UPPSALA
UNIVERSITET

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 701

Making Sense of Antisense

JOHAN REIMEGÅRD

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2010

ISSN 1651-6214
ISBN 978-91-554-7906-0
urn:nbn:se:uu:diva-131168

Dissertation presented at Uppsala University to be publicly examined in B41, BMC, Husargatan 3, Uppsala, Friday, November 5, 2010 at 13:30 for the degree of Doctor of Philosophy. The examination will be conducted in English.

Abstract
Reimegård, J. 2010. Making Sense of Antisense. Acta Universitatis Upsaliensis. *Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology* 701. 61 pp. Uppsala. ISBN 978-91-554-7906-0.

RNA is a highly versatile molecule with functions that span from being a messenger in the transfer from DNA to protein, a catalytic molecule important for key processes in the cell to a regulator of gene expression. The post-genomic era and the use of new techniques to sequence RNAs have dramatically increased the number of regulatory RNAs during the last decade. Many of these are antisense RNAs, as for example the miRNA in eukaryotes and most sRNAs in bacteria. Antisense RNAs bind to specific targets by basepairing and thereby regulate their expression. A major step towards an understanding of the biological role of a miRNA or an sRNA is taken when one identifies which target it regulates.

We have used RNA libraries to study the RNA interference pathway during development in the unicellular model organism *Dictyostelium discoideum*. We have also, by combining computational and experimental methods, discovered the first miRNAs in this organism and shown that they have different expression profiles during development.

In parallel, we have developed a novel approach to predict targets for sRNAs in bacteria and used it to discover sRNA/target RNA interactions in the model organism *Escherichia coli.* We have found evidence for, and further characterized, three of these predicted sRNA/target interactions. For instance, the sRNA MicA is important for regulation of the outer membrane protein OmpA, the sRNAs OmrA and OmrB regulate the transcription factor CsgD, which is important in the sessile lifestyle of *E. coli*, and MicF regulates its own expression in a feed forward loop via the regulatory protein Lrp.

In conclusion, we have discovered novel antisense RNAs, e.g. miRNAs in *D. discoideum*, developed an approach to identify targets for antisense RNAs, i.e. a target prediction program for sRNAs in bacteria, and verified and characterized some of the predicted antisense RNA interactions.

Keywords: sRNA, miRNA, RNA, target prediction, SOLiD siRNA, E.coli, D.discoideum

Johan Reimegård, Department of Cell and Molecular Biology, Microbiology, Box 596, Uppsala University, SE-75124 Uppsala, Sweden.

ISSN 1651-6214
ISBN 978-91-554-7906-0
urn:nbn:se:uu:diva-131168 (http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-131168)

Till min storebror Mattias och min dotter Fina Nelly Jasmine som båda påminde mig om att livet inte kan vänta

List of Papers

This thesis is based on the following papers, which are referred to in the text by their Roman numerals. The order is based on the order they are mentioned in the thesis

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

II Reimegård, J., Ardell, D., Wagner, E. G. H. (2010) AntisenseRNA: fast, specific target prediction for bacterial sRNAs. Manuscript

III Holmqvist, E. †, Unoson, C. †, Reimegård, J., Wagner, E. G. H., (2010) The sRNA MicF targets its own regulator Lrp and promotes a positive feedback loop. Manuscript

IV Holmqvist, E., Reimegård, J., Stark, M., Grantcharova, N., Römling, U., Wagner, E. G. H. (2010) Two antisense RNAs target the transcriptional regulator CsgD to inhibit curli synthesis. *EMBO Journal*, 29: 1840-50

V Hinas, A., Reimegård, J., Wagner, E. G. H., Nellen, W., Ambros, V. R., Söderbom F. (2007) The small RNA repertoire of Dictyostelium discoideum and its regulation by components of the RNAi pathway, *Nucleic Acids Research*, 35: 6714-26

VI Avesson, L.†, Reimegård, J. †, Wagner, E. G. H., Söderbom, F., 2010 The small RNA population in D. discoideum during growth and development. Manuscript

† These authors contributed equally

Contents

Abbreviations

aa	Amino acid
DNA	Deoxyribonucleic acid
miRNA	Micro RNA
mRNA	Messenger RNA
nt	Nucleotides
ORF	Open reading frame
RNA	Ribonucleic acid
siRNA	Small interfering RNA
sRNA	Small RNA in bacteria
rRNA	Ribosomal RNA
tRNA	Transfer RNA

Introduction

Before I started in Gerhart Wagner's research group, my knowledge about RNA was, to say the least, limited. I knew much about the genetic code written in DNA that was being passed on from generation to generation. I also knew quite a lot about the machines that made life possible, the proteins. The RNA's role was to me something rather dull, a passive player that was needed to be able to translate the genetic code – the DNA where all the information is, into the machines – the proteins that do all the work in the cell. Little did I know that this molecule, RNA, would make me cry, grit my teeth and keep me awake at night – but also make me laugh, spend countless hours talking about it and make me travel to new places and meet fascinating people, people I now consider my friends, hopefully for life. But before I can tell you about my story with what I now consider to be the most intriguing molecule in the cell and my contribution to the knowledge of RNA, I will have to introduce this topic of RNA and its role in the cell.

Life as we know it

Most scientists believe that life on planet earth can be divided into three different groups; archaea, bacteria and eukaryotes. In my research I have been working with both a prokaryote, called *Escherichia coli*, which is a bacterium that can be found in the intestine of all humans, and a eukaryote called *Dictyostelium discoideum*, a social amoeba that is found on the forest floor.

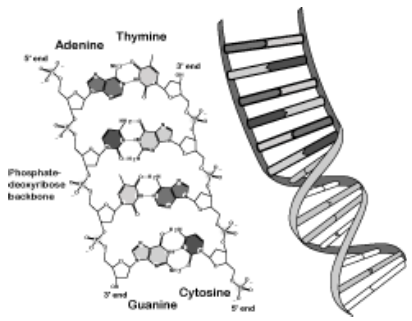
There are many similarities between bacteria, archaea and eukaryotes, suggesting that they all have a common ancestor. But there are also differences. In this thesis I will try to be as general as possible, but I will also stress differences between prokaryotes and eukaryotes when I feel that it is necessary. Hopefully I will also be able to convince you that some things are not as different as they might look at first glance. Finally I will stress that I have been working with RNA and, as you will see, the chapters will focus on RNA molecules.

Important molecules in the cell

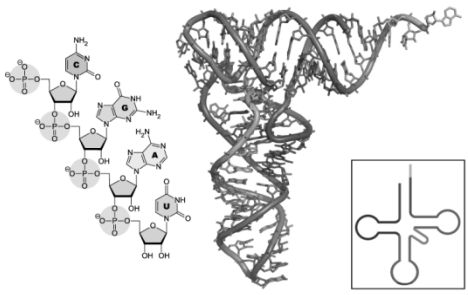
To survive in nature an organism has to be able to replicate and repair itself. In all living organisms that exist today, to the best of our knowledge, there are three different types of molecules that make this possible; DNA, RNA and proteins. To be able to understand molecular biology it is important to know some characteristics about these molecules. For how the molecules are built, see box 1.

Chemical properties of the molecules in the cell

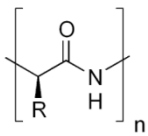
DNA is build up by nucleotides. The DNA-molecule has two different ends where one is referred to as the 5' end and the other the 3' end. Each nucleotide contains three different parts; a sugar (deoxyribose), a phosphate and a base. The base is what differs between the different nucleotides and the base in a DNA nucleotide can be an adenosine (**A**), a cytosine (**C**), a guanine (**G**) or a thymine (**T**).

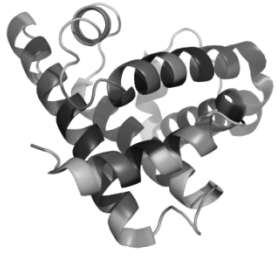


RNA is also build up by nucleotides (**nt**), like DNA. There are two differences between DNA and RNA. The sugar in the backbone is ribose instead of deoxyribose, and RNA has a uracil (**U**) instead of the base T.



Proteins are built up by amino acids (**aa**). Each aa is comprised of an amine group, a carboxylic acid and a side chain. There are 20 different side chains. The primary structure of a protein is a sequence of 20 letters representing the 20 different aa where the first letter is referred to as the N-terminal end and the last letter is the C-terminal end.





Figures were downloaded from the Wikipedia series on Gene expression and is distributed under the *GNU Free Documentation License, Version 1.2*

The second law of thermodynamics states that all things decays in a time dependent manner. At the same time molecules try to find their most stable

structure. The way by which DNA, RNA and proteins do this in the cell differs somewhat. The DNA molecule teams up with another DNA molecule and forms a double stranded DNA (**dsDNA**), a double helix (Watson and Crick 1953). The sugar-phosphate backbone of the two DNA molecules forms the two helices and the bases form basepairs between the two helices. Almost only basepairs of the type A-T and G-C are formed and they are called canonical Watson-Crick basepairs after the two persons who were the first to understand the structure of the dsDNA (Watson and Crick 1953). The reason why these structures are formed, i.e. why they are so stable, is because the A-T and G-C basepairs form planar structures that can stack on top of each other.

DNA is where the information is stored and the molecule is replicated from a cell to its daughter cells. The beauty of DNA as the genetic carrier lies in the strict selection of basepairs. If one of the bases in a basepair is C then the other base must be G and so on. This means that dsDNA can be replicated easily by splitting it into two single stranded DNA strands (**ssDNA**) and then rebuild two identical copies of the dsDNA molecule by just adding nucleotides that can form canonical Watson-Crick basepairs with the two ssDNA strands. A recently published article showed that it was possible to remove the DNA from a bacterial cell and replace it with an engineered dsDNA, thereby reprogramming the bacteria from one species to another (Lartigue et al. 2009).

The proteins are important for the metabolism in the cell, i.e. to build and repair molecules (anabolism) and to break down other molecules to produce energy (catabolism). What role a protein has in the cell is dependent on its structure, and the structure is dependent on the aa sequence. Since the backbone of an aa is small and there is a large diversity between the different aa, proteins can fold into stable and versatile structures making them perfect for performing all the different tasks in a cell.

An RNA molecule also tries to find its most stable structure but, unlike DNA, the RNA in the cell does not normally have a complementary RNA sequence to form a structure with. Instead, the RNA stabilizes its structure by intramolecular basepairs, i.e. basepairs between nucleotides in the same molecule. This means that the RNA forms a structure where some parts are double-stranded and some are single-stranded. Which regions that form basepairs are dependent on the sequence of the RNA. Therefore, different RNA molecules have different kinds of structures and, depending on the structure that they form, they can perform different tasks in the cell. Like in dsDNA, most of these basepairs are of the canonical Watson-Crick type, in RNAs these are A-U and C-G basepairs, but there are also other interactions between different bases.

A recent analysis of the structure of different rRNAs (for what an rRNA is, see next chapter) showed that 59 % of all nucleotides formed Watson-Crick basepairs but there were only 4 % of the nucleotides that did not inter-

act with any other nucleotide (Stombaugh et al. 2009). Apart from the A-U and C-G pairs, the non-canonical G-U basepair is also frequently used to stabilise the structure and during the rest of this thesis a "basepair" refers to an A-U, a C-G or a G-U basepair.

An RNA molecule can be depicted in two different ways. The first is the secondary structure version where only the basepairs are shown. A 2-D structure of a tRNA (for what a tRNA is, see next chapter) is shown in the right corner of the fact sheet for RNAs. The more accurate but a bit more complicated way is the structure where also tertiary interactions are shown. A 3-D model of a tRNA molecule is shown in the middle of the fact sheet for RNAs.

RNA's role in the coding world

If DNA is where the genetic information is kept and the proteins are the molecules important for metabolism, why does the cell need RNA? For many reasons, but the most well-known function of RNA is its role in the transfer of the information stored in DNA to proteins. This flow of information from DNA via RNA to proteins is the general idea in the "central dogma of molecular biology" (Crick 1970) and is accomplished through the two steps transcription and translation. Transcription implies that RNA is being created as a copy of the DNA. Translation means that the RNA is used by the ribosome, which is comprised of ribosomal RNAs (**rRNAs**) and proteins, as a blueprint to create the protein. The RNA that is used as a blueprint is the messenger RNA (**mRNA**) and the relationship between the mRNA sequence and the protein sequence is determined by the genetic code.

The genetic code is a set of rules that translates a triplet of nucleotides in the RNA to an aa. These triplets are called codons. There are a few exceptions (Jukes and Osawa 1990), but in general all living organisms use the same genetic code. Apart from having different codons for the 20 different aa, there are also start and stop codons that specify where the ribosome needs to start and stop translating. The sequence from the start codon to the stop codon on the RNA is called an open reading frame (**ORF**).

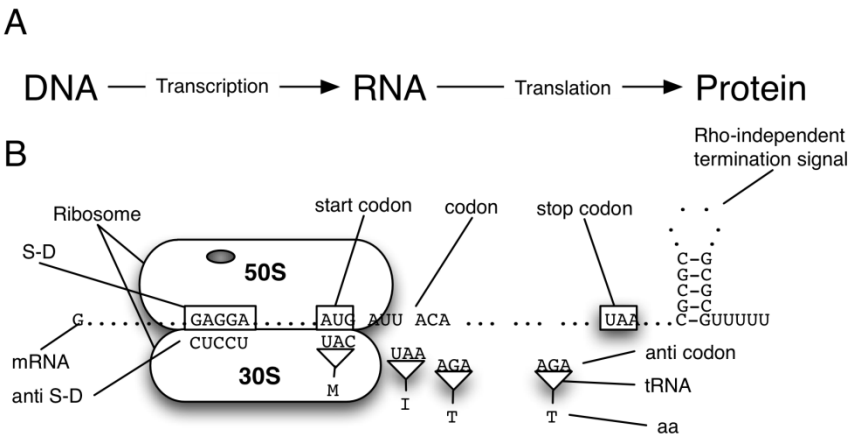


Figure 1: The flow of information from DNA to protein via RNA. The general idea of the central dogma of molecular biology (A). The molecules involved in translating the information from the RNA to the protein in bacteria (A). The mRNA with the ORF between the start codon and the stop codon and the S-D sequence. The tRNAs with the anticodon and the aa. The ribosome with its two subunits 50S and 30S. 30S contains the 16S rRNA with its anti S-D sequence which basepairs with the S-D sequence. The S-D sequence and the anti S-D sequences differ between species.

To translate mRNAs into proteins, the ribosome and the transfer RNAs (**tRNAs**) are needed. A tRNA is an RNA that carries an aa at the 3'-end of the RNA and a triplet of nucleotides that is complementary to the codon in the mRNA that carries the information of this aa. This triplet is called the anticodon of the tRNA. The ribosome binds the mRNA and uses the codon sequence to bind the appropriate aa-tRNA. A tRNA is "cognate" if the anticodon of the tRNA can form basepairs with the codon. When the tRNA has bound to the mRNA, the ribosome transfers the aa from the aminoacyl-tRNA to the growing aa chain on the so-called peptidyl-tRNA. The ribosome then moves to the next codon on the mRNA making it possible to select the next tRNA that matches that codon until it finally reaches the stop codon which terminates translation by releasing the newly formed protein and the mRNA.

