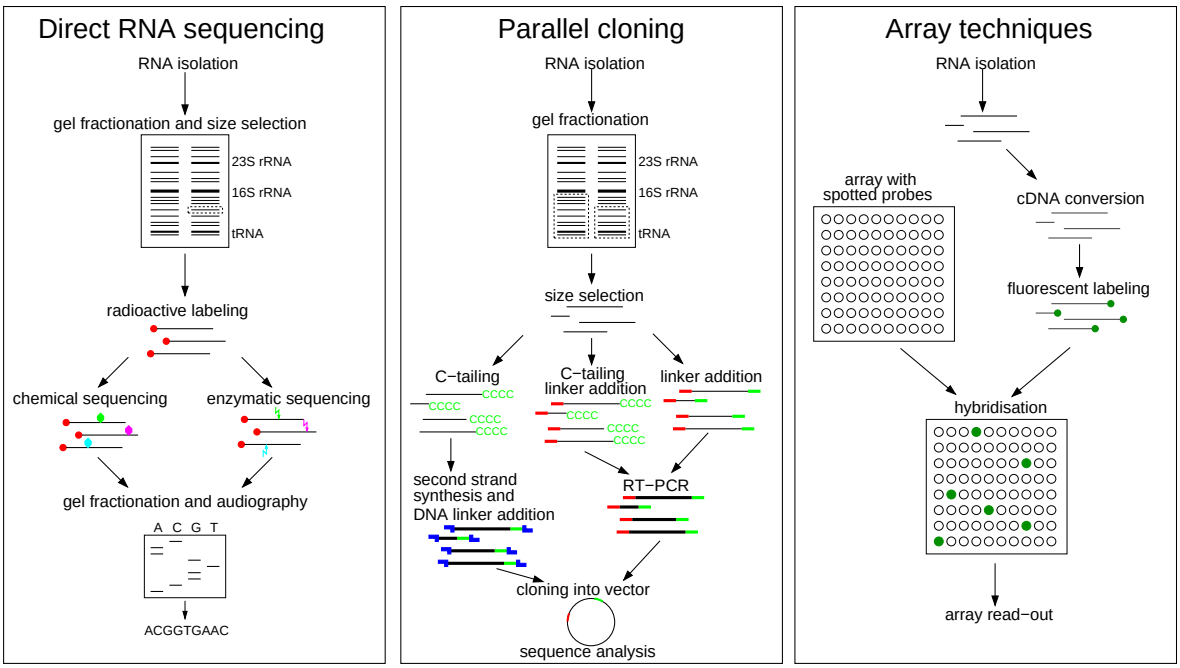


Figure 2.13. Experimental RNomics approaches to identify RNA transcripts. Identification of RNAs by chemical or enzymatic sequencing of size selected abundant RNAs (left). Copy numbers of isolated RNA molecules are multiplied during parallel cloning (middle). Three alternative methods are indicated to reverse transcribe ncRNAs into cDNA. Identification of expressed genomic regions by array analysis (right). DNA probes covering the entire genome or specific genomic regions are spotted onto the array, to which the labeled sample is hybridized. Adapted from [Hüttenhofer *et al.* 2002].

Using the direct sequencing methodology highly abundant sRNAs, such as tRNA, rRNA, tmRNA and 6S RNA have been identified.

Parallel cloning overcomes the major limitation of direct sequencing. The underlying idea of parallel cloning approaches and the resulting cDNA library construction is to multiply the copy number of isolated RNA molecules. Therefore, less abundant sRNA species can be enriched and identified. The experience of EST library preparation, which is used for mRNA sequencing, was utilized to develop parallel cloning methods for ncRNA identification. One major difference between mRNAs and ncRNAs is their lengths. While mRNAs are typically longer than 500 nts the length of most ncRNAs ranges between 20 and 500 nts. Another important feature of the majority of mRNAs is the presence of a poly(A) tail at their 3' end. EST library preparation uses oligo(dT) primers which bind preferentially to the poly(A) tail and initiate reverse transcription into cDNA (copy or complementary DNA). There are several approaches to reverse transcribe and amplify ncRNAs. The underlying idea is the

generation of RNA species with well defined 5' and 3' ends. The RNA sequence itself is not known and therefore an addition of oligo(C) or oligo(A) tails is performed and/or specific 5' and 3' linker sequences are ligated onto the RNA molecules. After addition of these sequences oligo(dG), oligo(dT) or primer sequences complementary to the linkers are used to reverse transcribe the RNA into cDNA. If linker sequences are added to both ends the cDNA can be amplified by several rounds of reverse transcription. Subsequent to the cDNA synthesis, the fragments generated are cloned into plasmids and transfected into bacteria. By use of antibiotic selection mechanisms bacteria containing a transfected plasmid are singled out. Finally, plasmid DNA extracted from grown bacteria is sequenced. Thus, the initial RNA pool is massively multiplied by amplification of cDNA fragments and the growth of selected bacteria. The big advantage of this method, compared to direct sequencing, is that less abundant RNA species can be detected as well.

Array techniques such as micro- and tilling-arrays have been developed to analyze expression levels of many molecules in parallel. A glass or silicon slide is used on which DNA probes are spotted in a well defined order. Thus each position of the slide holds a specific DNA sequence. During amplification steps, similar to that described for parallel cloning approaches, isolated RNA is labeled with modified nucleotides that carry fluorescence dyes. The sample prepared is mixed with hybridization buffer and applied onto the slide. Sample molecules now hybridize to the spotted DNA probes and the fluorescence signal is emitted. Signals are recorded by a scanner and their intensity corresponds to the amount of transcripts present in a cell. Sequences of the spotted probes are known and from that the sequence of the hybridized RNA fragment can be derived. The essential step of all array techniques is the probe design. Especially probe density and the avoidance of multiple probes with highly similar sequence, which would result in cross hybridization, have to be taken into account. However, especially tilling-arrays lead to the appealing opportunity of expression studies from a birds eye view. Not only genomic regions of interest are spotted onto the slide. Moreover, the complete genome with separated strands can be tilled and analyzed.

With the advent of next or second generation sequencing techniques a much higher throughput on a single base resolution of transcription can be achieved. The idea of RNA sequencing (RNA-seq) methods is different to that of first generation approaches. While sequences and regions to be probed are selected at the beginning of first generation experiments, RNA-seq methods map the generated data onto the genome and the underlying annotation at the end of the experiment. This difference results in several intrinsic advantages [Croucher & Thomson 2010]. Gene structure and novel genomic features are easily detectable since mapping of

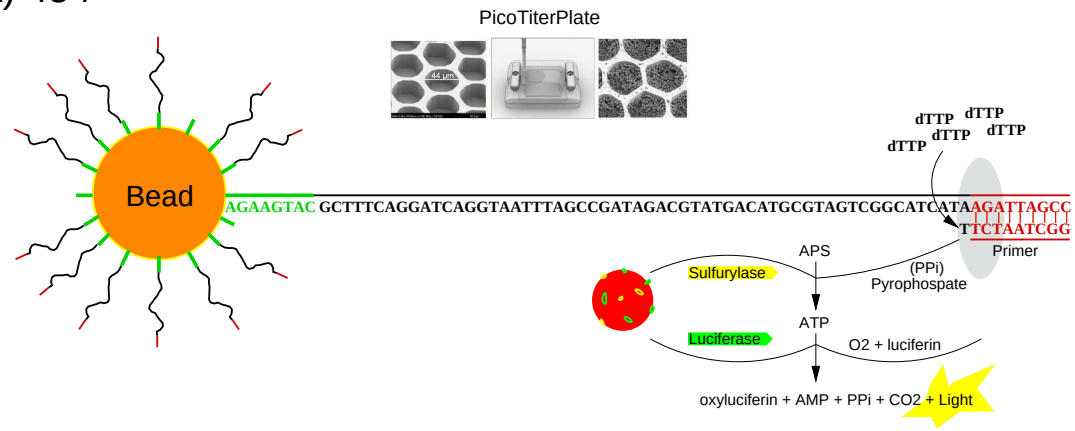
sequences is more accurate than binding of sequence pairs. While expression is measured as signal intensity in hybridization based approaches, RNA-seq determines expression in means of (normalized) read counts. In contrast to signal intensity measurements, the read counts have no upper bound and can not be saturated. This allows gene expression studies of highly and very low abundant transcripts at the same time. Three platforms are mainly used: i) 454 ii) Illumina/Solexa and iii) SOLiD sequencing. These methods are summarized in the next three paragraphs. As the platforms are constantly improved read length obtainable and technique details may have changed since this thesis has been written.

454 sequencing* Specific 5' and 3' adapter sequences are ligated to RNA fragments. The adapters are used for purification, fixation, amplification, and sequencing steps. First the 3' adapter is utilized to fixate fragments onto beads. Ideally, each bead caries a single fragment. Beads are separated into micro-reactors comprising a water oil mixture. All essential reagents, e.g. nutrients and polymerase are enclosed within each micro-reactor. A process called "emulsion PCR" amplifies the fragment to a number of several million identical copies on each bead. In the next step the emulsion is broken and beads with the amplified fragments are separated on a PicoTiter™Plate, which basically is a slide with wells. The size of the wells is optimized so that a single bead is isolated therein. Additionally, smaller beads carrying immobilized sequencing enzymes are inserted. The sequencing process is initiated by primer sequences that specifically bind to the free adapter end of the fragments. The fluidics subsystem of the sequencer flows individual nucleotides in a fixed order across the wells. Whenever a nucleotide is incorporated a luciferase light signal is emitted, Figure 2.14 A). The signal intensity is proportional to the number of simultaneously incorporated nucleotides for up to eight bases. This chemiluminescent signal is recorded by a CCD camera which tracks the location of bead loaded wells by their XY-coordinates on the plate. Finally a so called flowgram, which corresponds to the sequence of the fragments on one bead, is generated. Currently, read lengths of up to 400 bases are possible with the 454 sequencing technique.

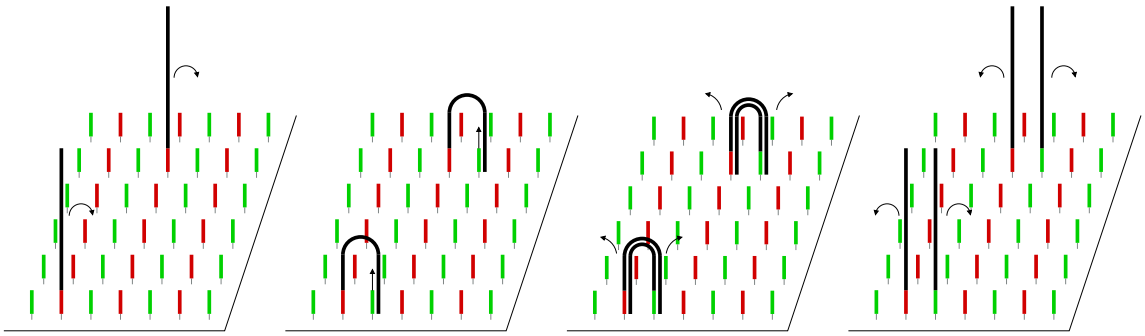
*Information adapted from `http://454.com`

Figure 2.14 (following page). Second generation sequencing approaches overview. The most discriminative step of each sequencing technique is shown. Ligated adapter sequences are indicated for 5' and 3' end in red and green, respectively. A) Sequencing process of the 454 method. The bead is separated in a well of the PicoTiterPlate™ (image taken from `http://www.454.com/downloads/news-events/how-genome-sequencing-is-done_FINAL.pdf`) and carries one fragment which is already amplified. All copies on the bead are sequenced simultaneously. For simplicity the process is only shown for the zoomed sequence. A primer complementary to the 5' adapter initiates polymerase binding (indicated in gray) and the elongation process of the complementary strand. Each time the polymerase adds a nucleotide (dNTP) to the growing strand, pyrophospate (PPi) is released. Sequencing enzymes, sulfurylase and luciferase, which are immobilized on an additional bead (indicated in red) transform the released PPi into a light signal. Sulfurylase converts PPi to ATP which is subsequently used by the luciferase enzyme to produce light from the luciferin substrate. B) In case of Illumina the bridge amplification is depicted in four steps. First ligated fragments randomly bind to spotted oligos on a slide. Next the bound fragment bridges with the free end to an compatible oligo on the slide. Polymerase binds and the complementary strand is produced. Denaturation of the bridge separates both strands, which are now free for further amplifications. C) Characteristic for SOLiD sequencing is that each nucleotide is queried two times. On the left hand side one ligation cycle is illustrated. After primer binding the ligase incorporates the oligo-nucleotide which carries the compatible di-nucleotide at its 3' end. The corresponding color signal is released and recorded. Finally the unspecific 5' end of the oligo-nucleotide is cleaved and the next ligation cycle starts. After a series of these cycles the extended sequence is removed and the first sequencing round is completed. On the right hand side the result of five sequencing rounds is shown. If the color signal and one base is known the unambiguous di-nucleotide can be determined. Since the annealed primer is shortened by one base each round the first three color signals can directly be decoded into the underlying di-nucleotides. This color signal into di-nucleotide decoding is propagated by the overlap of always two light signals in one base. Adapted from descriptions and images available online: `http://454.com`, `www.illumina.com`, `http://www.appliedbiosystems.com/absite/us/en/home.html`.

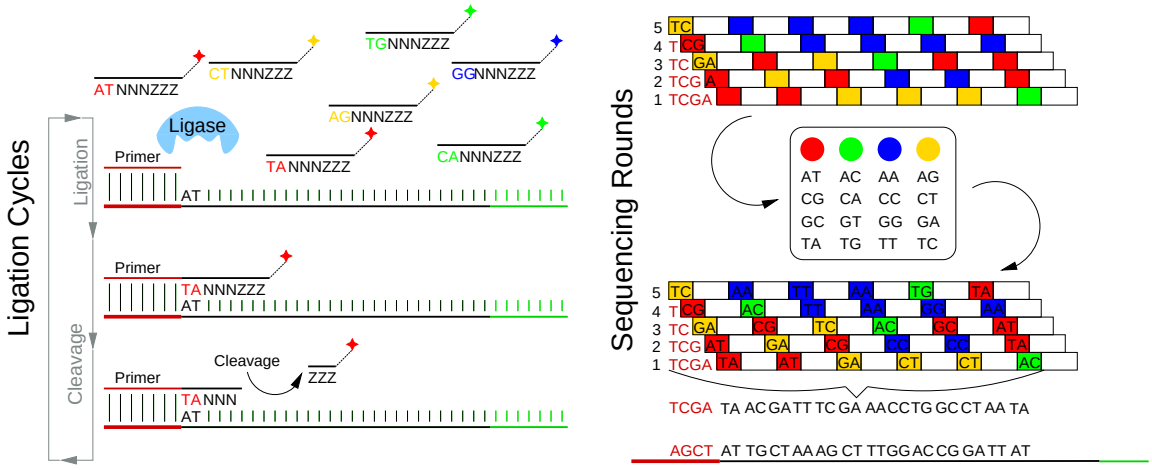
A) 454



B) Illumina



C) SOLiD



Illumina/Solexa sequencing* Well-defined 5' and 3' linker sequences are attached to isolated RNA fragments. Ligated fragments are size selected to a range of 150-200 bp. By complementary binding of the ligated linker sequences to single stranded oligo-nucleotides the fragments prepared are immobilized onto a slide. This slide is put into a flow cell and the so called "bridge amplification" is initiated, see Figure 2.14 B). Priming starts as the free linker of a bound fragment "bridges" to a complementary oligo on the surface. A double stranded bridge is generated and denaturation separates both strands. The duplicated fragments are now available for further bridge amplification steps. Repeated bridge amplification generates local clusters of identical sequences on the slide. Subsequently, the sequencing process is initiated. Specific primer sequences are annealed to the amplified fragments. Fluorescence labeled nucleotides are flowed one after an other over the slide. Whenever a base is incorporated local clusters emit a nucleotide specific color coded light signal and a high resolution camera records the signal distribution. The first nucleotides are used to calibrate the camera to cluster positions. Finally a time course of recorded images reveals the sequences of immobilized RNA fragments. At the moment Illumina reads range from 26-200 bases.

SOLiD sequencing† Standard library preparation begins with shearing of the RNA into small fragments and the ligation of unique adapter sequences to both 5' and 3' end, respectively. The 5' adapter is used for immobilization of the fragments onto beads. Each bead carries a single fragment which is amplified by emulsion PCR (see 454 sequencing). Subsequently, 3' modification of the fragments allow covalent binding of the beads onto a slide. Each fixated bead can be considered as a separated sequencing reaction, which is monitored in parallel with sequential digital imaging. The SOLiD sequencing technique incorporates oligo-nucleotides instead of single nucleotides. Each oligo is an 8-mer whose first two bases are one of 16 di-nucleotides (AC,CA, GT, TG,...). Position 3 to 5 are degenerated nucleotides (N's). The remaining bases are degenerated as well (Z's) and carry one of four fluorescent labeled dyes. Thus four di-nucleotides have to share the same fluorescent color. A universal sequence primer of length n initiates the first sequencing round. The first oligo-nucleotide is ligated to the 3' end of the primer. By that the di-nucleotide specific fluorescent signal is emitted. The first five bases (specific di-nucleotide and 3× N) are kept on the sequence and the rest is cleaved, Figure 2.14 C). This ligation process is repeated for 5-7 cycles. Then a reset is performed which removes the complement sequence of the fixated fragment. In the next round the universal sequence primer is shortened by one position to $n-1$ and again 5-7 cycles of oligo ligation and cleavage are performed. This procedure is repeated until a minimum length of $n-4$ of the universal primer is reached. Five rounds of primer ligation

*Information adapted from `www.illumina.com`

†Information adapted from `http://www.appliedbiosystems.com/absite/us/en/home.html`

with seven cycles of oligo ligation each result in 35 contiguous bases. The SOLiD approach queries each base by two different oligos. As mentioned above the recorded fluorescent signals do not directly tell which di-nucleotide has been incorporated. The knowledge of the annealed primer sequence which is shortened each round and the overlapping fluorescent signals are used to convert the recorded signal from the color into the sequence space, Figure 2.14 C). To date SOLiD sequencing achieves read lengths of 35 to 50 base pairs.

Each sequencing technique has its advantages and drawbacks. The SOLiD approach, for instance, is the only one which queries each base two times. Therefore it achieves a high sequencing accuracy and makes a precise identification of single nucleotide polymorphisms possible. However, the performance and accuracy of the method might be better if 16 di-nucleotides would not only be encoded by 4 fluorophores. Illumina and 454 techniques produce reads up to a length of 200 and 400 bases, respectively. While, Illumina results in several giga bases (Gb) of reads per run the 454 approach produces several million high-quality reads per run. It depends on the research focus which method one might prefer. In addition, different library preparation methods can be used to enrich different types of RNAs. Beside gel extraction methods as described above, (co-)immunoprecibitation, rRNA capturing and genomic SELEX (Systematic Evolution of Ligands by Exponential Enrichment) are commonly applied to specifically enrich or deplete RNA molecules (see Hüttenhofer & Vogel [2006], Sorek & Cossart [2010] and Croucher & Thomson [2010] for recent reviews, and Section 3.2).

2.3 Pathogenic bacteria analyzed

In order to survive bacteria have to monitor their surrounding and sense the existing environmental conditions. Fine tuning of the bacterial metabolism is especially important for pathogenic species. These bacteria have to rapidly activate expression of essential virulence genes if a host contact is sensed. In the next three paragraphs the pathogenic bacteria analyzed in this study are introduced.

2.3.1 *Pseudomonas aeruginosa* str. PAO1

Pseudomonas aeruginosa str. PAO1 (PAO1) is a Gram-negative ubiquitous microorganism which has been found in environments such as soil, water, humans, animals, plants, sewage, and hospitals [Hardalo & Edberg 1997]. The bacterium often infects patients whose immune system is already compromised (e.g. patients with cystic fibrosis, cancer, or AIDS) but seldom healthy individuals [Bodey *et al.* 1983]. PAO1 is therefore regarded as an opportunistic human pathogen. Intrinsic resistance to various different types of chemotherapeutic reagents and antibiotics, makes the bacteria a pathogen very hard to eliminate [Hancock 1998]. Furthermore, the bacterium is able to utilize a wide range of organic compounds as nutrient sources. This gives PAO1 an exceptional ability to colonize ecological niches where nutrients are limited. Analysis of its 6.4 million base pairs (Mbp) large genome sequence has identified genes involved in locomotion, attachment, transport and utilization of nutrients, antibiotic efflux, and systems involved in sensing and responding to environmental changes [Stover *et al.* 2000]. These features make PAO1 an interesting and important model organism which we analyzed in a cooperation with Udo Bläsi's group in Vienna (see Section 3.1).

2.3.2 *Helicobacter pylori* str. 26695

The human pathogen *H. pylori* is the major cause of chronic superficial gastritis as well as peptic ulcer disease. Approximately 50% of the worlds human population are infected with this Gram-negative bacterium of the ϵ-proteobacter group [Cover & Blaser 2009]. However, many people are carrier of the organism but are asymptomatic. The *Helicobacter* species colonize the stomach of their hosts and are highly adapted to this acidic environment. During infection *H. pylori* has to cope with different conditions such as pH and nutrient variations. Sequencing and annotation of the small and compact 1.67 Mbp genome of *H. pylori* revealed a very restricted repertoire of transcriptional regulators [Tomb *et al.* 1997]. Housekeeping RNAs such as tmRNA, RNaseP RNA and SRP RNA are known whereas a homolog of the 6S

RNA gene is still missing. Moreover, *H. pylori* lacks an homolog of Hfq [Valentin-Hansen *et al.* 2004]. However, the life style of the bacterium in combination with its compact genome necessitates additional regulatory mechanisms. Our cooperation with Jörg Vogel's lab revealed an unexpectedly complex transcript organization in *H. pylori*. Results are summarized in Section 3.2.

2.3.3 *Xanthomonas campestris* pv. *vesicatoria* str. 85-10

Many different pathovars of *Xanthomonas campestris* have been identified. These pathogens affect a wide range of plants including crops, cabbage, pepper, rice and paprika. *Xanthomonas campestris* pv. *vesicatoria* str. 85-10 (*XCV*) is a model system to elucidate the molecular communication between bacteria and plant. It is the causal agent of bacterial spot disease on pepper and tomato [Jones *et al.* 1998]. Essential for pathogenicity of *XCV* is the type III secretion (T3S) system, encoded by the hrp (hypersensitive response and pathogenicity) gene cluster [Bonas *et al.* 1991]. The main function of the T3S system is the trans-location of bacterial effector proteins into the host cell cytosol. In susceptible plants, effector proteins lead to disease symptoms such as typical black lesions that develop on plant surface. In resistant plants, effector proteins are recognized by plant resistant proteins which initiate the hypersensitive response. This rapid programmed cell death coincides with arrest of bacterial multiplication and stops infection. So far only protein-coding genes and their role during plant infection have been analyzed. Together with the group of Ulla Bonas from Halle, we investigated if sRNAs are involved as well. Results are shown in Section 3.3.

3

RNomics and Deep Sequencing

This chapter summarizes the results of three experimental surveys. First we analyzed the sRNA repertoire of *P. aeruginosa* in a joint project with Udo Bläsi's lab in Vienna. Here we focused on the detection of Hfq bound and evolutionary conserved sRNA sequences. The second part of this chapter describes the outcome of a cooperation with Jörg Vogel's lab in Berlin. Using a differential RNA sequencing approach we analyzed the primary transcriptome of *H. pylori*. An outline of the results of our joint project with the lab of Ulla Bonas in Halle is given in the third section of this chapter. Here the detection of virulence related sRNAs in *Xanthomonas campestris* was in the focus of research. Details on the experimental setup and used methods of all three studies can be found in our joint publications Sonnleitner *et al.* [2008], Sharma *et al.* [2010], Findeiß *et al.* [2010] and Schmidke *et al.* [in preparation].

3.1 Small RNA detection in *Pseudomonas aeruginosa* str. PAO1

Besides the housekeeping RNAs (i.e. 6S RNA, tmRNA) and RNA components of larger complexes (i.e. SRP and RNase P) the function of only a few additional PAO1 encoded sRNAs is known. The sRNAs RsmY [Valverde *et al.* 2003] and RsmZ [Heurlier *et al.* 2004] act by sequestration of the RsmA protein, a virulence gene regulator in PAO1 [Pessi *et al.* 2001]. This system seems to implement an analogon of the carbon storage regulator (csr) network found in *E. coli*, where two sRNAs (CsrB and CsrC) sequester the CsrA protein. [Babitzke & Romeo 2007, Timmermans & Melderen 2010]. Functional homologs of the *E. coli* encoded sRNA RyhB are PrrF1 and PrrF2 [Wilderman *et al.* 2004]. These sRNAs are $> 95\%$ identical to each other and have regulatory function in iron acquisition and storage. The expression of the tandem *prrF1* and *prrF2* genes is regulated by a Fur repressor in dependency on iron. If the essential nutrient becomes sparse (low iron concentration) the Fur blockade is removed and both sRNAs are transcribed. The sodB (superoxide dismutase) mRNA is a known target of PrrF1 and PrrF2.

In addition, 25 sRNAs have been computationally predicted and experimentally verified by Livny *et al.* [2006] and Gonzalez *et al.* [2008]. RgsA is one of the few examples which were predicted in both surveys. Expression of RgsA depends on the response regulator GacA and the stress sigma factor RpoS [Gonzalez *et al.* 2008]. Besides the 25 validated sRNAs dozens of intergenic candidates have been predicted in both studies.

In the contribution Sonnleitner *et al.* [2008] we used i) experimental RNomics approach and ii) a bioinformatic `RNAz` screen to identify sRNA candidates in PAO1. The aim of the RNomics approach was to identify sRNAs which interact with Hfq. The `RNAz` based analysis, on the other hand, aimed on the identification of evolutionary conserved RNA genes.

3.1.1 Identification of sRNA candidates with RNomics

Using an adopted version of the preparation protocol [Hüttenhofer & Vogel 2006] two cDNA libraries, enriched for Hfq bound RNA fragments, were generated by our collaborators (Udo Bläsi's lab, Vienna). Total RNA was isolated and centrifuged to deplete ribosomes. Subsequently, the size-fractionated (50-300 nt) RNA pool was co-immunoprecipitated with Hfq. Unbound RNA molecules were removed and the Hfq-bound fraction was sequenced. This procedure was repeated with cells grown to early stationary phase and cells exposed to hu-

Table 3.1. RNomics identified intergenic sRNA candidates. "Annotated transcripts" mapped to already known non-coding RNA genes and "experimentally verified" transcripts were independently validated either by Northern hybridization or RT-PCR experiments.

sRNA	Strand†	Growth‡	5'-end§	3'-end‖	Length	Identified sequence
annotated transcripts						
RsmY	←→←	S	586,867	586,990	124	586,867-586,690
tmRNA	←←←	LB				901,536-901,640
amiE leader	←←→	LB	3,778,134	3,778,034	100	3,778,054-3,778,098
RNase P	←←←	LB				4,956,348-4,956,591
experimentally verified						
$PhrD^{Northern}$	→→→	S/LB	758,497	785,570	72	785,498-785,547
$PhrS^{Northern}$	←←←	LB	3,705,522	3,705,309	212	3,705,342-3,705,515
$PhrX^{RT\text{-}PCR}$	→←→	S	5,836,429	5,836,579	151	5,836,450-5,836,479
$PhrY^{RT\text{-}PCR}$	←→→	LB	5,859,480	5,859,674	195	5,859,471-5,859,615
remaining candidates						
PhrC	←→→	LB				720,082-720,136
PhrR	→←→	LB				3,394,727-3,394,805
PhrU	→←→	S				4,332,627-4,332,676

†arrows indicate the genomicaly orientation of the sRNA candidate (middle arrow) and the adjacent genes

‡sRNA fragments were detected either in early stationary phase in LB medium (LB) and/or after human serum exposure (S)

§determined by primer extension

‖estimated from Northern blot experiments and rho-independent terminator

man serum. Using `blastn` the resulting sequencing clones were mapped onto the PAO1 genome. Details on the experimental setup and procedure can be found in our joint publication [Sonnleitner *et al.* 2008].

Although ribosomes were depleted from the RNA pool, 15% of both libraries mapped to rRNA genes. Additional 40% and 20% of the sequences originated from the sense or antisense strand of protein-coding genes of stationary phase and serum treated cells, respectively. Eleven sRNA candidates were predicted to have a closely located rho-independent terminator signal and are localized in intergenic regions (Table 3.1).

Three of the eleven candidate sRNAs show sequence similarity to sRNA genes already known. The presence of RsmY in the human serum exposed cDNA library verified the observed interaction of RsmY and Hfq [Sonnleitner *et al.* 2006]. Interestingly, a fragment at the 5'-

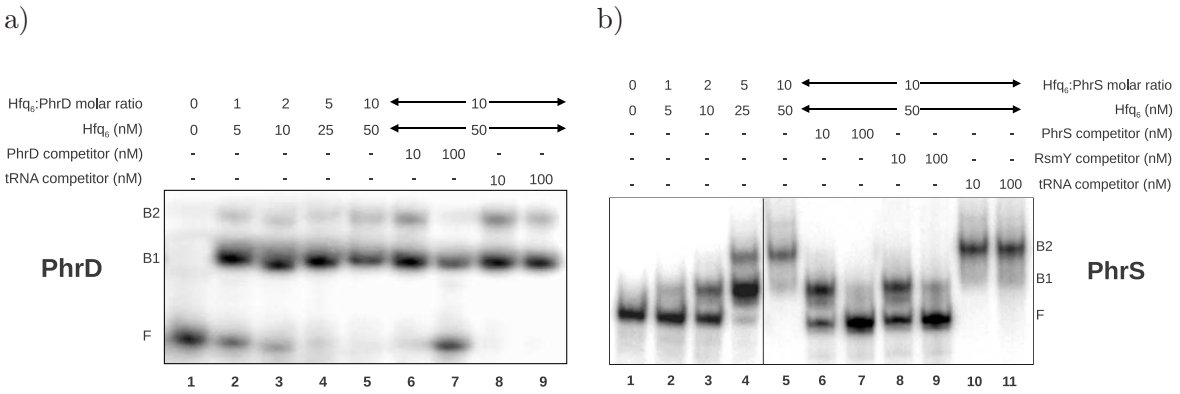


Figure 3.1. The Hfq hexamer protein (Hfq_6) binds PhrD and PhrS. The positions of free and Hfq-bound RNA are indicated by F and B1/B2, respectively. Labeled RNA (5 nM) was incubated in the absence (lane 1) or in the presence of 5 nM (lane 2), 10 nM (lane 3), 25 nM (lane 4) and 50 nM (lane 5) Hfq. The ratio of Hfq to RNA is indicated on top. Unlabeled PhrD and PhrS RNA (lane 6 and lane 7), PAO1 RsmY RNA (b, lane 8 and 9) and tRNA (a, lane 8 and 9; b, lane 10 and 11) were added as competitors.

end of tmRNA was found to co-immunoprecipitate with Hfq. The third Hfq-bound RNA overlapped with the annotated RNase P gene. Since tmRNA and RNase P were already annotated in PAO1 and experiments in *E. coli* showed that Hfq does not bind to these transcripts [Wassarman *et al.* 2001, Zhang *et al.* 2003] further analysis of these genes was postponed. Another transcript mapped to the upstream region of the amidase operon. This Hfq-binding RNA fragment corresponds to the already validated leader RNA of the *amiE* gene [Wilson & Drew 1995]. Like the other known RNAs, this locus was not further studied.

Among the remaining seven candidates, four Pseudomonas Hfq-binding RNA (Phr) transcripts could be validated by Northern hybridization or RT-PCR experiments (Table 3.1). Sequence length of these Phr transcripts varies between 72 and 212 nt. While transcripts of PhrS and PhrY were expressed in stationary phase only, PhrX was only detected after human serum exposure. PhrD was found to be expressed under both tested conditions. `Blast` analysis of all four Phr transcripts revealed that PhrD is only present in PAO1 and PhrS, PhrX as well as PhrY homologs were found in different *Pseudomonas* species (data not shown). Interestingly, the PhrS locus was already detected as candidate P20/1887 in the studies of Livny *et al.* [2006] and Gonzalez *et al.* [2008]. This together with the fact that PhrD as well as PhrS were abundant RNA species shifted the focus of all further experiments on these two transcripts.

First the binding of PhrS and PhrD to Hfq was verified by band-shift assays (Figure 3.1). Here the purified Hfq hexamer protein (Hfq_6) [Sonnleitner *et al.* 2006] was added in increasing

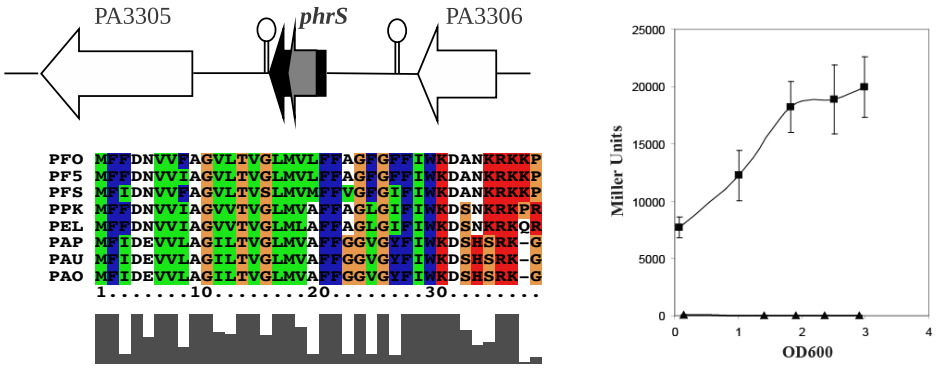


Figure 3.2. PhrS (black arrow) is genomicaly encoded between the genes PA3305 and PA3306 (white arrows). Terminator sequences are indicated by stem loop structures. The *phrS*-ORF is represented as gray arrow. Below, the multiple alignment of the amino acid sequences of pseudomonas homologs. Expression of the *phrS*-ORF was evidenced by the synthesis of the PhrSΦLacZ fusion protein (right hand side). The β-galactosidase activities were determined for the constructs *phrS*-ORF-*lacZ* (■) and *phrS*-ORF$_{AUG\rightarrow CUG}$-*lacZ* (▲) by triplicated experiments as described in [Miller 1972].

molar ratios to the 5'-end labeled RNA. A PhrD-Hfq complex has already been observed with a molar ratio of 1:1 (Figure 3.1a, lane 2 B1). An additional shift (B2) was only marginal and did not increase with high Hfq concentrations. The competition experiment (Figure 3.1a, lane 6-9) suggested that Hfq specifically binds to PhrD. Upon twofold molar excess of Hfq unlabeled PhrD RNA competed with the labeled RNA. As expected, no competition has been observed when *E. coli* tRNA was added. With an increasing molar excess of Hfq two bands were observed for PhrS (Figure 3.1b, lane 2-5). The additional shift (B2) indicates a 2:1 stoichiometry of Hfq and PhrS molecules. This is consistent with the observation that the second band was more prominent at higher molar excess of Hfq (Figure 3.1b, lane 5). Unlabeled PhrS RNA and RsmY RNA competed with the Hfq-PhrS complex (Figure 3.1b, lane 6-9), whereas the non-specific competitor *E. coli* tRNA did not (Figure 3.1b, lane 10 and 11).

Subsequent Northern blot experiments comparing wild-type PAO1 and a *hfq* deletion mutant revealed a 50% reduction of the steady state levels of both PhrS and PhrD. Half-lives of PhrD and PhrS were comparable in PAO1 and the *hfq* mutant strain. Hence, Hfq appears to regulate the expression of both sRNAs rather than to stabilize them. In order to identify possible target genes in proteome studies we attempted to transform both sRNA genes into a plasmid system. Unfortunately, several efforts to transform *phrD* into the plasmid failed. Plasmid-

directed over-expression of *phrS* followed by proteomics identified three possible targets: the heat-shock chaperon GroEL, the outer membrane porin OprD and the putative periplasmic binding protein PA51353. The proteome analysis of PhrS suggest that the sRNA could function as a riboregulator.

Inspection of the *phrS* gene revealed a coding capacity of a 37 amino acid long peptide. This short ORF was found to be conserved in all PhrS homologs (Figure 3.2). To test if the *phrS*-ORF is expressed a *phrS*-ORF-*lacZ* construct was engineered. In another construct (*phrS*-ORF$_{AUG\rightarrow CUG}$-*lacZ*) the canonical start codon was mutated to CUG. As expected expression of the *phrS*-ORF was evidenced by synthesis of the fusion protein PhrSΦLacZ. In contrast the *phrS*-ORF$_{AUG\rightarrow CUG}$-*lacZ* construct did not synthesize the fusion protein, see Figure 3.2. These experiments conducted by our collaborators indicate that PhrS is a new candidate for a dual-functional sRNA, acting on RNA and protein level.

3.1.2 PAO1 sRNAs predicted with `RNAz`

In addition to the RNomic approach, we made use of `RNAz` to scan PAO1 for additional sRNA candidates. Since the results of a `RNAz` screen depend on the quality of the analyzed multiple sequence alignments we used two approaches. Based on the less restrictive `multiz` [Blanchette *et al.* 2004] alignments 221 loci were predicted, of which 85 correspond to already annotated ncRNAs. Using `NcDNAlign` [Rose *et al.* 2008a] alignments as input we found 115 structured candidates, of which 101 correspond to known sRNA loci and 14 are novel candidates. Compared to `NcDNAlign`- the `multiz`-based screen resulted in significantly more `RNAz` hits. However, the `NcDNAlign` based screen identified more annotated hits at the cost of only a few novel candidates (Table 3.2*).

Besides tRNAs and rRNAs, housekeeping RNAs (6S, tmRNA) and RNA components of larger complexes (SRP and RNase P) were recovered in both screens. The FMN riboswitch is highly conserved structural element typically found within 5' UTRs of riboflavon biosynthesis (*rib*) genes. In PAO1 we found a homolog of this regulatory element upstream of *ribC*. The expression of a 180 nt long leader transcript has been verified by Gonzalez *et al.* [2008] (candidate 2315). Only 57 genomic loci were predicted in both `RNAz` screens, seven of which were novel. Indeed, expression in stationary phase of two of these seven loci has been demonstrated by RT-PCR. The remaining candidates may be expressed in different conditions.

Eight Hfq-binding sRNAs detected with the RNomics approach were neither detected in our

*Details of the `RNAz` screens of booth `multiz`- and `NcDNAlign`-generated alignments are provided at `http://www.bioinf.uni-leipzig.de/publications/supplements/07-023`

Table 3.2. Counts of RNAz loci for the NcDNAlign and multiz approaches. Hits are partitioned in two groups according to the prediction confidence ($p - value \geq 0.5$ and ≥ 0.9). Information from public databases (NCBI, Rfam, Noncode, tmRDB and ncRNAdb) was used to annotated RNAz loci. Additionally, RNAz loci were overlapped with our RNomics candidates and novel sRNAs identified by Livny *et al.* [2006] and Gonzalez *et al.* [2008]. The last row indicates all remaining RNAz loci of unknown function.

	NcDNAlign		multiz	
	$p \geq 0.5$	$p \geq 0.9$	$p \geq 0.5$	$p \geq 0.9$
RNAz loci	115	98	221	166
Annotated	101	89	85	74
Livny *et al.* [2006]	4	3	10	9
Gonzalez *et al.* [2008]	4	4	3	3
RNomics candidates	2	1	3	3
Unknown	14	9	136	92

RNAz screens nor by Livny *et al.* [2006] and Gonzalez *et al.* [2008]. Similar to PhrD, these candidates are likely to be PAO1 specific sRNAs. Among the remaining three candidates PhrS is the only novel and experimentally verified RNA. Since PhrS emerged from four different screens (P20 in Livny *et al.* [2006], 1887 in Gonzalez *et al.* [2008], Table 3.1 and Table 3.2) and *phrS* over-expression resulted in changes of protein patterns, it appears to be a prime candidate for a novel sRNA in *Pseudomonas*.

3.2 The primary transcriptome of *Helicobacter pylori*

The small and compact genome (1.67 Mbp) of *Helicopacter pylori* strain 26695 contains 1,576 ORFs, but relatively few genes encoding transcriptional regulators such as σ factors and two-component systems [Tomb *et al.* 1997, Alm *et al.* 1999]. Previous annotations of the *H. pylori* genome identified tRNAs, rRNAs, tmRNA, RNase P RNA and SRP RNA. Surprisingly, various attempts to identify 6S RNA homologs in *H. pylori* and the complete ϵ-proteobacteria subdivision failed [Weinberg *et al.* 2007, Barrick *et al.* 2005, Wassarman & Storz 2000]. Perhaps in accordance with the absence of Hfq in all ϵ-proteobacteria [Valentin-Hansen *et al.* 2004] no additional sRNAs have been identified in *H. pylori*.

In the cooperation with Jörg Vogel's lab a novel differential RNA sequencing (dRNA-seq) approach was established and used to identify the native 5' end of transcripts. This gave us the opportunity to analyze transcription initiation (the primary transcriptome) on a genome-wide scale and to unravel the unexpected complexity of transcript organization and RNA output of *H. pylori* [Sharma *et al.* 2010].

3.2.1 Differential RNA sequencing (dRNA-seq)

In order to cover the complete *H. pylori* transcriptome, RNA was extracted from standard growth, acid stress and infection samples (Figure 3.3). Primary bacterial transcripts (most mRNAs and sRNAs) have 5' characteristic tri-phosphate (5'PPP) ends. In contrast, processed RNA species (e.g. rRNA and tRNA) have 5' mono-phosphate (5'P) ends. To distinguish between primary and processed transcripts, two differential cDNA libraries of each sample were prepared by our collaborators: one library (−) from *H. pylori* total RNA containing both types of transcripts, and the other (+) following enrichment for primary transcripts by treatment with terminator exonuclease (TEX), which degrades 5'P but not 5'PPP RNA. Following 454 sequencing, a total of ∼217 million bases of sequenced cDNA reads were mapped to the *H. pylori* chromosome.

The combined cDNA reads of (−) or (+) libraries were observed to be distributed over the entire chromosome. Inspection of individual genomic loci such as the urease operon and the cag pathogenicity island, two key loci of *H. pylori* virulence, confirmed the expected expression of annotated ORFs. The TEX treatment revealed TSS by causing a characteristic change in the cDNA distribution over individual genes, resulting in a sawtooth-like profile with an elevated sharp 5' flank (Figure 3.4). In contrast, expression of processed transcripts peaks in the untreated library (Figure 3.4). As an example, the expression pattern matched the known TSS of cagA [Spohn *et al.* 1997] and discriminated the primary and processed 5'

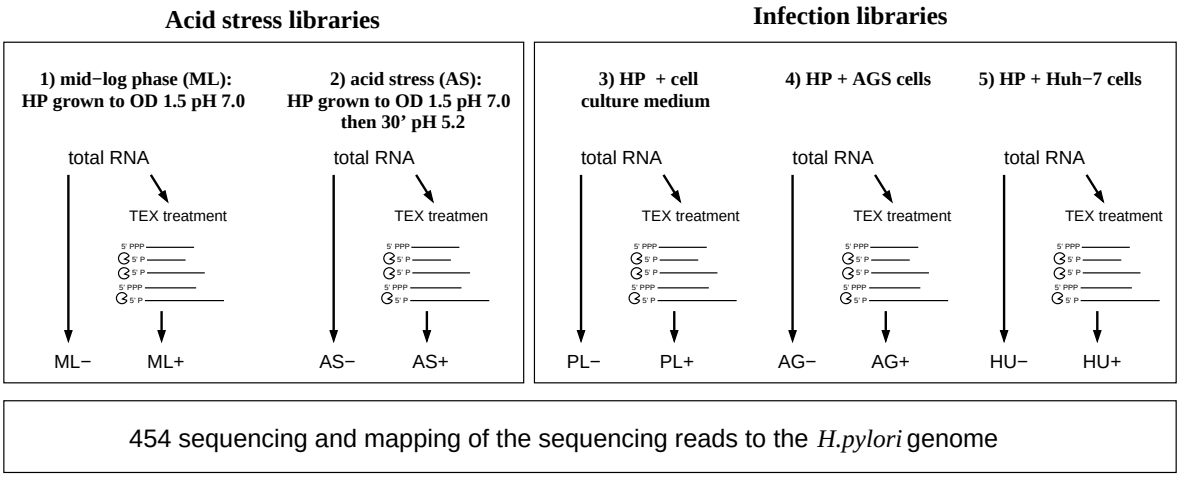


Figure 3.3. Total RNA extracted from five conditions: 1) culture in BHI media grown to mid-log phase (ML), and 2) following 30 min acid stress at pH 5.2 (AS), growth in cell culture medium in the 3) absence (PL), or 4) presence of responsive AGS human gastric epithelia cells (AG) and 5) non-responsive Huh-7 liver fibroblasts (HU). For each condition one library without (−) and one with terminator exonuclease (TEX) treatment (+) were prepared and sequenced on a Roche (454) FLX sequencer. After clipping of linker and 3' poly(A) sequences, reads were mapped to the *H. pylori* genome using the short read aligner `segemehl` [Hoffmann *et al.* 2009].

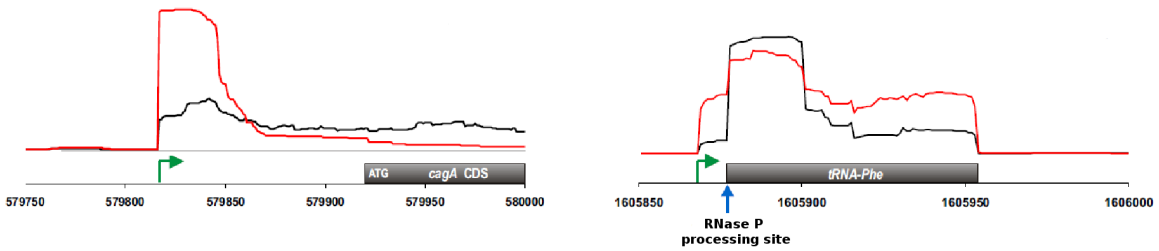


Figure 3.4. Schematic drawing of cDNA enrichment patterns. Annotated genes are indicated as gray bars and green arrows correspond to annotated TSS. Depletion of processed RNAs by TEX treatment leads to a characteristic change in the read distribution over transcripts, as shown for cagA and tRNA-Phe (expression profile from acid stress cDNA libraries). Exonuclease treatment (red) shifts the cagA cDNAs towards the nuclease-protected 5'-end, yielding a sawtooth-like profile with an elevated sharp 5' flank that matches the published TSS [Spohn *et al.* 1997]. In contrast, the mature (RNase P-cleaved) 5' end of tRNA-Phe is predominant in the untreated library (black curve) and the characteristic profile is missing.

ends of tRNA-Phe transcripts (Figure 3.4). For verification, 74 *H. pylori* TSS determined by independent experimental approaches were compared to TSS positions deduced from cDNA expression patterns. In fact 87% of the dRNA-seq determined TSS matched within a ±2 nt tolerance, demonstrating the high accuracy of our approach.

To build a genome-wide TSS map for *H. pylori* 5' ends that were enriched in the (+) versus (−) library in at least 2 of the 5 conditions and satisfied additional criteria such a plausible position relative to adjoining genes were manually annotated as TSS. This identified a total of 1,907 TSS which were grouped into five categories: 812 primary TSS (most cDNAs $\leq$ 500 bp upstream of annotated start codons or mature 5' ends of small RNAs), 119 secondary TSS (with fewer cDNA reads in the same gene), 439 internal TSS (internal to annotation on sense strand), 969 antisense TSS (antisense inside or within 100 bp of an annotation on the opposite strand) and 38 orphan TSS (no annotation in close proximity). Note that one TSS can be assigned to several categories.

3.2.2 Promoter and 5' UTR analysis

To identify promoter motifs of the housekeeping σ-factor of *H. pylori*, we used `MEME` [Bailey & Elkan 1995, Bailey & Gribskov 1996] to scan all 1,907 TSS for conserved upstream motifs. `MEME` searches for similarities among a given set of sequences and calculates descriptors

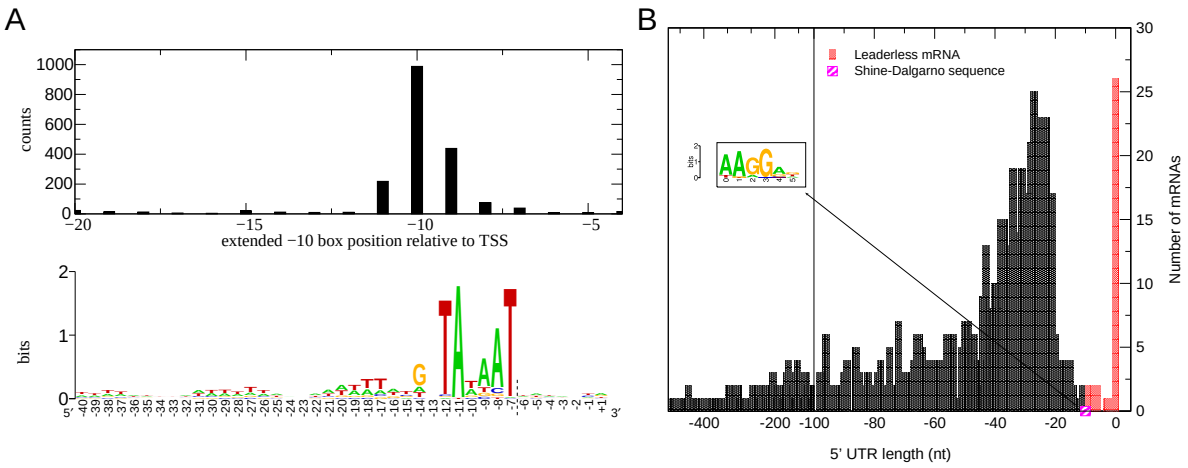


Figure 3.5. Regulatory motifs detected in promoter region (A) and 5' UTRs (B). Motif searches upstream of *H. pylori* TSS reveal extended Pribnow boxes (-10 signal: tgnTAtaAT) about 86% of which lie -11, -10 or -9 bp relative to TSS (top, histogram of distribution; bottom, logo of upstream TSS sequences). Frequency of individual 5' UTR length based on 825 TSS (primary and secondary) of mRNAs. 5' UTR lengths $\leq$ 10 nt (red bars) reveal 34 leaderless mRNAs. The inset shows the Shine-Dalgarno sequence motif of *H. pylori*.

for these motifs. Using 50 nt upstream sequence we discovered an extended Pribnow box, tgnTAtaAT, infront of all TSS (Figure 3.5). Instead of a -35 motif, we observed a periodic AT-rich signal upstream of -14 (position relative to TSS +1). This resembles the *in silico* predicted structure of σ^{80}-type promoters in *H. pylori* and the related Campylobacter jejuni bacteria [Vanet *et al.* 2000, Petersen *et al.* 2003]. Interestingly, promoter motifs of *H. pylori*'s alternative σ-factors (σ^{28} and σ^{54}) were not detected by the analysis of the complete TSS set. This indicates that the majority of transcriptional initiation in ϵ-subdivision bacteria is determined by the extended Pribnow box, relying on upstream periodic AT-rich stretches instead of a distinctive -35 motif.

The annotation of primary TSS for 825 mRNAs revealed the lengths of the 5' UTRs, i.e. the distance from TSS to the start codon. Approximately half of the 5' UTRs are 20-40 nt in length which is in accordance with the structurally determined mRNA contacts of initiating ribosomes [Ramakrishnan 2002]. Motif discovery in the 789 5' UTRs ($\geq$ 10 nt) and comparison against the entire genome confirmed the AAGGag motif [Vanet *et al.* 2000, Yada *et al.* 2001] as the consensus Shine-Dalgarno sequence of *H. pylori* mRNAs (Figure 3.5). The motif occurs with a median distance of 6 nt to the annotated start codons.

In Gram-negative species, leaderless mRNAs (UTR of <10 nt) are considered rare and primarily phage-associated. Our data indicates, however, that > 2% of all *H. pylori* proteins are synthesized from leaderless mRNAs (Figure 3.5). In case of 26 genes, including the important *dnaA*, *soj*, *recR*, and *hemH* housekeeping genes, transcription initiates exactly at the start codon, and all of these leaderless mRNAs possess an AUG start codon which is critical for stable ribosome binding [Brock *et al.* 2008].

Comparison of the mapped TSS with the position of the cognate ORFs revealed cases where the TSS was located downstream of the annotated start codon. We propose re-annotation of 18 additional *H. pylori* genes, and in most of these cases our experimental data support previous re-annotation of ORFs by genome comparison [Boneca *et al.* 2003].

3.2.3 An unexpected wealth of RNA regulators

Long 5' UTRs might contain post-transcriptional control elements such as riboswitches [Weinberg *et al.* 2007]. Comparison of the *H. pylori* 5' UTRs with the RNA families database Rfam confirmed the predicted thiamine pyrophosphate (TPP) riboswitch upstream of pnuC [Rodionov *et al.* 2002]. Both dRNA-seq and Northern blot analysis detected a ∼100 nt transcript from the 5' UTR of this gene, which likely results from transcriptional attenuation by the TPP riboswitch. Although no other known riboswitches were found there are 337 UTRs long enough (> 60 nt) to harbor regulatory RNA elements, see Figure 3.5.

All TSS that were not found in vicinity upstream of mRNAs represented sRNA candidates. We observed hundreds of sRNA candidates in intergenic regions, antisense to annotated ORFs, and less frequently, from the sense strand of ORFs. Northern blot experiments conducted by our collaborators validated the expression of about 60 new sRNAs. The sRNAs expressed from intergenic regions include a five-member family of ∼200 nt RNAs, whose apparent redundancy is reminiscent of the Qrr sRNA family acting in the control of quorum sensing and virulence in *Vibrio* species [Lenz *et al.* 2004].

The candidates also included the elusive 6S RNA (∼180 nt) which had notoriously failed to be identified in the ϵ-subdivision [Wassarman & Storz 2000, Barrick *et al.* 2005, Weinberg *et al.* 2007]. *H. pylori's* 6S RNA accumulates throughout growth and is located opposite to a non-conserved hypothetical ORF. Structural probing of *in vitro* synthesized *H. pylori* 6S RNA confirmed the characteristic long hairpin structure with a central asymmetric bulge that mimics DNA in an open promoter complex of RNAP (Figure 3.6). Interestingly we detected two distinct classes of pRNAs: one pRNA expressed from a similar region as in *E. coli* [Wassarman & Saecker 2006] and the other, pRNA*, originating from the opposite strand

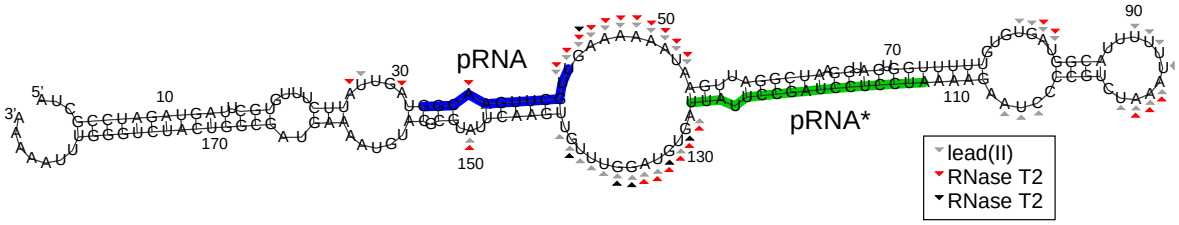


Figure 3.6. Structural probing confirmed the characteristic 6S RNA secondary structure [Barrick *et al.* 2005]. Rectangles indicate cleavage sites detected by *in vitro* probing (performed by our collaborators, Jörg Vogel's lab, Berlin) using RNase T1 (single stranded RNA; preferentially unpaired G), RNase T1 (single stranded RNA; preferentially unpaired A) and lead(II) (single stranded RNA). The template nucleotides of detected pRNA and pRNA* sequences are highlighted in blue and green, respectively. Secondary structure predicted by `RNAfold` according to the structural constraints.

(Figure 3.6).

According to cDNA coverage, some of the newly discovered *H. pylori* sRNAs are as abundant as the high-copy 6S RNA. Structural comparison of sRNA candidates and known *E. coli* as well as *Salmonella* sRNAs revealed only a few *H. pylori* transcripts that showed similarities to the repertoire of enterobacteria. In their majority, the new sRNAs are conserved in the *Helicobacter* species but rarely outside the ϵ-subdivision. Thus, except for 6S RNA and housekeeping RNAs, ϵ-proteobacteria including *H. pylori* might have evolved a unique sRNA repertoire.

3.2.4 New genes coding for short peptides

We constructed a pipeline to systematically analyze the coding capacity of all *H. pylori* sRNA candidates. An ORF was deemed valid if it started with the canonical start codon AUG, ended in frame with one of the three stop codons UAA, UAG or UGA and encoded at least 10 amino acids. Additionally, we enforced a 60% match including the sub-sequence AGG[A|G] of the SD-pattern within a distance of eleven nucleotides upstream of the ORF. Furthermore, this set of short peptides was reduced by a restrictive conservation filtering step. Using `Blast` homologs of the sRNAs (e-value 10; initial hit has to cover $\leq$ 50% of the *H. pylori* query) as well as the corresponding ORFs were searched within all fully sequenced ϵ-proteobacteria. Both truncated sRNAs and ORFs are maximally extended to the length of the initial *H. pylori* query sequence. Finally, valid ORF conserved in at least three ϵ-proteobacteria species passes our pipeline. In total 67 conserved short peptides were predicted.

We identified a family of six structurally related, 80 nt sRNAs expressed antisense to small ORFs of homologous 22-30 amino acid peptides, henceforth referred to as IsoA1-6 (RNA inhibitor of small ORF family A) and *aapA1-6* (antisense RNA-associated peptide family A), respectively (Figure 3.7). Five of the *aapA* genes produced stable mRNAs *in vivo*. *In vitro* translation assays yielded the expected small peptides, except for aapA2 mRNA whose Shine-Dalgarno sequence is mutated in strain 26695. Furthermore, translation of aapA1 or aapA3 was strongly and specifically inhibited in the presence of the cognate IsoA1 or IsoA3 sRNAs, thus revealing candidates of antisense regulation in *H. pylori*. The AapA peptides are conserved in other *H. pylori* strains and might interact with membranes, as suggested by their predicted high hydrophobicity, see Figure 3.7. Therefore, the *aapA-isoA* loci might be toxin-antitoxin systems that slow down growth of *H. pylori* or other organisms in the gastric mucosa.

We identified four additional antisense RNA/small ORF (asRNA/sORF) cassettes, see Figure 3.7. The aapB (antisense RNA associated peptide B) locus includes two experimentally validated sRNAs: one is a ∼100 nt long antisense RNA whereas the other encodes a 42 aa ORF conserved in many other *Helicobacter* strains. For the remaining three asRNA/sORF cassettes only the sRNAs are highly expressed. All three sRNAs, of which two are probably the result of a duplication event, contain a nearly perfect anti-Shine-Dalgarno sequence (TCTCCT). Thus, we termed these loci *aapC1/2* and *aapD*. Conservation analysis showed that the ShineDalgarno sequences, start codons and peptides are highly conserved in other *Helicobacter* strains. Moreover, the asRNA/sORF pairs are present in different copy numbers at the same genomic locations in different strains. Additional cassettes that are similar in sequence to both *aapC1/2* and *aapD* were identified in the other strains. Given the sequence similarity between the peptides of *aapD* and *aapC1/2*, these could constitute one large family. Furthermore, they have similar sequences to the short hydrophobic peptides of the Ibs family which has recently been identified in *E. coli* [Fozo *et al.* 2008a; b]. Overall, besides the AapA1-6 family, four additional small ORF/antisense RNA cassettes in the *H. pylori* genome were identified.

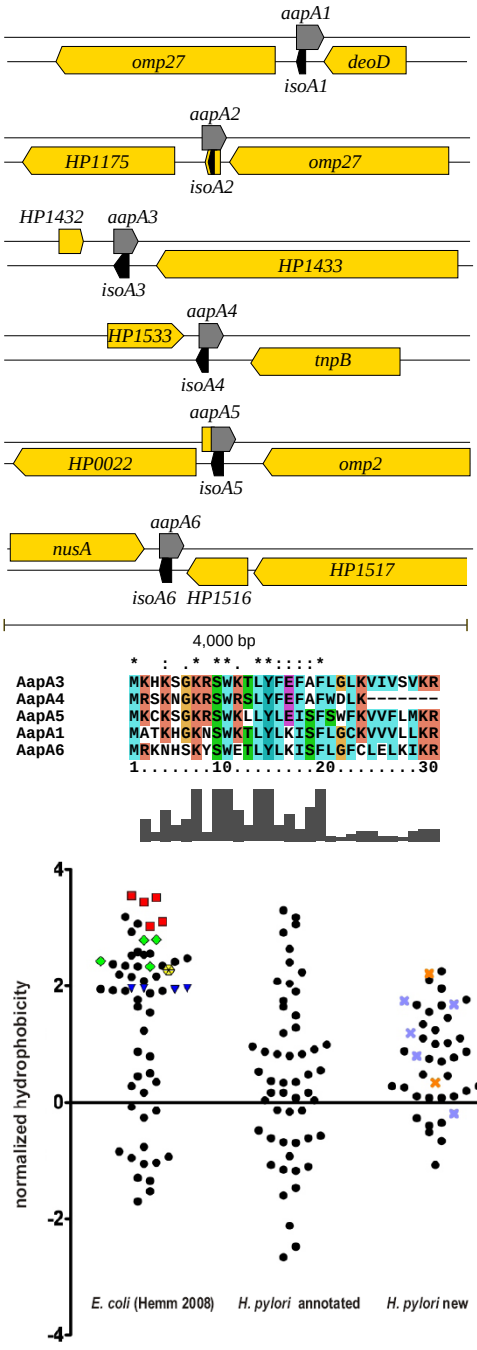
In order to estimate the hydrophobicity of the predicted short peptides, the Kyte-Doolittle scale [Kyte & Doolittle 1982] was used. A sliding window approach was applied to scan each peptide for the most hydrophobic sub-sequence of 11 amino acids. For each peptide scored one thousand random sequences out of the annotated and translated CDS were sampled and scored as well. This sample represents a normal distributed background. Thus the data-set could be normalized using the mean and the standard deviation of the background distribution. Using a confidence interval of 95% a normalized hydrophobicity value is significant if it is > 2.