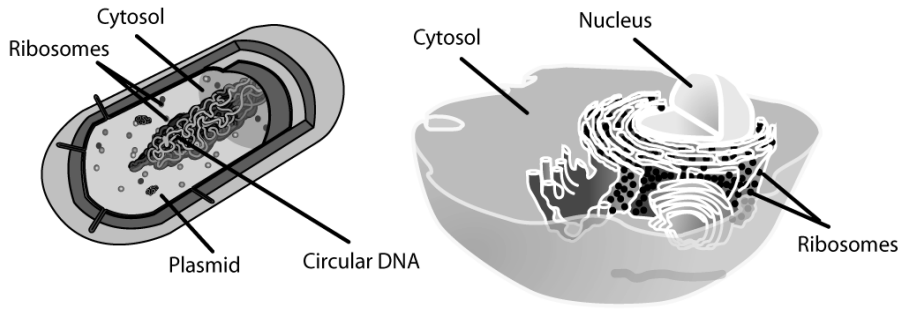


Figure 2. Schematic view of prokaryotic and eukaryotic cells: The prokaryotic cell has in most cases a circular DNA located in the cytosol, where also the ribosomes are. There can also be plasmids, which are DNA molecules independent of the chromosome. The eukaryotic cell has linear DNA on chromosomes located in the nucleus and here the ribosomes are located outside the nucleus, in the cytoplasmic compartment. Figures were downloaded from the Wikipedia. The prokaryotic cell picture is distributed under the public domain license and the eukaryotic cell picture is distributed under the *GNU Free Documentation License, Version 1.2*

In bacteria, DNA and ribosomes are in the same compartment in the cell. As soon as the start codon on the mRNA has been transcribed the ribosome can attach to the mRNA and start translating the mRNA into a protein, i.e. transcription and translation are coupled. Each mRNA has a 5' UTR, one or many ORFs in an operon, and a 3' UTR. The ribosome needs two sequence elements to recognize where to start translating. One is the start codon and the other is a sequence element called the Shine Dalgarno (**S-D**), which is ≈6 nt long and located approximately 6-8 nt upstream of the start codon. The S-D sequence basepairs with a region in the 16S rRNA referred to as the anti-S-D sequence (figure 1) (Shine and Dalgarno 1975), which is part of the 30S subunit of the ribosome.

In eukaryotes, the DNA is located in a compartment called the nucleus, and the ribosomes are found in the cytosol (figure 1). This means that transcription and translation is uncoupled. When the coding gene is being transcribed the so-called pre-mRNA sequence contains the coding regions but also regions that will be removed before it becomes a mature mRNA. The coding regions are called exons, and regions in between the exons are called introns. The introns are spliced out by an RNA- protein complex called the spliceosome. Before the mRNA is being transported out of the nucleus, a 5'cap is put on the 5' end of the mRNA and a poly-A tail on the 3' end. The cap and the poly-A tail are important for the ribosome to recognize and translate the mRNA.

RNA functions in many different ways. It can be passive like the mRNA, where the primary sequence is important. It can be the catalytic molecule

like in the ribosome, where the rRNAs are believed to be responsible for the peptidyl transfer of an aa from the tRNA to the growing peptide. It can also use its structure and sequence to create specific interactions. The tRNA uses this sequence specificity when the anticodon sequence basepairs with the codon sequence. Also the ribosome uses it when the anti-S-D sequence basepairs with the S-D sequence.

Regulation is the key

Even if a cell contains all genes that it needs to survive and carry out its biological role, it is also important that the appropriate proteins and RNAs at the appropriate levels are present in the cell at any given time. For example, a bacterium uses different sets of proteins and RNAs to adapt to an environment that is optimal for growing compared to a hostile environment. This is why every living organism has complex regulatory systems to turn on and off different processes in the cell, and it is in this context that the knowledge of RNA as a regulatory molecule has exploded over the last decade.

For this to be understandable we have to answer the question; at what levels can a cell determine the quantity of an active protein in the cell? The answer is; at the transcription level, the post-transcriptional level and the post-translational level. In most cases, proteins are responsible for regulation at all levels; the transcription factors (**TFs**) for promoting transcription, the degradosome, a protein complex, for degradation of RNA, and the proteosome, another protein complex, for degradation of proteins in the eukaryotic cell. But there are also RNAs that play important roles on all these levels.

At the transcription level, the cell can determine which genes should be transcribed by using RNAs to block or attract the transcription machinery to certain genes. For example, the RNAs Xist (Brown et al. 1991) and Tsix (Lee et al. 1999) are vital for ensuring that one of the two X-chromosomes in all female mammals is inactivated in all cells with two X-chromosomes, which in turn entails that the X-chromosome genes are expressed in appropriate amounts.

At the post-transcriptional level, i.e. when an mRNA is already transcribed, the number of proteins that are produced depends on how frequently the ribosomes can translate the mRNA before it is destroyed. It is at this level that antisense RNAs operate, which will be discussed in more detail in the next chapter.

At the post-translational level, i.e. after the protein has been produced, the protein function can be regulated by co-factors, for example different types of RNAs, which increase or decrease the efficiency of the protein. An example of a post-translationally regulating RNA is the 6S RNA, which is conserved in a wide range of bacteria. 6S RNA adopts a structure that resembles

a bacterial promoter and binds directly to the Sigma factor σ70, a specificity subunit of the RNA polymerase holoenzyme that is important for transcribing a set of genes in the cell. When 6S binds to σ70 it inhibits the transcription of σ70-dependent transcripts (Wassarman and Storz 2000).

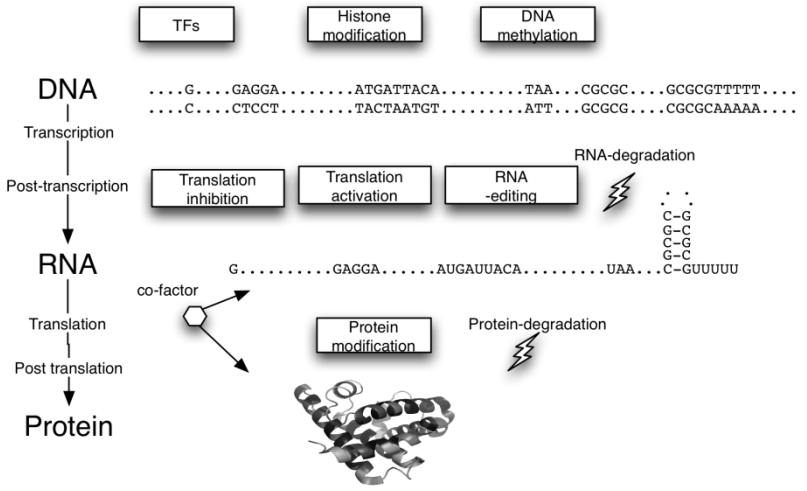


Figure 3: Different levels of regulation on the functional protein levels. There are many ways by which a cell can control the flow of information from of a gene to a protein. It can regulate the level of transcription by different TFs or by changing the structure of the DNA, making it more or less available for transcription. It can *edit* or degrade the mRNA, thereby increasing or decreasing the chance for the mRNA to be translated. It can modify and degrade the protein, thereby increasing or decreasing the functionality or level of the protein. Finally, different co-factors can also influence the regulation on all the three levels.

Non-coding regulatory RNAs are everywhere

Non-coding regulatory RNAs do not contain any ORF, i.e. they do not code for a protein. Their role is to regulate the expression of other genes in the cell. Many regulatory RNAs identify their targets by basepairing. This property is the same that is used in the interaction between the tRNA anticodon and the mRNA codon, and also in the interaction between the 16S rRNA anti-S-D and the mRNA S-D. The non-coding regulatory RNAs go under the name antisense RNAs. In this thesis an antisense RNA is defined as a regulatory RNA that via intermolecular basepairing to another RNA, a target RNA, changes that RNAs expression. Antisense comes from the notion that a piece of DNA that contains a coding gene has a sense strand and an antisense strand, where the sense strand is the one that contains the ORF and the antisense strand is complementary to the sense strand. This means that the

mRNA that is being transcribed from the DNA is identical to the sense strand and if RNA should be able to basepair to the mRNA it must contain part of the antisense sequence, hence antisense RNA.

The first non-coding regulatory RNAs found in prokaryotes and eukaryotes

The first two non-coding regulatory RNAs, CopA and RNAI, were independently found on two different plasmids, R1 and ColE1, in *E. coli* (Stougaard et al. 1981; Tomizawa et al. 1981). A plasmid is an extrachromosomal DNA element, mostly found in bacteria, that has its own replication control system. In ColE1 and R1, the cis-encoded antisense RNA is important for the regulation of the plasmid's copy number in the cell. Cis-encoded means that the antisense RNA is transcribed from the same DNA region as the target that it regulates but from the opposite strand. There is also trans-encoded antisense RNA, which implies that the antisense RNA is not transcribed from the same region on the DNA as its target

The first evidence of a chromosomal antisense RNA was found in *E. coli* and published in 1984 (Mizuno et al. 1984). The authors showed that a trans-encoded antisense RNA, MicF, basepaired with a region around the start codon of an mRNA, *ompF* mRNA, thereby inhibiting its translation. In eukaryotes, it was not until 1992 that the discovery of the first natural antisense RNA was published (Hildebrandt and Nellen 1992); a cis-encoded antisense RNA in *D. discoideum* . One year later, the first evidence of a trans-encoded antisense RNA in a multicellular organism was published (Lee et al. 1993; Wightman et al. 1993). This turned out to be the first case of a micro RNA (**miRNA**), which in the mature, active form is only 21 nt long. A few years later, Fire and Mello showed that the injection of double-stranded RNA (**dsRNA**) matching the sequence of a specific gene in the worm *Caenorhabditis elegans* caused silencing of that gene. For this mechanism, the authors coined the term RNA interference (**RNAi**) (Fire et al. 1998). This discovery awarded them the Nobel Prize in 2006. Even though the injected dsRNA was long, it turned out that the active RNA molecules had the same length as the miRNAs and were subsequently called short interfering RNAs (**siRNAs**).

The first systematic searches for antisense RNAs

It was not until the beginning of 2000 that there were tools available for systematic searches for non-coding RNAs (**ncRNAs**) (see chapter "How to find ncRNAs"). In 2001, the number of known ncRNAs in bacteria and eukaryotes increased from a few sporadic examples to hundreds. In *E. coli* alone, three different studies in 2001 experimentally verified 34 new ncRNAs. (Argaman et al. 2001; Rivas et al. 2001; Wassarman et al. 2001). This should

be compared to ten that were discovered before 2001 in *E. coli*. In bacteria these ncRNAs are referred to as small RNAs (**sRNA**). In the same year, three independent studies identified a total of 91 new miRNAs in worms, flies, and human cells (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001), followed by isolation of miRNA sequences in plants (Llave et al. 2002; Reinhart et al. 2002). The discoveries of sRNAs, miRNAs and other kinds of ncRNAs led to *Science* considering it the "breakthrough of the year" in 2002 (Couzin 2002).

Different flavours of regulatory ncRNAs

Since the discoveries of sRNAs and miRNAs, the interest in ncRNAs has increased dramatically. During the last eight years, a wide variety of ncRNAs has been found in bacteria, archaea and eukaryotes. Now we know that prokaryotes and eukaryotes use regulatory RNAs to protect themselves against viruses and plasmids. Archaea and bacteria use CRISPRs (Clustered Regularly Interspaced Short Palindromic Repeats) to fight viruses (Barrangou et al. 2007; Horvath and Barrangou 2010). Plants and invertebrates use siRNAs against viruses whereas vertebrates have evolved primarily other virus defences, like the inferon system. Conversely, viruses use miRNAs to affect the expression of the immune defence of the host (Stern-Ginossar et al. 2007). Another example is the Hepatitis C virus that invades the liver. It needs a human miRNA to replicate itself. This miRNA is only expressed in the liver (Jopling et al. 2005). Regulatory RNAs are also important for silencing of selfish DNA elements, like retrotransposable elements, to prevent their spreading within the genome (Aravin et al. 2003). A class called piwi-interacting RNA (**piRNA**) of a distinct length of 26-28 nt is expressed in the germ line cells in animals and is required for proper germ cell development. This was first described in fruit flies (Vagin et al. 2006).

The mechanisms for many of these processes are still not entirely understood. The common theme is that the ncRNA guides a complex of proteins, by sequence recognition, to the correct target. It is also be easy to imagine that one complex can regulate different targets depending on which ncRNA that is part of the complex (Ghildiyal and Zamore 2009).

Antisense RNAs: what is the mechanism?

One of the largest classes of ncRNAs found in both prokaryotes and eukaryotes are the antisense RNAs. These can be categorized into different classes, depending on their biogenesis, their regulatory mechanism and whether they are present in eukaryotes or prokaryotes.

The first antisense RNAs whose regulatory activity is mostly independent of proteins (except Hfq, see below) were discovered on plasmids and later on

bacterial chromosomes. For many of these antisense RNAs the mechanism has been studied in detail and is well understood (for summary see (Wagner et al. 2002). The exact mechanism for how the different antisense RNAs bind to their targets, like RNAI binds to RNAII, and CopA binds to CopT differs for these antisense RNAs, but still some general conclusions that can be drawn from these examples. The common nominator is that the interaction has at least two steps. The first one is the initial rate-limiting unstable interaction and the second one is the following fast propagation of basepairs that stabilizes the interaction. The first intermolecular basepairs must be formed from single-stranded sequences within in these RNAs, for initial contact to occur. After the initial binding, more basepairs are added in one direction from the initial basepairs. The rates of how fast these basepairs are formed are dependent on the intramolecular structures of the two RNAs and the intermolecular structure that is formed. If intramolecular structures have to change for the formation of the intermolecular structure, this will increase the time it takes for the intermolecular structure to be formed. Also, if there are bulges and internal loops in the intermolecular structure, this is expected to increase the time it takes to form the structure. Highly stable intramolecular structures or bulges and loops that disfavour the intermolecular structure drastically counteract the formation of a stable intermolecular structure (Simons 1997).

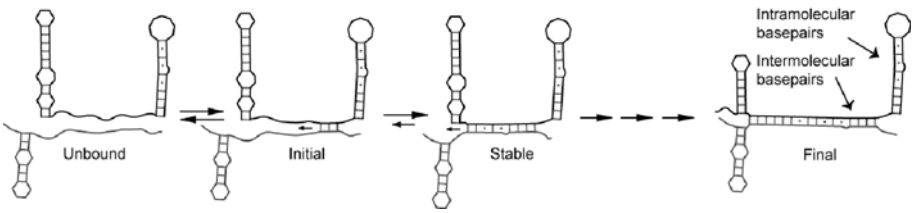


Figure 4 Common denominators for antisense RNA/target interactions. Antisense RNA interactions go through an initial binding that is then quickly stabilized. To reach the final structure, the RNAs can progress through many intermediate conformations.

Characteristics of sRNAs in *E. coli*

Bacterial sRNA genes have, in most cases, their own promoter and terminator, i.e. the RNAs do not have to be processed to become active (for reviews see (Romby et al. 2006; Waters and Storz 2009)). sRNAs can both up- and downregulate the expression of a gene but the majority of the reported targets are downregulated.