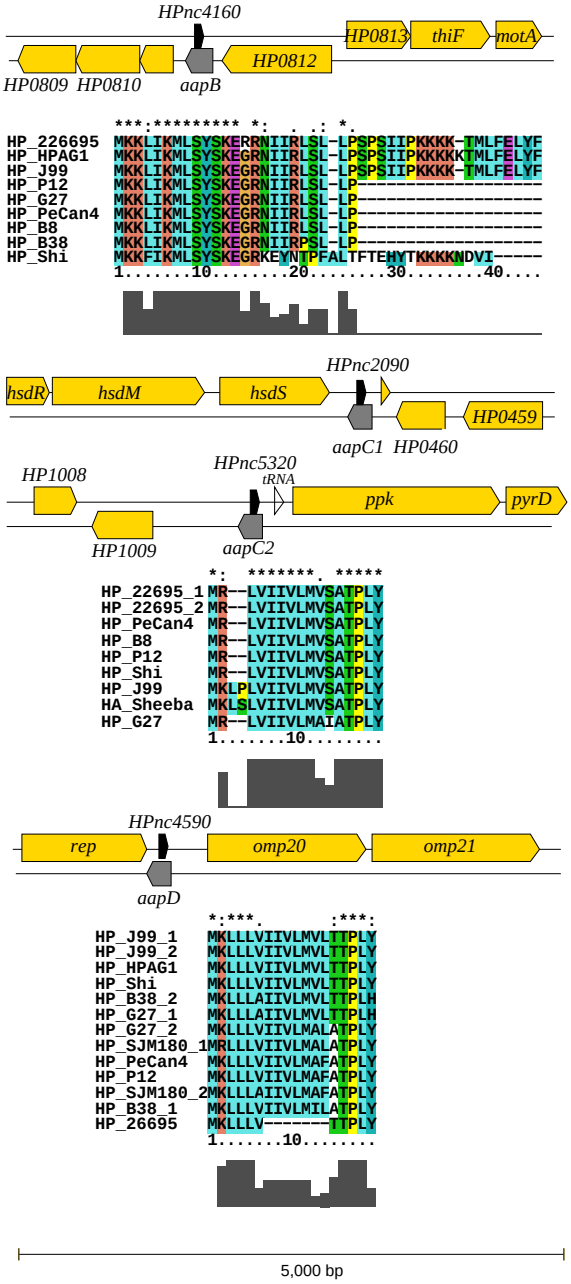
Figure 3.7 depicts the results for three independent data sets: i) experimentally verified short peptides from *E. coli* [Hemm *et al.* 2008]; ii) 53 annotated short ORFs $\leq$ 50 aa from the *H. pylori* NCBI annotation; iii) short peptides predicted in this study. Some of the *E. coli* peptides had already been predicted by others to be hydrophobic [Fozo *et al.* 2008a]. These molecules also have high hydrophobicity values as determined by the method described above. We note that the sets of previously annotated and predicted short peptides from our *H. pylori* study showed no general enrichment hydrophobicity values above the selected confidence interval of 95%, but that the majority of the newly predicted peptides from this study show positive hydrophobicity values.

Figure 3.7 (following page). The top left panel depicts the genomic locations of IsoA1-6 RNA (black boxes) and associated peptide-encoding *aapA* mRNAs (gray). IsoA2 overlaps with hypothetical ORF HP1176. In strain 26695, *aapA2* lacks a conserved ORF. IsoA5 and IsoA6 lie opposite to hypothetical ORFs, HP0024 (*aapA5*) and HP1515 (*aapA6*), respectively. Sequence alignment of AapA1 and AapA3-AapA6 is shown below (standard `ClustalX` color code).

The right panel shows additional short peptide coding RNAs and their associated antisense RNAs. Conservation of the short peptide is depicted below each genomic location (alignment of homologous peptides). Note that the peptides AapC1 and AapC2 are identical in sequence.

Hydrophobicity plot (bottom left) of known short peptides of *E. coli* (left column), of previously annotated small ORFs $\leq$ 50 amino acids in *H. pylori* (middle), and of small ORFs predicted in this study (right). The y-axis indicates a normalized hydrophobicity value for each protein. Positive values indicate increased hydrophobicity. The hydrophobicity of a protein is significantly higher than expected by chance if its value is above two standard deviations (normalized hydrophobicity $\geq$ 2). In the *E. coli* data set [Hemm *et al.* 2008] the Ibs (red squares), Hok (green diamonds) and Ldr (blue rectangles) clusters as well as the TisB peptide (yellow hexagon) are highlighted. In the right column, light blue crosses denote the Helicobacter AapA peptides, and orange crosses the AapC/D peptides.


aapA1
omp27
deoD
isoA1
aapA2
HP1175
omp27
isoA2
HP1432
aapA3
HP1433
isoA3
aapA4
HP1533
tnpB
isoA4
aapA5
HP0022
omp2
isoA5
aapA6
nusA
HP1517
isoA6 HP1516
4,000 bp
AapA3 MKHKSGKRSWKTLYFEFAFLGLKVIVSVKR
AapA4 MRSKNGKRSWRSLYFEFAFWDLK-------
AapA5 MKCKSGKRSWKLLYLEISFSWFKVVFLMKR
AapA1 MATKHGKNSWKTLYLKISFLGCKVVVLLKR
AapA6 MRKNHSKYSWETLYLKISFLGFCLELKIKR
normalized hydrophobicity
E. coli (Hemm 2008)
H. pylori annotated
H. pylori new


HPnc4160
HP0813
thiF
motA
HP0809 HP0810
aapB
HP0812
HP_226695 MKKLIKMLSYSKERRNIIRLSL-LPSPSIIPKKKK-TMLFELYF
HP_HPAG1 MKKLIKMLSYSKEGRNIIRLSL-LPSPSIIPKKKKKTMLFELYF
HP_J99 MKKLIKMLSYSKEGRNIIRLSL-LPSPSIIPKKKK-TMLFELYF
HP_P12 MKKLIKMLSYSKEGRNIIRLSL-LP----------
HP_G27 MKKLIKMLSYSKEGRNIIRLSL-LP----------
HP_PeCan4 MKKLIKMLSYSKEGRNIIRLSL-LP----------
HP_B8 MKKLIKMLSYSKEGRNIIRLSL-LP----------
HP_B38 MKKLIKMLSYSKEGRNIIRPSL-LP----------
HP_Shi MKKFIKMLSYSKEGRKEYNTPFALTFTEHYTKKKKNDVI-----
1.....10.....20.....30.....40....
HPnc2090
hsdR
hsdM
hsdS
HP0459
aapC1 HP0460
HP1008
HPnc5320
tRNA
ppk
pyrD
HP1009
aapC2
HP_22695_1 MR--LVIIVLMVSATPLY
HP_22695_2 MR--LVIIVLMVSATPLY
HP_PeCan4 MR--LVIIVLMVSATPLY
HP_B8 MR--LVIIVLMVSATPLY
HP_P12 MR--LVIIVLMVSATPLY
HP_Shi MR--LVIIVLMVSATPLY
HP_J99 MKLPLVIIVLMVSATPLY
HA_Sheeba MKLSLVIIVLMVSATPLY
HP_G27 MR--LVIIVLMAIATPLY
1.....10.....
HPnc4590
rep
omp20
omp21
aapD
HP_J99_1 MKLLLVIIVLMVLTTPLY
HP_J99_2 MKLLLVIIVLMVLTTPLY
HP_HPAG1 MKLLLVIIVLMVLTTPLY
HP_Shi MKLLLVIIVLMVLTTPLY
HP_B38_2 MKLLLAIIVLMVLTTPLH
HP_G27_1 MKLLLAIIVLMVLTTPLH
HP_G27_2 MKLLLVIIVLMALATPLY
HP_SJM180_1 MRLLLVIIVLMALATPLY
HP_PeCan4 MKLLLVIIVLMAFATPLY
HP_P12 MKLLLVIIVLMAFATPLY
HP_SJM180_2 MKLLLAIIVLMAFATPLY
HP_B38_1 MKLLLVIIVLMILATPLY
HP_26695 MKLLLV------TTPLY
1.....10.....
5,000 bp

The single nucleotide resolution TSS map compiled during our analysis constitutes the third global reference data set for *H. pylori*, complementing its genome sequence [Tomb *et al.* 1997, Alm *et al.* 1999] and global protein-protein interaction map [Rain *et al.* 2001]. Because our experimental conditions mimic the most prevalent environmental conditions encountered by *H. pylori* during infection, we are likely to have captured the vast majority of TSS. Our data provides new insight into *H. pylori* gene expression, and indicates that RNA-mediated control has been underestimated in the important ϵ-proteobacterial group of pathogens. The unexpectedly high numbers of antisense transcripts (969 antisense TSS) and sRNAs ($\sim$60 validated) can be assumed to have regulatory roles in this organism, because artificial antisense RNAs are functional in *H. pylori* [Croxen *et al.* 2007]. Altogether, $\sim$100 sRNAs are known in the primary model organism of sRNA research, *E. coli* [Waters & Storz 2009]. Corrected for its much smaller genome size ($\sim$33%), *H. pylori* expresses at least as many sRNAs, arguing for wide-spread riboregulation despite the lack of a conserved Hfq protein.

3.3 Genome-wide transcript analysis of *Xanthomonas campestris* pv. *vesicatoria* str. 85-10

One of the model systems to elucidate the molecular communication between plant pathogens and their hosts and to characterize bacterial virulence strategies is the Gram-negative γ-proteobacterium *Xanthomonas campestris* pv. *vesicatoria* str. 85-10 (*XCV*). Essential for pathogenicity of *XCV* is the type III secretion (T3S) system, encoded by the hrp (hypersensitive response and pathogenicity) gene cluster [Bonas *et al.* 1991]. The main function of the T3S is the trans-location of effector proteins into the plant cell where they manipulate host cellular processes to the benefit of the pathogen, e. g. by suppression of basal plant defense responses [Thieme *et al.* 2005, White *et al.* 2009, Szczesny *et al.* 2010, Büttner & Bonas 2010]. Expression of the hrp gene cluster, type III effector and putative virulence genes is mainly regulated by HrpG and HrpX proteins. HrpG is activated by a so far unknown plant signal and induces the expression of HrpX. The latter protein binds to a conserved motif (plant-inducible promoter; PIP box) in the promoters of target genes [Koebnik *et al.* 2006]. Genome analysis predicted 4,726 protein-coding genes [Thieme *et al.* 2005]. However the overall gene structure and especially the ncRNA output of this pathogen are poorly understood.

We used the dRNA-seq approach (see Section 3.2) to identify transcription start sites and sRNAs in *XCV*. With our analyses we provide a first insight into the transcriptional landscape of *XCV* and the involvement of sRNAs in virulence. For details on the experiments performed by our collaborators and used methods I refer to our joint publications Findeiß *et al.* [2010] and Schmidtke *et al.* [in preparation].

3.3.1 Transcription start site annotation

Focusing on the expression of virulence and especially sRNA genes total RNA of the *XCV* wild type strain 85-10 and a mutant strain 85* (a point mutation in HrpG renders the protein constitutively active) expressing the Hrp-regulon were mixed. cDNAs were synthesized from untreated total bacterial RNA (TEX− library) and terminator exonuclease treated RNA (TEX+ library). The terminator exonuclease treatment specifically depletes transcripts with 5' monophosphate (5'P) ends which are characteristic for processed RNA molecules. The TEX+ library is, therefore, enriched with primary transcripts which have a 5' tri-phosphate end (5'PPP). Sequencing on a 454 platform resulted in 149,596 reads for the TEX+ library and 160,349 reads for the TEX− library. A total of 84% of the reads were mapped to the *XCV* genome. As previously described, *XCV* contains two identical copies of the 5S, 23S and 16S rRNA clusters, respectively, and 56 tRNA loci [Thieme *et al.* 2005]. 68% of the reads

of the TEX+ library and 63% of the TEX− library mapped to these genes. The remaining 40,385 and 49,845 cDNA reads were analyzed in more detail.

While cDNA reads of the untreated TEX− library cover entire genes, the read starts of the TEX+ library are shifted towards the 5' end of primary transcripts (Figure 3.9, e.g., *XCV0520*). As described in Section 3.2 in means of dRNA-seq a TSS is (conservatively) defined as a position at which the observed number of read starts in the TEX+ library significantly exceeds the expected number of read starts inferred from the TEX− library.

So far, most of the published TSS maps are derived by tedious manual inspection of the sequencing data [Albrecht *et al.* 2009, Jäger *et al.* 2009, Sharma *et al.* 2010]. This is a time consuming, biased and not fully reproducible procedure. Furthermore, for larger eukaryotic genomes manual TSS map creation is infeasible. In the following I present how the problem of automated TSS annotation based on dRNA-seq data can be addressed using the Skellam distribution model.

In different fields of application the Poisson distribution and furthermore the Skellam distribution has been successfully applied [Hwang *et al.* 2007, Karlis & Ntzoufras 2009]. Moreover, enrichment patterns in ChiP-Seq data have already been analyzed using these statistical models [Kim *et al.* 2010].

General Skellam distribution model: Consider two discrete random variables X and Y and their difference $D = X - Y$. The resulting probability function of D is a discrete distribution defined for signed integers only. If X and Y are Poissonian distributed their difference D follows the Skellam distribution [Skellam 1946] which is defined by the probability mass function:

$$f(d, \lambda_1, \lambda_2) = e^{-(\lambda_1+\lambda_2)} \left(\frac{\lambda_1}{\lambda_2}\right)^{\frac{d}{2}} J_{|d|}(2\sqrt{\lambda_1\lambda_2}); \tag{3.1}$$

for all $d \in \mathbb{Z}$ and average arrival rates $\lambda_1, \lambda_2 > 0$. $J_{|d|}$ denotes the modified Bessel function of the first kind and order $|d|$. The expected value of the Skellam distribution is given by $E(D) = \lambda_1 - \lambda_2$ while the variance is defined as $Var(D) = \lambda_1 + \lambda_2$, see Figure 3.8. As the Skellam distribution is a discrete probability function its probability mass function is of course normalized:

$$\sum_{d=-\infty}^{\infty} f(d, \lambda_1, \lambda_2) = 1 \tag{3.2}$$

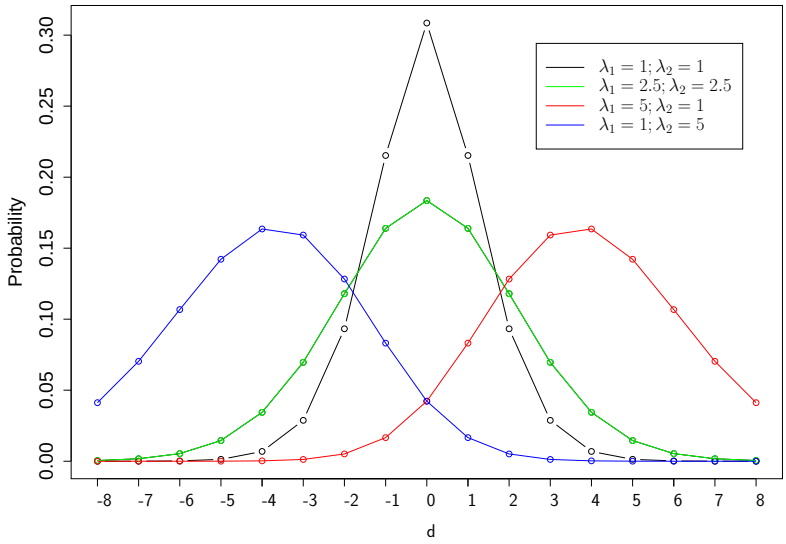


Figure 3.8. Examples of skellam probability mass functions. The x-axis shows difference d and the y-axis indicates the corresponding probability. The expected value of the skellam distribution is shifted in dependency on the difference of λ_1 and λ_2. Note that the Skellam distribution is a discrete probability distribution and is defined for signed integers only (connecting lines are drawn for visualization only).

The practical problem is that the expression (measured by the number of dRNA-seq reads) is not uniform across the genome but depends on the regional expression level of individual genes or operons (Figure 3.9). Depending on the sequencing depth and local expression levels an observation of a TSS is more or less likely.

The discrete Poisson distribution measures the significance of a number of events that occur within a fixed interval and a known average rate λ. For TSS annotation the fixed interval is a genomic region r of specified length and the observed event is the number of read starts within this region. Thus λ_{reg} can be estimated as the average rate of read starts s_r within r.

$$\lambda_{reg} = \frac{s_r}{length(r)} \tag{3.3}$$

This rate represents only the local expression of r. In order to normalize this rate with respect to the overall expression across the genome λ_{reg} is corrected by the average read start rate:

$$\lambda_{ave} = \frac{1}{n} \sum_{i=1}^{n} \lambda_i \tag{3.4}$$

where n is the number of possible regions of length r along the genome. Using equation (3.3) and (3.4) a normalized read start rate is define as $\lambda = \frac{\lambda_{reg}}{\lambda_{ave}}$. With this normalized read start rate the cumulative Poisson distribution function is defined as:

$$F(k, \lambda) = e^{-\lambda} \sum_{i=0}^{k} \frac{\lambda^i}{i!} \tag{3.5}$$

This function describes the probability that less than or equal to k read starts are observed at a genomic position. From that the significance of observed read starts at genomic positions can be estimated for both TEX+ and TEX− library.

The next challenge is to determine which of the genomic positions, supported by a number of read starts in the TEX+ library, are statistically significant TSS and which are explainable by the transcriptional background. The transcriptional background is given by the TEX− library.

For each genomic position the difference $D = P - M$ can be calculated. P and M are the number of locally observed read starts in the TEX+ and the TEX− library, respectively. As described above, the difference of two Poissonian distributed variables follows the Skellam distribution. Using equation 3.1 the cumulative distribution function of D is defined as:

$$F(D, \lambda_p, \lambda_m) = \sum_{d=-\infty}^{D} e^{-(\lambda_p+\lambda_m)} \left(\frac{\lambda_p}{\lambda_m}\right)^{\frac{d}{2}} J_{|d|}(2\sqrt{\lambda_p \lambda_m});\ d \in \mathbb{Z} \tag{3.6}$$

where λ_p and λ_m are the normalized read start rates of TEX+ and the TEX− library, respectively. Furthermore, $1 - F(D, \lambda_p, \lambda_m)$ represents the probability that a difference of more than D read starts is observed given the normalized read start rates λ_p and λ_m. Using the derived statistical model a probability value (p-value) can be assigned to any observed read start difference along the genome.

A sliding window approach can be used to measure the significance of one genomic position t times. In other words a sliding window of size x is shifted by an offset of y nucleotides along the genome and each position is scored $t = \frac{x}{y}$ times. Multiple testing of each position is corrected by:

$$p = \sqrt[t]{\prod_{i=1}^{t} p_i} \tag{3.7}$$

The sliding window approach stabilizes the determined p-values and ensures that the significance of a TSS is not directly dependent on the chosen window size.

For the analyzed *XCV* data set a TSS is defined as a genomic region of five nucleotides where at least three sequencing reads start. Furthermore, a gradient like behavior of the read endings is enforced so that not all reads stop at the same position. Using a window size of 500 nt, an offset of 50 nt all TSS with a p-value equal or less than 0.05 are reported.

In total, 1,372 chromosomal TSSs and 49 TSSs on the large plasmid pXCV183 of *XCV* were identified. The data confirm TSSs determined previously for selected pathogenicity genes, e.g., hrcU and hrpB1 [Fenselau & Bonas 1995, Koebnik *et al.* 2006]. TSSs were classified into four categories including i) primary TSSs that are located up to 300 bp 5' of an annotated translation start, ii) internal TSSs within an annotated CDS, iii) antisense TSSs that map to the opposite strand of CDSs ± 100 bp and iv) orphan TSSs that do not belong to the first three categories. The majority of the TSSs are primary TSSs (831) and probably correspond to the 5' end of mRNAs. As illustrated in Figure 3.9 B, TSSs can belong to more than one category, e.g., the primary TSS of XCV0523 is also antisense to XCV0522. Interestingly, 10% (86/831) of all primary TSSs are also classified as internal. Thus, some neighboring CDSs previously assumed to be co-transcribed as part of a polycistronic mRNA can also be transcribed from alternative promoters. As illustrated for XCV0522 (Figure 3.9A), we identified 71 TSSs which are located within the first 50 bp of annotated CDSs suggesting that previously annotated translation starts have to be revisited. Furthermore, 345 TSSs

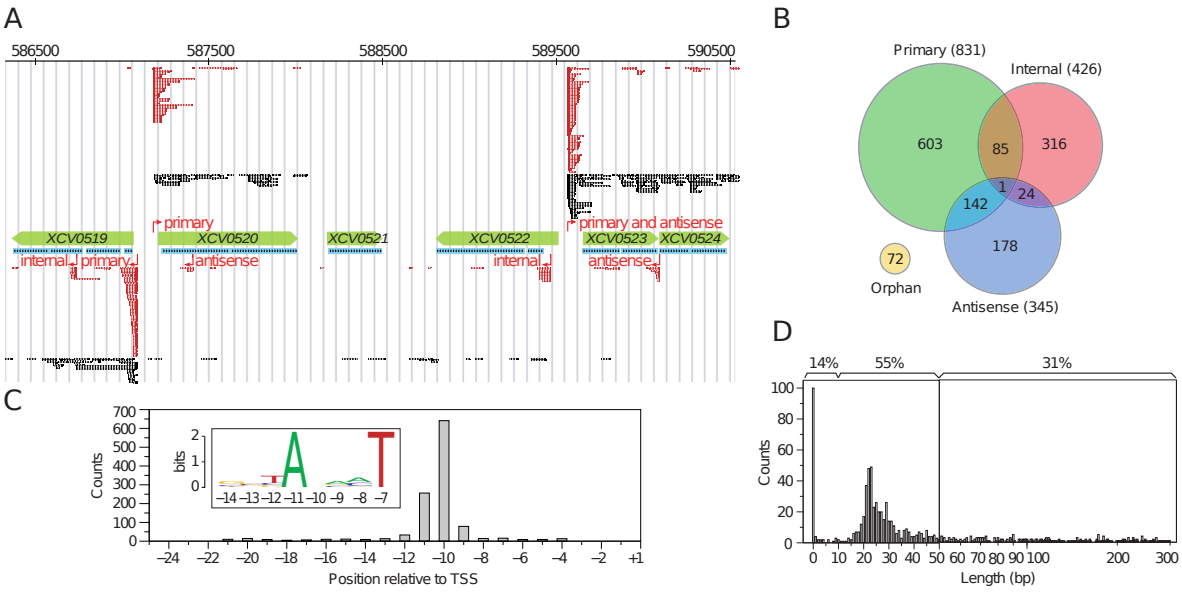


Figure 3.9. Identification of TSSs, promoter elements and analysis of 5' UTRs. A) Distribution of dRNA-seq reads at a chromosomal locus of *XCV*. Annotated CDSs and RNAcode high-scoring segments are highlighted in green and blue, respectively. Sequencing reads of TEX− (black) and TEX+ (red) are shown on top for the forward strand and below for the reverse strand. TSSs and corresponding classes are indicated in red. B) Venn diagram illustrating TSS classes. TSSs found upstream (max. 300 bp) of coding sequences are classified as primary. Internal TSSs are found within and antisense TSSs on the opposite strand of genes (± 100 bp). Orphan TSS do not belong to one of the other classes. C) TSS mapping identified a T/A-rich promoter element for 1,205 of 1421 TSSs. The histogram depicts the position of the conserved sequence pattern relative to the annotated TSSs at position +1. D) 5' UTR length distribution. The x-axis is split into linear (0-50) and logarithmic (51-300) scales. The top of the histogram gives the percentage of leaderless ($\leq$ 10 bp), short ($\leq$ 50 bp) and longer UTRs (between 50 and 300 bp).

are located antisense to annotated genes. Interestingly, 41% of these TSSs are also classified as primary TSSs, including 16 TSSs that correspond to overlapping mRNAs in an antisense orientation. 49 antisense TSSs are positioned in the 3' region (± 100 bp) of annotated sense genes.

3.3.2 Promoter and 5' UTR analysis

Promoter regions (50 nt upstream of all annotated TSS) and 5' UTRs (sequence between primary TSS and the corresponding CDS), were scanned with `MEME` for regulatory motifs. We found only a weakly conserved T/A-rich motif in the -10 region, see Figure 3.9. Surprisingly, other conserved promoter elements and a Shine-Dalgarno motif are missing. The later observation is consistent with a recent study by Nakagawa *et al.* [2010]. They analyzed the evolutionary process of translation initiation in prokaryotes and found that a SD-initiated translation in xanthomonads is rather unlikely.

Further analysis of 5' UTRs revealed an unexpected size diversity (Figure 3.9D). In *XCV*, the majority of 5' UTRs appears to be smaller than 50 bp which is characteristic for bacteria [Sorek & Cossart 2010]. Surprisingly, 14% of the mRNAs (118 of 831) are leaderless, their 5' UTR consisting less than 10 bp. Transcription of 82% of the leaderless mRNAs starts with AUG which was shown to be essential for stable ribosome binding to leaderless transcripts in *E. coli* [Brock *et al.* 2008]. This is in good agreement with the missing SD sequence and further supports the assumption of alternative mechanisms of ribosome guidance.

3.3.3 *XCV* encoded sRNAs

The *XCV* genome was scanned for known RNA regulators using the Rfam Database and the provided Perl script `rfam_scan.pl`. Seven riboswitches (FMN, SAH, Glycine, SAM, Cobalamin, TPP, yybP-ykoY) and ubiquitous RNAs (e.g. RNaseP, SRP, tmRNA, 6S-RNA) were identified. Based on the dRNA-seq data, the majority of these RNAs were strongly expressed. Opposite to the 6S RNA gene we found a highly expressed short region presumably corresponding to pRNA transcripts [Wassarman & Saecker 2006].

For experimental validation of sRNA candidates, our collaborators (Ulla Bonas's lab, Halle) performed Northern blot analysis. Using *hrpG* and *hrpX* (deletion) mutant strains the potential co-regulation of sRNAs with the T3S system was evaluated. Northern hybridizations confirmed 24 sRNAs, eight of which correspond to antisense RNAs, termed asX1-7 and PtaRNA1. The remaining 16 sRNAs mapped to intergenic regions and were termed sX1-15 and 6S. Intriguingly, three sRNAs (sX15, asX6, asX7) are encoded on the large plasmid, two

of which (asX7 and sX15) are in antisense orientation to each other. Interestingly, expression of eight sRNAs was affected by the key regulators of hrp gene expression, HrpG and HrpX, suggesting a role of these sRNAs or their targets in the interaction of *XCV* with the plant. HrpX-dependent induction of sRNA expression was observed for asX4, sX5, sX8 and sX12, whereas sX11 appeared to be HrpG/HrpX-dependently repressed.

In general, the dRNA-seq data and Northern blots suggest that *XCV* sRNAs do not accumulate as primary species but undergo growth-phase dependent processing. However, in most cases the apparent sizes of full-length and processed sRNAs in Northern blots were in agreement with the dRNA-seq data. In addition to full-length and processing products, Northern blots detected unexpectedly long signals, up to 900 nt, for the antisense RNAs asX1, asX2, asX3, asX6 and asX7. These signals may be caused by alternative termination of transcription. The sequencing data also suggest that sX7, sX13 and sX14 represent processing products of longer transcripts since reads mapping to these loci are predominantly found in the TEX− library, and no TSS was identified.

Phylogenetic distribution of sRNAs

While sX3 and asX5 are unique for *XCV*, homology searches revealed that 10 sRNA genes are exclusively found in sequenced *Xanthomonas* species that encode a hrp-T3S system. Four of the latter sRNAs and asX5 were co-regulated with the T3S system. Two intergenic sRNAs, sX1 and sX10, are highly similar in sequence and structure. Three additional homologous genes are present in the *XCV* genome, expressed and might therefore be considered as an sRNA family. As three to six copies of members of this gene family are found in other *Xanthomonas* species, we propose a functional redundancy of the respective sRNAs. A rather erratic phylogenetic distribution was observed for sX8 and PtaRNA1 since homologs are found in β- and γ-proteobacteria. Interestingly, this holds true also for the genes adjacent to both RNAs which suggests a common evolutionary origin of this region.

Plasmid transferred antisense RNA (PtaRNA1)

XCV encodes a constitutively expressed small RNA, which we designate PtaRNA1, "Plasmid transferred antisense RNA". The superposition of the individual reads revealed a small RNA encoded adjacent to the *trbL* (XCV2162) gene. Expression analysis revealed a constitutive expression with respect to the tested growth phases. Interestingly, two bands which indicate procession of the full length PtaRNA1 are detected in the exponential but not in the stationary growth.

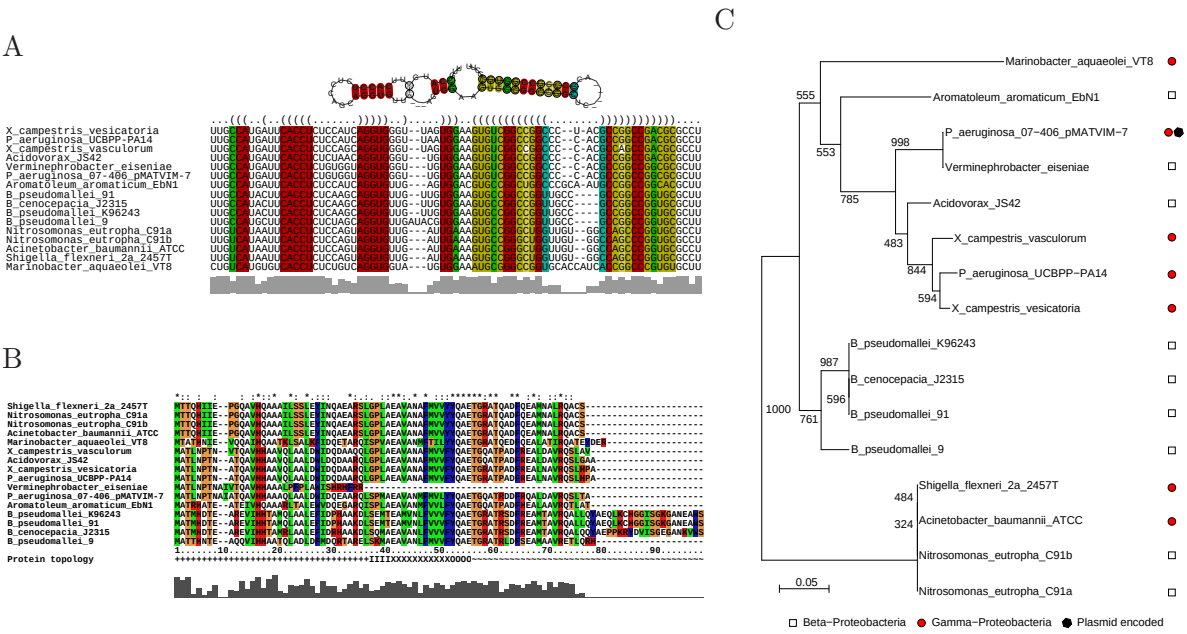


Figure 3.10. A) Consensus secondary structure model of PtaRNA1 based on the depicted seed alignment. The structure is highly stable (MFE -37.06 *kcal/mol*) and supported by various compensatory mutations within the stem on the right-hand side. B) Amino acid alignment of XCV2162 homologs. The alignment shows various totally (indicated by '∗') and by substitutions (indicated by ':' and '.') supported and therefore conserved columns. The protein topology of a trans-membrane domain, predicted by `MEMSAT3` [Nugent & Jones 2009] is indicated as well. '+' marks inside loop, '∼' outside loop, 'O' outside helix cap, 'X' central trans-membrane helix segment and 'I' inside cap. The truncated *Verminephrobacter* sequence was not used for the calculation of the conservation track. C) Phylogenetic tree based on PtaRNA1 alignment (similar for XCV2162 alignment, data not shown). Class of the "host" species is shown by the symbols on the right hand side. Numbers indicate bootstrap values of the inner nodes.