Most sRNAs, in *E. coli* at least, need a protein, called Hfq, for efficient regulation of their targets in vivo. Why Hfq is needed is not fully understood and the reason probably differs between different sRNAs and their targets. In *E.coli*, however, there are three characteristics that seem to be general for the

use of Hfq. First, Hfq stabilizes RNAs thus increasing their concentration in the cell (Udekwu et al. 2005; Holmqvist et al. 2010). Second, it may bind both an sRNA and an mRNA simultaneously, which means that Hfq increases the local concentration of the antisense RNA and its target RNA (Vecerek et al. 2008). Finally, this protein has been shown to increase the initiation rate of antisense-target RNA interaction (Vecerek et al. 2008), but the mechanism for this is still incompletely understood. There are also more specific examples, for instance the case of the sRNA RyhB. This RNA regulates the expression of the gene *sodB* by binding to the *sodB* mRNA, thereby inhibiting translation and promoting mRNA degradation. In this case, Hfq is needed to change the intramolecular structure of the *sodB* mRNA before RyhB can form a stable complex with it (Geissmann and Touati 2004).

The role of Hfq as a stabiliser of RNAs in the cell is somewhat counterintuitive since Hfq is part of a regulatory system that should adapt quickly. However, this is solved by the fact that Hfq actively exchanges on the mRNAs and sRNAs it is bound to, thereby ensuring that the pool of bound RNAs roughly reflects their abundance in the cell (Fender *et al*, Genes & Dev, in press).

Whether or not Hfq is important for maintaining a stable complex between an sRNA and an mRNA is still debated and could perhaps differ between different sRNAs. Studies of the sRNA MicA, which have been extensively performed in our research group, suggest that as soon as the two RNAs have bound, the RNA duplex is stable independent of Hfq (Fender *et al*, Genes & Dev, in press). However, results from studies of other sRNAs and their targets suggest that Hfq is important for the stability of the RNA complex (Soper et al. 2010). The sRNAs in Table 1 show the diversity of sRNA-target interactions in *E. coli* in terms of where they bind, how they regulate and whether or not Hfq is needed. For a complete list of sRNA targets, see sRNATarBase (Cao et al. 2010).

Most sRNA targets show changes both on the protein and the RNA level as a result of regulation, suggesting that the sRNA operates by initiating degradation of the target RNAs. This is supported by results showing that Hfq co-purifies with RNaseE and that two sRNAs, SgrS and RyhB, actively bring RNaseE to the target by binding to Hfq, thereby enhancing the degradation of their target RNAs, *ptsG* mRNA and *sodB* mRNA, respectively (Morita et al. 2005). However, SgrS can also regulate PtsG without degrading the mRNA, by inhibiting the translation of *ptsG* (Morita et al. 2006). Most sRNAs that basepair with an mRNA affect the rate with which the ribosome binds to that mRNA. So, even if it is hard to distinguish between direct and indirect sRNA-caused degradation of an mRNA, the indirect degradation, dependent on reduced translation initiation, is believed to be the most frequent. *In Salmonella*, the sRNA MicC and one of its target, *ompN* mRNA, represents the only case so far where the mode of regulation is not ambiguous. Here, the sRNA binds far from the start codon, see Table 1, and

is not affecting translation initiation. Instead MicC targets the coding sequence of the *ompN* mRNA by binding to it and targeting RNaseE for nearby cleavage followed by facilitated decay (Pfeiffer et al. 2009).

Table 1. sRNA/target interactions in *E. coli.*

sRNA-target RNA	Reference	sRNA-interaction region[1]	Target-interaction region[2]	Regulation	Hfq[3]
DsrA-*rpoS*	(Lease et al. 1998)	10 to 32	-97 to -119	up	yes
IstR-1-*tisB*	(Vogel et al. 2004)	1 to 22	-124 to -103	down	no
MicA-*ompA*	(Udekwu et al. 2005)	8 to 24	-21 to - 6	down	yes
MicC-*ompC*	(Chen et al. 2004)	1 to 16	-15 to -30	down	yes
MicC-*ompN*	(Pfeiffer et al. 2009)	1 to 12	+67 to +78	down	yes
MicF-*ompF*	(Schmidt et al. 1995)	1 to 28	-11 to +9	down	yes
OxyS-*fhlA*	(Argaman and Altuvia 2000)	98 to 104 23 to 30	-15 to -9 +33 to +40	down	yes
RprA-*rpoS*	(Majdalani et al. 2001)	33 to 61	-94 to -116	up	yes
RyhB-*sodB*	(Masse and Gottesman 2002; Geissmann and Touati 2004)	38 to 46	-4 to +5	down	yes
SgrS-*ptsG*	(Vanderpool and Gottesman 2004)(Kawamoto et al. 2006)	157 to 187	-28 to +4	down	yes
Spot42-*galK*	(Moller et al. 2002)	1 to 61	-20 to +55	down	yes
GcvB-*livJ*	(Sharma et al. 2007)	63 to 87	-45 to -22	down	yes
SroB[4]-*chbC*	(Figueroa-Bossi et al. 2009)	38 to 57	-68 to -49	down	yes
SroB[4]-*ybfM*	(Rasmussen et al. 2009)	45 to 56	-8 to -19	down	yes

[1]Which nt on the sRNA that is predicted or verified to interact. [2]Which nt on the mRNA that is predicted or verified to interact. Location is relative to the start codon. [3]Whether the sRNA needs Hfq to regulate. [4]SroB is also called called MicM

miRNA biogenesis and mechanism

The hallmark of the siRNA and the miRNA pathway is the involvement of two protein families, Dicer (Bernstein et al. 2001) and Argonaute. Dicer is an RNase III-type enzyme that cleaves dsRNA, and Argonaute promotes cleavage of the target RNA that the siRNA binds to. The siRNAs can have various sources, but the two most common ones are invading dsRNAs, like viruses, or RNAs derived from selfish elements in the cell, like retrotransposable elements. These are or generate often long dsRNA sequences that each can give rise to many different siRNAs. The complex that the siRNA

and the Argonaute is part of is called the RNA induced silencing complex (**RISC**) (Zamore et al. 2000). Since the siRNA in RISC binds to the same kind of RNA as the one that generated the siRNAs, the target RNA has full complementarity. This results in the RISC-dependent cleavage of the target RNA between the two nt that basepairs with the two nt 10 and 11 on the siRNA.

The miRNAs in animals are 21-22 nt long, but they are transcribed from the DNA as longer precursor transcripts. These can be transcribed as single genes but are more often transcribed in clusters. miRNA genes are mostly found in intergenic regions or in introns. The initial transcript of a miRNA is referred to as pri-miRNA (Lee et al. 2002). The first step of maturation is in the nucleus, where a complex called the microprocessor, which contains the two proteins Drosha and Pasha, processes the pri-miRNAs (Han et al. 2006). This generates a hairpin structure called pre-miRNA of ~70 nt that is exported out of the nucleus (Lee et al. 2002). There are however exceptions where miRNAs located in the introns of mRNAs use the splicing machinery instead of the microprocessor to form the pre-miRNA structure (Okamura et al. 2007; Ruby et al. 2007). After export, a Dicer protein further cuts the pre-miRNA into a 21-22 nt long dsRNA, with 2 nt 3' overhangs on both sides (Bernstein et al. 2001). One of the strands is degraded but the other one is preferentially incorporated into a protein complex called the microRNA ribonucleoprotein particle (**miRNP**), which contains the Argonaute protein. The incorporated strand is the (guide strand) miRNA and the degraded strand is often referred to as the (passenger strand) miRNA*. The miRNA guides the miRNP to the target RNA by forming an intermolecular structure with the complementary target sequence. Apart from a few exceptions, base-pairing occurs in the 3' UTR of the mRNA and inhibits translation. The nt 2-7 from the 5' end are the most important ones for determining which target the miRNA will bind to, and there is almost always full complementarity between the miRNA and the target in this region. It is therefore often referred to as the "seed" sequence. There are also other factors that are important: the location of the miRNA binding site, whether several miRNAs bind nearby, and what sequence environment the target site is embedded in (Grimson et al. 2007). Recently, a new type of interaction has been found in which it is not the seed sequence that determines the specificity of the miRNP but the region from nt 5 to 16. (Shin et al. 2010) Once the miRNP has bound to the mRNA via the basepairing of the miRNA, a protein called GW182 is important for the effect on target. GW182 has been reported to both degrade the poly-A tail and the cap.

In plants the biogenesis is almost the same with some minor differences. Most of the miRNAs are transcribed as single genes. The pre-miRNAs vary much more in length, from 60 to 300 nt, and the processing from the pre-miRNA to the 21-22 nt long miRNA is carried out in the nucleus before it is exported and incorporated into RISC. In plants, miRNA usually exhibit full

or almost full complementary to the target. Therefore, this results in the same RISC-dependent cleavage of the target RNA as for the siRNAs.

Early results suggested that the miRNAs in plants only caused degradation (Rhoades et al. 2002) of the target RNA, but recent data show that miRNAs in plants also inhibit translation of the target. In the case of animals, early studies in worms suggested that miRNAs regulate by inhibiting translation and do not promote degradation, whereas recent studies have shown that there is strong correlation between RNA degradation and translation inhibitions in mammals (Baek et al. 2008; Selbach et al. 2008; Guo et al. 2010).

Why regulatory RNAs?

An interesting question to ask is: what is the advantage of having regulatory RNAs in the cell? One rationale that has been promoted is speed, because a regulatory RNA is ready after transcription whereas a regulatory protein needs to be transcribed and subsequently translated. However, if a cell wants to adapt quickly it is even faster to have an inactive protein in the cell that turns active in the presence of a co-factor (Shimoni et al. 2007). Another argument in favour of regulatory RNAs compared to regulatory proteins is cost. It is much cheaper for the cell to produce a short ncRNA than a long mRNA that has to be translated into a protein. A counterargument is that one protein, if it is a TF, can inhibit the initiation of transcription, which is much cheaper than producing regulatory RNAs that operates after the target RNA already has been transcribed.

A more compelling argument why ncRNAs are found as regulatory molecules lies in the numbers game. Since sRNAs in bacteria and supposedly miRNAs in animals form stable complexes when binding to their target, an sRNA can only regulate one target RNA. This means that if there are many antisense RNAs in the cell and only a small number of target RNAs, the probability that the target RNA to become translated is small. Also, if there is a small number of sRNAs and an excess of target RNAs, the regulation of the sRNAs will be neglectable (Levine et al. 2007; Shimoni et al. 2007). This has some interesting implications. First, if expression of a gene at low level is disadvantageous for the cell, regulation at the post-transcriptional level can be more advantageous than regulation at the transcriptional level. The difference between the two regulatory systems lies in the number of proteins that are produced if regulation fails. If regulation at the transcriptional level fails, an mRNA is transcribed which can generate many proteins. However if the regulation is on the post-transcriptional level the regulatory RNA competes with the ribosome as long as the mRNA is present in the cell. As an outcome, there will be less fluctuation in the concentration level of a protein in a cell with a post-transcriptional regulating system than with a

transcriptional system (Levine et al. 2007). It is also quite easy to overrun the regulation of a regulatory RNA by either increasing the target mRNA or by titrating out the regulatory RNA by transcribing another target RNA that the sRNA binds to (Levine et al. 2007; Shimoni et al. 2007). There is a beautiful example of both these types of regulation; the case of the sRNA IstR-1 and its target *tisB* mRNA (Darfeuille et al. 2007). Since the protein TisB is toxic, its concentration levels must be kept low during normal growth (Unoson and Wagner 2007). To make sure that the *tisB* gene is turned off, IstR-1 is constitutively expressed in the cell. The concentration of IstR-1 is sufficient to make sure that *tisB* is almost never translated unless induced by DNA damage.

Interestingly, there are also other factors that determine the translation of TisB. The *tisB* mRNA S-D cannot initiate translation without a "standby" site that is located upstream of its start site. This is because the S-D site is, most of the time, in a conformation that is inaccessible for ribosome binding. It is believed that the ribosome, by "waiting" at the standby site, increases its probability to access the S-D site when it is in an open conformation. When the *tisB* gene is transcribed as a full-length, denoted +1, transcript, the standby site is inaccessible, and the *tisB* mRNA is not translated. When the mRNA is processed to a truncated version, called +42 mRNA, its structure changes, rendering the standby site accessible, and so the *tisB* mRNA can be translated. However, this active mRNA is also the preferred target of IstR-1. When IstR-1 binds to the +42 *tisB* mRNA, RNase III-mediated cleavage of the mRNA at position 106 from the original transcription start site (+106) is promoted. The +106 mRNA does not contain the standby site, and as a result the *tisB* mRNA cannot be translated. When the cell on the other hand is under SOS-response, which happens when there are breaks in the DNA, the concentration of IstR-1 remain the same but the +42 *tisB* mRNA accumulates to high levels. Thus, *tisB* is no longer regulated by IstR-1 since it is in molar excess over IstR-1 (Darfeuille et al. 2007).

A second example of RNA regulation inhibition is exemplified by another sRNA where SroB is out-titrated by the upstream region of *chbC* mRNA to relieve the inhibition of its normal target *YbfM* mRNA(Figueroa-Bossi et al. 2009). Similar phenomena have also been seen in plants, where an almost perfect target for a miRNA, which however cannot be cleaved due to a central mismatch, titrates the miRNA, thereby indirectly activating the translation of this miRNAs other targets (Franco-Zorrilla et al. 2007)

Methods to find ncRNAs

While mRNAs all have common signals for decoding, i.e. a start codon, an ORF and a stop codon, ncRNAs do have few general similarities. The most common theme is that the structure is important for the ncRNAs function. In

the year 2000, when comparing the minimum free energy (MFE) of secondary structures of scrambled sequences to the MFE of true ncRNAs, it was however shown that the MFE of the secondary structure of most ncRNAs does not contain sufficient information to identify them (Rivas and Eddy 2000). Nevertheless, there are ways to increase the success rate in attempts to discover novel ncRNAs, namely by increasing the information in the ncRNAs or reducing the number of RNAs that they are being compared against.

Computational approaches

Even though ncRNAs in general do not have enough information to be easily identified, each specific ncRNA gene class often has some distinguishing features. Pre-miRNA folds into a hairpin structure and tRNA has a cloverleaf like secondary structure, for example. The specific characteristics can also be found on the sequence level. For example, tRNAs display a structure that can be used to identify novel tRNAs, but it is the sequence in the anticodon that defines the specific species of tRNA. Therefore, for classification of a tRNA, the structure is used, but to determine what kind of tRNA it is, sequence information is needed. This kind of structure and sequence information has been used to develop programs that specifically identify one kind of ncRNA class. For example, tRNAscan-SE (Lowe and Eddy 1997) and ARAGORN (Laslett and Canback 2004) for detecting tRNAs or miR-abela (Sewer et al. 2005), mirfold (Billoud et al. 2005) and RNAmicro (Hertel and Stadler 2006) to discover miRNAs.

There are also programs that find novel ncRNAs from a class using homology from already known member of the class. One of the simplest one, but also the commonly used is blast (Altschul et al. 1990) where the model is just the primary sequence of a homologous RNA. But there are also programs that use more advanced models that include both structure and comparative genomics. For a overview and comparison on different homology search methods see Freyhult et al. 2007.

A non-coding RNA gene can be conserved over time, which makes it possible to find orthologs of the same gene in different species. If a characteristics of a potential ncRNA, a structure or a sequence, is conserved in different species, this scores as an increased likelihood that a true ncRNA is identified. Interestingly, in many cases the structure of the ncRNAs is more important that the sequence. For example, an ncRNA may form a basepair at a specific position whereas the nature of the basepair may be unimportant. If so, the corresponding gene in different genomes can contain different basepairs, i.e. A-U, C-G or G-U.

It is also possible to search for de-novo ncRNAs without prior knowledge about that RNA. Signals that are universal for all transcripts, like terminators and promoters, have successfully been used to identify potential ncRNA

transcripts (Argaman et al. 2001). Another approach to discover novel ncRNAs involves reduction of background noise by a-priori selection of genomic regions. For example, if the genome nt composition is biased from the beginning towards high AT content, it is possible to use this property to identify ncRNAs, which may be more GC-rich (Schattner, 2002; Klein et al., 2002, Larsson et al. 2008).

Experimental approaches

High throughput analysis techniques, like cDNA libraries, microarrays or deep sequencing (RNAseq), can also be used to identify novel ncRNAs, since they give information about which regions on the genome that are being transcribed (Huttenhofer and Vogel 2006). Due to the depth of the sequences from an RNAseq run, i.e. that even RNAs in very small amounts in the cell are found, it can be hard to distinguish an ncRNA gene from just random transcripts. Another caveat is that many genes are not transcribed under the experimental circumstances when the RNAs are extracted.

Combining computational and experimental results

Both computational and high throughput methods struggle with a quite high false positive rate. By combining the results of a computational approach with a RNA-sequence library the number of false positive can be reduced. This kind of approach has been successfully used to identify novel sRNAs in bacteria (Wassarman et al. 2001; Faucher et al. 2010) and miRNAs (Lu et al. 2005)

Methods to find antisense RNA targets

Finding an sRNA in a bacterium or a miRNA in an eukaryote does not immediately reveal anything about its biological role. Assuming that it is a regulatory RNA, one may want to initially search for its target(s). More specifically, assuming it uses an antisense mechanism, you would like to find the mRNA that it is basepairing to.

Computational approaches

Since an antisense RNA must basepair with its target, the first step towards identifying that target might be to search for an intermolecular structure that can be formed between the antisense RNA and the target. However, a major concern with computational methods is not finding potential targets as such, but to distinguish them from the many putative sequences that are not regu-

lated by the antisense RNA. This is analogous to the situation with identifying ncRNAs using the MFE of that ncRNA (Rivas and Eddy 2000). Still, for miRNAs in plants, where the interaction pattern is almost perfect, this approach has led to the correct identification of many miRNA targets (Jones-Rhoades and Bartel 2004).

By contrast, miRNA interactions in animals or sRNA interactions in bacteria display intermolecular antisense RNA-target RNA structures that are relatively short and often contain bulges and loops. In these cases, there is not enough information to separate true interactions from basepair patterns that occur by chance. However, there are other factors that can be used to identify true targets for antisense RNAs.

In animals, almost all known interactions with miRNAs occur in the 3'UTR of the mature mRNA. Also, almost all interactions have full complementarity between nt 2-7 in the miRNA and the target sequence. By assessing different miRNA/target interactions by microarray analysis and scoring for down regulation of target genes, some additional features have been determined (Grimson et al. 2007): the seed region can also include nt 1 and 8 and, if the miRNA has five consecutive basepairs starting from nt 12 to 14, the probability of a correct hit is increased. Also other features surrounding the interaction site on the mRNA affect the probability that the miRNA regulates the predicted target; if two or more adjacent binding sites are present, if the miRNA binds close to the beginning or the end of the 3'UTR, or if the interaction site is located in an AU-rich region, the probability of correct target assignment is higher. These criteria have been used to create a program called Content score, which evaluates a miRNA target. In a very recent study, another algorithm, mirSVR (SVR: Support Vector Regression), has been trained using predicted target feature outputs against the degree of regulation of genes (Betel et al. 2010). One of the features included in mirSVR but not in Content score is the structure accessibility of the miRNA on the target RNA. The researchers behind mirSVR also show that mirSVR is better than Content score in most cases when it comes to predicting the regulation of miRNA targets. If the interaction site was conserved in other species it further improved correct target prediction.

For bacterial sRNAs, the most common feature is that their target sites are located in the 5' UTR of mRNAs. Compared to miRNAs that carry defined "seeds", the region of interaction of an sRNA differs and can be located anywhere on the sRNA, from the immediate 5' part of the sRNA to the loop of the terminator hairpin. Nevertheless, the concept of a "seed" is used in some programs to reduce the number of false positives without reducing the number of true interactions (Tjaden et al. 2006; Busch et al. 2008). Another feature that increases the likelihood of predicting a true sRNA target is the inclusion of the intramolecular structures of the RNAs (Muckstein et al. 2006; Busch et al. 2008). Another approach is to select a-priori selected features and train a model based already known targets (Zhao et al. 2008). Fi-

nally, in one study, researchers have a-priori selected a set of nt in the sRNA that should interact based on their conservation in different species in which this sRNA was found (Richter et al. 2010).

Experimental approaches

Since a regulatory RNA changes gene expression, this can be assessed by genome-wide approaches using protein expression, RNA expression analyses (Vogel and Wagner 2007). However, there are some caveats associated with these experimental methods. First, the true targets may not be expressed under the conditions used when conducting the experiment. Second, the molecule whose abundance is monitored may not be affected. For example, if an sRNA regulates a target only at the protein level and not at the RNA level, RNA expression analyses will not detect these changes. Third, some changes may reflect secondary effects. For instance, an sRNA could regulate the expression of a second regulatory RNA, which in turn controls other genes. These would be incorrectly identified as direct targets of the sRNA

Targets identified using computational methods and/or experimental methods on a genome wide scale needs to be verified. To verify a predicted interaction site, a mutational analysis of the basepairs can be carried out. If a basepair is important for regulation, a mutation in one of the nts in either the antisense or the target RNA will disrupt the basepairing interaction between the two RNAs. Hence, the target RNA will not be regulated. If the complementarity between antisense and target RNA is restored, by mutating also the corresponding nt in the second RNA, regulation should occur again. This is a strong indication that the sRNA indeed regulates the target through the predicted interaction (Figure 5) (Vogel and Wagner 2007).

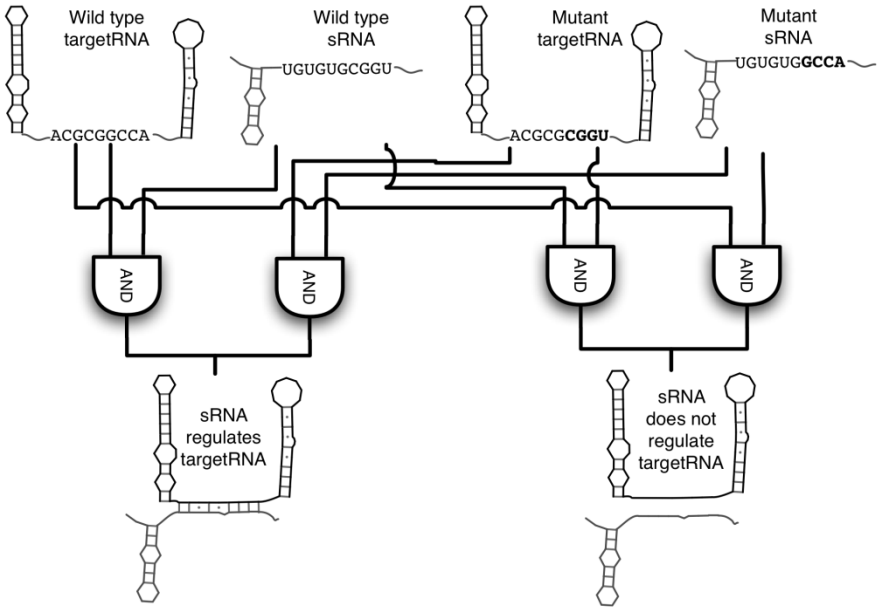


Figure 5 Mutational analysis. Wt sRNA regulates wt target RNA but not mutant target RNA. Mutant sRNA regulates mutant target RNA but not wt target RNA.

Present investigation

In the 21st century, our perception of the role of RNA has changed dramatically as the number of ncRNAs found has increased dramatically and as it was recognized that these play an important role in the cell. Still there are probably more ncRNA families to be found, and an even more cumbersome task is to try to understand what these RNAs are doing in the cell. The aim of my thesis was to identify novel antisense RNAs and their targets primarily by using computational methods. This section of the thesis will be divided into two different parts. In the first one we have tried to understand the mechanism for how sRNAs bind to their targets. We have then used these results to create a new target prediction program and identified several previously unknown targets for three sRNAs. The second part describes our attempt to discover novel small ncRNAs in *D. discoideum.*

Discovering sRNA targets in *E. coli* (Papers I –IV)

E. coli is arguably the most important prokaryotic model organism. Its genome sequence was published in 1997 (Blattner et al. 1997). It is a gram-negative bacterium that belongs to the Enterobacteriaceae family. Its habitat is usually the lower intestine of warm-blooded animals, and its preferred growth temperature is 37º C. However, it can survive outside the body for quite some time. Even though there were only ten sRNAs found before 2000, publications in the subsequent years reported on many new sRNAs in *E. coli*. In 2003, 55 sRNAs had been found in *E. coli*. Since then, this number has increased to 77 reported in Rfam (Griffiths-Jones et al. 2003) (2010-09-09). At the start of this study, a very small fraction of these sRNAs had a known function. To be able to determine the role of some of the sRNAs and to understand the mechanism for how they interact with target RNAs, we developed a computational target search program called AntisenseRNA. We have also evaluated the efficiency of the program and characterized two of the verified targets. I will here primarily focus on the computational parts in these papers for which I was responsible.

Clues to features incorporated into AntisenseRNA (Paper I)

One of the sRNAs identified in the first search for novel sRNAs was initially called SraD (Small RNA D) but was subsequently renamed to MicA. To identify the targets that the sRNA regulates, a 2D-protein gel approach was used (Udekwu et al. 2005). The cellular protein levels were compared in the presence of high, normal, or low MicA levels. The protein that showed the greatest MicA-dependent change was an outer membrane protein called OmpA. At the same time, an early version of AntisenseRNA, which searched for target sites around the start codon of all mRNAs in *E. coli* and closely related species, identified a segment of the 5' region of MicA as being complementary to the 5' UTR of the *ompA* mRNA. The interaction was comprised of 16 basepairs interrupted by a bulge between basepairs 4 and 5, located around the S-D sequence of OmpA, which was conserved in several enterobacterial species. An even more interesting observation indicated that even though the basepair pattern was the same in all species, the nucleotide sequence differed. This suggested that the intermolecular structure rather than the primary sequence was important for regulation.

Structural probing showed that the predicted region on *ompA* mRNA was single-stranded when probing with only *ompA* mRNA but when also MicA was added this region became double-stranded indicating intermolecular basepairs. To verify the predicted interaction, a mutational study of the interaction site was done as described in the introduction (Udekwu et al. 2005; Holmqvist et al. 2010). The mutant MicA had 6 nt changed in the predicted interaction site, and the *ompA* leader had six compensatory mutations corresponding to the mutations in MicA. In accordance with prediction, regulation of OmpA by MicA was lost when one of the RNAs carried the mutated sequence. Regulation was restored when both RNAs carried the complementary mutations in the interaction site (Udekwu et al. 2005). This indicated that the regulation of OmpA by MicA is direct, dependent on basepairing, and that the important region of interaction was the same as the predicted one.

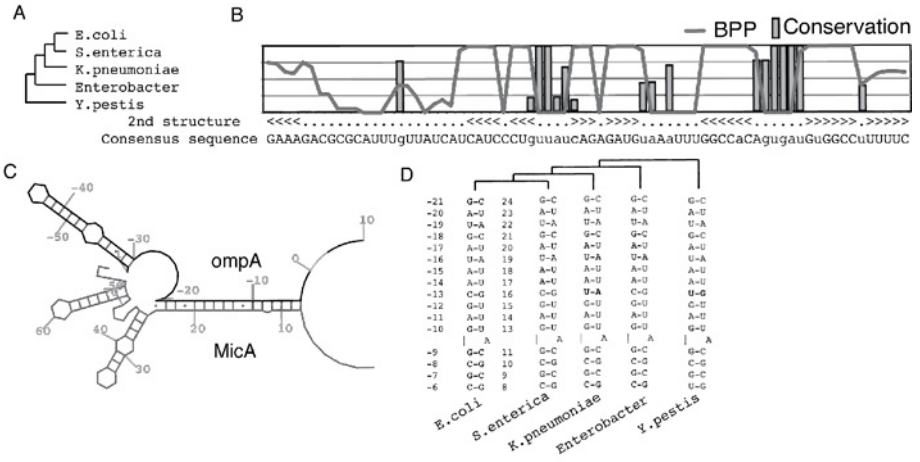


Figure 6: Different characteristics of MicA. (A)MicA is conserved in enterobacterial species. (B)Secondary structure and consensus sequence from RFam 9.0. Cumulative basepair probability using CONTRAfold for each nucleotide is depicted by solid line in. Conservation of nucleotides compared to E.coli. Nucleotides that are not conserved represented by bars. Nucleotides with a change in closely related species get a high bar compared to nucleotides with a change in more distantly related species. (C) The interaction between MicA and the *ompA* mRNA. (D) Different sequence but the same basepair pattern between MicA and *ompA* mRNA in different species.

Design of an sRNA prediction program (Paper II)

Clues from paper I and other published sRNA papers suggested that the interaction between an sRNA and a target RNA follows the principles in the chapter "Antisense RNA structure and function" in the book "RNA structure and function" (Brian N. Zeiler 1998). For instance, the first nt's that interact have to be in single-stranded regions in both RNAs, and other basepairs predicted to interact should be located immediately adjacent to the first basepairs on one side, and these basepairs should also preferentially be in less structured regions before they interact. We hypothesized that the initial basepairs and the adjacent basepairs next to it should be the ones under highest selective pressure. Thus, when comparing basepair pattern of the sRNA/target in different species it is more important that these basepairs are conserved than the rest of the basepairs.

Designing the model

We then designed a program that models the interaction between an sRNA and the target RNA using four rules. The first one states that there is a region of initiation where the first nt's of the sRNA and the target RNA form a short intermolecular structure. For these basepairs, the sRNA and target RNA need to be single-stranded. The second rule says that the interaction is direc-

tional. This means that, after the first initiating basepairs have been formed, the subsequent basepairs should follow on one side of the initiating basepairs. The third rule demands that until the intermolecular structure has reached a certain stability, energetically unfavorable bulges or internal loops are not allowed referred to as the propagating region. Also, if the region contains strong intramolecular structures it will inhibit the progression of the structure. The fourth rule is that the final structure should be the most stable RNA duplex, that is the lowest MFE, given that the first three rules are complied. To enforce that the initial basepairs are scored as particularly important, basepairs that are not conserved in the initial region or in the propagating region, are penalized more than those that merely stabilize the final structure.