PtaRNA1 shows an erratic phylogenetic distribution with occurrences on chromosomes in a few individual strains distributed across both β- and γ-proteobacteria, (Figure 3.10). Conspicuously, *ptaRNA1* was not found in other closely related genomes, e.g. other strains of *Burkholderia*, *Pseudomonas*, or *Xanthomonas*. The phylogeny of the PtaRNA1 sequences is not congruent with the phylogeny of their "host" species (Figure 3.10). This distinguishes PtaRNA1 from most other sRNAs. Moreover, a homologous gene located on plasmid pMATVIM-7 of *P. aeruginosa* is found.

PtaRNA1, therefore, exhibits all hallmarks of a novel RNA antitoxin that proliferates by frequent horizontal gene transfer. Although distinct toxin-antitoxin systems have been found in widely separated bacterial groups (e.g. *hok/sok* in *E. coli* and *txpA/ratA* in *B. subtilis* [Silvaggi *et al.* 2005]), each of the known examples exhibits a very narrow phylogenetic distribution. All *ptaRNA1* homologs are located antisense to a putative toxin, which in turn is never encountered without the small RNA, see Figure 3.10.

Possible targets of sX13

Expression of sX13 as investigated by means of cDNA reads is similar to that of housekeeping RNAs (e.g. tmRNA). Further analysis of the HrpG-dependently repressed sX13 RNA revealed a highly conserved and thermodynamically stable (MFE -56.09 $kcal/mol$; z-score -6.52) secondary structure, see Figure 3.11. sX13 exposes three highly conserved C-rich loops, that are likely to bind possible targets. The *Staphylococus* encoded RNAIII as well as *E. coli's* OxyS are known examples that form multiple contacts with their mRNA targets. Multiple interaction sites increase specificity of both binding partners and might have stabilizing effects for the sRNA-mRNA complex. Therefore we suggest a zipper-like interaction of sX13 with its mRNA targets, see Figure 3.11.

Target prediction approaches, which are applicable on a genome-wide scale (e.g. `RNAup` [Mückstein *et al.* 2006], `targetRNA` [Tjaden *et al.* 2006]) focus on the most stable interaction between ncRNA and the mRNA target. We, therefore implemented an iterative approach based on the thermodynamics of RNA-RNA interaction calculated by `RNAup`.

Among the high scoring (by means of binding energy) targets with three interaction sites we found: putative membrane binding proteins, transcriptional regulators (e.g. FurR), NAD/FAD binding protein and the algR mRNA, part of a two component system. Interestingly, the transcriptional regulator AlgR controls a variety of processes, including hydrogen cyanide production [Carterson *et al.* 2004], twitching motility [Whitchurch *et al.* 1996; 2002], biofilm formation and quorum sensing [Morici *et al.* 2007] in *P. aeruginosa*. The multiple

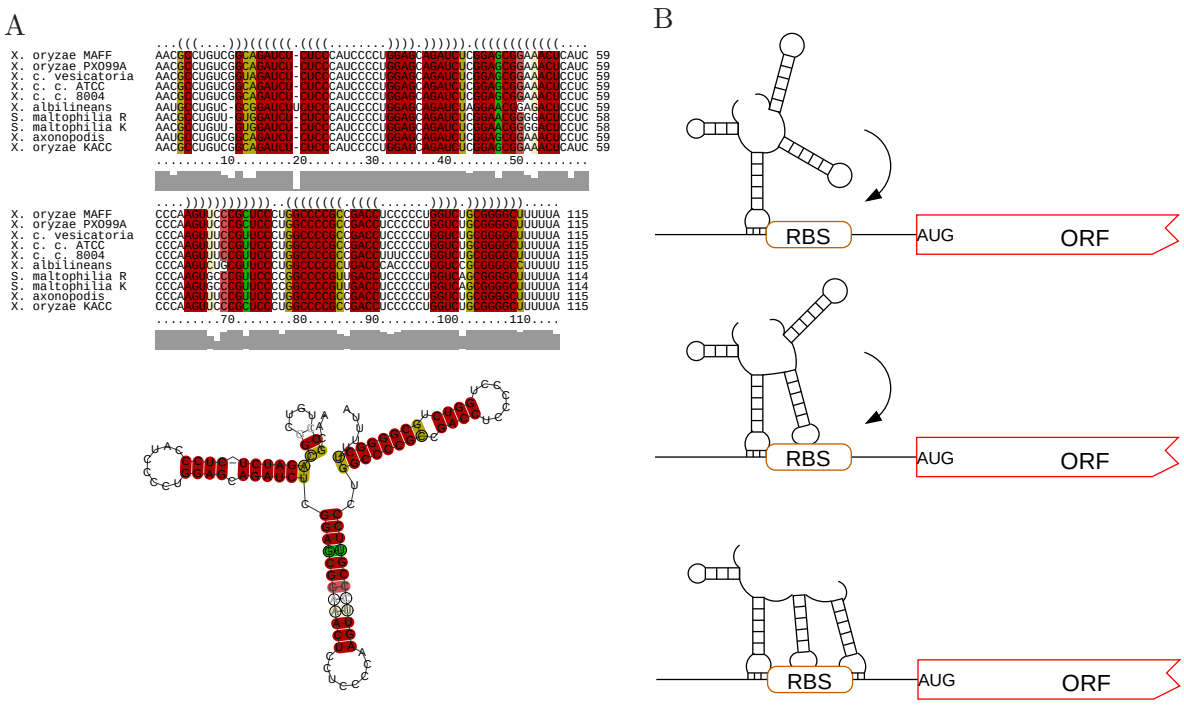


Figure 3.11. Structural conservation and a zipper-like interaction model of sRNA sX13. A) Multiple sequence alignment (top) and the resulting consensus secondary structure (bottom) of sX13. sX13 exposes three highly conserved C-rich loops. B) Proposed zipper-like interaction model of sX13 and its mRNA targets. Upon the first contact between sRNA and mRNA additional interactions stabilize the complex.

interaction model of sX13 and the predicted mRNA-targets need, of course, experimental validation. Furthermore it cannot be excluded that sX13 interacts with a protein instead of mRNAs.

3.3.4 Short peptides

To identify conserved short protein-coding genes in *XCV*, a multiple sequence alignment of closely related species was calculated with the `multiz` package. The alignments were analyzed for potential coding segments using `RNAcode` (see Section 4.2 for details). In total 24 potential short ORFs were predicted, twelve of which are further supported by dRNA-seq reads. One example is sX6 which has a predicted coding capacity of 80 amino acids. A protein of the predicted molecular mass (∼12 kDa) was detected in protein extracts of *XCV*. Besides sX6, TSSs for two of the predicted ORFs with a coding capacity of 36 and 67 amino acids were found. Interestingly, homologs of genes for the three small proteins are exclusively found in xanthomonads encoding a hrp-T3S system.

The dRNA-seq based analysis of *XCV* summarized here led to remarkable insights into the transcriptional landscape of this important model plant pathogen. In contrast to earlier dRNA-seq approaches, which were mostly based on laborious manual inspection of sequencing data [Jäger *et al.* 2009, Albrecht *et al.* 2009, Sharma *et al.* 2010], we established a fully-automated computational approach for the identification of TSSs that is based on the statistical differences of read starts in both sequenced cDNA libraries. Besides facilitating the analysis of TSSs, this approach provides a meaningful measure of statistical confidence and ensures that predictions are comparable between different studies. Globally fixing certain parameters, such as window size, minimum coverage, and a p-value, may lead to differences between the computational annotation and manual inspection of individual TSSs. Furthermore, TSSs that are inactive under the conditions interrogated will of course remain invisible. However, the analysis of the TSS map and the result of all subsequent analysis (e.g. massive antisense transcription, 24 validated sRNAs, detection of previously missed short peptides) revealed an unexpected complexity of *XCV* transcript organization.

3.4 Summary

All three studies recovered highly expressed housekeeping RNAs (e.g. tmRNA, RNase P RNA). Additional sRNA candidates were identified by the applied sequencing approach and validated by independent experiments. Most of the detected sRNAs show a rather narrow phylogenetic distribution and are group or species specific. Albeit *H. pylori* lacks a Hfq homolog we found ∼60 novel sRNAs. Inspection of sRNA candidates revealed the existence of short protein-coding genes in all three studies. These loci are likely to be small toxic proteins or could represent novel dual-functional RNAs.

Especially the two presented dRNA-seq analysis revealed an unexpected complexity of transcript organization in *XCV* and *H. pylori*. We found massive antisense transcription in both species. Northern blot analysis detected extremely long signals (up to 900 nt) of antisense transcripts in *XCV*. However, the regulatory function of these (long) antisense transcripts remains to be analyzed. Bacterial promoter elements usually contain specific sequences for binding of RNAP. Thus, we expected to detect similar sequence motifs (-35 TTGACA and -10 TATAAT boxes) as those known from *E. coli*. Intriguingly, neither in *H. pylori* nor *XCV* the expected promoter elements were detected. Furthermore, leaderless mRNAs are considered rare we found dozens of these transcripts in both studies.

Taken together, the results presented clearly indicate the necessity to revise our current understanding of prokaryotic transcriptional and translational processes.

4

How to assess (non-)coding potential

As shown in Chapter 3, experimental approaches are power full tools to detect and verify new RNA transcripts. Typical experimental settings are, however, not able to cover the complete transcriptome of the species under investigation. This is mainly caused by the expenditure of time and costs of experiments. Furthermore, RNAs which are only expressed in specific developmental stages or under certain stress conditions, are likely to be missed. Computational approaches represent an alternative. Once a method is developed and implemented it is (in principle) applicable to any organism. Our recent update of `RNAz` and the implementation of `RNAcode` represent good examples and their underlying methods are described in the following.

4.1 RNAz 2.0: improved non-coding RNA detection

RNAz is a widely used software package for *de novo* detection of structured RNA regulators in comparative genomic data. RNAz 1.0 has been used successfully to map structural ncRNAs in a wide variety of genomes [Washietl *et al.* 2005, Missal *et al.* 2005; 2006, Rose *et al.* 2007; 2008b, McGuire & Galagan 2008]. A large number of these predictions have also been verified experimentally [Weile *et al.* 2007, del Val *et al.* 2007, Sonnleitner *et al.* 2008, Mourier *et al.* 2008, Schilling *et al.* 2010]. Moreover, the generic approach and many algorithmic details developed for RNAz 1.0 have been re-used, extended, and adapted to other problems in the field of RNA gene-finding [Gardner *et al.* 2005, Hertel & Stadler 2006, Uzilov *et al.* 2006, Reiche & Stadler 2007, Sandmann & Cohen 2007, Hertel *et al.* 2008, Xu *et al.* 2009].

Four years of experience have not only demonstrated the applicability of the approach, but also helped us to identify limitations of the current implementation. In our contribution [Gruber *et al.* 2010], we described a major update of the RNAz program. It is based on the results of two follow-up studies [Gruber *et al.* 2008, Gesell & Washietl 2008], on our experiences gained during many real-life applications, and last but not least, on the received user feedback.

4.1.1 Methods

Overview of the RNAz algorithm

RNAz predicts functional RNA structures on two independent criteria: (i) thermodynamic stability and (ii) structural conservation.

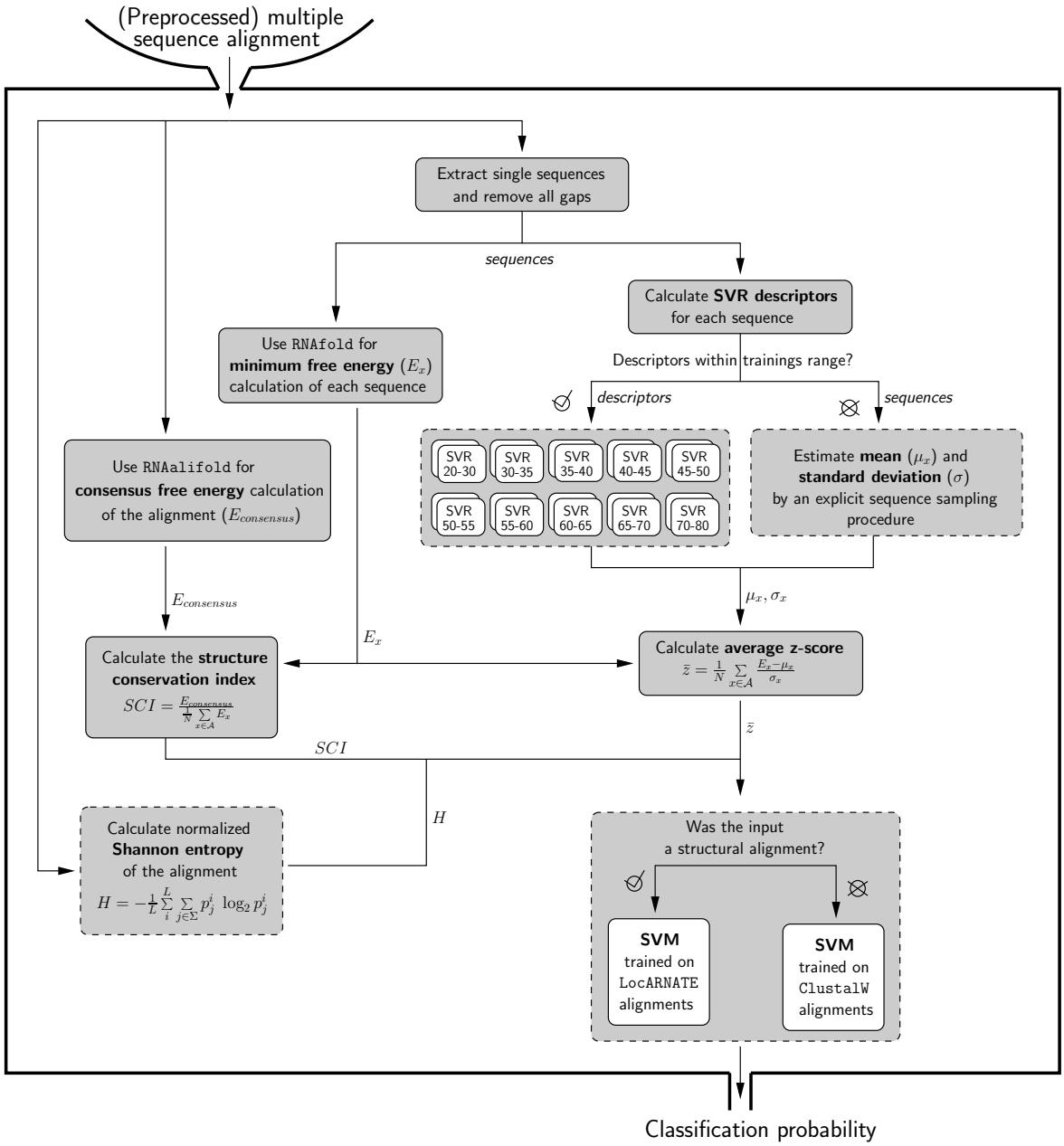
A common way to express thermodynamic stability is in terms of a z-score. This is simply the number of standard deviations by which the minimum free energy (MFE) deviates from the mean MFE of a set of randomized sequences with the same length and base composition. A negative z-score thus indicates that a sequence is more stable than expected by chance. As this procedure involves energy evaluation of a large set of random sequences it is not applicable for large-scale genomic screens. RNAz instead uses support vector regression (SVR) to estimate the mean and the standard deviation based on the nucleotide composition of a sequence.

RNAz evaluates evolutionary conservation of RNA structures in terms of the structure conservation index (SCI). RNAz measures structural conservation by calculating the ratio of the consensus folding energy to the unconstrained folding energies of the single sequences. Since the postulation of the SCI, it has been a major point of criticism that this measurement evaluates structural conservation on the energy level rather than on the RNA structures themselves.

However, it has been shown that the SCI is on average the most powerful method and that it is only outperformed by other approaches in the high sequence identity range [Gruber *et al.* 2008].

Both criteria are combined by a support vector machine model that classifies the input alignment as structural RNA or other. A graphical overview of the `RNAz` algorithm is depicted in Figure 4.1. In the following, independent refinements that significantly improved the overall prediction accuracy of the `RNAz` approach are described.

Figure 4.1 (following page). Outline of the `RNAz 2.0` work-flow and algorithm. Large genomic multiple alignments are processed into smaller alignments. This filtering procedure involves several steps: (i) overlapping windows given a fixed window and step size are created, (ii) sequences that contain too many gaps are removed and (iii) from the remaining sequences only those are kept that meet a predefined average pairwise identity threshold. The resulting alignments are then separately processed by `RNAz`. First, structure and energy predictions are performed for both the single sequences and the alignment. These results can be immediately combined to calculate the SCI as the measure of the evolutionary conservation of the RNA sequences in the alignment. In a second step, the mean free energy and the standard deviation used for the calculation of the z-score are estimated. For this purpose descriptors based on the nucleotide composition (G+C content, A/U ratio, C/G ratio, all 16 di-nucleotide frequencies and the length of the sequence) are calculated for each sequence. If descriptors are within the training boundaries they are passed to the corresponding support vector regression (SVR). Numbers in the SVR boxes indicate the G+C content the particular SVR is trained on. Otherwise, the mean and the standard deviation are evaluated explicitly by folding of 1,000 randomized sequences with the same di-nucleotide composition. In a final step the average z-score of the sequences, the SCI and the normalized Shannon entropy of the alignment are passed to the classification SVM, which returns a probability estimate that the given alignment harbors thermodynamically stable and/or evolutionary conserved RNA secondary structures. Parts that are highlighted in dashed boxes are new or modified components of `RNAz 2.0` implementation. `RNAfold` and `RNAalifold` are part of the Vienna RNA Package [Hofacker 2009]. See text for a detailed explanation of the formulas.


(Preprocessed) multiple sequence alignment
Extract single sequences and remove all gaps
sequences
Calculate SVR descriptors for each sequence
Use RNAfold for minimum free energy (E_x) calculation of each sequence
Descriptors within trainings range?
descriptors
sequences
Use RNAalifold for consensus free energy calculation of the alignment ($E_{consensus}$)
SVR 20-30
SVR 30-35
SVR 35-40
SVR 40-45
SVR 45-50
SVR 50-55
SVR 55-60
SVR 60-65
SVR 65-70
SVR 70-80
Estimate mean (μ_x) and standard deviation (σ) by an explicit sequence sampling procedure
$E_{consensus}$
μ_x, σ_x
E_x
Calculate the structure conservation index
$SCI = \frac{E_{consensus}}{\frac{1}{N}\sum_{x\in\mathcal{A}} E_x}$
Calculate average z-score
$\bar{z} = \frac{1}{N}\sum_{x\in\mathcal{A}} \frac{E_x-\mu_x}{\sigma_x}$
$\bar{z}$
SCI
H
Calculate normalized Shannon entropy of the alignment
$H = -\frac{1}{L}\sum_{i}^{L}\sum_{j\in\Sigma} p_j^i \log_2 p_j^i$
Was the input a structural alignment?
SVM trained on LocARNATE alignments
SVM trained on ClustalW alignments
Classification probability

z-score regression for di-nucleotide shuffled sequences

As in `RNAz 1.0`, we use support vector regression (SVR) to compute z-scores for folding energies because the direct approach via repeated shuffling and folding is too costly for genome-wide applications. The critical step is, of course, the adequate training of the SVR. Hence, the complete sequence range in means of nucleotide composition and lengths should be covered by the training data.

To estimate the mean and standard deviation of folding energies for mono-nucleotide shuffled sequences it is feasible to sample uniformly the four mono-nucleotide frequencies and to vary the length of the sequences. This method cannot be extended that easily for di-nucleotide shuffled sequences. One has to consider the much larger space of di-nucleotide compositions that is occupied by sequences of practical interest. In order to efficiently train the regression engine of `RNAz 2.0` we first apply uniform sampling to cover the mono-nucleotide space. Synthetic sequences of length 50 with G+C content, A/U ratio, and C/G ratio ranging from 0.20 to 0.80 in steps of 0.05 were generated. For each of these start sequences we then generated 500,000 mono-nucleotide shuffled sequence. This initial set can be generated very quickly and served as the basis for the selection of a much smaller, approximately evenly spaced, training set with representative di-nucleotide frequencies. Sequences of different lengths where then generated by concatenating the initial set. Additional filtering and the generation of representative sets resulted in a total of 1,155,737 training instances. For each of these instances, we generated 1,000 randomized sequences by the Altschul-Erikson algorithm [Altschul & Erickson 1985] with the same di-nucleotide composition and used `RNAfold` [Hofacker *et al.* 1994] to evaluate their folding free energy.

More than one million training instances are by far too many to be used in SVM training procedures in reasonable time. For this reason the training data was split into different ranges of the G+C content (see Figure 4.1) to guarantee efficient training and fast prediction. This comes at the price of increased memory consumption but keeps the number of support vectors comparable to the approach used in `RNAz 1.0`. The SVM library `LIBSVM` was used to train regression models for the mean and the standard deviation for each subsets. As input features we used the G+C content, the A/U ratio, the C/G ratio, all 16 di-nucleotide frequencies and the length of the sequence scaled to the interval [0,1]. The regression for estimating the mean free energy was trained to learn energy per nucleotide, while the standard deviation was not scaled. The standard grid search approach was used to find optimal combinations for SVM parameters.

Training data generation and training of the SVM classifier

At the time RNAz 1.0 was implemented only a few RNA families were known and well suited for the SVM training. This, of course, limits the predictive power of the approach. For the training and test sets of the new RNAz version 93 RNA families, available in the Rfam 9.1 database [Gardner *et al.* 2009], were selected based on their signals for thermodynamic stability and structural conservation. The RNAz 2.0 training set covers a broad range of different RNA families including major classes such as tRNAs, snoRNAs, microRNAs, riboswitches, and bacterial regulatory RNAs.

For each RNA family, a set of alignments with varying numbers of sequences and average pairwise identities was generated. Rfam full alignments were used if they contained less than 300 sequences, otherwise we used the seed alignments. For our purpose the use of at most 300 sequences proofed well to generate a set of alignments over the desired range of average pairwise identities. Rfam alignments were utilized as a source to retrieve family members of a particular ncRNA class and only extracted, ungapped RNA sequences were used for subsequent analysis.

First, Rfam alignments were filtered to remove nearly identical sequences, so that the training alignments contained sequences with at most 98% identity. The sequences were then re-aligned using ClustalW. For each of these ncRNA family alignments we then proceeded as follows: for each number of sequences from 2 to 15 we generated at most 10 alignments with a randomly chosen average pairwise identity between 50 and 98% and with a maximum relative difference in sequence lengths of 65% utilizing rnazWindow.pl which is part of the RNAz analysis pipeline [Washietl 2007].

To ensure that this set of positive training examples contained only instances with good structural conservation signals we filtered alignments by using tree editing distances between the structures of the sequences in the alignment. Ordered, rooted trees can be deduced from the dot-bracket notation of RNA secondary structures. Tree editing defines a metric in the space of trees by a set of operations (deletions, insertion and relabeling of nodes) and hence can be used to calculate distances between RNA secondary structures [Gruber *et al.* 2008]. For each alignment we extracted sequences, removed gaps and calculated the averaged pairwise tree editing distance using RNAdistance. We repeated this for a set of 100 randomized alignments and calculated an empirical p-value as a measure of structural conservation. Alignments with a p-value higher than 0.05 were removed from the training set. Alignments retained after this filtering procedure were re-aligned with ClustalW for application to sequence-based alignments. It is a well known fact that sequence-based alignment

methods fail to give high quality alignments regarding RNA secondary structures in low average pairwise identity ranges [Uzilov *et al.* 2006]. We therefore re-aligned the training set with `LocARNATE` to generate an additional set of sequence/structure-based alignments.

Negative instances of the training set were generated by shuffling using `multiperm` [Anandam *et al.* 2009] v. 0.9.3 if the normalized Shannon entropy of the alignment [Gruber *et al.* 2008] was less than 0.50. Otherwise, alignments were simulated using `SISSIz` [Gesell & Washietl 2008] to ensure full randomization for the more diverse alignments where shuffling can become inefficient. The final training set was composed of 10,538 alignments for each the positive and the negative class.

The `RNAz 2.0` SVM classifier uses three features to detect structured non-coding RNAs: (i) the average minimum free energy z-score ($\bar{z}$) estimated from a di-nucleotide shuffled background, (ii) the structure conservation index (SCI) and (iii) the normalized Shannon entropy (H) of the alignment as a measure for the content of evolutionary information.

Consider an alignment $\mathcal{A}$ consisting of N sequences. Let E_x denote the minimum free energy of sequence x, and let μ_x and σ_x be the mean and standard deviation, respectively, of the folding energies of a large number of random sequences of the same length and same di-nucleotide composition as x. The averaged z-score of the alignment $\mathcal{A}$ is defined as

$$\bar{z} = \frac{1}{N} \sum_{x \in \mathcal{A}} \frac{E_x - \mu_x}{\sigma_x}$$

The SCI of alignment $\mathcal{A}$ is given as the fraction of the consensus folding free energy ($E_{consensus}$) to the average of the folding free energies of the single sequences:

$$SCI = \frac{E_{consensus}}{\frac{1}{N} \sum\limits_{x \in \mathcal{A}} E_x}$$

The normalized Shannon entropy H of an alignment $\mathcal{A}$ of RNA sequences over the alphabet $\Sigma = \{\mathsf{A}, \mathsf{C}, \mathsf{G}, \mathsf{U}, \text{-}\}$ is defined as the sum of the Shannon entropy of the individual columns divided by the length of the alignment denoted by L:

$$H = -\frac{1}{L} \sum_{i}^{L} \sum_{\alpha \in \Sigma} p_\alpha^i \ \log_2 \ p_\alpha^i$$

The probability p_α^i is approximated by the observed frequency of character α in alignment column i (normalized by the number N of sequences in the alignment). All features were scaled to a range of [-1,1]. Standard grid search combined with a 10-fold cross validation was applied to find optimized SVM parameters. Among the models with the best cross-validation accuracy (top 20) we chose the model that showed best performance on an independent test

set created the same way as the training set. The output of the final classification SVM is a probability estimate that the input alignment contains thermodynamically stable and/or structurally conserved RNA sequences.

4.1.2 Results

Di-nucleotide based z-scores

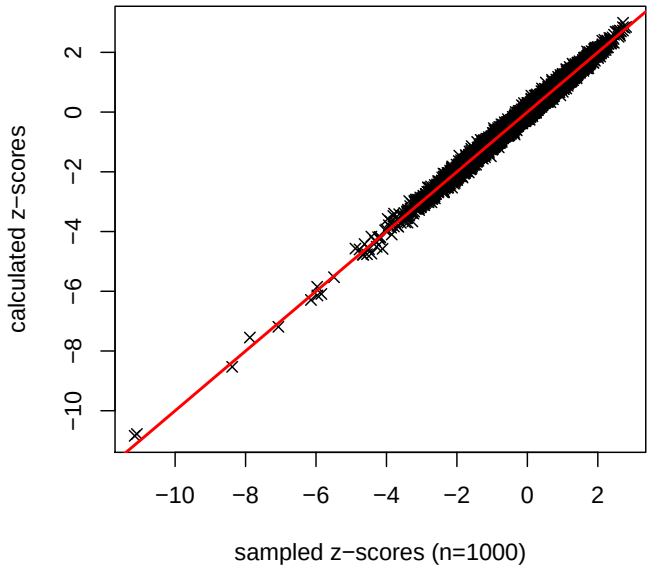


Figure 4.2. z-scores calculated by support vector regression in comparison with z-scores determined from 1,000 random samples preserving di-nucleotide frequencies for 10,000 sequences from the human ENCODE regions. Correlation of z-scores is 0.996 and the mean absolute error is 0.076.

Accuracy of the z-score regression for di-nucleotide shuffled sequences was evaluated on 10,000 randomly chosen sequences of variable length from 50 to 200 nt of the human ENCODE regions [ENCODE Consortium 2007] (Figure 4.2) and genomic sequences of *D. melanogaster* and *E. coli*. The mean absolute error (MAE) and the correlation (R) of z-scores calculated by SVR compared to z-scores determined from 1,000 random samples is 0.075 and 0.996, respectively. Comparisons of z-scores determined from 1,000 di-nucleotide shuffled sequences to 100 di-nucleotide shuffled sequences (MAE= 0.107, $R = 0.992$) and to 1,000 mono-nucleotide

shuffled samples (MAE= 0.420, $R = 0.916$) clearly demonstrate that our method is a suitable approach for fast and efficient estimation of di-nucleotide controlled z-scores. `RNAz 1.0` also showed restrictions on the base composition because of the limited training range of the SVR. This limitation is now overcome by explicit generation of shuffled sequences once the base composition of a sequence is out of the training range. Since boundaries have been chosen broadly (e.g. G+C content from 20 to 80%) this will only apply in a small minority of cases.

New training sets and improved classification model

The new `RNAz 2.0` algorithm now uses the average z-score of the sequences in the alignment based on a di-nucleotide background model, the SCI and the normalized Shannon entropy as features in the final classification model. Moreover, a much bigger training set covering a broader range of RNA families has been used for the revised version of `RNAz`.

To evaluate the predictive power of `RNAz 2.0` we chose a test set of 4,303 alignments of structural RNA families used in a previous study [Gesell & Washietl 2008]. This test set is especially well suited as it contains randomly chosen genomic regions from vertebrate alignments as negative controls. Although both versions perform well on this test set, `RNAz 2.0` clearly outperforms version `1.0` in the high specificity range (Figure 4.3). For example, at the generally used 0.01 false-positive cutoff, `RNAz 2.0` shows 0.899 sensitivity compared to 0.688 in the old version.

By using structural alignments one can expect an improvement in discrimination capability of the SCI for alignments with low sequence similarity [Uzilov *et al.* 2006]. We used a test set is composed of 2,455 alignments of various ncRNA families with an average pairwise identity between 30 and 70%, as well as a negative set consisting of 2,455 alignments derived by randomization of reference alignments with `multiperm` or `SISSIz`. As depicted in Figure 4.3 structural alignments improve the overall predictive power of `RNAz` especially in the high confidence interval.