Designing the AntisenseRNA program

To implement these rules for target prediction, we designed the program/ algorithm AntisenseRNA that includes both intermolecular and intramolecular structure information and takes into account the conservation of basepairs in related bacteria. AntisenseRNA is an adaptation of the dynamic program RNAhybrid (Rehmsmeier et al. 2004) with adaptations similar to IntaRNA (Busch et al. 2008). To include intramolecular structures, we used basepair probabilities, which by default are calculated by CONTRAfold (Do et al. 2006), and converted them into free energy scores using tabulated free energy values based on basepairs and dangling ends (Mathews et al. 1999). To include conservation of basepairs we asked how far from a genome of interest, in our case that of *E. coli*, critical basepairs are conserved. This approach considers both the structure of the phylogenetic tree and the branch lengths. This implies that for each basepair, bulge or internal loop, the score is dependent on three parameters. The first considers the change in free energy for the formation of a basepair in the intermolecular structure, the second the change in free energy for the intramolecular structures that have to be broken for the basepair to form, and the third the penalty that is determined by how far from the focal genome the basepair is conserved in closely related species.

Reducing the runtime

In IntaRNA the time to search all potential sRNA-target RNA interactions is dependent on the size of the sRNA, the size of the target RNA, and on how large internal loops the intermolecular structure can form. The latter is as default set to 15. However many of the intermolecular loops and bulges that are tested will never be used, because the change in free energy to create these bulges and loops is too high. We reasoned that instead of having a fixed length, we could instead have a fixed free energy cut-off. This means that, as soon as the change in free energy to create internal loops and bulges is too high, the program will stop searching for such elements. As an effect,

the program is dynamic in the sense that it searches for longer bulges and internal loops in regions of the sRNA and the target RNA where both are unstructured, whereas in structured regions of the sRNA and the target RNA, it only considers small bulges and internal loops, thus increasing the speed of the algorithm by an average factor of three.

Optimizing the parameters

Since the program is dependent on three functions, the intermolecular structure between the sRNA and the target, the intramolecular structure of the sRNA and the target RNA, and the conservation of basepairs, we put scalable factors in front of the intramolecular structure and the conservation of basepairs so that we could weight the three functions against each other.

To test the program and determine the best parameters, we ran it with some sRNAs against known targets and compared how these ranked compared to a set of randomly generated mRNA sequences with the same dinucleotide frequency as the real targets. We first optimized the program by only considering the focal genome, i.e. we omitted the assessment of conservation in other species, varied the intramolecular structure factor between 0 and 2, and calculated the mean and median specificity of the true targets compared to the generated targets. The best mean and median specificity was obtained when the intramolecular structure factor was set to 0.80. This implies that AntisenseRNA considers the intermolecular structure more than the intramolecular structures. To determine the basepair conservation parameter, we set the structure parameter to 0.8 and varied the conservation parameter to calculate the mean and median specificity. The best median value was acquired when the conservation of basepairs factor was set to 0.65. The median specificity was then 0.999.

Testing the program

Target prediction programs are needed for two reasons. First, we want to be able to predict new targets. When comparing AntisenseRNA to other established sRNA target prediction programs (Zhang et al. 2003; Muckstein et al. 2006; Tjaden et al. 2006; Busch et al. 2008; Tjaden 2008), AntisenseRNA gives a better mean and median value for both the training set and a set of interactions that it has not been trained on. This suggests that AntisenseRNA is a good choice if one needs to predict novel targets for an sRNA. Second, target prediction programs are used to determine the precise region of interaction. When comparing how well AntisenseRNA predicted the correct, that is experimentally validated, interaction regions, it performed as well as the best of the other target prediction programs. Among the programs that use both intermolecular structure and intramolecular structure information, AntisenseRNA is much faster.

Identifying novel targets for OmrA and MicF (Paper III and IV)

One of the reasons for developing AntisenseRNA was to aid us in the discovery of new targets for sRNAs that we worked on in our group. However, there is a concern with AntisenseRNA. For most sRNAs that it predicts, at least one of the known targets is among the top 20 candidates, whereas in some cases, AntisenseRNA is not well suited to find targets. To increase the chance of discovering novel targets for an sRNA, we focused on sRNAs where at least one target was already known, and for which AntisenseRNA had successfully predicted this target as one of the top candidates.

MicF regulates Lrp (Paper III)

The first RNA that we tried to find targets for with AntisenseRNA was MicF. OmpF, the known target (Mizuno et al. 1984) was ranked as number two among all candidates, and an orthologous *micF* gene was present in *Yersinia pestis*. This one differed substantially in primary sequence, which is an advantage when using comparative genomics. To test candidates target genes, we designed a translational fusion GFP reporter system with the 5' UTR region of the targets as described in (Urban and Vogel 2007). These experiments showed the strongest effect on *ompF*, which was known before and therefore gave proof-of-principle. A second high-scoring target derived from prediction was *lrp*. This gene encodes a transcription factor, Leucine responsive protein, a protein that responds to Leucine concentration and regulates many proteins in the cell (for a review see (Calvo and Matthews 1994)). There were two reasons why we considered this protein as interesting, apart from it being predicted as a good candidate by AntisenseRNA and that regulation was observed in the GFP reporter system. First, OmpF, the known target of MicF, also varies in abundance as an effect of the nutrition levels in the cell. Second, Lrp negatively regulating *micF* would suggest a regulatory feedforward loop involving MicF and Lrp by reciprocal repression.

In vitro translation studies showed that *lrp* mRNA was translated, but that addition of MicF and Hfq inhibited *lrp* mRNA translation. Mutations in MicF abolished this inhibition, and compensatory mutations restored regulation. This was in line with results from the GFP-reporter system *in vivo* and suggested that the effect of MicF on Lrp is direct.

When running AntisenseRNA with MicF against *lrp* mRNA, two regions of interaction were predicted as likely binding sites. One of the sites is located in the 5' region of MicF and interacts with a sequence around the start codon of *ompF* mRNA. The other site is located close to the rho independent terminator of MicF and 20 nt upstream of the start codon of the OmpF mRNA.

AntisenseRNA predicted that the 5'-most nucleotides in MicF should form the initial structure because the region around the start codon was con-

served in all species whereas the second region was not. A mutational study verified that the first region is the one important for regulation. Changing only one nucleotide in the predicted initial intermolecular structure was sufficient to abolish regulation of MicF on Lrp.

CsgD is a target of OmrA and OmrB (Paper IV)

When the paper describing the program targetRNA (Tjaden et al. 2006) was released, it reported on its performance by testing four different sRNAs, comparing prediction of targets with available microarray results. Two of the sRNAs tested were OmrA and OmrB. These are probably paralogues. Among the predicted targets by targetRNA only a few showed regulation based on the microarray data. A further study, published by the same group, reported another set of targets using microarrays (Guillier and Gottesman 2006) that had no overlap with the targets they predicted in the targetRNA paper. To test how AntisenseRNA prediction compared with the predictions of targetRNA, and the targets suggested in the subsequent paper, we ran our program against OmrA and OmrB. None of the targets predicted by targetRNA was among the top candidates of our program but from the other paper there were a few. Instead, the top target for OmrB, and rank number five for OmrA, was CsgD. CsgD is a transcriptional activator of genes encoding curli components, which build up surface structures required for a sessile life-style (Barnhart and Chapman 2006). The predicted interaction site of OmrA and OmrB on CsgD was located in a bulge and a stem in a hairpin located 70 nucleotides upstream of the CsgD start codon. The predicted initial and stabilizing basepairs where the ones that were located in the bulge of CsgD mRNA (Holmqvist et al. 2010).

Since CsgD is required for curli production in *E. coli*, phenotypic tests were available. When curli is expressed, bacteria form red colonies on Congo red indicator plates (Hammar et al. 1995). When we grew *E. coli* with an overexpression plasmid encoding *omrA* or *omrB*, the colonies remained white suggesting that OmrA and OmrB directly or indirectly downregulate CsgD.

By introducing a C-terminal 3xFLAG-tag sequence in the chromosomal *csgD* gene, we observed that both sRNAs reduced the synthesis of CsgD both at the RNA and protein level. To test if the regulation of CsgD by OmrA and OmrB was direct and occurred at the place we predicted, we mutated the predicted interaction site. The results showed that when challenging the wildtype *csgD* strain with a mutant version of OmrA or OmrB (four nt changes in the predicted interaction site), regulation was lost. Using the same mutant sRNAs against a *csgD* target mutant with compensatory changes restored basepairing and control. The same was found when the mutations were in the opposite interaction ligands. These results verified that the prediction was correct and that OmrA and OmrB regulate CsgD by basepairing 70 nt upstream of the start codon.

Additionally, a toeprint experiment showed that OmrA and OmrB directly inhibit binding of the ribosome to the translation start site. At this point, we do not yet understand the details of the inhibitory mechanism. A hint at what may be at play comes from analysis of the conservation of the 5'UTR structure of *csgD* mRNA. It turned out that not only the sRNA interaction site, but also the entire structure of the 5' UTR including a hairpin structure that contains the S-D and the start codon, was conserved in all species in which a *csgD* was found. Further experiments showed that the 5'-UTR of CsgD consists of two distinct modules. The first module is the stem loop structure to which OmrA and OmrB bind (SL1; (Holmqvist et al. 2010)) and the other one consists of the stem loop that includes the S-D sequence and the start codon (SL2; (Holmqvist et al. 2010)). SL2 has an inhibitory effect on the translation efficiency of the *csgD* mRNA because a mutation that disrupts the stem loop structure increases the translation of CsgD. An interesting feature of the SL1 module is that it can be placed in front of a heterologous S-D and start sequence, conferring regulation by OmrA and OmrB. This means that the regulation on the SL2 module is independent of SL1. It also suggests that the answer to how the binding of OmrA and OmrB to the 5' UTR affects the binding of the ribosome the start region can be found within SL2.

Unraveling the RNAi dependent RNA repertoire in *Dictyostelium discoideum* (Papers V -VI)

The slime mold *D. discoideum* is a single cell eukaryote that lives on the forest floor feeding on bacteria. When food supplies are running low, up to 100 000 cells aggregate to enter a multi-cellular development. About 80% of them differentiate into pre-spore cells, and the remaining 20% into pre-stalk cells. This "multicellular organism" can form a motile slug that can move around in response to light. The final stage of development is a stalk with a ball of spores on top. After dispersion, the spore cells can germinate when conditions improve to enter a vegetative life cycle again. The stalk cells, on the other hand, are dead and have sacrificed themselves for the benefit of the other cells in the multicellular community. *D. discoideum* belongs to the amoebozoa lineage, which is on the same branch as animals, but has branched off after the plant-animal split, and before the split between animals and fungi (Baldauf and Doolittle 1997; Bapteste et al. 2002). This makes *D. discoideum* an interesting model organism not least from an evolutionary point of view. The genome, which comprises six chromosomes and one tandem repeat that contains the ribosomal RNA genes, has been sequenced. *D. discoideum* has predicted homologs of many of the proteins important in the RNAi machinery. There are genes encoding two Dicer-like

proteins, *drnA* and *drnB*, three RNA-dependent RNA polymerases (RdRPs), *rrpA*, *rrpB* and *rrpC,* and five Argonaute-like proteins (Martens et al. 2002; Kuhlmann et al. 2005). An RNA library with longer ncRNAs had previously revealed two novel classes of small RNAs, class I and class II RNAs, but until then the entire small RNA repertoire had not been assessed (Aspegren et al. 2004). The aim of this project was to identify the small RNA population by massively parallel sequencing in order to understand the different pathways, with the additional hope of discovering miRNAs in a single-cell organism.

Dissecting the RNAi-associated RNAs in *D. discoideum (Paper V)*

When I started on this project, two cDNA libraries, derived from 18-30 nucleotides long RNAs, had already been sequenced. However, since these libraries where large (approximately 5000 sequences each) my experience was solicited to characterize and sort these sequences. The difference between the two libraries, referred to as the 5'-ligation-dependent – requiring a mono-phosphate at the 5' end, and the 5'-ligation-independent – which is not sensitive to the number of phosphates, is that they reflect different kind of RNAs (Hinas et al. 2007).

The project on the RNAi-dependent RNAs in *D. discoideum* was conducted as a collaboration between four groups, Fredrik Söderbom's group interested in ncRNAs in *D. discoideum*, Wolfgang Nellen's group interested in the proteins responsible for RNAi, Victor Ambros with an interest in miRNAs and with expertise in creating small RNA libraries, and me from Gerhart Wagner's group with experience in primarily bacterial sRNAs and the underlying bioinformatics. Since I work with antisense RNAs, I wish to mention that this collaboration included scientists who pioneered work on antisense/ target RNA interactions (Wagner, Söderbom), the first published natural antisense RNA in eukaryotes (Nellen), the first microRNA in eukaryotes (Ambros) and the first published papers for systematically looking for miRNAs in *C. elegans* (Ambros) and sRNAs in *E. coli* (Wagner). Based on this, the chance of finding something interesting was definitely in our favour.

After mapping the cDNA library sequences to the genome we were left with 2387 sequences in the 5'-ligation dependent library and 1432 in the 5'-ligation independent library. We already knew from a prior study that many of the small RNAs would originate from the retrotransposon element called DIRS-1, and these were highly represented in both libraries. There were also many sequences that corresponded to fragments of rRNAs and tRNAs. There were two differences observed between the two highly abundant groups. First, the DIRS-1-derived small RNAs were more abundant in the 5' ligation dependent library, whereas the tRNA and rRNAs were more highly represented in the 5'ligation-independent library. Also, most of the small RNAs

from DIRS-1 were 21 nt long, whereas the other small ncRNAs varied in length. Both these results suggest that DIRS-1-derived small RNAs are Dicer-dependent, whereas the other shorter RNA fragments more likely are degradation products.

Since Nellen's group had constructed single knockout mutants in the RNAi pathway genes, we used these to analyze whether DIRS-1 small RNAs were affected in mutant strains. Surprisingly, we did not see any differences in expression levels of the DIRS-1 small RNAs in any of the Dicer knockouts. These results suggested that the small RNAs generated by Dicer where either not part of the RNAi pathway, or that DIRS-1 small RNAs can be generated by either of the two homologs (*drnA* and *drnB*), i.e. indicating redundancy. A previous report had also shown that an *rrpC mutant* highly up-regulates the long DIRS-1 transcripts (Kuhlmann et al. 2005), but no effect could be seen on the DIRS-1 small RNAs that we probed against.

After setting apart all DIRS-1 and ncRNA transcripts, we were left with 212 sequences in the 5' dependent-, and 230 sequences in the 5'-independent ligation library. Most of these sequences were unique hits. Three findings caused continued investigations. First, we found transcripts that were antisense to coding genes. Since the first antisense RNA was found in *D. discoideum,* we investigated whether these came from longer transcripts. For three of the genes, we found longer antisense RNA transcripts, suggesting that there were indeed more antisense RNAs in this organism. Second, in contrast to previous results, we found eight RNAs that originated from another retrotransposon element called skipper. They were all found in the ligation-dependent library. Six of these, all 21 nt long, originated from a region of 44 nt. When examining this region, two potential double-stranded RNA structures were predicted, meaning that the 21 nt long RNAs could come from either of these sources. Using Northern blots, probing for sequences immediately upstream of the six cloned RNAs, showed a band representing a 21 nt long RNA. Unfortunately, this probe failed to distinguish the two potential sources from each other.

Of the remaining cloned small RNAs, we selected the RNAs that were mapped to intergenic regions, introns, and sequences antisense to coding genes. For these we extracted 150 nt upstream and downstream of the mapped location of the small RNA clone. We then used mirfold (Billoud et al. 2005), a program that is used to find miRNAs in plants, to identify putative pre-miRNA structures, and further refinement was obtained by a second run using miR-abela (Sewer et al. 2005), a program developed to identify human pre-miRNA-like structures. After all filtrations, five candidates matched the criteria of a potential miRNA. For two of these, the mature miRNA was confirmed by Northern blot. One of these was upregulated like skipper in an *rrpC* mutant strain. More importantly, unlike the skipper small RNAs, this RNA was undetectable in a *drnB* mutant. Unfortunately, homologs of the two miRNAs were not found in any other species. Neverthe-

less, validation by Northern blot and the presence of the pre-miRNA structure qualified for registration of these miRNA candidates as ddi-mir-1176 and ddi-mir-1177 in miRBase (Ambros et al. 2003; Griffiths-Jones 2006).

Discovering more miRNAs in different developmental stages of *D. discoideum* (Paper VI)

> "Many of the miRNAs were identified by single cDNA sequences, so it is clear that none of these screens are near saturation " Sean R. Eddy

This quote was written by Sean R. Eddy in a review article in 2001 (Eddy 2001) as a comment on the three articles that reported 70 miRNAs in *C. elegans,* 14 in *Drosophila melanogaster,* and 19 in human cells. Now, more than 940 different miRNAs deposited in miRbase for humans, and 171 for *C. elegans* (as of 2010-09-09) (Griffiths-Jones 2006). Following the lead of S. Eddy, we set out to create a larger library to discover novel miRNAs, but also to find miRNA* sequences for the miRNAs we already had reported. For this, we used next generation sequencing (SOLiD) to increase the number of RNAs in the library 1000-fold. Since miRNAs have been shown to be important in the development of multicellular organisms, we also searched for miRNAs that were differentially expressed in different stages of the developmental cycle of *D. discoideum*.

Three new libraries of small RNAs, of sizes between 10 and 40 nt, from different developmental stages: growing cells (0h), slugs (16h), and fruiting bodies (24h), were created. The library construction requires a single phosphate at the 5' end of the RNA for ligation and subsequent reverse transcription and amplification. Since we did not pre-treat our RNA samples, the libraries should not contain primary transcripts. The sequencing resulted in 50 nt long sequences that contained the *D. discoideum* sequences and part of the 3' adapter sequence used to sequence the small RNAs.

The first step involved determination of the length of the sequences and removal of the adapter sequence. This also served as quality control because the presence of a correct 3' adapter sequence in a read makes it likely that the preceding sequence also is correctly sequenced. The false rate was very low (less than 1/100 000), and adapter sequences were read up to 15 nt, so that *D. discoideum* sequences could be determined spanning lengths between 0 and 34 nt. Their length distribution, with a distinct peak around 21 nt, resembled that of the 5' dependent ligation step sequence library in Paper V (Hinas et al. 2007).

To discover potential miRNAs, we chose a different approach than previously. Instead of relying on the annotation of the genome, we filtered out sequences based on the knowledge acquired from previous work. This means we only considered reads of 21 nt and that they should not come from re-

peats or regions where transcripts from both strands can be found. Finally, we removed all locations where there less than 5 sequences that mapped to the same location. After filtration, we had between ~4000 and ~13000 reads, depending on developmental stage, that mapped to approximately 250 locations in the genome independent of the library analyzed. To identify pre-miRNA like structures, we used mirfold (Billoud et al. 2005) with very strict criteria.

Among the locations that remained were those that encoded the two miRNAs ddi-mir-1176 and ddi-mir-1177. It was exciting that we also found sequences that were perfect matches to the corresponding miRNA* sequences. One observation was that the distribution of reads between the different libraries was not identical, suggesting differential expression patterns during development for mir-ddi1176 and mir-ddi-1177. We could also get the miRNAs expressed from a vector that included the pre-miRNA sequence. That a miRNA can be expressed from a plasmid has been used to verify that a potential miRNA actually is a miRNA (Chiang et al. 2010). Also under overexpressing conditions, miRNA processing was dependent on DrnB.

We found an additional 20 miRNA candidates, three of which however were perfect miRNA* sequences of three other candidates. Of these, 14 fitted the criteria of human pre-miRNA structures when using miR-abela (Sewer et al. 2005). The hits on the miRNA sequences gave two distinct subgroups. The first class comprised other pre-miRNAs like ddi-mir-1176 and ddi-mir-1177 where most of the sequences mapped to one or two distinct locations in the genome. But there were also miRNAs that was represented in many locations in the genome. It turns out that those miRNA candidates come from degenerated inverted repeats of a retrotransposable element.

miRNAs derived from retrotransposable elements

The miRNA candidates that came from the degenerated inverted repeat can be mapped back to derivatives of Thug-S. Thugh-S is a repeat element that shares some features with so-called miniature inverted-repeat transposable elements (MITEs) (Glockner et al. 2001). In humans, MITEs have been reported to be the source of most transposon-derived miRNAs (Piriyapongsa et al. 2007). It has also been proposed that MITEs are an evolutionary link between siRNAs and miRNAs (Piriyapongsa and Jordan 2007).

miRNAs expressed from intergenic regions and introns

Considering their loci, most of the miRNAs mapped to the intergenic regions, some mapped to Thug-S fragments, and one mapped to an intron.

To verify the expression of these reads we selected two candidates, mir-can1 and mir-can2. Mir-can1 is located in an intergenic region on chromosome 2 and was selected because of its differential expression during devel-

opment, with many reads in the 0h library but only a few in the other libraries. Mir-can2 was located in an intron and did not show differential expression. When probing for the miRNA sequences, mir-can1 could be detected only in the *rrpC*- strain while mir-can2 was below detection level in all strains. No miRNA* sequences could be found. To verify that the miRNAs where expressed (Chiang et al. 2010) we cloned also these pre-miRNA hairpins into the extra chromosomal expression vector and transformed into *D. discoideum.* For mir-can1, the miRNA but not the miRNA* sequences were detected. When expressed in a *drnB*$^{-}$ strain, neither miRNA nor miRNA* was produced. The results for mir-can2 were more ambiguous since both 20 and 21 nt species were seen for both the miRNA and the miRNA*. In the *drnB*- strain, only the 21 nt band was absent suggesting that this hairpin generates two kinds of small RNAs. One (20 nt) is independent of *drnB,* and one (21 nt) that is *drnB* dependent.

Discussion and future perspectives

In one of my first oral presentations on antisense RNA I chose the title "Making Sense of Antisense" because it is catchy and it makes sense in that if you know what the antisense RNA binds to you know the function of that sRNA. I still think that the title makes sense but not in the same sense as before. Instead, making sense of antisense has been my struggle of coming to terms with the fact that all antisense RNAs do not behave in the same way. I have also come to realize that even if one *does* know the target, this still does not tell you much about the RNA. It only opens the door to the next question.

I will also point out that RNA is not "alive". It does not have a choice but instead follows the rules of nature (or chemistry or physics I guess). These are the rules that we as natural scientists try to understand. During my time with Gerhart, we have developed a program that uses thermodynamic criteria to find potential targets for sRNAs. What separates AntisenseRNA from other programs that also search for sRNA targets is that I have expanded the algorithm to include that AntisenseRNA interactions have an initiation and propagation step. That is, rates and accessibility are additional important criteria, in addition to stability. I also included comparative genomics and separated the importance of basepair conservation depending on the position within the interaction region. The results show, not surprisingly, that thermodynamic rules for intramolecular and intermolecular structures do not give all the answers to the question which targets an sRNA regulates. Yet, the results show that it can be useful to predict novel targets. Apart from developing this program, I had the chance to participate in the subsequent characterization of some of the targets identified by AntisenseRNA. MicA regulates OmpA, OmrA and OmrB regulates CsgD, and MicF regulates Lrp.

At the same time, our collaboration with Söderbom's group has given me a chance to also discover novel antisense RNAs in *D. discoideum*. With the help of RNA libraries we had the possibility to test the interesting hypothesis that there might be miRNAs in *D. discoideum.* At the same time, we discovered other RNAi-related pathways. In the course of this work, we have gathered enough information to support that, as in multicellular organisms, the standard miRNA biogenesis pathway operates, with hairpin structure precursors, and the need for a Dicer-like enzyme for its processing. We also found a repeat element that could be a source for generating new miRNAs in D. discoideum.

Even though I have had the chance to participate in many projects, there are still many interesting things to learn about antisense RNAs in general, and the specific antisense RNAs that I have been working with. Some of these open questions are addressed briefly below.

Even though MicF regulates Lrp, this RNA is not being degraded and would not have been discovered if using an experimental approach that compares RNA levels when MicF is upregulated compared to when MicF is downregulated. This raises the question of whether there are more RNAs that are regulated only or at least primarily on the translation level. It would be interesting to do a similar study to the ones carried out to correlate the relationship and interdependence of RNA degradation and translation inhibition for miRNAs (Baek et al. 2008). Not only could we identify new targets for an sRNA, but it would perhaps also give a clue on what (if there are many that are just being regulated at the translation level) differs between the two groups.

In the case of regulation of CsgD by OmrA and OmrB, a major remaining question is its elusive mechanism of inhibition. In spite of many hours of discussion and a lot of more or less sane hypotheses regarding the mechanism, we still do not understand it. By using mutational studies using upregulation of *csgD* expression in the presence of OmrA or OmrB, and mutations that downregulate the expression of *csgD* in the absence of the sRNAs we hope to identify which nucleotides are important for regulation. This might lead us to identify a sequence or a structure element that is important for the regulation and help us unravel the mystery.

Many, if not most sRNA targets in enterobacteria encode outer membrane proteins, surface proteins and membrane-bound transporters (Guillier and Gottesman 2006). Is this so because post-transcriptional regulation is the best way to accomplish the rapid remodeling of the outer membrane, or does this reflect particular features in the sequence of the targets? If one could identify other signals from the interaction site, the prediction of targets would be improved giving a much lower false discovery rate. One example that suggests such a signal is the case of SgrS and PtsG. Here, the regulation is dependent on a sequence on PtsG, which is important for targeting PtsG to the membrane, which is distinct from the antisense interaction (Kawamoto et al. 2005). Maybe by selecting highly ranked AntisenseRNA targets that also have the same signal could help us identify more targets.

Since miRNAs in plants and in animals are proposed to have evolved separately (Grimson et al. 2008; Shabalina and Koonin 2008), the discoveries of miRNAs in *D. discoideum* ask the question of whether miRNAs were present before the ancestors of *D. discoideum* branched off from the animal branch, or whether these miRNAs represent a third independently evolved system of the miRNA pathway. We may also have to reconsider that miRNAs could have been present even before the animal-plant split.

For the miRNAs in *D. discoideum,* we do not know by which mechanism they regulate their targets. Is it like the miRNAs in plants where they bind with almost full complementarity in the coding sequence? When predicting plant like miRNA interaction sites using the same approach as in (Zhao et al. 2007) against the coding sequences of *D. discoideum* the number of potential targets was the same as when running the miRNAs against a set of scrambled sequences. This suggests that there is no selection for plant-like miRNA targets. We have also done initial experiments mapping 5' ends for some of the plant-like predicted targets of the miRNAs. So far we have not found any 5' end product that match to the predicted interaction site between the miRNA and the targets. We are also in the process of developing a 3' UTR GFP reporter system. In this system it is possible to add a designed 3' UTR to a gene expressing GFP on an expression vector. We are currently designing 3' UTRs that has different kind of miRNA target like properties. Both 3' UTRs that have regions with full complementarity to one of the miRNA and 3' UTRs that are designed with multiple animal-like miRNA target sites against the same miRNA. Depending on which, if any, of these reporter systems that the miRNA can regulate we can get one step closer in understanding the mechanism of miRNA regulation in *D. discoideum.*

Finally, many of the newly found miRNAs were only represented by a few numbers of sequences in the RNA libraries, suggesting that these are not highly expressed under the conditions we tested. It would be interesting to find the conditions, if any, when these candidates are being expressed at higher levels, giving a hint at what role they play in the cell.

Conclusion

To work with ncRNAs in both bacteria and eukaryotes has been a cumbersome but valuable experience. Sometimes this broad view has helped me see things that I would probably not have seen if I just had been working with one organism or one kind of ncRNA. Also, my time with Gerhart has made me realize the importance of details that have to be understood by decisive experiments if one wants to understand a biological phenomenon in detail. This has made me recognize that it is important to sometimes take a step back and try to find analogies of your question in other topics to help you bring your research forward. At the same time be careful not to draw conclusions based solely on these analogies but also pursue it with experiments that strengthen or nullify your newly formed hypothesis. This is probably the experience that I will take with me for the rest of my life, not only in science but also in life.

Svensk sammanfattning

DNA, RNA och proteiner är tre olika typer av makromolekyler som är helt avgörande för att liv på jorden ska kunna finnas till, växa och föröka sig. Sedan länge har det varit känt att arvsmassan som förs vidare från generation till generation består av DNA, att RNA är kopior av DNA som används för att skapa proteiner och att cellerna, livets byggstenar, behöver proteinerna för att fungera. Dessa RNA kallas budbärarRNA. Men för omkring 30 år sedan gjordes de första upptäckterna som visade på att RNA inte bara har i uppgift att föra instruktionerna för hur olika proteiner ska se ut vidare från DNA till det maskineri i cellerna som bygger proteinerna. Idag vet man att så kallade regulatoriska RNA även spelar en viktig roll för regleringen av olika proteiners koncentration i cellen.

År 2001 hittade tre forskargrupper, bland annat den jag har ingått i, oberoende av varandra flera små regulatoriska RNA i tarmbakterien *E. coli*. Dessa fick benämningen småRNA och är mellan 50 och 200 nukleotider (nt) långa. I masken *C. elegans*, i bananflugan *D. melanogaster* och i människan hittade man samma år väldigt många regulatoriska RNA som kom att kallas mikroRNA p.g.a. att de var mycket korta - bara 21 nt. Året därefter upptäckte man även mikroRNA i växter.