Recent studies (e.g. Washietl *et al.* [2007]) have shown that `RNAz` suffers from a high false discovery rate (FDR). We therefore evaluated the performance of both versions on 193,634 alignments retrieved from the human ENCODE regions. A di-nucleotide background model was generated with `SISSIz` [Gesell & Washietl 2008] and all hits detected by `RNAz` on this data set were considered to be false positives. While `RNAz 1.0` shows a very high FDR of around 80%, the FDR of `RNAz 2.0` is much lower being around 54% for high confident hits, see Table 4.1.

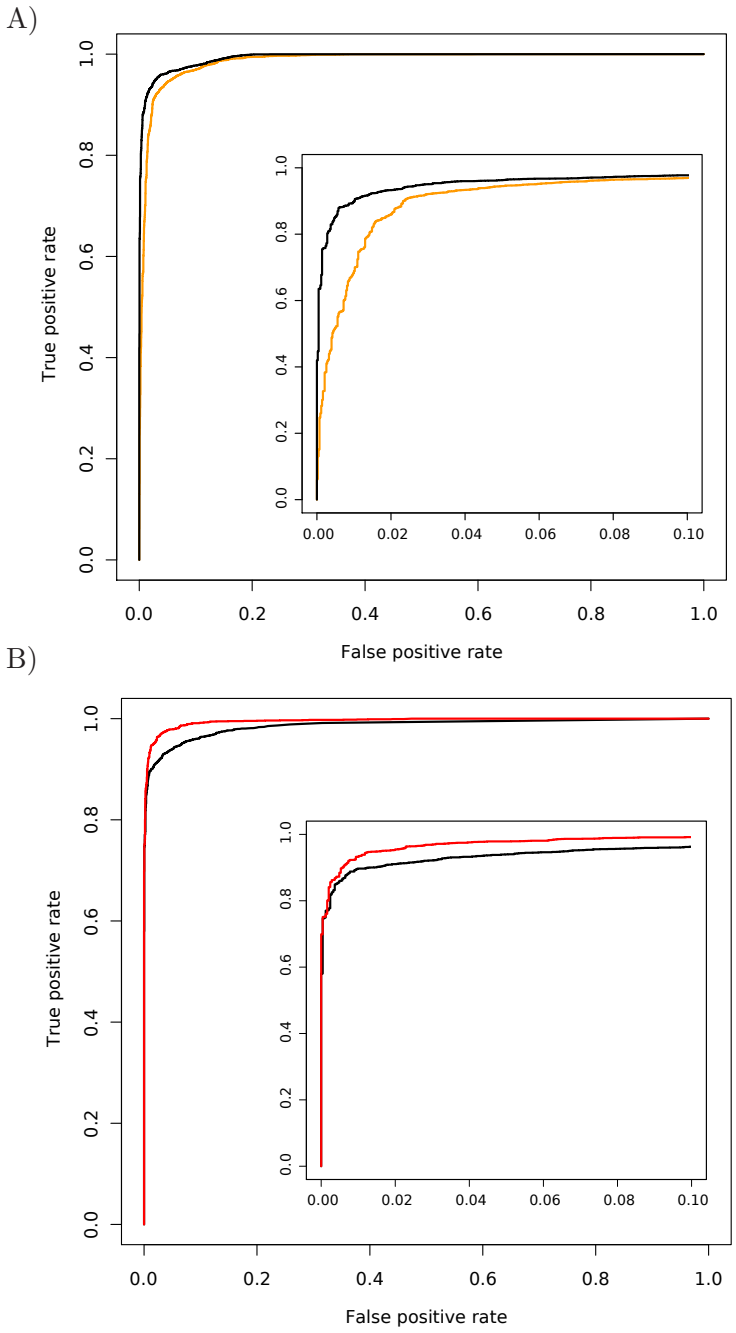


Figure 4.3. ROC curves for the RNAz prediction accuracy. A) Comparison of RNAz 2.0 classification (black) vs. RNAz 1.0 classification (orange) on a previously published data set for the evaluation of non-coding RNA gene finders [Gesell & Washietl 2008]. B) RNAz 2.0 prediction accuracy on sequence-based alignments (black) vs. structural alignments (red). A significant improve of the overall predictive power of RNAz 2.0 is achieved by use of structural alignments. Sequence-based alignments were generated with ClustalW, while structural alignments were generated with LocARNATE. Insets highlight the main differences within the high specificity regions.

Table 4.1. Comparison of the false discovery rate (FDR) based on ENCODE regions and a di-nucleotide background model for low (p-value > 0.5) and high (p-value > 0.9) confidence hits. A hit corresponds to a single alignment.

	`RNAz 1.0`		`RNAz 2.0`	
	# low conf.	# high conf.	# low conf.	# high conf.
ENCODE regions	17,814	6,854	6,880	2,259
background	14,489	5,596	4,090	1,219
estimated FDR	81%	82%	59%	54%

Computational speed

The performance of `RNAz 2.0` in comparison to `RNAz 1.0` was tested on 50,000 randomly chosen alignments from the ENCODE data set. Alignment length was 120 nucleotides and alignments contained at most six sequences. Experiments were conducted on an Intel Xeon 2.40GHz CPU. For each alignment both reading directions were examined, resulting in a total of 100,000 alignments that had to be scored. The execution time required by `RNAz 1.0` was 202 min, `RNAz 2.0` without explicit shuffling was 252 min and `RNAz 2.0` using explicit shuffling was 1,230 min. Although explicit shuffling had to be used for only 1% of the sequences (5,524/549,210), it comes with an tremendous overhead increasing the run time of `RNAz 2.0` almost 5-fold. We extracted those alignments where explicit shuffling was used and compared the classification probability to the one derived from calling `RNAz` without explicit shuffling. For 96% of the cases the change in classification probability was less than 1%. For this data set the maximal observed difference was 0.21. In general, we observed larger differences in the probability range from 0.2 to 0.8 than in the regions close to 0 or 1.

4.1.3 Conclusion

The presented major updates of the `RNAz` algorithm clearly improved the prediction accuracy of this widely used approach. Evaluation of thermodynamic stability has been changed from a mono- to a di-nucleotide background model. This directly translates into a significantly lower false discovery rate. In addition to the di-nucleotide z-score, the overall prediction accuracy is improved by a combination of the use of a new training set and the normalized Shannon entropy as a measure of sequence variation. Furthermore, the updated version is not any more restricted to limitations concerning the base composition or number of sequences in the input alignment.

The generation of structural alignments is computationally expensive but we showed that they can improve the RNAz classification power. This is true in particular for alignments of low average pairwise identity. Given that the overall computational complexity of LocARNATE is $O(n^4)$, the routine use of structural alignments on a genome-wide scale is still out of reach, at least when off-the shelf hardware is used. The re-scoring of positively classified hits of a sequence-based RNAz screen after re-aligning them with a structural aligner may help to increase the overall accuracy, in particular for relatively poorly conserved alignment slices. One could also use RNALfold [Hofacker *et al.* 2004] or the alignment version RNALalifold [Athanasius F Bompfünewerer Consortium *et al.* 2007] to preselect genomic loci that show signature of increased thermodynamic stability. This would significantly reduce the number of alignments to be screened by RNAz and therefore makes sequence/structure based classification possible.

4.2 RNAcode: robust discrimination of coding and non-coding RNAs

Distinguishing protein-coding from non-coding sequence is the first and most crucial step in genome annotation. While the coding regions are subsequently investigated for properties of their protein products, a completely different toolkit is applied to the nucleic acid sequences of the non-coding regions. The quality of the analysis of coding potential therefore also affects the annotation of putative ncRNA genes.

The detection of protein-coding genes in genomic DNA data is a well studied problem in computational biology [Burge & Karlin 1998]. Using machine learning techniques, sophisticated models of genes have been built that can be used to annotate whole genomes [Brent 2008] and that have been constantly improved over the years [Brent 2008, Flicek 2007]. However, the repeated detection of unannotated short peptides in our transcriptome studies (see Chapter 3) and the fact that classical gene finders suffer from the lack of training data of verified short peptides, pointed us to new challenges beyond classical gene finding. A reliable analysis of the coding potential of (expressed) genomic regions is an essential step preceding any downstream analysis.

In this section I introduce RNAcode, a program to detect local protein-coding segments in multiple sequence alignments [Washietl *et al.* 2011]. In a cooperation with Martin von Bergens group at the UFZ Leipzig we showed how RNAcode in combination with a newly developed protocol for mass spectrometry experiments [Müller *et al.* 2010] can improve protein annotation even in model organisms like *E. coli*.

4.2.1 Methods

Algorithm

Evolutionary changes in the nucleotide sequence of coding genes typically preserve the encoded protein. This type of negative (stabilizing) selection leads to frequent synonymous and conservative amino acid mutations, insertions/deletions preserving the reading frame, and the absence of premature stop codons. Our algorithm integrates this information in a unified scoring scheme. It takes as input a multiple nucleotide sequence alignment including a *reference* sequence, which is the one we wish to search for potential coding regions and predicts local segments that show statistically significant protein-coding potential. Figure 4.4 shows an overview of the algorithm that is described in more detail in the following sections.

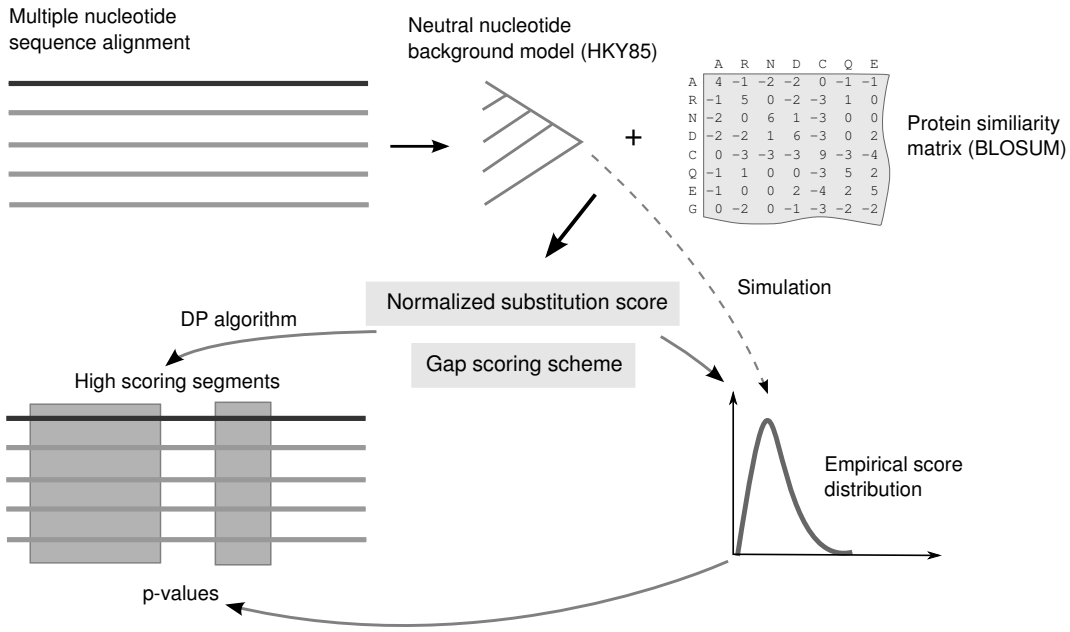


Figure 4.4. Overview of the `RNAcode` algorithm. First, a phylogenetic tree is estimated from the input alignment including a reference sequence (darker line) under a non-coding (neutral) nucleotide model. From this background model and a protein similarity matrix, a normalized substitution score is derived to evaluate observed mutations for evidence of negative selection. This substitution score and a gap scoring scheme is the basis for a dynamic programming (DP) algorithm to find local high scoring coding segments. To estimate the statistical significance of these segments a background score distribution is estimated from randomized alignments that are simulated along the same phylogenetic tree. The parameters of the extreme value distributed random scores are estimated and used to assign p-values to the observed segments in the native alignment.

Amino acid substitutions

Consider two aligned nucleotide triplets a and b that correspond to two potential codons. To see if they encode synonymous or biochemically similar amino acids, we can translate the triplets and use amino acid similarity matrices such as the widely used BLOSUM series of matrices [Henikoff & Henikoff 1992]. Let A_a and A_b be the translated amino acids of the triplets a and b, respectively, and $s(A_a, A_b)$ their BLOSUM score. In absolute terms this score is of little value: highly conserved nucleotide sequences will get high amino acid similarity scores upon translation even when they are non-coding.

We need to ask, therefore, what is the *expected* amino acid similarity score assuming that the

two triplets evolve under some non-coding (neutral) model. Deviations from this expectation will be evidence of coding potential. To this end, we estimate a phylogenetic tree for the input alignment using a maximum-likelihood method under the well-known HKY85 nucleotide substitution model [Hasegawa *et al.* 1985]. Further, we note that two aligned triplets can have zero, one, two or three differing positions, i.e., they can have a Hamming distance $h(a,b) \in \{0,1,2,3\}$. It is straightforward to calculate the expected score for a given protein matrix, a parametrized HKY85 background model and a given Hamming distance x:

$$\langle s \rangle_{h=x} = \sum_{\substack{a,b \\ h(a,b)=x}} s(A_a, A_b)\pi_{a_1}\pi_{a_2}\pi_{a_3} Prob(a \to b|t) \tag{4.1}$$

Here a_1, a_2, and a_3 denote the first, second, and third nucleotide in triplet a, π is the stationary frequency in the HKY85 model, and $Prob(a \to b|t)$ is the probability that triplet a changes to b after some time t. The analytic expression for this probability is given by Hasegawa *et al.* [1985]. The pairwise evolutionary distance t between two sequences is calculated as the sum of all branch lengths separating the two sequences in the estimated phylogenetic tree.

Put in simple terms, the score $\langle s \rangle$ is the average score over all possible pairs weighted by the probability to observe such a pair under our background assumption. We condition on the observed Hamming distance $h(a,b)$ as this reduces the effect of implicit information on average amino acid frequencies contained in the BLOSUM matrix, and was found to give better results. We can use this expected score $\langle s \rangle$ to normalize our observed scores s arriving at the final protein-coding score σ for an aligned triplet:

$$\sigma = s - \langle s \rangle \,. \tag{4.2}$$

To illustrate this with an example, consider the aligned triplets GAA and GAT. The triplets encode glutamic acid and aspartic acid, respectively, and score $s = +3$ in the BLOSUM62 matrix. Further, assume that under some background model the expected score for pairs with one difference is $\langle s \rangle_{h=1} = -1$. The overall score is thus $\sigma = 3 - (-1) = +4$. The positive score reflects the conservative mutation between the biochemically similar amino acids. A synonymous mutation usually gives the strongest support for negative selection. Since it also gives the highest scores in any protein matrix there is no need to treat it differently from conservative mutations and we can score both types of mutations using the same rules. Under this simple scoring scheme, the average triplet score in a coding alignment under negative selection will be positive, while in non-coding alignments it will be 0 on average. We found that the HKY85 substitution model accurately models non-coding regions for this particular purpose.

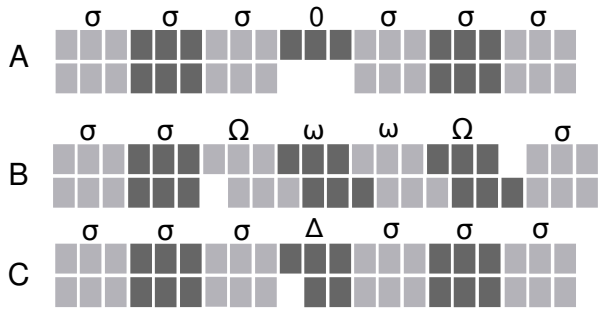


Figure 4.5. Examples of typical gap patterns and scoring paths in a pairwise alignment assumed to be coding. Nucleotides are shown as blocks, codons as three consecutive blocks of the same shading. A) A gap of length three does not change the reading frame and in frame-aligned codons are scored with the normalized substitution score σ. B) A single gap destroys the reading frame but gets corrected downstream by another gap. The triplets that are out of phase because of this obvious alignment error are penalized by the two frame-shift penalties Ω and ω. C) A single gap that, in principle, destroys the reading frame, is interpreted as a sequence error. Penalized by a high negative score Δ, this frame-shift is ignored and downstream codons are considered to be in phase.

Reading frames and gaps

It is straightforward to score an alignment that does not contain gaps. The alignment can simply be translated in all reading frames and the resulting triplets assigned a substitution score σ as described above. Real alignments, however, usually contain gaps. For the purpose of finding coding regions, gap patterns contribute valuable information [Kellis *et al.* 2004]. Negative selection not only acts on the type of amino acid but also on the reading frame which is generally preserved when insertions/deletions occur. The `RNAcode` algorithm incorporates this information into the scoring scheme and, in addition, also deals with practical problems that occur in real-life data such as alignment and sequencing errors.

The basic idea is that insertions or deletions in coding regions affect complete triplets and, therefore, result in gap patterns with a multiple of three that do not break the coding frame (Figure 4.5 A). We treat this kind of gap neutrally and give it a score 0. The aligned triplets before and after the gap are in the same phase and thus can be assigned a score σ.

In real data frameshifts can also be observed in coding regions because of alignment errors. Any gap not a multiple of three will result in a frameshift and the sequences are out of phase. We assign a penalty score $\Omega < 0$ for the frameshift event and each subsequent aligned triple that is out of phase receives an additional smaller penalty $\omega < 0$. However, in real

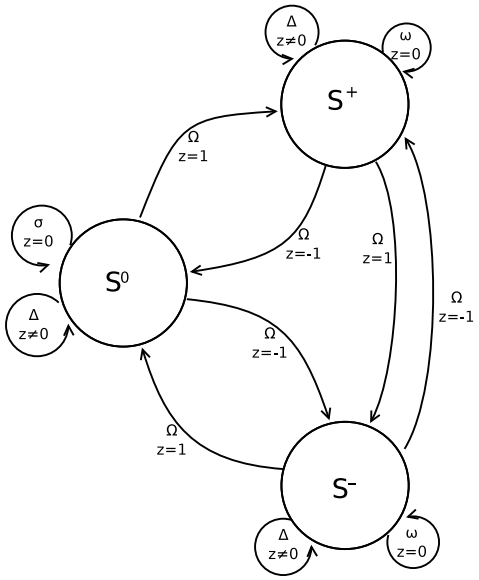


Figure 4.6. Finite state automaton representing the scoring of pairwise alignments. The three states correspond to the relative phases of the sequences. The gap pattern of each triplet along the pairwise alignment is analyzed. If both sequences stay in the same phase ($z = 0$; both triplets have the same number of gaps or one of the triplets completely consists of gaps while the other has no gap) the state remains constant. Insertions and deletions with $z \neq 0$ lead to local changes in phase that are penalized by Ω. While extensions in the "in-frame" state S^0 is scored with the normalized protein-coding score σ, extension in each of the two "out-of-frame" states S^+ and S^- is penalized by ω. In/dels interpreted as sequencing errors or true frameshifts are penalized by Δ.

coding regions such frameshifts usually get reverted soon by another gap. Consequently, only relatively short regions are out of frame. Changing the frame back is also penalized, again by Ω (Figure 4.5 B).

Gaps in coding regions that are not a multiple of three can also be the result of sequence errors. This is particularly problematic for low coverage sequencing. In order not to miss substantial parts of true coding regions that appear to be out of frame because of a single sequence error, we allow change of the phase and penalize this event with a negative score Δ (Figure 4.5 C). Clearly, this event should be rare and hence the penalty must be high; the condition $\Delta < 2\Omega$ must be met at least, or otherwise a sequence error event would always be chosen as a more favorable explanation than the frameshifting gaps in the optimization algorithm. The finite state automaton depicted in Figure 4.6 summarizes all possible transitions and the corresponding penalties for two sequences.

Stop codons

Under normal conditions a reading frame cannot go beyond a stop codon. To reflect this in our algorithm stop codons in the reference sequence get a score of $-\infty$. We allow relaxation of this for stop codons in the other sequences because if they are of low quality erroneous stop codons might be observed. These should not automatically destroy a potentially valid coding region but rather be penalized with a relatively large negative score.

Calculating the optimal score for a pairwise alignment

Using the scoring scheme introduced above, we need to find the interpretation of a given alignment as aligned codons in a particular reading frame, out-of-frame codons, and sequence errors that maximizes the score. This is achieved by a dynamic programming algorithm which is described in full detail in the Appendix.

Finding maximum scoring segments in a multiple alignment

To find regions of high coding potential in a multiple sequence alignment we first consider the pairwise combinations of the reference sequence with each other sequence. In these pairwise alignments, we calculate the optimal score of each alignment block delimited by two columns i and j using the dynamic programming algorithm. Once the maximum scores have been found for each pairwise alignment, we take the average of all pairs and store the optimal scores for the blocks between any two columns i and j of the multiple alignment in a matrix S_{ij} (see Appendix for details). In this matrix we identify maximal scoring segments, i.e., segments with a positive score that cannot be improved by elongating the segment in any direction.

Statistical evaluation

To assess the statistical significance of high scoring segments we empirically estimate the score distribution of neutral alignments conditional on the phylogeny derived from the alignment under consideration. Again, we use the phylogenetic tree estimated under the HKY85 model as our null model. We simulate neutral alignments along this tree and calculate high scoring segments in exactly the same way as for the native alignment. The score distribution follows an extreme value distribution and we found that it is well approximated by the Gumbel variant with two free parameters (Figure 4.7). Fitting this distribution allows us to calculate a p-value for every high scoring segment actually observed. This p-value expresses the probability that a segment with equal or higher score would be found in the given alignment by chance.

4.2.2 Results

Classification accuracy

RNAcode's algorithm is built on a direct statistical model that deliberately ignores any species-specific information and does not need any training. RNAcode is thus not optimized for the genome-wide annotation of protein-coding genes in well known model organisms. We tested RNAcode on six different comparative test sets. These test sets were created from genome-wide alignments [Blanchette *et al.* 2004, Schneider *et al.* 2006, Kuhn *et al.* 2009]. The set consisted of alignments of *E. coli* with 9 enterobacteria, *Methanocaldococcus jannaschii* with 10 methanogen Archaea, *Saccharomyces cerevisae* with 6 other Saccharomyces strains, *Drosophila melanogaster* with 11 drosophilid species and three other insects, *Caenorhabditis elegans* with 5 other nematode species and *Homo sapiens* aligned to 16 vertebrate genomes. From these alignments, we extracted both annotated coding regions/exons and randomly chosen regions without coding annotation. We then calculated the maximum coding potential score and its associated p-value for each alignment. We did not include explicit information on the reading direction, i.e., the coding regions were randomly either in forward or reverse complement direction and both directions were scored.

A typical score distribution (Figure 4.7 A) shows that random non-coding regions generally do not contain maximal scoring segments with scores higher than 15, whereas coding regions show a wide range of maximal scoring segments of much higher scores. The score efficiently discriminates coding and non-coding regions. Receiver operating curves (ROC) show the sensitivity and specificity of the classification at different score cutoffs (Figure 4.7 B). In general, we observe the area under the curves (AUC) of the ROCs to be close to 1, i.e. close to perfect discrimination. Usually, the high specificity range (Figure 4.7 B, insets) is of particular interest for large scale analysis. At a false positive rate of 0.05%, for example, we can detect approximately 90% of coding regions in all six test sets.

Accuracy of p-value estimates

The fact that the amino acid similarity scores used in our scoring scheme are adjusted by the expected score under a neutral null model ensures that the RNAcode score is properly normalized with respect to base composition and sequence diversity (phylogeny). In other words, the RNAcode score is independent of sequence conservation and G+C content. Unlike other abstract classifiers, it is therefore possible to interpret and compare scores in absolute terms. However, even more important is an accurate estimate of the statistical significance

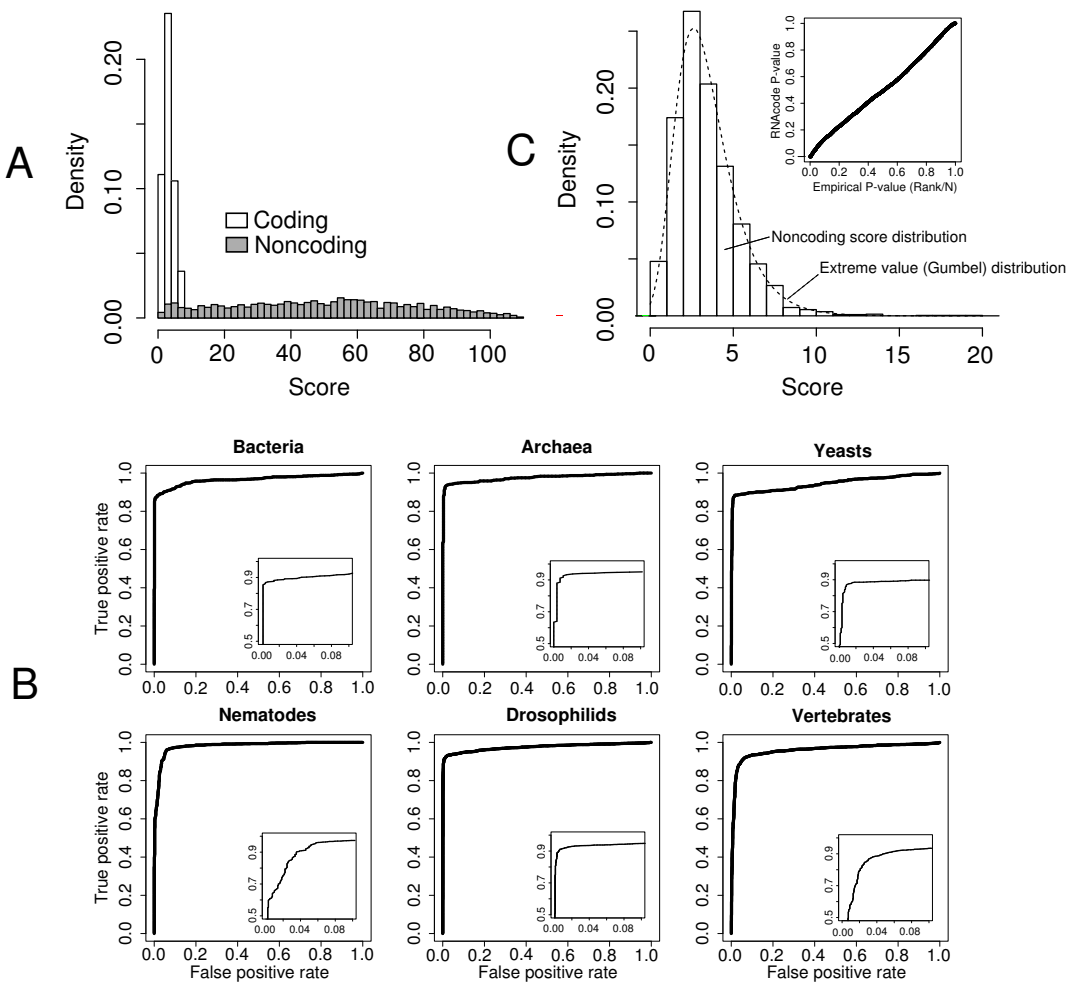


Figure 4.7. `RNAcode` results on comparative test sets from various species. A) Score distributions of annotated coding regions and randomly chosen non-coding regions in the *Drosophila* test set. B) ROC curves for six test sets. The full curve for all ranges of sensitivity/specificity from 0 to 1 is shown in the main diagrams. The inset depicts the high specificity range with false positive rates from 0 to 0.1. C) Score distribution of non-coding alignments. The same distribution of the *Drosophila* test set as shown in A) is shown in more detail. The fitted Gumbel distribution is indicated in red. The upper right diagram compares the calculated *p*-values (via simulation and fitting of the Gumbel distribution) to the empirical *p*-values, i.e. the actual observed frequencies in the test set.

of a prediction. `RNAcode` scores follow an extreme value distribution (Figure 4.7 C) which allows the calculation of *p*-values.

To test the accuracy of this approach, we compared *p*-values calculated by this procedure to empirically determined *p*-values on a set of non-coding *Drosophila* alignments. To this end,

we calculated the p-value for each alignment in the set and compared each to the proportion of alignments with better scores than the given one (Figure 4.7 C, inset). The excellent agreement of the p-values calculated by RNAcode and the actual observed frequencies confirms that the Gumbel distribution is an accurate approximation of the background scores. In addition, it also confirms that the HKY85 nucleotide substitution model and our simulation procedure accurately model real non-coding data.

Novel peptides in *Escherichia coli*

The main purpose of RNAcode is to classify conserved regions of unknown function, to discriminate coding from non-coding transcripts and to analyze the coding potential in non-standard genes (e.g. short ORFs or dual-function RNAs).

The *E. coli* genome was one of the first completely sequenced genomes and is generally well annotated. However, even in this compact and extensively studied genome the protein annotation is far from perfect. Protein gene annotation is largely based on compositional analysis and homology to known protein domains. The statistical power of these criteria is limited for small proteins. Standard gene finding software is usually run with an arbitrary cutoff of 40–50 amino acids to avoid an excess of false positives and suffers from the lack of training data of verified short peptides.

We attempted to produce a set of predictions based on evolutionary signatures only. We created alignments of the *E. coli* reference strain K12 MG1655 to 53 other completely sequenced enterobacteria strains including *Erwinia*, *Enterobacter* and *Yersinia*. A screen of these alignments with RNAcode and a p-value cutoff of 0.05 resulted in 6,542 high scoring coding segments. We discarded all predictions that overlapped annotated proteins. For the remaining RNAcode predictions, we tried to identify a complete ORF (starting with AUG and ending in a stop codon) in the *E. coli* reference sequence. This step is necessary because the boundaries of high scoring segments usually do not correspond exactly to the ORF (a main problem here is the relatively short alignment blocks produced by multiz, which do not always cover an ORF over its full length). This procedure gave 35 potential new protein-coding genes with a coding capacity between 11 and 73 amino acids.

To assess the quality of these predictions we first looked at the overall sensitivity of our screen on already annotated proteins. Of the 4,267 RefSeq proteins, 3,987 overlapped with a RNAcode prediction (sensitivity 93.4%). Hemm *et al.* [2008] revisited the annotation of small proteins in *E. coli* and found 18 novel examples using a combination of different bioinformatics and experimental methods. In a set of 18 new and 42 literature-curated proteins between 16–50

Table 4.2. Protein counts of the LMW (low molecular weight) proteome registered in SwissProt database and the proteins recovered by the developed MS (mass spectrometry) protocol, which is specifically optimized for small proteins, are summarized by their SwissProt evidence and RNAcode support. A detailed description of the MS protocol is published in [Müller *et al.* 2010]

	SwissProt LWM			Optimized LWM MS protocol		
		RNAcode			RNAcode	
	total	found	missed	total	found	missed
Total Number	1605	1401	204	455	449	6
Evidence on protein level	843	805	38	359	359	0
Evidence on transcript level	34	28	6	6	6	0
Inferred from homology	272	245	27	36	34	2
Predicted	378	288	90	54	50	4
Uncertain	78	35	43	0	0	0

amino acids compiled by Hemm *et al.*, 30 (50.0%) overlap with `RNAcode` predictions.