När jag började som doktorand hos professor Gerhart Wagner hade han och hans kollegor upptäckt nya småRNA i *E. coli.* Däremot var deras funktion i cellen fortfarande okänd. Under min tid som doktorand har jag utvecklat ett program som försöker identifiera så kallade målRNA-molekyler, det vill säga de RNA-molekyler som ett småRNA binder till. Programmet utgår primärt från tre kriterier: småRNA-molekylen och målRNA-molekylen måste innehålla nt-sekvenser som kan bilda par med varandra, dessa regioner måste vara fysiskt tillgängliga och samma nt-sekvenser ska dessutom finnas bevarade i andra bakteriearter. Det senare är ett tydligt tecken på att regionen är viktig för att ett småRNA ska kunna hitta sin partner. Med hjälp av mitt program har vi lyckats identifiera flera potentiella målRNA. Bland dem har vi vidare karakteriserat tre sådana RNA-RNA-interaktioner. Den första är interaktionen mellan MicA och budbärarRNAt till OmpA. OmpA är ett yttre mebranprotein som MicA reglerar ner när *E. coli* utsätts för membranstress. Den andra interaktionen är den mellan småRNA-molekylerna OmrA eller OmrB och deras målRNA CsgD. *E. coli* behöver CsgD för att kunna bilda biofilm, vilket innebär att flera bakterier samlas genom att bygga upp en matris runt sig. Om det finns mycket OmrA eller OmrB i cellen kan *E. coli*

inte längre skapa någon biofilm. Den tredje småRNA/målRNA-interaktionen vi har karakteriserat är den mellan småRNAt MicF och ett målRNA som heter Lrp och kodar för ett styrprotein som bland annat reglerar hur mycket MicF det ska finnas i cellen. Detta samband skapar ett intressant regulatoriskt system där MicF genom att reglera ner Lrp reglerar upp sin egna aktivitet i en så kallad ”feed-forward loop”.

Parallellt med arbete i Gerhart Wagners forskargrupp har jag även fått möjlighet att samarbeta med docent Fredrik Söderbom och hans kollegor, som forskar på den encelliga eukaryoten *D. Discoideum,* även kallad slemsvamp, vilket är en encellig social amöba som växer på marken i skogar. Det som är intressant med *D. discoideum* är bland annat att när tillgången på föda tryter så går många celler ihop och börjar bilda en stjälke och en sporboll. Sporbollens celler sprids och väntar tills omständigheterna blir bättre för att då återgå till sin encelliga livsstil medan cellerna i stjälken dör. *D. discoideum* är alltså ett exempel på en encellig organism med förmågan att övergå till en "flercellig organism" där delar av populationen offrar sig för de andra cellernas överlevnad. *D. Discoideum* är också intressant ur ett evolutionärt perspektiv då man tror att den delar gemensamt ursprung med djur efter det att djur och växter delade upp sig i det evolutionära trädet. En annan intressant sak med . *D. Discoideum* är att den innehåller alla komponenter som behövs för att generera små interferensRNA. Samma komponenter som används för att generera små interferensRNA är också viktiga för att generera mikroRNA.

Min forskning i samarbete med Fredrik Söderboms grupp har gått ut på att identifiera olika små interferensRNA och mikroRNA i *D. discoideum.* Det har vi gjort genom att analysera olika sekvensbibliotek av småRNA som är mellan 15 och 30 nt långa. Sekvenserna kommer från olika utvecklingsstadier när *D. discoideum* går från en encellig till en flercellig organism. Vi har identifierat små interferensRNA som är viktiga för att se till att själviska DNA-element i cellen inte sprider sig i genomet. Vi har även hittat de första mikroRNA i *D. discoideum* där vi har kunnat visa att uttrycksmönstret av dessa skiljer sig åt i de olika utvecklingsstadierna och mellan varandra.

Under framtagandet av egna datorbaserade metoder för att identifiera nya mikroRNA i *D. discoideum* och målRNA för småRNA i *E. coli* har jag fått möjligheten att utveckla ett kritiskt synsätt till metoder och analyser av stora datamängder. Samtidigt har jag haft förmånen att i ett nära samarbete med mina kollegor kunna följa upp mina analyser experimentellt. Detta har lett till att vår forskning har förts framåt på ett sätt som inte skulle ha varit möjligt om vi hade arbetat var för sig.

Acknowledgements

To those who made sense:

This is for the people that in one way or the other have helped me going from a student to a scientist.

First and fore most I would like to thank my supervisor **Gerhart** for loving science so much it is contagious. I do not think I would have become a PhD student if it was not for you and for that I am forever grateful. I will also remember, with joy, all the personal stories that you shared over the years.

I would also like to thank **David Ardell** who has tried, repeatedly, to knock some ”from a bioinformaticians point of view” into my head. I also would like to thank you for putting my ideas into words that makes sense.

My special gratitude goes to **Fredrik Söderbom** who got me ad**dicty**ed on a slime mold. There are few people who I would rather share my results with than with you. Your joy and enthusiasm is what has kept me going from time to time. The same gratitude goes to **Andrea** for a great collaboration. That night when I first realized that the surrounding sequence folds into a pre-miRNA structure I will never forget. I will also never forget our extra trip around France and Germany. Thank you **Åsa** for optimism. Most of all I must thank **Lotta** who has lightened up my life inside and outside the lab for the last couple of years. Thanks for all the help and support during the last weeks. I would not have made it without you.

A special thanks goes to **Sandra** who introduced me to miRNAs and plants (not flowers). I really enjoyed the time we worked together. It would have been fun if we could have pursued it further.

To the people in my group, former and present, I must start with **Erik** with whom I have had such great collaboration. The OmrA /OmrB story has been so much fun and I really enjoyed all our discussions. Big thanks **Klas** for all the fun moments, especially for the great time in Kassel, **Fabien** for being such an inspiration and **Aurelie** for being different from Fabien but just as great. I would like to acknowledge **Cia** who just make things happen, I wish I had more of that, and **Maaike** for being happy all the time. I must not forget to thank **Magnus L** for being a “skåning” and such a good person. I also would like to thank **Salme** for being so nice to me when I started in the group. In the CRISPR group I would like to thank **Magnus** and **Amanda** for some good times. And of course I want to acknowledge **Nadja** with whom I have had so many wonderful moments while smoking and discussing science and life.

I would like to thank **Leif** for being nice to me, **Mattias** for the great parties and **Fredrik** for the time in Iceland. Most I would like to thank **Shiying** for putting a smile on my face every time we meet. It has been a pleasure knowing you. While I am at it I must acknowledge **Diarmaid** and **Santanu,** you really got me hooked on genetic switches. The best course I ever took so far. **Marie** thanks for sacrificing your helium balloon so we could sing Billy Idol with a high pitch and **Disa** for making the every day life at work just a little bit more interesting.

I would like to thank **Karin Carlsson** for the help when I have been teaching and **Pernilla** for the time in the course lab. **Nora**, you are such a great person and your work has really broadened my view on bacteria. **Bhupi** and **Sonchita**, you really make a great couple.

I would also like to thank **Staffan** not only for being a great person but also for bringing so many wonderful people to the microbiology group. **Jon** for being interested in biotechnology like me, **Johan** for just being a great guy, **Emma** whose laughter I miss so much and **Karin**, thank you for being a friend. The time at ICM would not have been the same without you guys.

Finally I would like to thank **Kurt Nordström** whose eagerness to learn more got me thinking that this is something that I also would like to do. I hope that I will stay as enthusiastic about science as you for the rest of my life.

To those who made antisense:

This is for all you guys (and girls) who has not helped me become a scientist but has regulated my time at work making sure that I survive during all these years.

First I must thank **Pontus.** You could have been on the sense list but you are so much more a friend than you ever was my colleague. Thanks for all the great moments and I am looking forward to the next ones. I would also like to thank **Per, Niklas** for all the lunches and discussions. A big thanks to the **ICM football team**, I really wished I had more time to play football. A big hurray for the **ICM-band** which made me feel like a rock star for the first time ever and probably for the last time also. Thank you **Mats** and **Prune** for all the nice talks and for the great music. Thanks **Christofer B** for your dry sense of humor I really enjoy it.

Also among former ICM members I must thank **Helene B** for all the fun time we have had together. I am so glad that you have never beaten me in badminton. **Ulrika L** for being such a good sport and having such a wonderful smile. **Hava**, I really miss you, I think that you are the only one who understood all my not always so funny jokes and **Henrik** who has an even more twisted sense of humor than me.

Outside the lab my gratitude goes to my friends **Skogis**, **Kalle**, **Greken**, **Marre** and **Klabbe** for just being friends for life. I also like to acknowledge my childhood friends for the once a year weekend were we meet up and just

hang out. Here I would also like to send my warm thoughts to my family-in-law who always makes me feel like the perfect son-in-law.

I like to acknowledge my family, especially my mother and my father for loving me no matter what, my younger brother and sisters for always being there for me and my elder brother whom I miss so much it hurts. **Bettan** and **Patrik** with family for sticking close to the family and my **grandmother** who always helped me financially when economy was not that great. I love you all so much.

Finally, to the biggest antisense in terms of science but the biggest sense in terms of life I must acknowledge my children **Emil**, **Nohmi** and **Nelly** who makes sure that I never work to much and fills my life with meaning and of course my wife **Lisa**. You are the most amazing woman I have ever met. I hope that you stick on to me for the rest of our life. Nothing else would make sense.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J Mol Biol* **215**: 403-410.

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M. et al. 2003. A uniform system for microRNA annotation. *RNA* **9**: 277-279.

Aravin, A.A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. 2003. The small RNA profile during Drosophila melanogaster development. *Dev Cell* **5**: 337-350.

Argaman, L. and Altuvia, S. 2000. fhlA repression by OxyS RNA: kissing complex formation at two sites results in a stable antisense-target RNA complex. *J Mol Biol* **300**: 1101-1112.

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

Aspegren, A., Hinas, A., Larsson, P., Larsson, A., and Soderbom, F. 2004. Novel non-coding RNAs in Dictyostelium discoideum and their expression during development. *Nucleic Acids Res* **32**: 4646-4656.

Baek, D., Villen, J., Shin, C., Camargo, F.D., Gygi, S.P., and Bartel, D.P. 2008. The impact of microRNAs on protein output. *Nature* **455**: 64-71.

Baldauf, S.L. and Doolittle, W.F. 1997. Origin and evolution of the slime molds (Mycetozoa). *Proc Natl Acad Sci U S A* **94**: 12007-12012.

Bapteste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Durufle, L., Gaasterland, T., Lopez, P., Muller, M. et al. 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. *Proc Natl Acad Sci U S A* **99**: 1414-1419.

Barnhart, M.M. and Chapman, M.R. 2006. Curli biogenesis and function. *Annu Rev Microbiol* **60**: 131-147.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D.A., and Horvath, P. 2007. CRISPR provides acquired resistance against viruses in prokaryotes. *Science* **315**: 1709-1712.

Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**: 363-366.

Betel, D., Koppal, A., Agius, P., Sander, C., and Leslie, C. 2010. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol* **11**: R90.

Billoud, B., De Paepe, R., Baulcombe, D., and Boccara, M. 2005. Identification of new small non-coding RNAs from tobacco and Arabidopsis. *Biochimie* **87**: 905-910.

Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F. et al. 1997. The complete genome sequence of Escherichia coli K-12. *Science* **277**: 1453-1462.

Brian N. Zeiler, R.W.S. 1998. Antisense RNA Structure and Function. in *RNA Structure and Function*. Cold Spring Harbor Monograph Archive.

Brown, C.J., Ballabio, A., Rupert, J.L., Lafreniere, R.G., Grompe, M., Tonlorenzi, R., and Willard, H.F. 1991. A gene from the region of the human X inactivation centre is expressed exclusively from the inactive X chromosome. *Nature* **349**: 38-44.

Busch, A., Richter, A.S., and Backofen, R. 2008. IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. *Bioinformatics* **24**: 2849-2856.

Calvo, J.M. and Matthews, R.G. 1994. The leucine-responsive regulatory protein, a global regulator of metabolism in Escherichia coli. *Microbiol Rev* **58**: 466-490.

Cao, Y., Wu, J., Liu, Q., Zhao, Y., Ying, X., Cha, L., Wang, L., and Li, W. 2010. sRNATarBase: A comprehensive database of bacterial sRNA targets verified by experiments. *RNA*.

Chen, S., Zhang, A., Blyn, L.B., and Storz, G. 2004. MicC, a second small-RNA regulator of Omp protein expression in Escherichia coli. *J Bacteriol* **186**: 6689-6697.

Chiang, H.R., Schoenfeld, L.W., Ruby, J.G., Auyeung, V.C., Spies, N., Baek, D., Johnston, W.K., Russ, C., Luo, S., Babiarz, J.E. et al. 2010. Mammalian microRNAs: experimental evaluation of novel and previously annotated genes. *Genes Dev* **24**: 992-1009.

Couzin, J. 2002. Breakthrough of the year. Small RNAs make big splash. *Science* **298**: 2296-2297.

Crick, F. 1970. Central dogma of molecular biology. *Nature* **227**: 561-563.

Darfeuille, F., Unoson, C., Vogel, J., and Wagner, E.G. 2007. An antisense RNA inhibits translation by competing with standby ribosomes. *Mol Cell* **26**: 381-392.

Do, C.B., Woods, D.A., and Batzoglou, S. 2006. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**: e90-98.

Eddy, S.R. 2001. Non-coding RNA genes and the modern RNA world. *Nat Rev Genet* **2**: 919-929.

Faucher, S.P., Friedlander, G., Livny, J., Margalit, H., and Shuman, H.A. 2010. Legionella pneumophila 6S RNA optimizes intracellular multiplication. *Proc Natl Acad Sci U S A* **107**: 7533-7538.

Figueroa-Bossi, N., Valentini, M., Malleret, L., Fiorini, F., and Bossi, L. 2009. Caught at its own game: regulatory small RNA inactivated by an inducible transcript mimicking its target. *Genes Dev* **23**: 2004-2015.

Fire, A., Xu, S., Montgomery, M.K., Kostas, S.A., Driver, S.E., and Mello, C.C. 1998. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**: 806-811.

Franco-Zorrilla, J.M., Valli, A., Todesco, M., Mateos, I., Puga, M.I., Rubio-Somoza, I., Leyva, A., Weigel, D., Garcia, J.A., and Paz-Ares, J. 2007. Target mimicry provides a new mechanism for regulation of microRNA activity. *Nat Genet* **39**: 1033-1037.

Freyhult, E.K., Bollback, J.P., and 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.

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

Ghildiyal, M. and Zamore, P.D. 2009. Small silencing RNAs: an expanding universe. *Nat Rev Genet* **10**: 94-108.

Glockner, G., Szafranski, K., Winckler, T., Dingermann, T., Quail, M.A., Cox, E., Eichinger, L., Noegel, A.A., and Rosenthal, A. 2001. The complex repeats of Dictyostelium discoideum. *Genome Res* **11**: 585-594.

Griffiths-Jones, S. 2006. miRBase: the microRNA sequence database. *Methods Mol Biol* **342**: 129-138.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., and Eddy, S.R. 2003. Rfam: an RNA family database. *Nucleic Acids Res* **31**: 439-441.

Grimson, A., Farh, K.K., Johnston, W.K., Garrett-Engele, P., Lim, L.P., and Bartel, D.P. 2007. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27**: 91-105.

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B.J., Chiang, H.R., King, N., Degnan, B.M., Rokhsar, D.S., and Bartel, D.P. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* **455**: 1193-1197.

Guillier, M. and Gottesman, S. 2006. Remodelling of the Escherichia coli outer membrane by two small regulatory RNAs. *Mol Microbiol* **59**: 231-247.

Guo, H., Ingolia, N.T., Weissman, J.S., and Bartel, D.P. 2010. Mammalian microRNAs predominantly act to decrease target mRNA levels. *Nature* **466**: 835-840.

Hammar, M., Arnqvist, A., Bian, Z., Olsen, A., and Normark, S. 1995. Expression of two csg operons is required for production of fibronectin- and congo red-binding curli polymers in Escherichia coli K-12. *Mol Microbiol* **18**: 661-670.

Han, J., Lee, Y., Yeom, K.H., Nam, J.W., Heo, I., Rhee, J.K., Sohn, S.Y., Cho, Y., Zhang, B.T., and Kim, V.N. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell* **125**: 887-901.

Hertel, J. and Stadler, P.F. 2006. Hairpins in a Haystack: recognizing microRNA precursors in comparative genomics data. *Bioinformatics* **22**: e197-202.

Hildebrandt, M. and Nellen, W. 1992. Differential antisense transcription from the Dictyostelium EB4 gene locus: implications on antisense-mediated regulation of mRNA stability. *Cell* **69**: 197-204.

Hinas, A., Reimegard, J., Wagner, E.G., Nellen, W., Ambros, V.R., and Soderbom, F. 2007. The small RNA repertoire of Dictyostelium discoideum and its regulation by components of the RNAi pathway. *Nucleic Acids Res* **35**: 6714-6726.

Holmqvist, E., Reimegard, J., Sterk, M., Grantcharova, N., Romling, U., and Wagner, E.G. 2010. Two antisense RNAs target the transcriptional regulator CsgD to inhibit curli synthesis. *EMBO J* **29**: 1840-1850.

Horvath, P. and Barrangou, R. 2010. CRISPR/Cas, the immune system of bacteria and archaea. *Science* **327**: 167-170.

Huttenhofer, A. and Vogel, J. 2006. Experimental approaches to identify non-coding RNAs. *Nucleic Acids Res* **34**: 635-646.

Jones-Rhoades, M.W. and Bartel, D.P. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell* **14**: 787-799.

Jopling, C.L., Yi, M., Lancaster, A.M., Lemon, S.M., and Sarnow, P. 2005. Modulation of hepatitis C virus RNA abundance by a liver-specific MicroRNA. *Science* **309**: 1577-1581.

Jukes, T.H. and Osawa, S. 1990. The genetic code in mitochondria and chloroplasts. *Experientia* **46**: 1117-1126.

Kawamoto, H., Koide, Y., Morita, T., and Aiba, H. 2006. Base-pairing requirement for RNA silencing by a bacterial small RNA and acceleration of duplex formation by Hfq. *Mol Microbiol* **61**: 1013-1022.

Kawamoto, H., Morita, T., Shimizu, A., Inada, T., and Aiba, H. 2005. Implication of membrane localization of target mRNA in the action of a small RNA: mechanism of post-transcriptional regulation of glucose transporter in Escherichia coli. *Genes Dev* **19**: 328-338.

Kuhlmann, M., Borisova, B.E., Kaller, M., Larsson, P., Stach, D., Na, J., Eichinger, L., Lyko, F., Ambros, V., Soderbom, F. et al. 2005. Silencing of retrotransposons in Dictyostelium by DNA methylation and RNAi. *Nucleic Acids Res* **33**: 6405-6417.

Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294**: 853-858.

Lartigue, C., Vashee, S., Algire, M.A., Chuang, R.Y., Benders, G.A., Ma, L., Noskov, V.N., Denisova, E.A., Gibson, D.G., Assad-Garcia, N. et al. 2009. Creating bacterial strains from genomes that have been cloned and engineered in yeast. *Science* **325**: 1693-1696.

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

Lau, N.C., Lim, L.P., Weinstein, E.G., and Bartel, D.P. 2001. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294**: 858-862.

Lease, R.A., Cusick, M.E., and Belfort, M. 1998. Riboregulation in Escherichia coli: DsrA RNA acts by RNA:RNA interactions at multiple loci. *Proc Natl Acad Sci U S A* **95**: 12456-12461.

Lee, J.T., Davidow, L.S., and Warshawsky, D. 1999. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat Genet* **21**: 400-404.

Lee, R.C. and Ambros, V. 2001. An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294**: 862-864.

Lee, R.C., Feinbaum, R.L., and Ambros, V. 1993. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**: 843-854.

Lee, Y., Jeon, K., Lee, J.T., Kim, S., and Kim, V.N. 2002. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J* **21**: 4663-4670.

Levine, E., Zhang, Z., Kuhlman, T., and Hwa, T. 2007. Quantitative characteristics of gene regulation by small RNA. *PLoS Biol* **5**: e229.

Llave, C., Kasschau, K.D., Rector, M.A., and Carrington, J.C. 2002. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14**: 1605-1619.

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

Lu, C., Tej, S.S., Luo, S., Haudenschild, C.D., Meyers, B.C., and Green, P.J. 2005. Elucidation of the small RNA component of the transcriptome. *Science* **309**: 1567-1569.

Majdalani, N., Chen, S., Murrow, J., St John, K., and Gottesman, S. 2001. Regulation of RpoS by a novel small RNA: the characterization of RprA. *Mol Microbiol* **39**: 1382-1394.

Martens, H., Novotny, J., Oberstrass, J., Steck, T.L., Postlethwait, P., and Nellen, W. 2002. RNAi in Dictyostelium: the role of RNA-directed RNA polymerases and double-stranded RNase. *Mol Biol Cell* **13**: 445-453.

Masse, E. and Gottesman, S. 2002. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. *Proc Natl Acad Sci U S A* **99**: 4620-4625.

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

Mizuno, T., Chou, M.Y., and Inouye, M. 1984. A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA). *Proc Natl Acad Sci U S A* **81**: 1966-1970.

Moller, T., Franch, T., Udesen, C., Gerdes, K., and Valentin-Hansen, P. 2002. Spot 42 RNA mediates discoordinate expression of the E. coli galactose operon. *Genes Dev* **16**: 1696-1706.

Morita, T., Maki, K., and Aiba, H. 2005. RNase E-based ribonucleoprotein complexes: mechanical basis of mRNA destabilization mediated by bacterial noncoding RNAs. *Genes Dev* **19**: 2176-2186.

Morita, T., Mochizuki, Y., and Aiba, H. 2006. Translational repression is sufficient for gene silencing by bacterial small noncoding RNAs in the absence of mRNA destruction. *Proc Natl Acad Sci U S A* **103**: 4858-4863.

Muckstein, U., Tafer, H., Hackermuller, J., Bernhart, S.H., Stadler, P.F., and Hofacker, I.L. 2006. Thermodynamics of RNA-RNA binding. *Bioinformatics* **22**: 1177-1182.

Okamura, K., Hagen, J.W., Duan, H., Tyler, D.M., and Lai, E.C. 2007. The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila. *Cell* **130**: 89-100.

Pfeiffer, V., Papenfort, K., Lucchini, S., Hinton, J.C., and Vogel, J. 2009. Coding sequence targeting by MicC RNA reveals bacterial mRNA silencing downstream of translational initiation. *Nat Struct Mol Biol* **16**: 840-846.

Piriyapongsa, J. and Jordan, I.K. 2007. A family of human microRNA genes from miniature inverted-repeat transposable elements. *PLoS One* **2**: e203.

Piriyapongsa, J., Marino-Ramirez, L., and Jordan, I.K. 2007. Origin and evolution of human microRNAs from transposable elements. *Genetics* **176**: 1323-1337.

Rasmussen, A.A., Johansen, J., Nielsen, J.S., Overgaard, M., Kallipolitis, B., and Valentin-Hansen, P. 2009. A conserved small RNA promotes silencing of the outer membrane protein YbfM. *Mol Microbiol*.

Rehmsmeier, M., Steffen, P., Hochsmann, M., and Giegerich, R. 2004. Fast and effective prediction of microRNA/target duplexes. *Rna* **10**: 1507-1517.

Reinhart, B.J., Weinstein, E.G., Rhoades, M.W., Bartel, B., and Bartel, D.P. 2002. MicroRNAs in plants. *Genes Dev* **16**: 1616-1626.

Rhoades, M.W., Reinhart, B.J., Lim, L.P., Burge, C.B., Bartel, B., and Bartel, D.P. 2002. Prediction of plant microRNA targets. *Cell* **110**: 513-520.

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

Rivas, E. and Eddy, S.R. 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* **16**: 583-605.

Rivas, E., Klein, R.J., Jones, T.A., and Eddy, S.R. 2001. Computational identification of noncoding RNAs in E. coli by comparative genomics. *Curr Biol* **11**: 1369-1373.

Romby, P., Vandenesch, F., and Wagner, E.G. 2006. The role of RNAs in the regulation of virulence-gene expression. *Curr Opin Microbiol* **9**: 229-236.

Ruby, J.G., Jan, C.H., and Bartel, D.P. 2007. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448**: 83-86.

Schmidt, M., Zheng, P., and Delihas, N. 1995. Secondary structures of Escherichia coli antisense micF RNA, the 5'-end of the target ompF mRNA, and the RNA/RNA duplex. *Biochemistry* **34**: 3621-3631.

Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. 2008. Widespread changes in protein synthesis induced by microRNAs. *Nature* **455**: 58-63.

Sewer, A., Paul, N., Landgraf, P., Aravin, A., Pfeffer, S., Brownstein, M.J., Tuschl, T., van Nimwegen, E., and Zavolan, M. 2005. Identification of clustered microRNAs using an ab initio prediction method. *BMC Bioinformatics* **6**: 267.

Shabalina, S.A. and Koonin, E.V. 2008. Origins and evolution of eukaryotic RNA interference. *Trends Ecol Evol* **23**: 578-587.

Sharma, C.M., Darfeuille, F., Plantinga, T.H., and 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.

Shimoni, Y., Friedlander, G., Hetzroni, G., Niv, G., Altuvia, S., Biham, O., and Margalit, H. 2007. Regulation of gene expression by small non-coding RNAs: a quantitative view. *Mol Syst Biol* **3**: 138.

Shin, C., Nam, J.W., Farh, K.K., Chiang, H.R., Shkumatava, A., and Bartel, D.P. 2010. Expanding the microRNA targeting code: functional sites with centered pairing. *Mol Cell* **38**: 789-802.

Shine, J. and Dalgarno, L. 1975. Determinant of cistron specificity in bacterial ribosomes. *Nature* **254**: 34-38.

Simons, R.W. 1997. Antisense RNA Structure and Function. in *RNA Structure and Function* (ed. R.W. Simons and M. Grunberg-Manago). Cold Sprin Harbor Labaratory Press.

Soper, T., Mandin, P., Majdalani, N., Gottesman, S., and Woodson, S.A. 2010. Positive regulation by small RNAs and the role of Hfq. *Proc Natl Acad Sci U S A* **107**: 9602-9607.

Stern-Ginossar, N., Elefant, N., Zimmermann, A., Wolf, D.G., Saleh, N., Biton, M., Horwitz, E., Prokocimer, Z., Prichard, M., Hahn, G. et al. 2007. Host immune system gene targeting by a viral miRNA. *Science* **317**: 376-381.

Stombaugh, J., Zirbel, C.L., Westhof, E., and Leontis, N.B. 2009. Frequency and isostericity of RNA base pairs. *Nucleic Acids Res* **37**: 2294-2312.

Stougaard, P., Molin, S., and Nordstrom, K. 1981. RNAs involved in copy-number control and incompatibility of plasmid R1. *Proc Natl Acad Sci U S A* **78**: 6008-6012.

Tjaden, B. 2008. TargetRNA: a tool for predicting targets of small RNA action in bacteria. *Nucleic Acids Res* **36**: W109-113.

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

Tomizawa, J., Itoh, T., Selzer, G., and Som, T. 1981. Inhibition of ColE1 RNA primer formation by a plasmid-specified small RNA. *Proc Natl Acad Sci U S A* **78**: 1421-1425.

Udekwu, K.I., Darfeuille, F., Vogel, J., Reimegard, J., Holmqvist, E., and Wagner, E.G. 2005. Hfq-dependent regulation of OmpA synthesis is mediated by an antisense RNA. *Genes Dev* **19**: 2355-2366.

Unoson, C. and Wagner, E.G. 2007. Dealing with stable structures at ribosome binding sites: bacterial translation and ribosome standby. *RNA Biol* **4**: 113-117.

Urban, J.H. and Vogel, J. 2007. Translational control and target recognition by Escherichia coli small RNAs in vivo. *Nucleic Acids Res* **35**: 1018-1037.

Vagin, V.V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., and Zamore, P.D. 2006. A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313**: 320-324.

Vanderpool, C.K. and Gottesman, S. 2004. Involvement of a novel transcriptional activator and small RNA in post-transcriptional regulation of the glucose phosphoenolpyruvate phosphotransferase system. *Mol Microbiol* **54**: 1076-1089.

Vecerek, B., Rajkowitsch, L., Sonnleitner, E., Schroeder, R., and Blasi, U. 2008. The C-terminal domain of Escherichia coli Hfq is required for regulation. *Nucleic Acids Res* **36**: 133-143.

Vogel, J., Argaman, L., Wagner, E.G., and Altuvia, S. 2004. The small RNA IstR inhibits synthesis of an SOS-induced toxic peptide. *Curr Biol* **14**: 2271-2276.

Vogel, J. and Wagner, E.G. 2007. Target identification of small noncoding RNAs in bacteria. *Curr Opin Microbiol* **10**: 262-270.

Wagner, E.G., Altuvia, S., and Romby, P. 2002. Antisense RNAs in bacteria and their genetic elements. *Adv Genet* **46**: 361-398.

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

Wassarman, K.M. and Storz, G. 2000. 6S RNA regulates E. coli RNA polymerase activity. *Cell* **101**: 613-623.

Waters, L.S. and Storz, G. 2009. Regulatory RNAs in bacteria. *Cell* **136**: 615-628.

Watson, J.D. and Crick, F.H. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. *Nature* **171**: 737-738.

Wightman, B., Ha, I., and Ruvkun, G. 1993. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75**: 855-862.

Zamore, P.D., Tuschl, T., Sharp, P.A., and Bartel, D.P. 2000. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**: 25-33.

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

Zhao, T., Li, G., Mi, S., Li, S., Hannon, G.J., Wang, X.J., and Qi, Y. 2007. A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii. *Genes Dev* **21**: 1190-1203.

Zhao, Y., Li, H., Hou, Y., Cha, L., Cao, Y., Wang, L., Ying, X., and Li, W. 2008. Construction of two mathematical models for prediction of bacterial sRNA targets. *Biochem Biophys Res Commun* **372**: 346-350.

Acta Universitatis Upsaliensis

Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology 701

Editor: The Dean of the Faculty of Science and Technology

A doctoral dissertation from the Faculty of Science and Technology, Uppsala University, is usually a summary of a number of papers. A few copies of the complete dissertation are kept at major Swedish research libraries, while the summary alone is distributed internationally through the series Digital Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology. (Prior to January, 2005, the series was published under the title "Comprehensive Summaries of Uppsala Dissertations from the Faculty of Science and Technology".)

Distribution: publications.uu.se
urn:nbn:se:uu:diva-131168

ACTA
UNIVERSITATIS
UPSALIENSIS
UPPSALA
2010