We furthermore compared the `RNAcode` predictions with the low molecular weight (LWM < 25 kDa) *E. coli* proteome registered in the SwissProt database [UniProt Consortium 2010]. For each protein the type of evidence and the amino acid sequence was extracted and mapped to the *E. coli* genome. Eighty percent of the 1605 mapped LMW SwissProt protein loci overlapped with `RNAcode` predicted high scoring segments. Interestingly, 95% (833/868) of the proteins with either proteome or transcriptome evidence listed in the SwissProt database are positively classified by RNAcode (Table 4.2). This indicates a strong enrichment of experimentally supported proteins in `RNAcode` predictions. On the other hand, of the proteins which are not validated experimentally or inferred by sequence homology, only 70% (323/456) were supported by `RNAcode` predictions (Table 4.2). This difference suggests that many but probably not all of the as-yet unverified reading frames in the SwissProt database are real protein-coding segments.

These results show that our screen not only gives almost perfect results on typical *E. coli* proteins, but also recovers a substantial fraction of small proteins which are particularly difficult to detect. Moreover, our final list of 35 candidates for novel proteins is rather short and shows the high specificity in this screen.

For additional support, we compared our list of predicted candidates with publicly available transcriptome data [Tjaden *et al.* 2002, Cho *et al.* 2009]. These data sets cover a broad range of experimental conditions and therefore reflect a comprehensive genome-wide transcription

map of *E. coli*. Eight candidates (23%) overlap with regions that show clear evidence for transcription.

To further substantiate our predictions, we used mass spectrometry (MS) as a direct experimental test for the existence of the novel peptides in *E. coli* cells. MS is particularly well suited to screen simultaneously for a large set of proteins without resorting to cloning or recombinant expression [Aebersold & Mann 2003]. Many, but by no means all, proteins of an organism are expressed and detectable under the actual applied conditions by current MS-based proteomics. Detecting small peptides in complex protein mixtures is particularly challenging for various reasons. Compared to the overall protein expression level, short peptides often show low abundance, they are easily lost using standard proteomic protocols, and only a limited number of proteolytic peptides can be obtained [Klein *et al.* 2007]. To meet these challenges, our collaborators (Martin von Bergen's lab, Leipzig) developed a protocol which is specifically optimized for small proteins by avoiding sample loss by a simple extraction method and a combined purification and enrichment step using filtration [Müller *et al.* 2010]. In order to improve the reliability of our results two different buffer systems are applied for extractions and for an improved coverage of peptides two different proteases are used. This strategy led to an increased detection rate as well as to higher confidence in the hits by confirmation in independent experiments.

Using this protocol, we were able to identify 455 LWM proteins representing 27% of the 1672 known *E. coli* proteins below this size listed in the SwissProt protein database. Among the 455 proteins 449 (99%) show a clear evolutionary signal for conservation at the nucleic acid level, as measured by `RNAcode` (Table 4.2). Proteome or transcriptome evidence is also reported in the SwissProt database for 81% (365/449) of these. Thus, the proteins identified with the LMW optimized MS protocol and the `RNAcode` predictions are highly correlated.

In a search against the list of 35 newly predicted proteins, we obtained evidence for the expression of 7 candidates (20%). For the rest of the candidates we cannot distinguish whether they are false positive `RNAcode` predictions or false negatives in the MS experiment. However, considering that the success rate of the MS experiments is roughly the same on known and predicted proteins (27% and 20%, respectively), we would expect a good fraction of our candidates to be true proteins not detectable by this particular growth conditions and MS approach.

Although it is not possible to give a conclusive statement on all predictions without additional experiments, compelling evidence from evolutionary analysis, transcriptomics data, and the MS experiments strongly suggest that several of the candidates are *bona fide* proteins. Figure 4.8 shows two examples in more detail. In both cases `RNAcode` reported short but

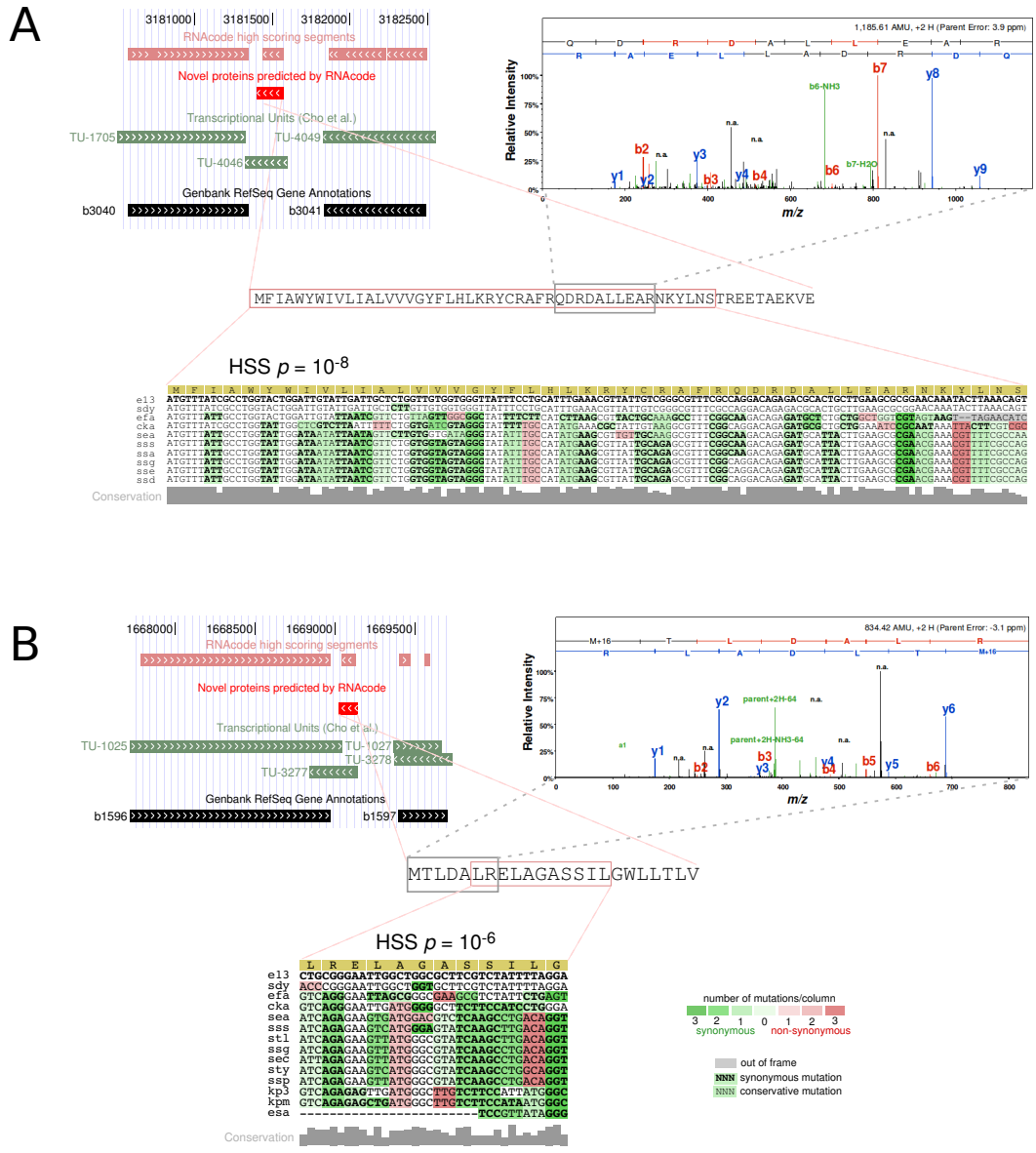


Figure 4.8. Examples of novel short proteins in *E. coli*. Sequence, genomic context, the high scoring `RNAcode` segment and fragment ion mass spectra are shown. Genome browser screen-shots were made at `http://archaea.ucsc.edu` [Schneider *et al.* 2006]. Arrows within annotated elements indicate their reading direction. The shading of mutational patterns was directly produced by the `RNAcode` program. The mass spectra are shown for two selected proteolytic peptides which were scored with 80% probability and used in combination with the detection of additional peptides to confirm the expression of the candidates.

statistically highly significant ($p \approx 10^{-8}$ and $p \approx 10^{-6}$, respectively) signals between two well-annotated proteins. The loci overlap with transcribed regions as determined by Cho *et al.* [2009]. In addition, our MS experiments detected several proteolytic fragments that can be assigned to these proteins.

The coding potential of "non-coding" RNAs

In addition to assisting and complementing classical protein gene annotation strategies, a major area of application of RNAcode is the functional classification of individual conserved or transcribed regions. As an illustrative example we analyzed the bacterial RNA C0343 which is listed in the Rfam database [Gardner *et al.* 2009] as non-coding RNA (ncRNA) of unknown function. The RNA originally detected by Tjaden *et al.* [2002] is also detected as transcript in the study of Cho *et al.* [2009]. In our screen of the *E. coli* genome, we found a high scoring coding segment with $p \approx 10^{-9}$ overlapping the C0343 ncRNA. The prediction corresponds to a potential ORF encoding 57 amino acids (Figure 4.9 A). Analysis of the secondary structure using RNAz does not give any evidence for a functional RNA. Given the strong coding signal, we conclude that the "non-coding RNA" C0343 is in fact a small protein. This is also confirmed by our MS experiments that detected proteolytic fragments of this protein in *E. coli* cells.

To test RNAcode on another example from Rfam, we analyzed RNAIII, a ncRNA known to regulate the expression of many genes in *S. aureus* [Boisset *et al.* 2007]. In addition to its role as regulatory RNA, the RNAIII transcript also contains an ORF coding for the 26 amino acid long delta-haemolysin gene (*hld*). We ran RNAcode with standard parameters on the Rfam seed alignment. It reports one high scoring segment below a p-value cutoff of 0.05 which corresponds to the *hld* gene (Figure 4.9 B). The annotated alignment shows that the ORF is highly conserved with only few mutations. Nevertheless, these few mutations are sufficient to yield a statistically significant signal that allows RNAcode to locate the correct ORF. Again, we also ran RNAz on the alignment, which reports a conserved RNA secondary structure with a probability of 0.99. The combination of RNAcode and RNAz clearly shows the dual function of RNAIII. This example demonstrates how RNAcode can assist the classification of ncRNAs in particular for non-standard and ambiguous cases [Dinger *et al.* 2008].

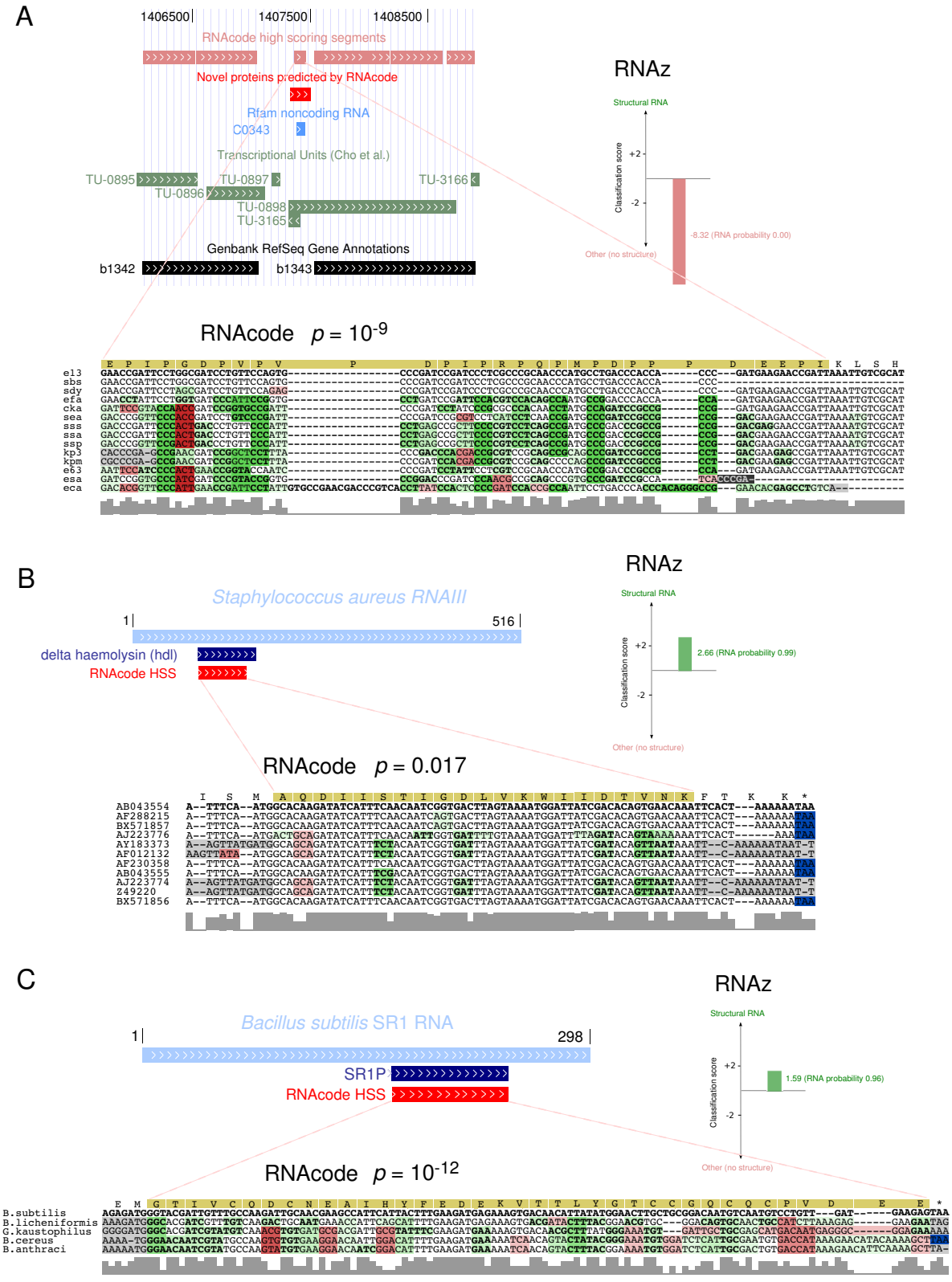
As another example, we analyzed the SR1 RNA of *B. subtilis* that was originally found by Licht *et al.* [2005] (Figure 4.9 C). Although the authors noticed a potential short ORF in the transcript, the corresponding peptide could not be detected. Further experiments [Heidrich *et al.* 2006; 2007] clearly showed a function of SR1 in the arginine catabolism pathway by RNA/RNA interaction with the ahrC mRNA, thus confirming its nature as functional non-

coding RNA. Using RNAcode, we found clear evolutionary evidence for a well-conserved small peptide deriving from the SR1 ($p \approx 10^{-12}$), arguing for a role as dual function RNA. Only recently, Gimpel *et al.* [2010] showed that gapA operon is regulated by a short peptide encoded in SR1, which exactly corresponds to the high scoring coding segment found by RNAcode (Figure 4.9 C).

4.2.3 Conclusion

RNAcode was designed to fill a specific gap in the current repertoire of comparative sequence analysis software. The classification of RNAcode relies on evolutionary signatures only and is based on a direct statistical model. No machine learning or training is involved and it can thus be applied in a generic way to data from all species. The presented consistency of current protein-coding gene annotation and RNAcode predictions clearly shows that the implemented algorithm gives almost perfect results on data of all domains of life. We furthermore identified new small protein-coding genes in *E. coli* and gave convincing evidence that several of these genes are indeed expressed. In addition we showed that the combined usage of RNAcode and RNAz discriminates coding from non-coding RNAs and identifies the known dual-function of RNAIII and SR1.

Figure 4.9 (following page). Examples of ambiguities between coding and non-coding nature of three RNAs. A) The RNA C0343 from *E. coli* is listed as non-coding RNA in Rfam. However, it overlaps with an RNAcode predicted coding segment. While there is no evidence for an RNA secondary structure according to the RNAz classification value, the highly significant RNAcode prediction and MS experiments suggest that C0343 is a mRNA and not a ncRNA. B) RNAIII of *S. aureus* (Rfam RF00503) contains a short ORF of a heamolysin gene. RNAcode predicts the open reading frame at the correct position, while RNAz clearly detects a structural signal. These results are consistent with the well established dual nature of this molecule. C) The *B. subtilis* RNA SR1 is known to be functional on the RNA level by targeting a mRNA. RNAcode detects a short ORF that was shown by Gimpel *et al.* [2010] to produce a small peptide and is thus another example of a dual-function RNA.


A
1406500
1407500
1408500
RNAcode high scoring segments
Novel proteins predicted by RNAcode
Rfam noncoding RNA
C0343
Transcriptional Units (Cho et al.)
TU-0895
TU-0897
TU-3166
TU-0896
TU-0898
TU-3165
Genbank RefSeq Gene Annotations
b1342
b1343
RNAz
Structural RNA
Classification score
+2
-2
-8.32 (RNA probability 0.00)
Other (no structure)
RNAcode p = 10-9
el3
sbs
sdy
efa
cka
sea
sss
ssa
ssp
kp3
kpm
e63
esa
eca
B
Staphylococcus aureus RNAIII
1
516
delta haemolysin (hdl)
RNAcode HSS
RNAz
Structural RNA
Classification score
+2
-2
2.66 (RNA probability 0.99)
Other (no structure)
RNAcode p = 0.017
AB043554
AF288215
BX571857
AJ223776
AY183373
AF012132
AF230358
AB043555
AJ223774
Z49220
BX571856
C
Bacillus subtilis SR1 RNA
1
298
SR1P
RNAcode HSS
RNAz
Structural RNA
Classification score
+2
-2
1.59 (RNA probability 0.96)
Other (no structure)
RNAcode p = 10-12
B.subtilis
B.licheniformis
G.kaustophilus
B.cereus
B.anthraci

5

Summary

RNA sequencing methods are powerful to identify and analyze RNA transcripts on a genome-wide scale. We found dozens of small RNA candidates and their validation by independent methods is only a first step towards the functional characterization of these transcripts. The high amount of putative RNA transcripts resulting from RNA sequencing experiments demands the development of new bioinformatic methods. Starting from the single sequence of a transcribed region the assignment of orthologs in other species is an essential first task. If an ORF is found within a single sequence the abstraction of the translated amino acid sequence is typically used to infer orthologous genomic regions by `Blast`. In case of non-coding regions classical sequence based methods often fail. Since ncRNAs typically vary in sequence but still fold into the same secondary structure, sophisticated methods (e.g. `RNAmotif` and `Infernal`) are used to identify structured sequences. Anyway, these methods result in long candidate lists that encompass predictions of different size and quality. Measurements are needed to estimate how well a new candidate fits to an established set of sequences. The similarity between protein-coding sequences is typically assigned using variation patterns of synonymous and non-synonymous substitutions. For structured non-coding RNA genes the base pair difference between a candidate structure and the established set or the ratio of the consensus energy with and without the new candidate might help to distinguish true from false predictions. The co-occurrence of candidate sequences with adjacent genes can be used to filter the initial list. This so called synteny information can be inferred by manual inspection of the given genome annotation. Recently, we implemented `Proteinortho` [Lechner *et al.*, submitted] an efficient reciprocal best blast approach that reliably detects orthologous proteins in a given set of sequences. Thus, synteny of adjacent genes can be inferred from `Proteinortho` re-

sults and subsequent visualization of genomic regions surrounding the predicted candidates. Furthermore, the structural clustering approach implemented in `RNAclust.pl` can help to identify related RNA sequences (of different RNA sequencing analysis) that share structural motifs even if common homology search tools like `Blast` fail.

As indicated for the sX13 sRNA of *Xanthomonas campestris* pv. *vesicatoria* (*XCV*) structural models and interaction prediction methods like `RNAup` can help to identify plausible mRNA targets. Obviously, such predictions can direct experimental setups for extensive validations. Only a few sRNA examples are currently known that directly mediate protein activity. How these sRNA-protein complexes are formed is mainly unknown. These examples might be exceptions to the rule, but if more sRNA-protein complexes are identified the uncharted territory of protein-target prediction will be moved into the focus of research.

The presented dRNA-seq approach enabled us to generate genome-wide TSS maps of *H. pylori* and *XCV*. Other research groups already applied this technique to analyze the primary transcriptome of different bacteria [Albrecht *et al.* 2009, Mitschke *et al.* 2011] and archaea [Jäger *et al.* 2009]. This clearly indicates the general applicability of the dRNA-seq method. Most studies that apply dRNA-seq in prokaryotes generate the TSS map by manual inspection. To get in touch with data such pilot studies are essential. Nevertheless, this is a time consuming biased and not fully reproducible procedure. Furthermore, for larger eukaryotic genomes manual TSS map creation is not feasible. I presented a novel approach that addresses the problem of automated TSS annotation based on dRNA-seq data and subsequent statistical analysis. This method gave convincing results for the analyzed *XCV* data set. The application of the approach to examine other data sets is one next step that remains to be done in the future. Used parameters (e.g. window size, p-value cutoff) were selected upon inspection of the *XCV* data and are therefore specific. Systematic analysis of parameter dependencies to the sequencing depth of the library are compulsory. Issues such as the correction for multiple testing have to be taken into account as well. The predictive power of the statistical approach will be analyzed based on independently validated transcription start sites and manually curated TSS maps of other studies.

The above mentioned and many other transcriptome studies (e.g. Wurtzel *et al.* [2009], Toledo-Arana *et al.* [2009], Güell *et al.* [2009]) revealed a far more complex transcript organization than previously appreciated. An unexpected high number of transcript was found to originate from the opposite strand of annotated protein-coding genes. Its likely that these antisense transcripts directly regulate the expression of the protein-coding gene on the opposite strand. It remains to be seen if the formation of antisense RNA-mRNA complexes or transcriptional interference is the common regulatory mode of these antisense configura-

tions. The high amount of transcribed leaderless mRNAs and the lack of an SD sequence in *XCV* supports the assumption of alternative mechanisms of ribosome guidance. These and other findings indicate the necessity to revise our current understanding of prokaryotic transcriptional and translational processes.

With the availability of genome-wide transcription data and massive comparative sequencing the discrimination of coding from non-coding RNAs in evolutionarily conserved regions arose as a core analysis task. Thus I addressed the question "How to assess protein-coding and non-protein-coding potential of genomic regions?" in the second part of my thesis. We significantly improved the predictive power of the widely used ncRNA gene finding tool `RNAz`. Limitations of the old version have been eliminated by the usage of a di-nucleotide background model, a much larger training set and the introduction of the normalized Shannon entropy as a measure of sequence variation. We furthermore trained an additional classification model for structural alignments which provides an additional possibility to increase the prediction accuracy of `RNAz`.

With the implementation of `RNAcode` we filled a specific gap in the current repertoire of protein-coding gene finding software. `RNAcode` makes use of all evolutionary signatures that are known to be relevant rather than focusing on one particular feature. The statistical model of `RNAcode` relies on evolutionary signals only and no training on known protein-coding genes is involved. Thus, `RNAcode` is applicable to identify protein-coding regions in any species. Our analysis of the *E. coli* genome showed that `RNAcode` predictions are in good agreement with the current annotation. We furthermore detected novel small protein-coding gene candidates in this exhaustively studied species. Comparison with transcriptomic and proteomic data provide ample evidence that a considerable amount of these genes is indeed functional.

Outlook

High-throughput RNA sequencing is still in its infancy and further improvements are necessary. Ideally transcriptome analysis would capture full length RNAs of an individual cell. However, current technologies are limited to read lengths of a few hundred nucleotides which is far from full length mRNA or operon spanning transcripts. The presented dRNA-seq approach enriches the 5'-end of primary transcripts but the analysis of transcription units suffers from the limited read lengths and therefore correct 3'-ends of individual transcripts are missing. Technical problems during library construction (e.g. premature stops of the RT-PCR in the cDNA conversion step) and the nature of RNA molecules (e.g. stable secondary structure) cannot be addressed by increased read lengths only. The implementation of an RNA sequencing protocol that specifically enriches the 3'-end of transcripts in combination with

the presented dRNA-seq approach could help to generate complete transcription unit maps.

To produce sufficient starting material for RNA sequencing analysis the RNA output of a bacterial culture is required instead of a single cell. Thus, a mixture of species which might be in different developmental stages is sequenced simultaneously. Only recently Tang *et al.* [2010] presented an RNA sequencing strategy to capture the transcriptome landscape of an individual cell. Their current approach is limited to mRNAs with a poly(A) tail and a maximum length of 3 kb. Obviously, this method has to be improved and adapted to be applicable for prokaryotes, where mRNAs typically lack a poly(A) tail.

Due to cost reasons no replicates of individual experiments are available. Thus, the reproducibility of observations and the error rate of RNA sequencing experiments has not been examined in detail. Many potential biases may be introduced during cDNA construction, adapter ligation, amplification and sequencing that have to be analyzed in the future. With the emergence of novel sequencing technologies such as FRT-seq [Mamanova *et al.* 2010], Nanopore [Clarke *et al.* 2009] and direct RNA sequencing [Ozsolak *et al.* 2009] prices of current technologies will drop and replicates become affordable.

Transcription-profiling in diverse growth conditions also suffers from the lack of replicated RNA sequencing experiments. To overcome this limitation statistical tools have been implemented to model the expected distribution of sequencing reads that map to a genomic locus in different samples [Wang *et al.* 2010, Robinson & Oshlack 2010]. In contrast the quantification of differential gene expression has been reported using the less expensive array techniques [Toledo-Arana *et al.* 2009, Cho *et al.* 2009]. It has been observed that biological replicates often reflect non-trivial differences in molecular binding activity and that averaging can abolish strong enrichment signals or indicate binding events that are not supported by any individual replicate. Hence, even if replicates of individual experiments are available the evaluation and correction of these data sets is a non-trivial task.

Comparative genomic methods can be used as an alternative to detect novel protein-coding and non-protein-coding genes. Approaches like `RNAz` and `RNAcode` predict conserved sequences based on (genome-wide) multiple sequence alignments. In contrast to experimental surveys these methods miss species specific genes and their results strongly depend on the alignment quality and the selected set of species.

Even the updated version of `RNAz` has an estimated false discovery rate of $\sim$50%. Since, RNA secondary structure prediction is sensitive to the length of the given sequence alignment the used fixed window approach is still a major source of prediction errors. Alignment slices of a fixed length are classified as structural RNA or other. Hence, boundaries of known

ncRNAs which are shorter than the selected window size are overestimated. The LocARNA-P approach, overcomes this limitation [Will *et al.* 2011]. The implemented method is based on efficiently computed reliability profiles that take structure and sequence information of the given alignment into account. In a case study LocARNA-P has been successfully applied to refine RNAz predictions. The detected boundaries of known ncRNAs were in good agreement with the given annotation and the false discovery rate of RNAz has been significantly reduced. The pre-selection of locally structured regions from the typically large set of genome-wide alignments with RNALalifold could replace the currently used sliding window approach. To filter genomic regions that already show signatures of increased thermodynamic stability would also reduce the number of alignments to be screened by RNAz and therefore makes sequence/structure based classification more feasible.

RNAcode predicts local high-scoring coding segments within multiple sequence alignments. In the analyzed pro- and eukaryotic data sets we found that RNAcode predictions are in good agreement with the given annotation of protein-coding genes. However, the implemented method does not recover the complete gene structure. Hence, adjacent high scoring segments (typically the result of fragmented genome-wide alignment blocks) have to be merged. In prokaryotes the detection of a complete ORF covering merged RNAcode predictions enabled us to detect new short protein-coding genes that have evaded previous annotation. The protein-coding gene structure in eukaryotes comprises additional intronic sequences that have to be spliced out before translation. Since, RNAcode recovers only the coding regions (exons) additional information (e.g. conservation of splice sites) has to be used to reveal the typical exon-intron structure of eukaryotic protein-coding genes. Using a few post-processing steps to refine the initial RNAcode predictions will help to identify novel short (spliced) protein-coding RNAs in eukaryotes as well.

The recently published NAPP (Nucleic acid phylogenetic profiling) approach represents an alternative to the usage of standard genome-wide alignments [Marchais *et al.* 2009]. NAPP systematically infers the distribution of intergenic regions of a reference species across all fully sequenced bacterial genomes. An empirical conservation index is assigned to each intergenic position of the reference genome. Subsequently, conserved non-coding elements (CNEs) are defined as segments with a continuous conservation index above a certain threshold. The detected CNEs covered ∼80% of all known *E. coli* and *B. subtilis* sRNAs. The authors introduced a very elegant way to further investigate those CNEs which do not belong to known RNA genes. They used phylogenetic profiling to cluster vector representations of CNEs and CDSs. Some of the resulting clusters clearly showed an enrichment of known ncRNA genes with novel CNEs. Furthermore, CDS with a similar phylogenetic distribution are found within

these clusters and represent potential targets of sRNAs and CNEs. Compared to RNAz the NAPP approach neither requires extensive training data nor involves the analysis of sequence composition. While RNAz is designed to detect structured ncRNAs only NAPP is able to detect both structured and unstructured RNAs. However, the huge amount of predicted CNEs needs further investigations. At this point RNAz and RNAcode could help to classify CNEs according to their thermodynamic stability and protein-coding potential. Since, the complete set of sequenced bacterial genomes is used as input for the NAPP analysis sRNAs with an erratic distribution (e.g. PtaRNA1 described in Section 3.3) are detectable as well. The current NAPP implementation uses Blast to search for intergenic regions of the reference species in all other bacterial genomes. Novel short read aligner (e.g. segemehl) which are initially designed to map RNA sequencing data to a reference sequence could replace Blast. These novel methods are designed to efficiently map huge amounts of short sequences even if they have mismatches, insertions or deletions with respect to the reference genome. These short read aligner not only allow an error tolerant comparison of the intergenic regions but also could be used to flip database and query of the analysis. In other words genomic sequences of all bacteria could be fractionated into shorter sequences and used as "read library" that is mapped against the reference genome. Consequently, the conservation index could be assigned to each genomic position and the analysis could be extended to all conserved elements (CEs) of the reference genome.

The initial motivation of RNAcode and RNAz was the functional classification of conserved and transcribed genomic regions. As shown in our studies each tool can assist individually to detect novel protein-coding and non-protein-coding genes. As exemplified on a few known dual-functional RNAs (e.g. RNAIII and SR1) both tools in combination can help to analyze ambiguities between the coding and non-coding nature of RNA transcripts. This particular issue has only been addressed by the RNA-DECODER approach [Pedersen *et al.* 2004, Meyer & Miklós 2005]. This tool explicitly models RNA structures that overlap protein-coding regions, as are frequently observed in RNA viruses. The implemented method employs a stochastic context-free grammar in combination with a set of carefully devised phylogenetic substitution models that reflect both the coding and non-coding property of functional RNA structures within protein-coding regions. The development of RNA-DECODER was challenged by a limited set of well curated coding RNA structures. Although RNAz was specifically trained on known ncRNAs its application to identify structured RNA elements within protein-coding regions is possible. We are currently investigating the combined usage of RNAz and RNAcode in order to analyze structured RNA elements within protein-coding regions and to detect dual-functional RNAs on a genome-wide scale.

Taken together, this thesis demonstrates the power of combining experimental approaches with computational predictions. The huge amount of data generated during RNA sequencing experiments requires the development of new algorithms and tools. On the other hand comparative methods, like `RNAz` and `RNAcode`, predict large numbers of novel protein-coding and non-coding RNA candidates, that need to be experimentally analyzed on genome-wide scale. Only if we use the combination of experimental and computational approaches we will be able to explore the fascinating world of RNA.

A

Supporting Material

Data Links

In the following, we list corresponding web links to the supplemental material hosted by the University of Leipzig or the publisher's page for all publications on which this thesis is based.

- Sonnleitner E, Sorger-Domenigg T, Madej MJ, Findeiß S, Hackermüller J, Hüttenhofer A, Stadler PF, Bläsi U and Moll I. *Detection of small RNAs in Pseudomonas aeruginosa by RNomics and structure-based bioinformatic tools.* Microbiology, 2008 Oct; 154(10): 3175–3187
  Publication: `http://mic.sgmjournals.org/cgi/content/abstract/154/10/3175`
  Supplemental: `http://mic.sgmjournals.org/cgi/content/full/154/10/3175/DC1` and `http://www.bioinf.uni-leipzig.de/publications/supplements/07-023`

- Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiß S, Sittka A, Chabas S, Reiche K, Hackermüller J, Reinhardt R, Stadler PF and Vogel J. *The primary transcriptome of the major human pathogen Helicobacter pylori.* Nature, 2010 Mar; 464(7286): 250-5.
  Publication: `http://www.nature.com/nature/journal/v464/n7286/full/nature08756.html`
  Supplemental: `http://www.nature.com/nature/journal/v464/n7286/suppinfo/nature08756.html`

- Schmidtke C, Findeiß S, Sharma CM, Kuhfuß J, Hoffmann S, Vogel J, Stadler PF and Bonas U. *Genome-wide automated differential transcriptome analysis of the plant pathogen Xanthomonas identifies sRNAs with putative virulence functions.*
  Publication: in preparation
  Supplemental: `http://www.bioinf.uni-leipzig.de/publications/supplements/10-035`

- Findeiß S, Schmidtke C, Stadler PF and Bonas U. *A novel family of plasmid-transferred anti-sense ncRNAs*. RNA Biol, 2010 Mar; 7(2): 120-4.
  Publication: `http://www.bioinf.uni-leipzig.de/Publications/PREPRINTS/09-041.pdf` (preprint)
  Supplemental: `http://www.bioinf.uni-leipzig.de/publications/supplements/09-041`

- Gruber AR, Findeiß S, Washietl S, Hofacker IL and Stadler PF. *RNAZ 2.0: Improved noncoding RNA detection*. Pac Symp Biocomput, 2010 Jan; 15:69-79.
  Publication: `http://www.bioinf.uni-leipzig.de/Publications/PREPRINTS/09-026.pdf` (preprint)
  Supplemental: `http://www.bioinf.uni-leipzig.de/publications/supplements/09-026`
  Software: `http://www.tbi.univie.ac.at/~wash/RNAz/`

- Washietl S, Findeiß S, Müller S, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF and Goldman N. *RNAcode: robust discrimination of coding and noncoding RNAs in comparative sequence data*. RNA, accepted
  Publication: `http://www.bioinf.uni-leipzig.de/Publications/PREPRINTS/10-001.pdf` (preprint)
  Supplemental: `http://www.bioinf.uni-leipzig.de/publications/supplements/10-001`
  Software: `http://wash.github.com/rnacode/`

- Müller SA, Kohajda T, Findeiß S, Stadler PF, Washietl S, Kellis M, von Bergen M and Kalkhof S. *Optimization of Parameters for Coverage of Low Molecular Weight Proteins.* Anal Bioanal Chem, 2010 Dec; 398(7-8): 2867–2881.
  Publication: `http://www.springerlink.com/content/q73l761011386212/fulltext.pdf`

Detailed RNAcode dynamic programming algorithm

The core algorithm of RNAcode is a dynamic programming algorithm to find the optimal score for a pairwise alignment from all possible interpretations of the aligned sites as in-frame codons, codons, out-of-frame codons or sequence errors. The scores from pairwise alignments are then combined to find optimal scoring segments in a multiple alignment.

We start from a fixed multiple sequence alignment $\mathbb{A}$ and assume that the first row is the *reference sequence*. The projected pairwise alignment of the reference sequence with sequence k is denoted by $\mathbb{A}^k$. Now consider a position i in the reference sequence. It corresponds to a uniquely determined alignment column $\alpha(i)$, which in turn determines i_k, the last position of sequence k that occurs in or before alignment column $\alpha(i)$.

Suppose i is a third codon position. Then the alignment block $\mathbb{A}[\alpha(i-3)+1, \alpha(i)]$ corresponds to the (potential) codon ending in i. We define a score

$$\sigma_i^k = \text{score}\left(\mathbb{A}^k[\alpha(i-3)+1, \alpha(i)]\right) \tag{A.1}$$

.

In the ungapped case σ_i^k is the normalized BLOSUM score that was introduced in the main text. Let g_i^k denote the number of gaps in sequence k in this block. We observe that sequences 1 (reference) and k stay in frame if and only if $g_i^k - g_i^1 \equiv 0 \mod 3$. Otherwise, the two sequences change their phase within this interval. The local shift in frame between sequence k and the reference sequence is therefore

$$z_i^k = \begin{cases} 0 & \text{if } g_i^k - g_i^1 \equiv 0 \mod 3 \\ +1 & \text{if } g_i^k - g_i^1 \equiv 1 \mod 3 \\ -1 & \text{if } g_i^k - g_i^1 \equiv 2 \mod 3 \end{cases} \tag{A.2}$$

As discussed in the main text, alignment errors or sequence errors may destroy coherence between aligned codons and give $z_i^k \neq 0$. Therefore, we introduce the penalties (negative scores) Ω for switching from in-frame to out-frame or back, as well as ω for every out-of frame codon in between, and Δ for silently changing the phase and assuming subsequent codons are still in frame (sequencing error). All penalties are negative; in particular $\frac{1}{2}\Delta < \Omega < \omega < 0$. Furthermore, we set $\sigma_i^k = -\infty$ if $z_i^k \neq 0$ to mark the fact that we lose coherence of the frame and force the algorithm to select a frameshift or sequence error penalty and not a substitution score that would be meaningless for out-of-frame triples.

Having defined all possible states and the associated scores, we now describe a dynamic programming algorithm to calculate the optimal score for a pairwise alignment. Let $S_{b,i}^{0,k}$ be

the optimal score of the pairwise alignment $\mathbb{A}^k[\alpha(b), \alpha(i)]$ subject to the condition that i is a third codon position and sequence k ends in frame, i.e., also with a third codon position. Analogously, we define $S^{+,k}_{b,i}$ and $S^{-,k}_{b,i}$ for those alignments where sequence k ends in the 1st and 2nd codon position, respectively. Clearly we initialize $S^{\chi,k}_{b,b} = 0$ for $\chi \in \{0, +, -\}$.

The entries in these matrices satisfy the following recursions:

$$S^{0,k}_{b,i} = \begin{cases} S^{0,k}_{b,i-3} + \sigma^k_i & \text{if } z^k_i = 0 \\ \max \begin{cases} S^{0,k}_{b,i-3} + \Delta, \\ S^{-,k}_{b,i-3} + \Omega \end{cases} & \text{if } z^k_i = +1 \\ \max \begin{cases} S^{0,k}_{b,i-3} + \Delta, \\ S^{+,k}_{b,i-3} + \Omega \end{cases} & \text{if } z^k_i = -1 \end{cases} \tag{A.3}$$

The expressions for the two out-of-frame scores are analogous. We show only one of them explicitly:

$$S^{+,k}_{b,i} = \begin{cases} S^{+,k}_{b,i-3} + \omega & \text{if } z^k_i = 0 \\ \max \begin{cases} S^{0,k}_{b,i-3} + \Omega \\ S^{+,k}_{b,i-3} + \Delta \end{cases} & \text{if } z^k_i = +1 \\ \max \begin{cases} S^{+,k}_{b,i-3} + \Delta \\ S^{-,k}_{b,i-3} + \Omega \end{cases} & \text{if } z^k_i = -1 \end{cases} \tag{A.4}$$

As presented here, the algorithm assumes that any sequence errors (penalized by Δ) occur in sequence k, not in the reference.

Now we determine the optimal score S_{bi} of the multiple alignment $\mathbb{A}[\alpha(b), \alpha(i)]$, subject to the condition that b is a 1st codon position and i is a third codon position.

$$S_{bi} = \max \begin{cases} \sum_{k>1} \max_{\chi \in \{0,+,-\}} S^{\chi,k}_{b,i} \\ S_{b,i-1} + \Delta \\ S_{b,i-2} + \Delta \end{cases} \tag{A.5}$$

The second and third terms here correspond to frameshifts in the reference sequence.

It is easy now to determine the best scoring segment(s) of $\mathbb{A}$ from the maximal entries in the matrix (S_{bi}). If we were to score only pairwise alignments it would be possible to use a local alignment-like algorithm that does not keep track of the beginning of the segment, b. In the multiple alignment, however, the individual pairwise alignments are constrained by the

requirement that a coding segment starts in the same column for all sequences, forcing us to keep track of b explicitly. The algorithm scales as $\mathcal{O}(N \cdot n^2)$ in time and space, where n is the length of the reference sequence and N the number of rows in the alignment.

List of Figures

List of Tables

Bibliography

Aebersold, R. & Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, **422**, 198–207.

Albrecht, M., Sharma, C. M., Reinhardt, R., Vogel, J. & Rudel, T. (2009). Deep sequencing-based discovery of the Chlamydia trachomatis transcriptome. *Nucleic Acids Res*.

Alm, R. A., Ling, L. S., Moir, D. T., King, B. L., Brown, E. D., Doig, P. C., Smith, D. R., Noonan, B., Guild, B. C., deJonge, B. L., Carmel, G., Tummino, P. J., Caruso, A., Uria-Nickelsen, M., Mills, D. M., Ives, C., Gibson, R., Merberg, D., Mills, S. D., Jiang, Q., Taylor, D. E., Vovis, G. F. & Trust, T. J. (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. *Nature*, **397**, 176–180.

Altschul, S. F. & Erickson, B. W. (1985). Significance of nucleotide sequence alignments: a method for random sequence permutation that preserves dinucleotide and codon usage. *Mol Biol Evol*, **2**, 526–538.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, **215**, 403–410.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**, 3389–3402.

Anandam, P., Torarinsson, E. & Ruzzo, W. L. (2009). Multiperm: shuffling multiple sequence alignments while approximately preserving dinucleotide frequencies. *Bioinformatics*, **25**, 668–669.

Andersen, E. S., Rosenblad, M. A., Larsen, N., Westergaard, J. C., Burks, J., Wower, I. K., Wower, J., Gorodkin, J., Samuelsson, T. & Zwieb, C. (2006). The tmRDB and SRPDB resources. *Nucleic Acids Res*, **34**, D163–D168.

Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E. G., Margalit, H. & Altuvia, S. (2001). Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Curr Biol*, **11**, 941–950.

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

Axmann, I. M., Holtzendorff, J., Voss, B., Kensche, P. & Hess, W. R. (2007). Two distinct types of 6S RNA in Prochlorococcus. *Gene*, **406**, 69–78.

Babitzke, P. & Romeo, T. (2007). CsrB sRNA family: sequestration of RNA-binding regulatory proteins. *Curr Opin Microbiol*, **10**, 156–163.

Bailey, T. L. & Elkan, C. (1995). The value of prior knowledge in discovering motifs with MEME. *Proc Int Conf Intell Syst Mol Biol*, **3**, 21–29.

Bailey, T. L. & Gribskov, M. (1996). The megaprior heuristic for discovering protein sequence patterns. *Proc Int Conf Intell Syst Mol Biol*, **4**, 15–24.

Barrick, J. E. & Breaker, R. R. (2007). The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome Biol*, **8**, R239.

Barrick, J. E., Sudarsan, N., Weinberg, Z., Ruzzo, W. L. & Breaker, R. R. (2005). 6S RNA is a widespread regulator of eubacterial RNA polymerase that resembles an open promoter. *RNA*, **11**, 774–784.

Bernhart, S. H. & Hofacker, I. L. (2009). From consensus structure prediction to RNA gene finding. *Brief Funct Genomic Proteomic*, **8**, 461–471.

Bernhart, S. H., Hofacker, I. L. & Stadler, P. F. (2006). Local RNA base pairing probabilities in large sequences. *Bioinformatics*, **22**, 614–615.

Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R. & Stadler, P. F. (2008). RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.

Bindewald, E. & Shapiro, B. A. (2006). RNA secondary structure prediction from sequence alignments using a network of k-nearest neighbor classifiers. *RNA*, **12**, 342–352.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F. A., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., Haussler, D. & Miller, W. (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res*, **14**, 708–715.

Bodey, G. P., Bolivar, R., Fainstein, V. & Jadeja, L. (1983). Infections caused by Pseudomonas aeruginosa. *Rev Infect Dis*, **5**, 279–313.

Boisset, S., Geissmann, T., Huntzinger, E., Fechter, P., Bendridi, N., Possedko, M., Chevalier, C., Helfer, A. C., Benito, Y., Jacquier, A., Gaspin, C., Vandenesch, F. & Romby, P. (2007). Staphylococcus aureus RNAIII coordinately represses the synthesis of virulence factors and the transcription regulator Rot by an antisense mechanism. *Genes Dev*, **21**, 1353–1366.

Bonas, U., Schulte, R., Fenselau, S., Minsavage, G. V., Staskawicz, B. & Stall, R. E. (1991). Isolation of a gene-cluster from Xanthomonas-campestris pv. vesicatoria that determines pathogenicity and the hypersensitive response on pepper and tomato. mol. plant-microbe interact. 4:81-88. *Molecular Plant-Microbe Interactactions*, **4**, 81–88.

Boneca, I. G., de Reuse, H., Epinat, J.-C., Pupin, M., Labigne, A. & Moszer, I. (2003). A revised annotation and comparative analysis of Helicobacter pylori genomes. *Nucleic Acids Res*, **31**, 1704–1714.

Bonnet, E., Wuyts, J., Rouzé, P. & de Peer, Y. V. (2004). Evidence that microRNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences. *Bioinformatics*, **20**, 2911–2917.

Brantl, S. (2002). Antisense RNAs in plasmids: control of replication and maintenance. *Plasmid*, **48**, 165–173.

Brennan, R. G. & Link, T. M. (2007). Hfq structure, function and ligand binding. *Curr Opin Microbiol*, **10**, 125–133.

Brent, M. R. (2008). Steady progress and recent breakthroughs in the accuracy of automated genome annotation. *Nat Rev Genet*, **9**, 62–73.

Brock, J. E., Pourshahian, S., Giliberti, J., Limbach, P. A. & Janssen, G. R. (2008). Ribosomes bind leaderless mRNA in Escherichia coli through recognition of their 5'-terminal AUG. *RNA*, **14**, 2159–2169.

Brown, M. P. (2000). Small subunit ribosomal RNA modeling using stochastic context-free grammars. *Proc Int Conf Intell Syst Mol Biol*, **8**, 57–66.

Brownlee, G. G. (1971). Sequence of 6S RNA of E. coli. *Nat New Biol*, **229**, 147–149. No paper available.

Burge, C. B. & Karlin, S. (1998). Finding the genes in genomic DNA. *Curr Opin Struct Biol*, **8**, 346–354.

Büttner, D. & Bonas, U. (2010). Regulation and secretion of Xanthomonas virulence factors. *FEMS Microbiol Rev*, **34**, 107–133.

Carterson, A. J., Morici, L. A., Jackson, D. W., Frisk, A., Lizewski, S. E., Jupiter, R., Simpson, K., Kunz, D. A., Davis, S. H., Schurr, J. R., Hassett, D. J. & Schurr, M. J. (2004). The transcriptional regulator AlgR controls cyanide production in Pseudomonas aeruginosa. *J Bacteriol*, **186**, 6837–6844.

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

Cho, B.-K., Zengler, K., Qiu, Y., Park, Y. S., Knight, E. M., Barrett, C. L., Gao, Y. & Palsson, B. O. (2009). The transcription unit architecture of the Escherichia coli genome. *Nat Biotechnol*, **27**, 1043–1049.

Clarke, J., Wu, H.-C., Jayasinghe, L., Patel, A., Reid, S. & Bayley, H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol*, **4**, 265–270.

Clote, P., Ferré, F., Kranakis, E. & Krizanc, D. (2005). Structural RNA has lower folding energy than random RNA of the same dinucleotide frequency. *RNA*, **11**, 578–591.

Cochrane, J. C., Lipchock, S. V. & Strobel, S. A. (2007). Structural investigation of the GlmS ribozyme bound to Its catalytic cofactor. *Chem Biol*, **14**, 97–105.

Coppins, R. L., Hall, K. B. & Groisman, E. A. (2007). The intricate world of riboswitches. *Curr Opin Microbiol*, **10**, 176–181.

Cover, T. L. & Blaser, M. J. (2009). Helicobacter pylori in health and disease. *Gastroenterology*, **136**, 1863–1873.

Croucher, N. J. & Thomson, N. R. (2010). Studying bacterial transcriptomes using RNA-seq. *Curr Opin Microbiol*, **13**, 619–624.

Croxen, M. A., Ernst, P. B. & Hoffman, P. S. (2007). Antisense RNA modulation of alkyl hydroperoxide reductase levels in Helicobacter pylori correlates with organic peroxide toxicity but not infectivity. *J Bacteriol*, **189**, 3359–3368.

Deana, A. & Belasco, J. G. (2005). Lost in translation: the influence of ribosomes on bacterial mRNA decay. *Genes Dev*, **19**, 2526–2533.

del Val, C., Rivas, E., Torres-Quesada, O., Toro, N. & Jimnez-Zurdo, J. I. (2007). Identification of differentially expressed small non-coding RNAs in the legume endosymbiont Sinorhizobium meliloti by comparative genomics. *Mol Microbiol*, **66**, 1080–1091.

DiChiara, J. M., Contreras-Martinez, L. M., Livny, J., Smith, D., McDonough, K. A. & Belfort, M. (2010). Multiple small RNAs identified in Mycobacterium bovis BCG are also expressed in Mycobacterium tuberculosis and Mycobacterium smegmatis. *Nucleic Acids Res*, **38**, 4067–4078.

Dinger, M. E., Pang, K. C., Mercer, T. R. & Mattick, J. S. (2008). Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol*, **4**, e1000176.

Douchin, V., Bohn, C. & Bouloc, P. (2006). Down-regulation of porins by a small RNA bypasses the essentiality of the regulated intramembrane proteolysis protease RseP in Escherichia coli. *J Biol Chem*, **281**, 12253–12259.

Dulebohn, D., Choy, J., Sundermeier, T., Okan, N. & Karzai, A. W. (2007). Trans-translation: the tmRNA-mediated surveillance mechanism for ribosome rescue, directed protein degradation, and nonstop mRNA decay. *Biochemistry*, **46**, 4681–4693.

Eddy, S. R. (2004). How do RNA folding algorithms work? *Nat Biotechnol*, **22**, 1457–1458.

Eddy, S. R. & Durbin, R. (1994). Rna sequence analysis using covariance models. *Nucleic Acids Res*, **22**, 2079–2088.

ENCODE Consortium (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Fenselau, S. & Bonas, U. (1995). Sequence and expression analysis of the hrpB pathogenicity operon of Xanthomonas campestris pv. vesicatoria which encodes eight proteins with similarity to components of the Hrp, Ysc, Spa, and Fli secretion systems. *Mol Plant Microbe Interact*, **8**, 845–854.

Findeiß, S., Schmidtke, C., Stadler, P. F. & Bonas, U. (2010). A novel family of plasmid-transferred anti-sense ncRNAs. *RNA Biol*, **7**.

Flamm, C., Hofacker, I. L. & Stadler, P. F. (2004). Computational chemistry with RNA secondary structures.

Flicek, P. (2007). Gene prediction: compare and CONTRAST. *Genome Biol*, **8**, 233.

Fozo, E. M., Hemm, M. R. & Storz, G. (2008a). Small toxic proteins and the antisense RNAs that repress them. *Microbiol Mol Biol Rev*, **72**, 579–89, Table of Contents.

Fozo, E. M., Kawano, M., Fontaine, F., Kaya, Y., Mendieta, K. S., Jones, K. L., Ocampo, A., Rudd, K. E. & Storz, G. (2008b). Repression of small toxic protein synthesis by the Sib and OhsC small RNAs. *Mol Microbiol*, **70**, 1076–1093.

Freyhult, E. K., Bollback, J. P. & Gardner, P. P. (2007). Exploring genomic dark matter: a critical assessment of the performance of homology search methods on noncoding RNA. *Genome Res*, **17**, 117–125.

Gardner, P. P., Daub, J., Tate, J., Moore, B. L., Osuch, I. H., Griffiths-Jones, S., Finn, R. D., Nawrocki, E. P., Kolbe, D. L., Eddy, S. R. & Bateman, A. (2010). Rfam: Wikipedia, clans and the "decimal" release. *Nucleic Acids Res*.

Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., Finn, R. D., Griffiths-Jones, S., Eddy, S. R. & Bateman, A. (2009). Rfam: updates to the RNA families database. *Nucleic Acids Res*, **37**, D136–D140.

Gardner, P. P., Wilm, A. & Washietl, S. (2005). A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res*, **33**, 2433–2439.

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

Gerdes, K. & Wagner, E. G. H. (2007). RNA antitoxins. *Curr Opin Microbiol*, **10**, 117–124.

Gesell, T. & Washietl, S. (2008). Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics*, **9**, 248.

Gilbert, W. (1986). The RNA world. *Nature*, **319**.

Gimpel, M., Heidrich, N., Mäder, U., Krügel, H. & Brantl, S. (2010). A dual-function sRNA from B. subtilis: SR1 acts as a peptide encoding mRNA on the gapA operon. *Mol Microbiol*.

Gonzalez, N., Heeb, S., Valverde, C., Kay, E., Reimmann, C., Junier, T. & Haas, D. (2008). Genome-wide search reveals a novel GacA-regulated small RNA in Pseudomonas species. *BMC Genomics*, **9**, 167.

Görke, B. & Vogel, J. (2008). Noncoding RNA control of the making and breaking of sugars. *Genes Dev*, **22**, 2914–2925.

Gruber, A. R., Bernhart, S. H., Hofacker, I. L. & Washietl, S. (2008). Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, **9**, 122.

Gruber, A. R., Findeiß, S., Washietl, S., Hofacker, I. L. & Stadler, P. F. (2010). RNAZ 2.0: IMPROVED NONCODING RNA DETECTION. *Pac Symp Biocomput*, **15**, 69–79.

Güell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., Rode, M., Suyama, M., Schmidt, S., Gavin, A.-C., Bork, P. & Serrano, L. (2009). Transcriptome complexity in a genome-reduced bacterium. *Science*, **326**, 1268–1271.

Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. (1983). The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, **35**, 849–857.

Guillier, M., Gottesman, S. & Storz, G. (2006). Modulating the outer membrane with small RNAs. *Genes Dev*, **20**, 2338–2348.

Han, K., Kim, K.-S., Bak, G., Park, H. & Lee, Y. (2010). Recognition and discrimination of target mRNAs by Sib RNAs, a cis-encoded sRNA family. *Nucleic Acids Res*.

Hancock, R. E. (1998). Resistance mechanisms in Pseudomonas aeruginosa and other nonfermentative gram-negative bacteria. *Clin Infect Dis*, **27 Suppl 1**, S93–S99.

Hardalo, C. & Edberg, S. C. (1997). Pseudomonas aeruginosa: assessment of risk from drinking water. *Crit Rev Microbiol*, **23**, 47–75.

Hasegawa, M., Kishino, H. & Yano, T. (1985). Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*, **22**, 160–174.

Havgaard, J. H., Lyngsø, R. B., Stormo, G. D. & Gorodkin, J. (2005). Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics*, **21**, 1815–1824.

Heidrich, N., Chinali, A., Gerth, U. & Brantl, S. (2006). The small untranslated RNA SR1 from the Bacillus subtilis genome is involved in the regulation of arginine catabolism. *Mol Microbiol*, **62**, 520–536.

Heidrich, N., Moll, I. & Brantl, S. (2007). In vitro analysis of the interaction between the small RNA SR1 and its primary target ahrC mRNA. *Nucleic Acids Res*, **35**, 4331–4346.

Hemm, M. R., Paul, B. J., Schneider, T. D., Storz, G. & Rudd, K. E. (2008). Small membrane proteins found by comparative genomics and ribosome binding site models. *Mol Microbiol*, **70**, 1487–1501.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89**, 10915–10919.

Hertel, J., de Jong, D., Marz, M., Rose, D., Tafer, H., Tanzer, A., Schierwater, B. & Stadler, P. F. (2009). Non-coding RNA annotation of the genome of Trichoplax adhaerens. *Nucleic Acids Res*.

Hertel, J., Hofacker, I. L. & Stadler, P. F. (2008). SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, **24**, 158–164.

Hertel, J. & Stadler, P. F. (2006). Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics*, **22**, e197–e202.

Heurlier, K., Williams, F., Heeb, S., Dormond, C., Pessi, G., Singer, D., Cmara, M., Williams, P. & Haas, D. (2004). Positive control of swarming, rhamnolipid synthesis, and lipase production by the posttranscriptional RsmA/RsmZ system in Pseudomonas aeruginosa PAO1. *J Bacteriol*, **186**, 2936–2945.

Hindley, J. (1967). Fractionation of 32P-labelled ribonucleic acids on polyacrylamide gels and their characterization by fingerprinting. *J Mol Biol*, **30**, 125–136. No paper available.

Hofacker, I. L. (2007). RNA consensus structure prediction with RNAalifold. *Methods Mol Biol*, **395**, 527–544.

Hofacker, I. L. (2009). RNA secondary structure analysis using the Vienna RNA package. *Curr Protoc Bioinformatics*, **Chapter 12**, Unit12.2.

Hofacker, I. L., Bernhart, S. H. F. & Stadler, P. F. (2004). Alignment of RNA base pairing probability matrices. *Bioinformatics*, **20**, 2222–2227.

Hofacker, I. L., Fekete, M. & Stadler, P. F. (2002). Secondary structure prediction for aligned RNA sequences. *J Mol Biol*, **319**, 1059–1066.

Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S. L., Tacker, M. & Schuster, P. (1994). Fast Folding and Comparison of RNA Secondary Structures. *Monatsh. Chem.*, **125**, 167–188.

Hofacker, I. L. & Stadler, P. F. (2007). Bioinformatics: From Genomes to Therapies. T Lengauer(Ed.), Wiley-VCH, Weinheim, Germany.

Hoffmann, S., Otto, C., Kurtz, S., Sharma, C. M., Khaitovich, P., Vogel, J., Stadler, P. F. & Hackermüller, J. (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput Biol*, **5**, e1000502.

Horler, R. S. P. & Vanderpool, C. K. (2009). Homologs of the small RNA SgrS are broadly distributed in enteric bacteria but have diverged in size and sequence. *Nucleic Acids Res*, **37**, 5465–5476.

Hsu, L. M., Zagorski, J., Wang, Z. & Fournier, M. J. (1985). Escherichia coli 6S RNA gene is part of a dual-function transcription unit. *J Bacteriol*, **161**, 1162–1170. Not printed.

Hüttenhofer, A., Brosius, J. & Bachellerie, J. P. (2002). RNomics: identification and function of small, non-messenger RNAs. *Curr Opin Chem Biol*, **6**, 835–843.

Hüttenhofer, A. & Vogel, J. (2006). Experimental approaches to identify non-coding RNAs. *Nucleic Acids Res*, **34**, 635–646.

Hwang, Y., Kim, J.-S. & Kweon, I.-S. (2007). Sensor noise modelling using the Skellam distribution: Application to the color edge detection. *Computer Vision and Pattern Recognition*, 1–8.

Jäger, D., Sharma, C. M., Thomsen, J., Ehlers, C., Vogel, J. & Schmitz, R. A. (2009). Deep sequencing analysis of the Methanosarcina mazei gö1 transcriptome in response to nitrogen availability. *Proc Natl Acad Sci U S A*, **106**, 21878–21882.

Johansson, J., Mandin, P., Renzoni, A., Chiaruttini, C., Springer, M. & Cossart, P. (2002). An RNA thermosensor controls expression of virulence genes in Listeria monocytogenes. *Cell*, **110**, 551–561.

Jones, J. B., Stall, R. E. & Bouzar, H. (1998). Diversity among xanthomonads pathogenic on pepper and tomato. *Annu Rev Phytopathol*, **36**, 41–58.

Jühling, F., Mörl, M., Hartmann, R. K., Sprinzl, M., Stadler, P. F. & Pütz, J. (2009). `tRNAdb` 2009: compilation of tRNA sequences and tRNA genes. *Nucleic Acids Res.*, **37**, D159–D162.

Karlis, D. & Ntzoufras, I. (2009). Bayesian modelling of football outcomes: Using the Skellams distribution for the goal difference. *IMA Journal of Management Mathematics*, **20**, 133–145.

Kellis, M., Patterson, N., Birren, B., Berger, B. & Lander, E. S. (2004). Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *J Comput Biol*, **11**, 319–355.

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., Harmin, D. A., Laptewicz, M., Barbara-Haley, K., Kuersten, S., Markenscoff-Papadimitriou, E., Kuhl, D.,

Bito, H., Worley, P. F., Kreiman, G. & Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**, 182–187.

Klein, C., Aivaliotis, M., Olsen, J. V., Falb, M., Besir, H., Scheffer, B., Bisle, B., Tebbe, A., Konstantinidis, K., Siedler, F., Pfeiffer, F., Mann, M. & Oesterhelt, D. (2007). The low molecular weight proteome of Halobacterium salinarum. *J Proteome Res*, **6**, 1510–1518.

Klein, D. J. & Ferré-D'Amaré, A. R. (2006). Structural basis of glmS ribozyme activation by glucosamine-6-phosphate. *Science*, **313**, 1752–1756.

Klein, R. J., Misulovin, Z. & Eddy, S. R. (2002). Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc Natl Acad Sci U S A*, **99**, 7542–7547.

Knight, R., Birmingham, A. & Yarus, M. (2004). BayesFold: rational 2 degrees folds that combine thermodynamic, covariation, and chemical data for aligned RNA sequences. *RNA*, **10**, 1323–1336.

Knudsen, B. & Hein, J. (2003). Pfold: RNA secondary structure prediction using stochastic context-free grammars. *Nucleic Acids Res*, **31**, 3423–3428.

Koebnik, R., Krger, A., Thieme, F., Urban, A. & Bonas, U. (2006). Specific binding of the Xanthomonas campestris pv. vesicatoria AraC-type transcriptional activator HrpX to plant-inducible promoter boxes. *J Bacteriol*, **188**, 7652–7660.

Kuhn, R. M., Karolchik, D., Zweig, A. S., Wang, T., Smith, K. E., Rosenbloom, K. R., Rhead, B., Raney, B. J., Pohl, A., Pheasant, M., Meyer, L., Hsu, F., Hinrichs, A. S., Harte, R. A., Giardine, B., Fujita, P., Diekhans, M., Dreszer, T., Clawson, H., Barber, G. P., Haussler, D. & Kent, W. J. (2009). The UCSC Genome Browser Database: update 2009. *Nucleic Acids Res*, **37**, D755–D761.

Kyte, J. & Doolittle, R. F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol*, **157**, 105–132.

Lapouge, K., Schubert, M., Allain, F. H.-T. & Haas, D. (2008). Gac/Rsm signal transduction pathway of gamma-proteobacteria: from RNA recognition to regulation of social behaviour. *Mol Microbiol*, **67**, 241–253.

Laslett, D. & Canback, B. (2004). ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res*, **32**, 11–16.

Laslett, D., Canback, B. & Andersson, S. (2002). BRUCE: a program for the detection of transfer-messenger RNA genes in nucleotide sequences. *Nucleic Acids Res*, **30**, 3449–3453.

Le, S. V., Chen, J. H., Currey, K. M. & Maizel, J. V. (1988). A program for predicting significant RNA secondary structures. *Comput Appl Biosci*, **4**, 153–159.

Lee, T. & Feig, A. L. (2008). The RNA binding protein Hfq interacts specifically with tRNAs. *RNA*, **14**, 514–523.

Lenz, D. H., Mok, K. C., Lilley, B. N., Kulkarni, R. V., Wingreen, N. S. & Bassler, B. L. (2004). The small RNA chaperone Hfq and multiple small RNAs control quorum sensing in Vibrio harveyi and Vibrio cholerae. *Cell*, **118**, 69–82.

Licht, A., Preis, S. & Brantl, S. (2005). Implication of CcpN in the regulation of a novel untranslated RNA (SR1) in Bacillus subtilis. *Mol Microbiol*, **58**, 189–206.

Livny, J., Brencic, A., Lory, S. & Waldor, M. K. (2006). Identification of 17 Pseudomonas aeruginosa sRNAs and prediction of sRNA-encoding genes in 10 diverse pathogens using the bioinformatic tool sRNAPredict2. *Nucleic Acids Res*, **34**, 3484–3493.

Livny, J., Teonadi, H., Livny, M. & Waldor, M. K. (2008). High-throughput, kingdom-wide prediction and annotation of bacterial non-coding rnas. *PLoS One*, **3**, e3197.

Loh, E., Dussurget, O., Gripenland, J., Vaitkevicius, K., Tiensuu, T., Mandin, P., Repoila, F., Buchrieser, C., Cossart, P. & Johansson, J. (2009). A trans-acting riboswitch controls expression of the virulence regulator PrfA in Listeria monocytogenes. *Cell*, **139**, 770–779.

Lowe, T. M. & Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, **25**, 955–964.

Machado-Lima, A., del Portillo, H. A. & Durham, A. M. (2008). Computational methods in noncoding RNA research. *J Math Biol*, **56**, 15–49.

Maeda, H., Fujita, N. & Ishihama, A. (2000). Competition among seven Escherichia coli sigma subunits: relative binding affinities to the core RNA polymerase. *Nucleic Acids Res*, **28**, 3497–3503. Not printed.

Mamanova, L., Andrews, R. M., James, K. D., Sheridan, E. M., Ellis, P. D., Langford, C. F., Ost, T. W. B., Collins, J. E. & Turner, D. J. (2010). FRT-seq: amplification-free, strand-specific transcriptome sequencing. *Nat Methods*, **7**, 130–132.

Marchais, A., Naville, M., Bohn, C., Bouloc, P. & Gautheret, D. (2009). Single-pass classification of all noncoding sequences in a bacterial genome using phylogenetic profiles. *Genome Res*, **19**, 1084–1092.

Massé, E., Salvail, H., Desnoyers, G. & Arguin, M. (2007). Small RNAs controlling iron metabolism. *Curr Opin Microbiol*, **10**, 140–145.

Mathews, D. H., Disney, M. D., Childs, J. L., Schroeder, S. J., Zuker, M. & Turner, D. H. (2004). Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc Natl Acad Sci U S A*, **101**, 7287–7292.

Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol*, **288**, 911–940.

Mathews, D. H. & Turner, D. H. (2002). Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J Mol Biol*, **317**, 191–203.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers*, **29**, 1105–1119.

McGuire, A. M. & Galagan, J. E. (2008). Conserved secondary structures in Aspergillus. *PLoS One*, **3**, e2812.

Meyer, I. M. & Miklós, I. (2005). Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Res*, **33**, 6338–6348.

Meyer, M. M., Ames, T. D., Smith, D. P., Weinberg, Z., Schwalbach, M. S., Giovannoni, S. J. & Breaker, R. R. (2009). Identification of candidate structured RNAs in the marine organism Candidatus Pelagibacter ubique. *BMC Genomics*, **10**, 268.

Miller, J. H. (1972). Experiments in Molecular Genetics. Cold Spring Harbor Laboratory Press.

Missal, K., Rose, D. & Stadler, P. F. (2005). Non-coding RNAs in Ciona intestinalis. *Bioinformatics*, **21 Suppl 2**, ii77–ii78.

Missal, K., Zhu, X., Rose, D., Deng, W., Skogerbo, G., Chen, R. & Stadler, P. F. (2006). Prediction of structured non-coding RNAs in the genomes of the nematodes Caenorhabditis elegans and Caenorhabditis briggsae. *J Exp Zool B Mol Dev Evol*, **306**, 379–392.

Mitschke, J., Georg, J., Scholz, I., Sharma, C. M., Dienst, D., Bantscheff, J., Vo, B., Steglich, C., Wilde, A., Vogel, J. & Hess, W. R. (2011). An experimentally anchored map of transcriptional start sites in the model cyanobacterium Synechocystis sp. PCC6803. *Proc Natl Acad Sci U S A*.

Moll, I., Leitsch, D., Steinhauser, T. & Bläsi, U. (2003). RNA chaperone activity of the Sm-like Hfq protein. *EMBO Rep*, **4**, 284–289.

Morici, L. A., Carterson, A. J., Wagner, V. E., Frisk, A., Schurr, J. R., zu Bentrup, K. H., Hassett, D. J., Iglewski, B. H., Sauer, K. & Schurr, M. J. (2007). Pseudomonas aeruginosa AlgR represses the Rhl quorum-sensing system in a biofilm-specific manner. *J Bacteriol*, **189**, 7752–7764.

Mosig, A., Zhu, L. & Stadler, P. F. (2009). Customized strategies for discovering distant ncRNA homologs. *Brief Funct Genomic Proteomic*, **8**, 451–460.

Mourier, T., Carret, C., Kyes, S., Christodoulou, Z., Gardner, P. P., Jeffares, D. C., Pinches, R., Barrell, B., Berriman, M., Griffiths-Jones, S., Ivens, A., Newbold, C. & Pain, A. (2008). Genome-wide discovery and verification of novel structured RNAs in Plasmodium falciparum. *Genome Res*, **18**, 281–292.

Mückstein, U., Tafer, H., Hackermüller, J., Bernhart, S. H., Stadler, P. F. & Hofacker, I. L. (2006). Thermodynamics of RNA-RNA binding. *Bioinformatics*, **22**, 1177–1182.

Müller, S. A., Kohajda, T., Findeiß, S., Stadler, P. F., Washietl, S., Kellis, M., von Bergen, M. & Kalkhof, S. (2010). Optimization of parameters for coverage of low molecular weight proteins. *Anal Bioanal Chem*, **398**, 2867–2881.

Nakagawa, S., Niimura, Y., ichiro Miura, K. & Gojobori, T. (2010). Dynamic evolution of translation initiation mechanisms in prokaryotes. *Proc Natl Acad Sci U S A*, **107**, 6382–6387.

Narberhaus, F. (2010). Translational control of bacterial heat shock and virulence genes by temperature-sensing mRNAs. *RNA Biol*, **7**, 84–89.

Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. (2009). Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.

Nielsen, J. S., Bøggild, A., Andersen, C. B. F., Nielsen, G., Boysen, A., Brodersen, D. E. & Valentin-Hansen, P. (2007). An Hfq-like protein in archaea: crystal structure and functional characterization of the Sm protein from Methanococcus jannaschii. *RNA*, **13**, 2213–2223.

Nugent, T. & Jones, D. T. (2009). Transmembrane protein topology prediction using support vector machines. *BMC Bioinformatics*, **10**, 159.

Nussinov, R., Piecznik, G., Griggs, J. R. & Kleitman, D. J. (1978). Algorithms for Loop Matching. *SIAM J Appl Math*, **35**, 68–82.

Opdyke, J. A., Kang, J.-G. & Storz, G. (2004). GadY, a small-RNA regulator of acid response genes in Escherichia coli. *J Bacteriol*, **186**, 6698–6705.

Otto, W., Will, S. & Backofen, R. (2008). Structure local multiple alignment of RNA. In *Proceedings of the German Conference on Bioinformatics (CGB 08)*, volume P-136 of *LNI*. GI, pp. 178–188.

Ozsolak, F., Platt, A. R., Jones, D. R., Reifenberger, J. G., Sass, L. E., McInerney, P., Thompson, J. F., Bowers, J., Jarosz, M. & Milos, P. M. (2009). Direct RNA sequencing. *Nature*, **461**, 814–818.

Pedersen, J. S., Meyer, I. M., Forsberg, R., Simmonds, P. & Hein, J. (2004). A comparative method for finding and folding RNA secondary structures within protein-coding regions. *Nucleic Acids Res*, **32**, 4925–4936.

Pessi, G., Williams, F., Hindle, Z., Heurlier, K., Holden, M. T., Cmara, M., Haas, D. & Williams, P. (2001). The global posttranscriptional regulator RsmA modulates production of virulence determinants and N-acylhomoserine lactones in Pseudomonas aeruginosa. *J Bacteriol*, **183**, 6676–6683.

Petersen, L., Larsen, T. S., Ussery, D. W., On, S. L. W. & Krogh, A. (2003). RpoD promoters in Campylobacter jejuni exhibit a strong periodic signal instead of a -35 box. *J Mol Biol*, **326**, 1361–1372.

Pichon, C. & Felden, B. (2007). Proteins that interact with bacterial small RNA regulators. *FEMS Microbiol Rev*, **31**, 614–625.

Pichon, C. & Felden, B. (2008). Small RNA gene identification and mRNA target predictions in Bacteria. *Bioinformatics*.

Rain, J. C., Selig, L., Reuse, H. D., Battaglia, V., Reverdy, C., Simon, S., Lenzen, G., Petel, F., Wojcik, J., Schächter, V., Chemama, Y., Labigne, A. & Legrain, P. (2001). The protein-protein interaction map of Helicobacter pylori. *Nature*, **409**, 211–215.

Ramakrishnan, V. (2002). Ribosome structure and the mechanism of translation. *Cell*, **108**, 557–572.

Redko, Y., de Lasierra-Gallay, I. L. & Condon, C. (2007). When all's zed and done: the structure and function of RNase Z in prokaryotes. *Nat Rev Microbiol*, **5**, 278–286.

Reiche, K. & Stadler, P. F. (2007). RNAstrand: reading direction of structured RNAs in multiple sequence alignments. *Algorithms Mol Biol*, **2**, 6.

Repoila, F. & Darfeuille, F. (2009). Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects. *Biol Cell*, **101**, 117–131.

Repoila, F., Majdalani, N. & Gottesman, S. (2003). Small non-coding RNAs, co-ordinators of adaptation processes in Escherichia coli: the RpoS paradigm. *Mol Microbiol*, **48**, 855–861.

Richards, J., Sundermeier, T., Svetlanov, A. & Karzai, A. W. (2008). Quality control of bacterial mRNA decoding and decay. *Biochim Biophys Acta*, **1779**, 574–582.

Rivas, E. (2005). Evolutionary models for insertions and deletions in a probabilistic modeling framework. *BMC Bioinformatics*, **6**, 63.

Rivas, E. & Eddy, S. R. (2000). Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics*, **16**, 583–605.

Rivas, E. & Eddy, S. R. (2001). Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.

Robinson, M. D. & Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*, **11**, R25.

Rodionov, D. A., Vitreschak, A. G., Mironov, A. A. & Gelfand, M. S. (2002). Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. *J Biol Chem*, **277**, 48949–48959. Not printed.

Rose, D., Hackermüller, J., Washietl, S., Reiche, K., Hertel, J., Findeiß, S., Stadler, P. F. & Prohaska, S. J. (2007). Computational RNomics of drosophilids. *BMC Genomics*, **8**, 406.

Rose, D., Hertel, J., Reiche, K., Stadler, P. F. & Hackermüller, J. (2008a). NcDNAlign: plausible multiple alignments of non-protein-coding genomic sequences. *Genomics*, **92**, 65–74.

Rose, D., Jöris, J., Hackermüller, J., Reiche, K., Li, Q. & Stadler, P. F. (2008b). Duplicated RNA genes in teleost fish genomes. *J Bioinform Comput Biol*, **6**, 1157–1175.

Sakakibara, Y., Brown, M., Hughey, R., Mian, I. S., Sjlander, K., Underwood, R. C. & Haussler, D. (1994). Stochastic context-free grammars for tRNA modeling. *Nucleic Acids Res*, **22**, 5112–5120.

Sandmann, T. & Cohen, S. M. (2007). Identification of novel Drosophila melanogaster microRNAs. *PLoS One*, **2**, e1265.

Sankoff, D. (1985). Simultaneous solution of the RNA folding, alignment, and proto-sequence problems. *SIAM J Appl Math*, **45**, 810–825.

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

Schneider, K. L., Pollard, K. S., Baertsch, R., Pohl, A. & Lowe, T. M. (2006). The UCSC Archaeal Genome Browser. *Nucleic Acids Res*, **34**, D407–D410.

Sharma, C. M., Darfeuille, F., Plantinga, T. H. & Vogel, J. (2007). A small RNA regulates multiple ABC transporter mRNAs by targeting C/A-rich elements inside and upstream of ribosome-binding sites. *Genes Dev*, **21**, 2804–2817.

Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiß, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F. & Vogel, J. (2010). The primary transcriptome of the major human pathogen Helicobacter pylori. *Nature*, **464**, 250–255.

Silvaggi, J. M., Perkins, J. B. & Losick, R. (2005). Small untranslated RNA antitoxin in Bacillus subtilis. *J Bacteriol*, **187**, 6641–6650.

Skellam, J. G. (1946). The frequency distribution of the difference between two Poisson variates belonging to different populations. *J R Stat Soc Ser A*, **109**, 296.

Song, T., Mika, F., Lindmark, B., Liu, Z., Schild, S., Bishop, A., Zhu, J., Camilli, A., Johansson, J., Vogel, J. & Wai, S. N. (2008). A new Vibrio cholerae sRNA modulates colonization and affects release of outer membrane vesicles. *Mol Microbiol*.

Sonnleitner, E., Schuster, M., Sorger-Domenigg, T., Greenberg, E. P. & Bläsi, U. (2006). Hfq-dependent alterations of the transcriptome profile and effects on quorum sensing in Pseudomonas aeruginosa. *Mol Microbiol*, **59**, 1542–1558.

Sonnleitner, E., Sorger-Domenigg, T., Madej, M. J., Findeiß, S., Hackermüller, J., Hüttenhofer, A., Stadler, P. F., Bläsi, U. & Moll, I. (2008). Detection of small RNAs in Pseudomonas aeruginosa by RNomics and structure-based bioinformatic tools. *Microbiology*, **154**, 3175–3187.

Sorek, R. & Cossart, P. (2010). Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet*, **11**, 9–16.

Spohn, G., Beier, D., Rappuoli, R. & Scarlato, V. (1997). Transcriptional analysis of the divergent cagAB genes encoded by the pathogenicity island of Helicobacter pylori. *Mol Microbiol*, **26**, 361–372.

Sridhar, J., Narmada, S. R., Sabarinathan, R., Ou, H.-Y., Deng, Z., Sekar, K., Rafi, Z. A. & Rajakumar, K. (2010). sRNAscanner: A Computational Tool for Intergenic Small RNA Detection in Bacterial Genomes. *PLoS One*, **5**.

Storz, G., Opdyke, J. A. & Zhang, A. (2004). Controlling mRNA stability and translation with small, noncoding RNAs. *Curr Opin Microbiol*, **7**, 140–144.

Stoughton, R. B. (2005). Applications of DNA microarrays in biology. *Annu Rev Biochem*, **74**, 53–82.

Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., Brinkman, F. S., Hufnagle, W. O., Kowalik, D. J., Lagrou, M., Garber, R. L., Goltry, L., Tolentino, E., Westbrock-Wadman, S., Yuan, Y., Brody, L. L., Coulter, S. N., Folger, K. R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G. K., Wu, Z., Paulsen, I. T., Reizer, J., Saier, M. H., Hancock, R. E., Lory, S. & Olson, M. V. (2000). Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature*, **406**, 959–964.

sun Kim, K. & Lee, Y. (2004). Regulation of 6S RNA biogenesis by switching utilization of both sigma factors and endoribonucleases. *Nucleic Acids Res*, **32**, 6057–6068.

Sundermeier, T., Ge, Z., Richards, J., Dulebohn, D. & Karzai, A. W. (2008). Studying tmRNA-mediated surveillance and nonstop mRNA decay. *Methods Enzymol*, **447**, 329–358.

Sutherland, J. D. (2010). Ribonucleotides. *Cold Spring Harb Perspect Biol*, **2**, a005439.

Svenningsen, S. L., Tu, K. C. & Bassler, B. L. (2009). Gene dosage compensation calibrates four regulatory RNAs to control Vibrio cholerae quorum sensing. *EMBO J*, **28**, 429–439.

Szczesny, R., Büttner, D., Escolar, L., Schulze, S., Seiferth, A. & Bonas, U. (2010). Suppression of the AvrBs1-specific hypersensitive response by the YopJ effector homolog AvrBsT from Xanthomonas depends on a SNF1-related kinase. *New Phytol*, **187**, 1058–1074.

Tang, F., Barbacioru, C., Nordman, E., Li, B., Xu, N., Bashkirov, V. I., Lao, K. & Surani, M. A. (2010). RNA-Seq analysis to capture the transcriptome landscape of a single cell. *Nat Protoc*, **5**, 516–535.

Thieme, F., Koebnik, R., Bekel, T., Berger, C., Boch, J., Büttner, D., Caldana, C., Gaigalat, L., Goesmann, A., Kay, S., Kirchner, O., Lanz, C., Linke, B., McHardy, A. C., Meyer, F., Mittenhuber, G., Nies, D. H., Niesbach-Klösgen, U., Patschkowski, T., Rückert, C.,

Rupp, O., Schneiker, S., Schuster, S. C., Vorhölter, F.-J., Weber, E., Pühler, A., Bonas, U., Bartels, D. & Kaiser, O. (2005). Insights into genome plasticity and pathogenicity of the plant pathogenic bacterium Xanthomonas campestris pv. vesicatoria revealed by the complete genome sequence. *J Bacteriol*, **187**, 7254–7266.

Thompson, J. D., Higgins, D. G. & Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22**, 4673–4680.

Timmermans, J. & Melderen, L. V. (2010). Post-transcriptional global regulation by CsrA in bacteria. *Cell Mol Life Sci*, **67**, 2897–2908.

Tinoco, I., Uhlenbeck, O. C. & Levine, M. D. (1971). Estimation of secondary structure in ribonucleic acids. *Nature*, **230**, 362–367.

Tjaden, B., Goodwin, S. S., Opdyke, J. A., Guillier, M., Fu, D. X., Gottesman, S. & Storz, G. (2006). Target prediction for small, noncoding RNAs in bacteria. *Nucleic Acids Res*, **34**, 2791–2802.

Tjaden, B., Saxena, R. M., Stolyar, S., Haynor, D. R., Kolker, E. & Rosenow, C. (2002). Transcriptome analysis of Escherichia coli using high-density oligonucleotide probe arrays. *Nucleic Acids Res*, **30**, 3732–3738.

Toledo-Arana, A., Dussurget, O., Nikitas, G., Sesto, N., Guet-Revillet, H., Balestrino, D., Loh, E., Gripenland, J., Tiensuu, T., Vaitkevicius, K., Barthelemy, M., Vergassola, M., Nahori, M.-A., Soubigou, G., Regnault, B., Coppee, J.-Y., Lecuit, M., Johansson, J. & Cossart, P. (2009). The Listeria transcriptional landscape from saprophytism to virulence. *Nature*, **459**, 950–956.

Toledo-Arana, A., Repoila, F. & Cossart, P. (2007). Small noncoding RNAs controlling pathogenesis. *Curr Opin Microbiol*, **10**, 182–188.

Tomb, J. F., White, O., Kerlavage, A. R., Clayton, R. A., Sutton, G. G., Fleischmann, R. D., Ketchum, K. A., Klenk, H. P., Gill, S., Dougherty, B. A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E. F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H. G., Glodek, A., McKenney, K., Fitzegerald, L. M., Lee, N., Adams, M. D., Hickey, E. K., Berg, D. E., Gocayne, J. D., Utterback, T. R., Peterson, J. D., Kelley, J. M., Cotton, M. D., Weidman, J. M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W. S., Borodovsky, M., Karp, P. D., Smith, H. O., Fraser, C. M. & Venter, J. C. (1997). The complete genome sequence of the gastric pathogen Helicobacter pylori. *Nature*, **388**, 539–547.

Tramonti, A., Canio, M. D. & Biase, D. D. (2008). GadX/GadW-dependent regulation of the Escherichia coli acid fitness island: transcriptional control at the gadY-gadW divergent promoters and identification of four novel 42 bp GadX/GadW-specific binding sites. *Mol Microbiol*, **70**, 965–982.

UniProt Consortium (2010). The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res*, **38**, D142–D148.

Urban, J. H. & Vogel, J. (2008). Two seemingly homologous noncoding rnas act hierarchically to activate glms mrna translation. *PLoS Biol*, **6**, e64.

Urbanowski, M. L., Stauffer, L. T. & Stauffer, G. V. (2000). The gcvB gene encodes a small untranslated RNA involved in expression of the dipeptide and oligopeptide transport systems in Escherichia coli. *Mol Microbiol*, **37**, 856–868.

Uzilov, A. V., Keegan, J. M. & Mathews, D. H. (2006). Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, **7**, 173.

Valentin-Hansen, P., Eriksen, M. & Udesen, C. (2004). The bacterial Sm-like protein Hfq: a key player in RNA transactions. *Mol Microbiol*, **51**, 1525–1533.

Valle, M., Gillet, R., Kaur, S., Henne, A., Ramakrishnan, V. & Frank, J. (2003). Visualizing tmRNA entry into a stalled ribosome. *Science*, **300**, 127–130.

Valverde, C., Heeb, S., Keel, C. & Haas, D. (2003). RsmY, a small regulatory RNA, is required in concert with RsmZ for GacA-dependent expression of biocontrol traits in Pseudomonas fluorescens CHA0. *Mol Microbiol*, **50**, 1361–1379.

Valverde, C., Lindell, M., Wagner, E. G. H. & Haas, D. (2004). A repeated GGA motif is critical for the activity and stability of the riboregulator RsmY of Pseudomonas fluorescens. *J Biol Chem*, **279**, 25066–25074.

Vanet, A., Marsan, L., Labigne, A. & Sagot, M. F. (2000). Inferring regulatory elements from a whole genome. An analysis of Helicobacter pylori sigma(80) family of promoter signals. *J Mol Biol*, **297**, 335–353.

Varadarajan, A., Bradley, R. K. & Holmes, I. H. (2008). Tools for simulating evolution of aligned genomic regions with integrated parameter estimation. *Genome Biol*, **9**, R147.

Vecerek, B., Moll, I. & Bläsi, U. (2005). Translational autocontrol of the Escherichia coli hfq RNA chaperone gene. *RNA*, **11**, 976–984.

Viegas, S. C. & Arraiano, C. M. (2008). Regulating the regulators: How ribonucleases dictate the rules in the control of small non-coding RNAs. *RNA Biol*, **5**.

Vogel, J. & Sharma, C. M. (2005). How to find small non-coding RNAs in bacteria. *Biol Chem*, **386**, 1219–1238.

Voss, B., Gierga, G., Axmann, I. M. & Hess, W. R. (2007). A motif-based search in bacterial genomes identifies the ortholog of the small RNA Yfr1 in all lineages of cyanobacteria. *BMC Genomics*, **8**, 375.

Wadler, C. S. & Vanderpool, C. K. (2007). A dual function for a bacterial small RNA: SgrS performs base pairing-dependent regulation and encodes a functional polypeptide. *Proc Natl Acad Sci U S A*, **104**, 20454–20459.

Waldminghaus, T., Gaubig, L., Klinkert, B. & Narberhaus, F. (2009). The Escherichia coli ibpA thermometer is comprised of stable and unstable structural elements. *RNA Biol*, **6**.

Wang, L., Feng, Z., Wang, X., Wang, X. & Zhang, X. (2010). DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, **26**, 136–138.

Wang, Z., Gerstein, M. & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, **10**, 57–63.

Washietl, S. (2007). Prediction of structural noncoding RNAs with RNAz. *Methods Mol Biol*, **395**, 503–526.

Washietl, S., Findeiß, S., Müller, S., Kalkhof, S., von Bergen, M., Hofacker, I. L., Stadler, P. F. & Goldman, N. (2011). RNAcode: robust discrimination of coding and noncoding RNAs in comparative sequence data. *RNA*. Accepted.

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

Washietl, S., Hofacker, I. L. & Stadler, P. F. (2005). Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A*, **102**, 2454–2459.

Washietl, S., Pedersen, J. S., Korbel, J. O., Stocsits, C., Gruber, A. R., Hackermüller, J., Hertel, J., Lindemeyer, M., Reiche, K., Tanzer, A., Ucla, C., Wyss, C., Antonarakis, S. E., Denoeud, F., Lagarde, J., Drenkow, J., Kapranov, P., Gingeras, T. R., Guigó, R., Snyder, M., Gerstein, M. B., Reymond, A., Hofacker, I. L. & Stadler, P. F. (2007). Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res*, **17**, 852–864.

Wassarman, K. M. (2007). 6S RNA: a small RNA regulator of transcription. *Curr Opin Microbiol*, **10**, 164–168.

Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G. & Gottesman, S. (2001). Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev*, **15**, 1637–1651.

Wassarman, K. M. & Saecker, R. M. (2006). Synthesis-mediated release of a small RNA inhibitor of RNA polymerase. *Science*, **314**, 1601–1603.

Wassarman, K. M. & Storz, G. (2000). 6S RNA regulates E. coli RNA polymerase activity. *Cell*, **101**, 613–623. Not printed.

Waters, L. S. & Storz, G. (2009). Regulatory RNAs in bacteria. *Cell*, **136**, 615–628.

Weile, C., Gardner, P. P., Hedegaard, M. M. & Vinther, J. (2007). Use of tiling array data and RNA secondary structure predictions to identify noncoding RNA genes. *BMC Genomics*, **8**, 244.

Weinberg, Z., Barrick, J. E., Yao, Z., Roth, A., Kim, J. N., Gore, J., Wang, J. X., Lee, E. R., Block, K. F., Sudarsan, N., Neph, S., Tompa, M., Ruzzo, W. L. & Breaker, R. R. (2007). Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Res*, **35**, 4809–4819.

Whitchurch, C. B., Alm, R. A. & Mattick, J. S. (1996). The alginate regulator AlgR and an associated sensor FimS are required for twitching motility in Pseudomonas aeruginosa. *Proc Natl Acad Sci U S A*, **93**, 9839–9843.

Whitchurch, C. B., Erova, T. E., Emery, J. A., Sargent, J. L., Harris, J. M., Semmler, A. B. T., Young, M. D., Mattick, J. S. & Wozniak, D. J. (2002). Phosphorylation of the Pseudomonas aeruginosa response regulator AlgR is essential for type IV fimbria-mediated twitching motility. *J Bacteriol*, **184**, 4544–4554.

White, F. F., Potnis, N., Jones, J. B. & Koebnik, R. (2009). The type iii effectors of Xanthomonas. *Mol Plant Pathol*, **10**, 749–766.

Wilderman, P. J., Sowa, N. A., FitzGerald, D. J., FitzGerald, P. C., Gottesman, S., Ochsner, U. A. & Vasil, M. L. (2004). Identification of tandem duplicate regulatory small RNAs in Pseudomonas aeruginosa involved in iron homeostasis. *Proc Natl Acad Sci U S A*, **101**, 9792–9797.

Will, S., Joshi, T., Hofacker, I. L., Stadler, P. F. & Backofen, R. (2011). LocARNA-P: Prediction of Accurate ncRNA-Boundaries using Reliability Pro
les of Sequence-Structure Alignment. Submitted.

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

Wilson, S. A. & Drew, R. E. (1995). Transcriptional analysis of the amidase operon from Pseudomonas aeruginosa. *J Bacteriol*, **177**, 3052–3057.

Workman, C. & Krogh, A. (1999). No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res*, **27**, 4816–4822.

Wurtzel, O., Sapra, R., Chen, F., Zhu, Y., Simmons, B. A. & Sorek, R. (2009). A single-base resolution map of an archaeal transcriptome. *Genome Res*.

Xu, X., Ji, Y. & Stormo, G. D. (2007). RNA Sampler: a new sampling based algorithm for common RNA secondary structure prediction and structural alignment. *Bioinformatics*, **23**, 1883–1891.

Xu, X., Ji, Y. & Stormo, G. D. (2009). Discovering cis-regulatory RNAs in Shewanella genomes by Support Vector Machines. *PLoS Comput Biol*, **5**, e1000338.

Yada, T., Totoki, Y., Takagi, T. & Nakai, K. (2001). A novel bacterial gene-finding system with improved accuracy in locating start codons. *DNA Res*, **8**, 97–106.

Yao, Z., Weinberg, Z. & Ruzzo, W. L. (2006). CMfinder–a covariance model based RNA motif finding algorithm. *Bioinformatics*, **22**, 445–452.

Yusuf, D., Marz, M., Stadler, P. F. & Hofacker, I. L. (2010). Bcheck: a wrapper tool for detecting RNase P RNA genes. *BMC Genomics*, **11**, 432.

Zhang, A., Wassarman, K. M., Rosenow, C., Tjaden, B. C., Storz, G. & Gottesman, S. (2003). Global analysis of small RNA and mRNA targets of Hfq. *Mol Microbiol*, **50**, 1111–1124.

Eigenständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder Dienstleistungen als solche gekennzeichnet.

Sven Findeiß

Leipzig den, 6. Februar 2011

Curriculum vitae

Education:

Since February 2008	Scientific assistant at the University of Leipzig (Germany), Department of Computer Science, Bioinformatics Group
October 2002 - February 2008	Study of Computer Science with focus on Bioinformatics at the University of Leipzig (Germany); Degree: Diploma
September 1993 - July 2001	Secondary school, Gerhart-Hauptmann Gymnasium (Germany); Degree: Abitur, general qualification for university entrance
September 1989 - June 1993	Elementary school, Sigmund-Jähn (Germany)

Selected Scientific Cooperations:

- Stefan Washietl
  (MIT, Computer Science and Artificial Intelligence Laboratory, Boston.)
  Assessment of (non-)coding potential in comparative sequence data

- Cynthia M. Sharma and Jörg Vogel
  (Institut für Molekulare Infektionsbiologie, Würzburg.)
  RNA sequencing analysis

- Cornelius Schmidtke and Ulla Bonas
  (Institute of Biology, Department of Genetics, Halle.)
  transcript analysis in *Xanthomonas campestris* pv. *vesicatoria*

- Andreas R. Gruber
  (Biozentrum, University of Basel.)
  non-coding RNA detection

- Stefan Kalkhof
  (Helmholtz-Zentrum für Umweltforschung - UFZ, Leipzig.)
  experimental validation of short peptides

(Programming) Languages:

Perl, Commandline based scripting (e.g. awk, sed, cut) LaTeX, Postscript, HTML, PHP

German (native), English, Latin